

Addressing the Challenge of Ambiguous Data in Deep Learning

A Strategy for Creating High-quality Image
Annotations with Human Reliability and
Judgement Enhancement

vorgelegt von
M.Sc. Lars Schmarje
aus
Schleswig

Dissertation
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
(Dr.-Ing.)
der Technischen Fakultät
der Christian-Albrechts-Universität zu Kiel
eingereicht im Jahr 2024

Kiel Computer Science Series (KCSS) 2024/06 dated 2024-06-06

ISSN 2193-6781 (print version)

ISSN 2194-6639 (electronic version)

Electronic version, updates, errata available via <https://www.informatik.uni-kiel.de/kcss>

The author can be contacted via science@schmarje-sh.de

Published by the Department of Computer Science, Kiel University

Multimedia Information Processing Group

Please cite as:

▷ Lars Schmarje. *Addressing the Challenge of Ambiguous Data in Deep Learning* Number 2024/3 in Kiel Computer Science Series. Department of Computer Science, 2024. Dissertation, Faculty of Engineering, Kiel University.

```
@book{schmarje2024dis,  
  author    = {Schmarje, Lars},  
  title     = {Adressing the Challenge of Ambiguous Data in Deep Learning},  
  publisher = {Department of Computer Science, Kiel University},  
  year      = {2024},  
  number    = {2024/3},  
  doi       = {10.21941/kcss/2024/3},  
  series    = {Kiel Computer Science Series},  
  note      = {Dissertation, Faculty of Engineering,  
               Kiel University.}  
}
```

© 2024 by Lars Schmarje

About this Series

The Kiel Computer Science Series (KCSS) covers dissertations, habilitation theses, lecture notes, textbooks, surveys, collections, handbooks, etc. written at the Department of Computer Science at Kiel University. It was initiated in 2011 to support authors in the dissemination of their work in electronic and printed form, without restricting their rights to their work. The series provides a unified appearance and aims at high-quality typography. The KCSS is an open access series; all series titles are electronically available free of charge at the department's website. In addition, authors are encouraged to make printed copies available at a reasonable price, typically with a print-on-demand service.

Please visit <http://www.informatik.uni-kiel.de/kcss> for more information, for instructions how to publish in the KCSS, and for access to all existing publications.

1. Gutachter: Prof. Dr. Ing. Reinhard Koch
Christian-Albrechts-Universität
Kiel
2. Gutachter: Prof. Dr. rer. nat. Carsten Meyer
Zweitmitglied Christian-Albrechts-Universität
Kiel

Datum der mündlichen Prüfung: 03.05.2024

Zusammenfassung

Im Bereich des maschinellen Lernens ist die Verfügbarkeit von qualitativ hochwertigen, annotierten Daten für das Training hochleistungsfähiger neuronaler Netze unerlässlich. Menschen sind sich jedoch oft nicht einig, wenn es um die Annotation von Bildklassifikationsdaten geht, was zu mehrdeutigen Daten führt, die eine große Herausforderung für Deep Learning darstellen. Diese Dissertation konzentriert sich auf dieses Problem, indem sie einen Überblick über die Definition des Problems, die benötigten Daten, die Bewertungsmetriken und die Methoden zur Lösung des Problems gibt. Das Ziel ist es, eine effektive Annotationsstrategie für die Bildklassifikation zu entwickeln, die auf mehreren tausend Experimenten basiert, um die Datenqualität zu verbessern und gleichzeitig die Kosten zu senken.

Die Hauptidee dieser Arbeit besteht darin, sich nicht auf eine Annotation pro Bild zu verlassen, sondern den Durchschnitt mehrerer Annotationen zu betrachten. In der Realität ist dies jedoch kaum durchführbar, ohne die Kosten für die Annotation zu reduzieren. Daher werden in dieser Dissertation Vorschläge als Orientierungshilfe während des Annotationsprozesses verwendet, um die Konsistenz der Annotationen zu steigern und die Annotationszeit und die damit verbundenen Kosten zu minimieren. Die Verwendung von Vorschlägen kann jedoch zu einer Verzerrung der Datenverteilung führen, da sich Menschen von den Vorschlägen beeinflussen lassen. Diese Verzerrung wurde analysiert und darauf aufbauend eine Möglichkeit entwickelt, sie zu minimieren. Die Untersuchungen wurden schließlich in einer Annotationsstrategie in Form eines Flussdiagramms zusammengefasst. Dieses Flussdiagramm ermöglicht es zukünftigen Forschern, mit minimalem Aufwand qualitativ hochwertige Daten für ihren spezifischen Anwendungsfall zu erzeugen. Die Strategie wurde erfolgreich an einer realen biomedizinischen Aufgabenstellung verifiziert.

Zusammenfassend stellt diese Arbeit eine umfassende Untersuchung der Herausforderungen dar, die sich aus der Annotation mehrdeutiger

Daten im Rahmen des maschinellen Lernens ergeben. Diese Forschungsarbeit bietet eine effektive Strategie, um das grundlegende Problem der Datenqualität anzugehen und dadurch die Qualität und Zuverlässigkeit von Modellen zu verbessern und das Gebiet des maschinellen Lernens insgesamt voranzubringen.

Abstract

In machine learning, the availability of high-quality labeled data is essential for training accurate models. However, humans often disagree among themselves or over time when labeling or annotating image classification data. As a result, they create ambiguous data that poses a significant challenge to deep learning. This research focuses on addressing this issue by proposing an overview that defines the problem, provides the necessary data for research, establishes evaluation metrics and presents methods for solving the problem. The ultimate goal is to develop an effective annotation strategy for image classification based on several thousand different experiments to improve data quality while reducing cost.

The main idea of this dissertation is not to rely on one annotation per image, but to consider the average of multiple annotations. However, this would not be feasible in reality without reducing the cost of acquiring such annotations. This dissertation uses proposals as a guide during the annotation process to improve the consistency of the annotations and to reduce the annotation time and thus the associated costs. The use of proposals could introduce a bias into the data. Analysis of this bias and its introduction led to possible methods to minimize or reverse it. All this research is finally unified in an annotation strategy in the form of a flowchart that allows future research to easily understand the necessary steps needed to produce high quality data for their specific use case with as little effort as possible. This strategy is successfully verified on a real biomedical task.

In conclusion, this thesis presents a comprehensive investigation of the challenges posed by annotating ambiguous data in deep learning. This research provides an effective strategy to address the fundamental issue of data quality, thereby improving the quality and reliability of models and advancing the field of deep learning as a whole.

Acknowledgements

This dissertation has been influenced and guided by many people in my life. While this small piece of text cannot reflect all of my gratitude, I will try to acknowledge the overwhelming support I have received.

First of all, I have to thank my wonderful wife for supporting me all these years during my Ph.D. and before. She keeps my life in balance and I am truly grateful for her. She put up with every over-complicated explanation of mine and helped me find a simpler way of describing it. She listened to my frustrations and complaints about work and encouraged me to keep going and see the bright side. Not only did she marry me, but she reduced her work hours to raise our two wonderful children, allowing me to start and continue my path of this dissertation. I can truly say that this work would not have been possible without her.

The colleagues and fellow researchers I have met over the years have helped me in many ways. It can be as simple as looking at some data together and trying to make sense of it. On other occasions, we have hunted for bugs together in source code or in video games in our spare time. We helped each other either in person at the office or later in virtual chats and video calls. We created valuable memories at joint conferences, trips, or seminars. Sharing data and knowledge with researchers across Europe allowed me to expand and improve my research. I hope that the impact of my shared expertise was as valuable to the others as theirs was to me. Overall, I appreciate the broad and unconditional support from colleagues and fellow researchers for my ideas and ultimately this dissertation.

My friends and family have been the stabilizing force in the background. They always showed me that life is so much more than the next paper, while still reading or checking it on often very short notice. A special thanks goes to my parents. They created an environment in my childhood where I could grow to my full potential and supported me along the way

with love and encouragement. Without them, I most likely would not have taken the path that has led me to this dissertation.

Lastly, but perhaps most importantly, as with all of my work, I thank my advisor for his guidance throughout my Ph.D. program. He had the confidence in me to find my own way and let me discover on my own terms where I wanted to go with this dissertation. He provided guiding boundaries within which I could explore. This guidance could come in the form of good advice, the acquisition of interesting projects and the necessary financial support. While this may sound easy, and he certainly made it seem easy, I have heard enough testimonials from other graduate students to realize that what I often took for granted was actually a great accomplishment on his part.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Problem statement	2
1.3	Research questions	3
1.4	Structure of dissertation	5
1.5	List of published papers	7
I	Background	11
2	Data issues and their definitions	13
2.1	Ambiguity	13
2.2	Other definitions	15
2.2.1	Labels & Annotations	15
2.2.2	Noisy & Ambiguity	18
2.2.3	Aleatoric Uncertainty & Epistemic Uncertainty . . .	18
2.2.4	Soft Labels & Label Smoothing	21
3	Quantity & Quality of Data	25
3.1	Semi-, Self- and Unsupervised Learning	27
3.1.1	Extended common ideas	29
3.1.2	Extended methods	31
3.1.3	Extended comparison	35
3.2	Proposal guided annotations	40
II	Own methodical and data contributions	43
4	Overview about the methodology	45

Contents

5	How to measure data quality?	51
5.1	Benchmark	51
5.2	Metrics	56
5.2.1	Measuring distributional differences	56
5.2.2	Associated metrics	57
5.2.3	Special benchmark metrics	58
5.2.4	Costs	58
5.3	Human user studies	59
5.3.1	Evaluation study of annotation strategy	61
5.4	Simulated Proposal Acceptance (SPA)	64
6	Improving data quality	67
6.1	Fuzzy Overclustering (FOC)	68
6.1.1	Inverse cross-entropy (CE^{-1})	72
6.2	Data-Centric Classification & Clustering (DC3)	72
6.3	Cost-effective labeling using validated proposals and re-paired labels (CleverLabel)	76
6.4	Overclustering Stochastic Proposal (OSP)	79
6.5	Strategy for creating high-quality image annotations (SMART)	80
III	Analysis	83
7	Unified Evaluation	85
7.1	Evaluation on benchmark	85
7.2	Benefit of using Overclustering Stochastic Proposal	90
7.3	Verification of annotation strategy	92
8	Discussion	95
8.1	Research question discussion	95
8.1.1	Research Question 1: “What are the characteristics of ambiguous labels in relation to other data quality issues?” (RQ1)	95
8.1.2	Research Question 2: “How can improved data quality for image classification be quantified?” (RQ2)	96

8.1.3	Research Question 3: “What are the implications of using proposals to guide the annotation process for image classification?” (RQ3)	97
8.1.4	Research Question 4: “How should ambiguous data be annotated based on the results of RQ1, RQ2 and RQ3?” (RQ4)	98
8.2	Limitations & Future Work	99
9	Conclusion	103
IV	Appendix	107
A	Own previous papers	109
A.1	Long papers	109
A.1.1	2D and 3D Segmentation of uncertain local collagen fiber orientations in SHG microscopy	109
A.1.2	A Survey on Semi-, Self- and Unsupervised Learning for Image Classification	124
A.1.3	Fuzzy Overclustering: Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy	148
A.1.4	A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering	164
A.1.5	Is one annotation enough? - A data-centric image classification benchmark for noisy and ambiguous label estimation	183
A.1.6	Label Smarter, Not Harder: CleverLabel for Faster Annotation of Ambiguous Image Classification with Higher Quality	202
A.1.7	Annotating Ambiguous Images: General Annotation Strategy for Image Classification with Real-World Biomedical Validation on Vertebral Fracture Diagnosis	222
A.2	Short papers	237

Contents

A.2.1	Life is not black and white – Combining Semi-Supervised Learning with fuzzy labels	237
A.2.2	A Data-Centric Image Classification Benchmark . . .	246
A.2.3	Beyond hard labels: investigating data label distributions	254
A.3	Miscellaneous papers	260
A.3.1	Artificial intelligence-driven hip fracture prediction based on pelvic radiographs exceeds performance of DXA: the Study of Osteoporotic Fractures (SOF) . . .	260
A.3.2	Opportunistic hip fracture risk prediction in Men from X-ray: Findings from the Osteoporosis in Men (MrOS) Study	262
A.3.3	Dulling while judging? Veränderte Beurteilung von Fotos zu Pickverletzungen bei Puten durch Wiederholungen	275

Bibliography	277
---------------------	------------

Abbreviations

ACC	Accuracy	57
b	budget	54
BC	Bias Correction	77
CB	Class Blending	77
CE	Cross-Entropy	27
CE*	Cross-Entropy (used not as common supervised loss)	27
CL	Clustering Loss	27
CleverLabel	Cost-effective LabEling using Validated proposal-guidEd annotations and Repaired LABELs	47
CV	Computer Vision	1
DC3	Data-Centric Classification & Clustering	xii
δ	dataset specific offset	65
ECE	Expected calibration error	57
EM	Entropy Minimization	27
DL	Deep Learning	1
FOC	Fuzzy Overclustering	xii
GT	Ground-Truth	14
<i>in. sup.</i>	initial supervision	59
CE^{-1}	Inverse cross-entropy	xii
κ	Cohen’s Kappa Score	58
KL	Kullback-Leibler divergence	27
MAE	Masked Autoencoder	29
MI	Mutual Information	27
MSE	Mean Squared Error	27
MU	Mix up	27
NLP	Natural Language Processing	1
OOD	Out of distribution	20

Contents

OSP	Overclustering Stochastic Proposal	xii
OC	Overclustering	27
PL	Pseudo-Labeling	27
PT	Pretext Task	27
RQ1	Research Question 1: “What are the characteristics of ambiguous labels in relation to other data quality issues?”	xii
RQ2	Research Question 2: “How can improved data quality for image classification be quantified?”	xii
RQ2.1	Research Question 2.1: “What data are required to quantify improved data quality?”	4
RQ2.2	Research Question 2.2: “What metrics and algorithms are needed to quantify improved data quality?”	4
RQ3	Research Question 3: “What are the implications of using proposals to guide the annotation process for image classification?”	xiii
RQ3.1	Research Question 3.1: “What positive effects can be achieved by using proposals to guide the annotation process for image classification?”	4
RQ3.2	Research Question 3.2: “How can negative effects be minimized when using proposals to guide the annotation process for image classification?”	4
RQ4	Research Question 4: “How should ambiguous data be annotated based on the results of RQ1, RQ2 and RQ3?”	xiii
RQ4.1	Research Question 4.1: “Is one annotation enough to capture the ambiguity in image classification tasks?”	4
RQ4.2	Research Question 4.2: “Are proposals based on classification or overclustering better for the annotation process?”	4
RQ4.3	Research Question 4.3: “How does an increased or decreased budget during the annotation process affect the resulting data quality?”	4
S	Speedup	59
SMART	Strategy for creating high-quality iMAge Annotations with human Reliability and judgement enhancementT	48

Contents

SPA	Simulated Proposal Acceptance	xii
SSL	Semi-Supervised Learning	1
VAT	Virtual Adversarial Training	27
ViT	Vision Transformer	29
VLM	Vision Language Model	30

Introduction

1.1 Motivation

Modern machine learning, especially Deep Learning (DL), has solved many previously seemingly impossible tasks, for example, in the fields of Computer Vision (CV) or Natural Language Processing (NLP). In recent years, it has also had a major impact on other fields such as medicine, biology, engineering and even art [37, 40, 58, 126, 145]. Major advances are already being made in the field of autonomous driving [61] or improved healthcare systems [39, 77].

The fuel for these rapid and major developments is data [88, 198]. This dissertation examines data as a combination of images and labels, whereas raw data refers only to image data. In general, vast amounts of human-labeled data are required to train neural networks for DL. This is obviously a problem for several reasons. First, it can be difficult to obtain raw data, for example, in long-term studies such as MrOs¹ or SOF². Second, even if the raw data is available, we may not be allowed to use it due to copyright, privacy, or financial interests. Third, and most often, the human labor required to label the data is a limitation [81, 136]. This may be due to the limited amount of money available or the limited time of domain experts.

The amount of data required for image classification could be reduced by replacing labeled data with unlabeled data which is often readily available. This line of research is called Semi-Supervised Learning (SSL). In recent years, SSL research has achieved significant reductions in the amount of labeled data required, up to a factor of 100, with almost the same performance compared to full supervision [165, 174]. However,

¹<https://mrosonline.ucsf.edu/>

²<https://sofonline.ucsf.edu/>

1. Introduction

current research is often conducted on large and heavily curated datasets such as ImageNet [88], and thus is not easily applicable to smaller datasets with realistic noise and quality issues.

In recent years, many researchers have also looked at these problems from a data or data quality perspective. This perspective is called data-centric DL and tries to improve the data in contrast to just the model [115, 142]. The focus on the model is called model-centric DL. Data-centric research includes ideas such as considering the disagreement between annotations for the data quality [59], identifying over- or under-performing data splits, selecting the relevant image data for training of a larger set [55] or the issue of (adversarial) augmentation to ensure good models [11]. See a graphic representation of these key terms model- and data-centric in Figure 1.1.

Proposals, pseudo-labels or network predictions can be thought of as the 0th annotator suggesting a class to all other annotators during the image classification annotation process [41, 105, 130, 159]. While the use of such proposals has been reported to increase annotation speed and resulting quality [42], they also introduce bias into the data because people are more likely to accept the proposed class than without [78, 155].

1.2 Problem statement

As shown by concurrent research [15, 38, 193] and my own [149, 152, 154], it is often not enough to use only one human judgment during the labeling process for image classification, because human labor is costly and not entirely reliable. Humans suffer from intra- and interobserver variance [2, 20, 36, 48, 126, 149, 170] which is the variance of annotations over time for the same person or between people for the same task. This variance impacts the quality of a rating and can lead to noise or ambiguity. Due to the fact that classifications can be subjective due to difficult image quality [133, 149], imprecise definitions [111] or varying levels of experience among annotators, one cannot rely on one annotation per image [152]. However, more annotations increase the already high cost, and we found in [152] that common semi- and self-supervised approaches have problems with ambiguous data in real-world datasets. While proposals could minimize

1.3. Research questions

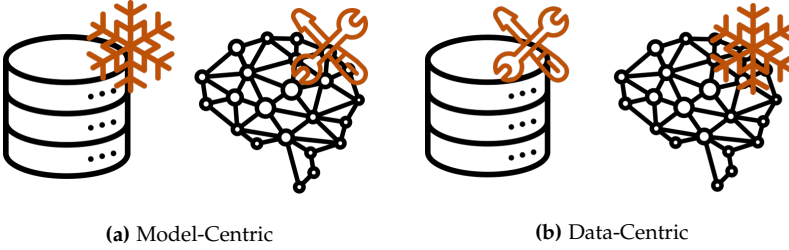


Figure 1.1. Illustration of the key concepts of model- and data-centric – The cylinder / database icon represents the data analyzed in this dissertation, including images and label information. The brain icon represents the model used, typically a neural network, which is trained on the provided data to make predictions. The snowflake indicates a less prominent or inactive part, whereas the tools icon represents the focus of the engineering/research effort. Model-centric focuses on the model while considering the given data fixed, whereas data-centric takes the opposite approach and focuses on improving data. Model-centric and data-centric are two approaches to improving image classification task performance. It is important to note that model-centric still may address data issues and data-centric employs models in its efforts to improve the data. The primary difference lies in the intention or motivation for improving the classifying model or the data used to train such a model. Both approaches are equally suitable for improving downstream image classification task performance. As much research currently emphasizes the model-centric approach, this dissertation delves into the data-centric perspective. The images are created with icons from Flaticon.com.

the issue of increased cost, they introduce a bias into the data which is often not considered or investigated [78, 155]. All in all, the aim of this dissertation is to produce high-quality data for image classification. Specifically, the data quality should be enhanced compared to previously reported methods. Therefore, the phrases “produce high-quality data” and “improve the data” will be used interchangeably throughout this study.

1.3 Research questions

The aforementioned problems of data quantity and quality in real-world datasets for image classification are addressed in this dissertation. To

1. Introduction

this end, research questions are formulated for the various aspects of these issues. These research questions define the main structure of this dissertation and are intended to guide the reader through this work.

- ▷ Research Question 1: “What are the characteristics of ambiguous labels in relation to other data quality issues?” (RQ1)
- ▷ Research Question 2: “How can improved data quality for image classification be quantified?” (RQ2)
 - ▷ Research Question 2.1: “What data are required to quantify improved data quality?” (RQ2.1)
 - ▷ Research Question 2.2: “What metrics and algorithms are needed to quantify improved data quality?” (RQ2.2)
- ▷ Research Question 3: “What are the implications of using proposals to guide the annotation process for image classification?” (RQ3)
 - ▷ Research Question 3.1: “What positive effects can be achieved by using proposals to guide the annotation process for image classification?” (RQ3.1)
 - ▷ Research Question 3.2: “How can negative effects be minimized when using proposals to guide the annotation process for image classification?” (RQ3.2)
- ▷ Research Question 4: “How should ambiguous data be annotated based on the results of RQ1, RQ2 and RQ3?” (RQ4)
 - ▷ Research Question 4.1: “Is one annotation enough to capture the ambiguity in image classification tasks?” (RQ4.1)
 - ▷ Research Question 4.2: “Are proposals based on classification or overclustering better for the annotation process?” (RQ4.2)
 - ▷ Research Question 4.3: “How does an increased or decreased budget during the annotation process affect the resulting data quality?” (RQ4.3)

The research questions build on each other starting from the definition over how to measure and improve the quality to general implications.

1.4. Structure of dissertation

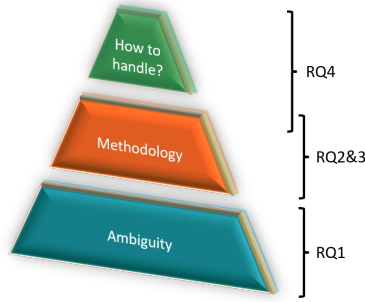


Figure 1.2. Major parts of the structure of this dissertation – The first part will discuss the background of this work and in particular will define the term ambiguity by answering RQ1. The second part is my methodology where my own contributions will be listed and which will answer RQ2 and RQ3. The unified strategy for dealing with ambiguous data is also proposed in this part. The third part is an analysis and discussion especially of the previously proposed strategy to answer the question of how to handle ambiguous data (see RQ4).

1.4 Structure of dissertation

This dissertation is divided into three main parts, except for the introduction and conclusion, as shown in Figure 1.2. This is a cumulative work of my related scientific research, so the details of the individual papers [37, 63, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 185] are listed in Section 1.5 and reprinted in Appendix A. The focus of this dissertation is to highlight the connections and conclusions between my individual papers based on the research questions defined above. The three main parts are described in detail below.

Background Part I describes the relevant literature and background information necessary to understand this dissertation in a broader context. It also shows the connection to previous methods and what makes this research different.

Chapter 2 describes common data issues in terms of quality and quantity. This motivates the definition of ambiguity for image classification

1. Introduction

labels, which is in line with my previous papers [149, 152, 154, 155]. Based on this definition, the similarity and difference to other definitions such as uncertainty or label noise is shown, which answers Research Question 1: “What are the characteristics of ambiguous labels in relation to other data quality issues?” (RQ1).

In Chapter 3, a review of the literature on Semi-Supervised Learning (SSL) and proposal systems for improving quantity and quality, respectively, is presented. These approaches are contrasted with others, which are beyond the scope of this dissertation. Much of the relevant SSL literature has been published in [151] and is expanded and updated in this part.

Own Contributions Part II summarizes my own methodological and data research. It gives an overview of the created benchmark [152], the evaluated methods [149, 154], the label improvement [155] and the resulting strategy [156]. The included contributions are

- ▷ the construction of a multi-domain benchmark [152] with multiple annotations per image to allow the quantification of improved data quality. Most of the image data come from real-world classification problems [20, 94, 125, 149, 154, 157, 166, 184, 186] of fellow researchers, which emphasizes its closeness to the real world and answers Research Question 2.1: “What data are required to quantify improved data quality?” (RQ2.1) in Chapter 5.
- ▷ a summary of successfully applied metrics and algorithms to quantify different aspects of improved data quality [149, 152, 154, 155] in Chapter 5. The summary includes a comparison between the metrics, the description of the used user studies and our simulation for proposal acceptance to allow larger scale evaluation, which answers Research Question 2.2: “What metrics and algorithms are needed to quantify improved data quality?” (RQ2.2).
- ▷ an overview of my developed methods [149, 154, 155] to improve the data quality in Chapter 6. This chapter only describes the approaches used and gives an overview of the results, which partially answers Research Question 3: “What are the implications of using proposals to

1.5. List of published papers

guide the annotation process for image classification?” (RQ3). A unified answer will be given in the next part in Chapter 7.

- ▷ a strategy about “how to handle ambiguous data” is provided in Section 6.5, which was proposed in [156]. This partially answers Research Question 4: “How should ambiguous data be annotated based on the results of RQ1, RQ2 and RQ3?” (RQ4), but the complete evaluation and discussion is given in the next part. It is important to emphasize that this strategy combines all the research efforts and thus can be considered as the main result of this dissertation.

Analysis & Discussion Part III presents a unified comparison of my obtained results in Chapter 7. The combination of the previously defined strategy and this evaluation allows answering Research Question 4: “How should ambiguous data be annotated based on the results of RQ1, RQ2 and RQ3?” (RQ4) including all subquestions. The raised research questions from Section 1.3 will be answered in Section 8.1. Furthermore, the limitations and future work will be discussed.

1.5 List of published papers

This section contains a list of my published or only peer-reviewed papers. The papers are divided into long, short and miscellaneous papers. All papers are peer-reviewed and have been presented at a conference or published in a journal. However, the short papers are often not published in a dedicated proceedings, but are published non-exclusively on the respective websites. The miscellaneous papers are first authorships or co-authorships of side projects with research related to this dissertation. Further unrelated papers are not described here. The focus lies on the long papers in the rest of the dissertation, but all are reprinted at the end of this work in Appendix A. An asterisk (*) indicates equal contribution of first authorship. The papers are sorted by date of publication per subsection. The citation and link to the corresponding sections are given where applicable.

1. Introduction

Long papers 2D and 3D Segmentation of Uncertain Local Collagen Fiber Orientations in SHG Microscopy, *41th German Conference on Pattern Recognition (DAGM GCPR 2019)* Lars Schmarje, Claudius Zelenka, Ulf Geisen, Claus-C. Glüer, Reinhard Koch, [148]

A Survey on Semi-, Self- and Unsupervised Learning for Image Classification, *IEEE Access*, vol. 9, pp. 82146-82168, 2021 Lars Schmarje, Monty Santarossa, Simon-Martin Schröder and Reinhard Koch, [151], see Section 3.1

Fuzzy Overclustering: Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy, *Sensors* 2021, 21(19), 6661 Lars Schmarje, Johannes Brünger, Monty Santarossa, Simon-Martin Schröder, Rainer Kiko, Reinhard Koch, [149], see Section 6.1

A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering, *Proceedings of the European Conference on Computer Vision (ECCV 2022)* Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, Claudius Zelenka, Rainer Kiko, Jenny Stracke, Nina Volkmann, Reinhard Koch, [154], see Section 6.2

Is one annotation enough? A data-centric image classification benchmark for noisy and ambiguous label estimation, *36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks* Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, Reinhard Koch, [152], see Section 5.1

ACCEPTED AT GCPR 2023, Label Smarter, Not Harder: CleverLabel for Faster Annotation of Ambiguous Image Classification with Higher Quality, 2023 Lars Schmarje, Vasco Grossmann, Tim Michels, Jakob Nazarenius, Monty Santarossa, Claudius Zelenka, Reinhard Koch, [155], see Section 5.4 and Section 6.3

UNDER REVIEW AT ICDE 2024, Annotating Ambiguous Images: General Annotation Strategy for Image Classification with Real-World Biomedical

1.5. List of published papers

Validation on Vertebral Fracture Diagnosis, 2023 Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Reinhard Koch, [156], see Section 5.3 and Section 6.5

Short papers Life is not black and white – Combining Semi-Supervised Learning with fuzzy labels, *Proceedings of LWDA 21: Lernen, Wissen, Daten, Analysen September 2021, Munich, Germany* Lars Schmarje and Reinhard Koch, [147]

A Data-Centric Image Classification Benchmark, *NeurIPS Data-Centric AI Workshop, 2021* Lars Schmarje, Yuan-Hong Liao and Reinhard Koch, [150]

Beyond hard labels: investigating data label distributions, *ICML 2022, Workshop DataPerf: Benchmarking Data for Data-Centric AI* Vasco Grossmann*, Lars Schmarje*, Reinhard Koch, [63]

Miscellaneous papers Artificial intelligence-driven hip fracture prediction based on pelvic radiographs exceeds performance of DXA: the Study of Osteoporotic Fractures(SOF), *Journal of bone and mineral research, Vol. 37, p193-193* Timo Damm, Lars Schmarje, Niklas Koser, Stefan Reinhold, Eren Yilmaz, Nicolai Krekieln, Li-Yung Lui, John T Schousboe, Steven R Cummings, Reinhard Koch, Claus-C Glueer, [37]

Opportunistic hip fracture risk prediction in Men from X-ray: Findings from the Osteoporosis in Men (MrOS) Study, *MICCAI 2022, PRIME Workshop* Lars Schmarje, Stefan Reinhold, Timo Damm, Eric Orwoll, Claus-C. Glüer, Reinhard Koch, [153]

Dulling while judging? Veränderte Beurteilung von Fotos zu Pickverletzungen bei Puten durch Wiederholungen, *Aktuelle Arbeiten zur artgemäßen Tierhaltung 2022, 24.-26.11. 2022, Freiburg/online: 54. Tagung Angewandte Ethologie bei Nutztieren der DVG, 24.-26.11. 2022, Freiburg/online, p. 282-284* Nina Volkmann, Lars Schmarje, Reinhard Koch, Nicole Kemper, [185]

Part I

Background

Data issues and their definitions

The data used is a central component of modern DL, and the literature provides several examples where the quality and quantity of these data directly affect the results [14, 82, 143, 174, 193]. In this and the next chapter (Chapter 3) common data problems and their possible solutions from the literature will be described. This chapter will define my understanding of ambiguity and then compare and contrast it with other possible definitions. Therefore, it answers RQ1.

2.1 Ambiguity

The term ambiguity is motivated by the examples in Figure 2.1 and Figure 2.2. In both examples, the left and right images are easily identifiable but the center image is not. In Figure 2.1, the image quality does not allow a clear distinction between the classes cat and dog. Peterson et al. showed



Figure 2.1. Ambiguity in dogs and cats, image source: [87]

2. Data issues and their definitions

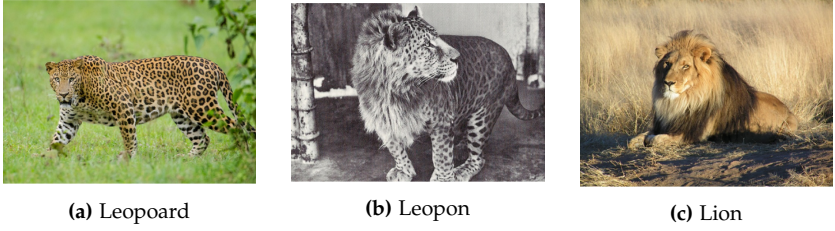


Figure 2.2. Ambiguity in the wild, image sources: leopard¹, lion², leopon³

that for this particular image (center), about half of the 50 participants assumed the class cat and the other half assumed dog [133]. In Figure 2.2, the distinction between a leopard (left) and a lion (right) is not possible for the center because it is a leopon a crossbreed between the two animals. In nature, there can often be intermediate stages between two different classes that do not fit exactly into the used taxonomy, for example, living plankton gradually degenerating into unidentifiable dirt in the ocean [158].

This mixture of classes occurs in many real-world datasets [20, 36, 44, 126, 143, 149, 157, 158, 181] and motivates our definition of ambiguous labels for a classification of image x with $K \in \mathbb{N}$ classes. This is the case because subjective biases drive the decisions of annotators in cases of uncertainty [160]. Let $L^x : S \rightarrow \{1, \dots, K\}$ be a random variable over the K classes for the image x , P_x the discrete probability measure over the finite and discrete label space S for the image x , and P the short version of P_x , then the induced probability distribution P_{L^x} or $P(L^x = \cdot)$ can be used to describe an ambiguous Ground-Truth (GT) label distribution of x . A major difference from other definitions of GT labels is that no hard-encoded probability distributions are enforced, where only one class of the K classes has a probability different from zero. Ambiguous data or images may have a mixture of hard-encoded and soft-encoded GT probabilities. Hard encoded GT means that there exists one class k with $P(L^x = k) = 1$ and for all other classes k' $P(L^x = k') = 0$. Soft encoded GT means that for all classes k $P(L^x = k) \in [0, 1]$.

¹<https://en.wikipedia.org/wiki/Leopard>

²<https://en.wikipedia.org/wiki/Lion>

³<https://www.boredpanda.com/strange-hybrid-animals-that-are-hard-to-believe-actually-exist/>

2.2 Other definitions

The literature describes this or similar topics using slight variations of our term ambiguity. This section uses other terms from the literature to show differences and similarities.

2.2.1 Labels & Annotations

The definition of ambiguity is based on the label of an image x and the goal is to improve the annotations for better quality of the labels. In most cases, the terms “label” and “annotation” are used synonymously, since labels can be considered as aggregated annotations. However, it is important to point out some minor differences especially when compared to the literature.

Considering an image classification problem with the GT probability distribution or label of P_{L^x} , this label or distribution can be an arbitrary discrete probability distribution over the K classes. As assumed in [152, 155] this distribution is generally unknown. This assumption is reasonable because in general no external information is available to calculate the label. For example, information lost due to downsampling cannot be recovered. If P_{L^x} is to be used as supervision to train a network with cross-entropy, an approximation of the label is needed. This approximation of the true GT distribution is also called P_{L^x} because one can never really work with the unknown probability, but only with its approximation. It is important to note that in this dissertation only the approximation is used, which means that different people may come up with a different GT for the same data.

This dissertation focuses on approximations based on human inputs which are called annotations. Other approaches would include automatically generating data using an algorithm such as pseudo-labeling [95] or raw data from sensors. Annotations $a \in \{0, 1\}^K$ are defined as hard-encoded estimates of the expected human label for the classification task. While this approach is common and used in many image classification benchmarks [32, 87, 88], it is not the only option. Soft annotations could also be used, meaning that an annotator could also quantify the uncertainty for each class, as in [111]. The drawback, however, is that humans already have a problem with ambiguity in images with respect to class. Adding

2. Data issues and their definitions

another dimension (the uncertainty per class) would further increase the ambiguity and complicate the task. The assumption is that with enough annotations, one could also filter or quantify this ambiguity with respect to the uncertainty per class, which would drastically increase the annotation cost. For this reason, this dissertation only considers the simpler and more readily available case of hard annotations.

The approximation of the label $P_{L^x} \in [0, 1]^K$ is the average $(\sum_{i=0}^N \frac{a_i}{N})$ of N hard-encoded annotations $a_1, \dots, a_N \in \{0, 1\}^K$ for each image. This approximation is motivated by the notion of collective intelligence where one annotation might be wrong, but the average of multiple annotations is a good approximation of the label [51]. An illustration of this approximation is given in Figure 2.3 and it also shows the convergence of the approximation of the label against a fixed distribution or vector. This does not prove that our approximation is the perfect solution, but that it can be reasonably assumed to be so. Another advantage of this approach is that the vector representing the label, which is used as input by the neural network has the same dimension as the commonly used hard labels. This means that the encoding within the vector may be different, but architectures or loss function definitions can be easily adapted since no dimensions need to be changed.

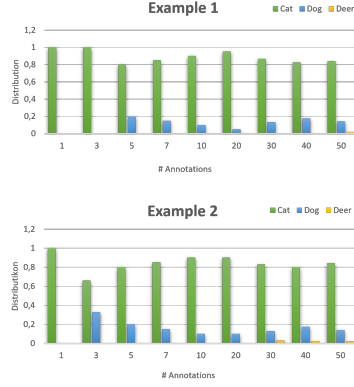
Another common aggregation for multiple annotations is majority vote where the majority class across all annotations is used as the label. However, this means that the resulting label is always an one hot-encoded label. The main advantage of this approach is that it can remove noise or annotation errors from the data. There is a clear difference between noise and the previously defined notion of ambiguity which will be discussed in Section 2.2.2. The use of hard labels is not well suited to capture this ambiguity, as argued in [152], and the argument will be repeated in Section 2.2.4. Davani et al. showed in [38] that majority voting is not well suited for cases where annotators systematically disagree or in other words where the data is ambiguous.

The use of multi-label classification as in [31, 138, 164] would also be possible. This means that multiple classes are correct instead of just one with a hard-coded label, which is of course appropriate for situations where multiple objects are present or where two answers are equally likely or correct, such as different names for the same object. However, the limi-

2.2. Other definitions



(a) source image



(b) example distributions

Figure 2.3. Approximation of a label for a random cat from the CIFAR10H dataset [133] - The right side displays two possible random simulations of the same example cat with varying numbers of annotations from one to 50, approximating the distribution. Since each annotation affects the overall distribution up to that point and we only know the final distribution, multiple distributions are possible during the annotation process, which ultimately converge to the known final distribution. It is evident that with more annotations, the resulting distributions become comparable since the effect of each additional random annotation diminishes with more total annotations. This tendency towards a stable distribution with increased annotations supports the use of the average across multiple annotations as our soft ground truth label distribution.

tation of this approach compared to soft labels is that it cannot represent different levels of agreement with different classes. For example, if annotators are asked to classify a large car, some might say it is a truck because they would call it a truck rather than a car. The agreement for different images would most likely vary. Using a probability distribution for each class would elevate this problem, as in [35, 111]. This would drastically increase the cost by requiring not just one, but K GT distributions, which is often simply not feasible.

2. Data issues and their definitions

2.2.2 Noisy & Ambiguity

Ambiguous could have been also called noisy instead, since the two words have similar meanings. However, many people understand under noise as human errors during the annotation process, at least based on the definitions in the literature [5, 6, 82, 92, 96, 122]. As defined above, ambiguity includes these errors as a reason, but it is not limited to this problem. Some tasks are subjective as illustrated in Figure 2.4. Methods to counteract these labeling errors often rely on robust loss functions [103, 121] that are inherently tolerant of noisy labels, separating noisy and clean data during training [96, 200, 201] or correcting the labels on the fly with better estimates [66, 95]. Moreover, these methods often assume a synthetic noise pattern such as balanced or skewed label flips [96, 102] during the evaluation. Gao et al. proposed a learning algorithm that specifically analyzes the individual labeling behavior of annotators to achieve more realistic patterns and also showed that human labeling errors are indeed not just random [52]. Wei et al. showed that realistic noisy data is different from commonly assumed noise patterns such as class assignment errors [193]. This can be explained by unrealistic assumptions for the synthetic noise but this work hypothesizes that the effect of the other source of ambiguity also impacts these results, leading to this difference.

2.2.3 Aleatoric Uncertainty & Epistemic Uncertainty

Our definition of ambiguity is also often confused or conflated with uncertainty. The literature mainly distinguishes between aleatoric and epistemic uncertainty [33, 163], while acknowledging that it is difficult to distinguish between them in DL [1, 83, 144, 178]. Aleatoric uncertainty is a statistical uncertainty that is inherent in the data and cannot be influenced by the model. If the uncertainty is homogeneous across the data space, it is called homoscedastic, and if it varies, it is called heteroscedastic. Epistemic uncertainty is a systemic uncertainty that is contained in the model itself, for example, because it does not have enough information. In addition, Malinin et al. in [109] introduced the term distributional uncertainty, which describes the uncertainty introduced in a model prediction by a shift in

2.2. Other definitions

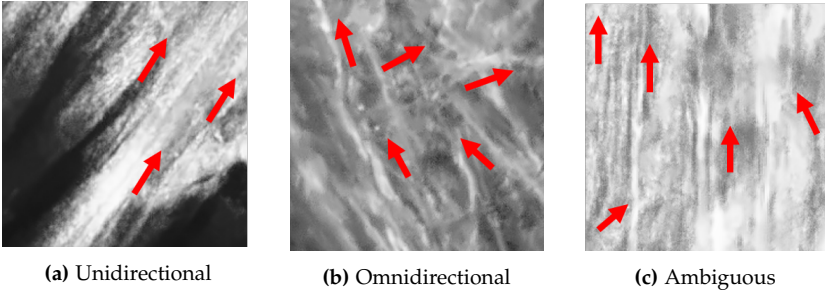


Figure 2.4. Three example images of collagen fibers in mice from [148] with fiber orientations indicated by red arrows – The left example has only fiber orientations in one direction. The middle image has multiple fibers in different directions. The right image has dominant orientation from bottom to top, but some fibers also go crosswise. The first two images should each be given the same label on an individual basis, regardless of who is annotating them. However, some annotators might label them incorrectly if they are careless. This would result in label noise, because a correct class could have been chosen with more time or care. The right image shows a mixture of straight and intertwined collagen fibers. This case is inherently ambiguous because it is not clear what the class orientation is. It depends on the annotator’s interpretation which label is chosen for the image.

the data. For example, when new classes are introduced, the distribution between classes changes or the characteristics of classes shift.

All of these definitions can be explained in a unified example by trying to guess the roll of a dice. If you know that a perfect six-sided dice is being used, then all answers from one to six are equally likely. In this case, there is only homoscedastic aleatoric uncertainty, because the uncertainty is only outside of a model for estimating it and is the same for all possibilities. With a manipulated dice that more often shows a particular side, heteroscedastic aleatoric uncertainty is described because the uncertainty is still generated only from the data, but is now biased toward one outcome. If the model does not know what type of dice (number of sides) is being used, then there is a form of epistemic uncertainty because the model cannot know whether the outcome 10 is a valid outcome or not. Distributional uncertainty can be represented by changing the dice

2. Data issues and their definitions

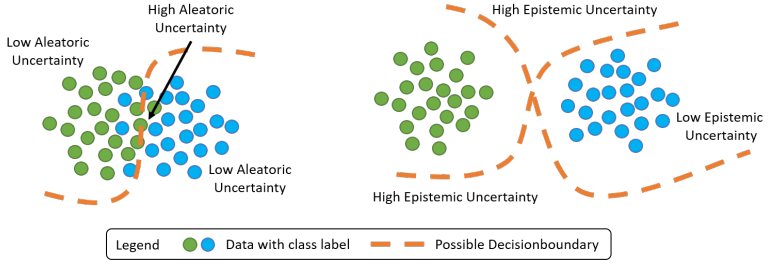


Figure 2.5. Visualization of aleatoric and epistemic uncertainty inspired by [53]

during the game. According to these definitions, ambiguity is a form of heteroscedastic aleatoric uncertainty.

Uncertainty is a keyword in many publications but they often refer to the uncertainty of a model [50, 64, 75, 93, 134]. In these cases, the authors often do not discuss whether they mean aleatoric or epistemic but state that they are considering model uncertainty or predictive uncertainty. Gawlikowski et al. state in [53] that predictive uncertainty is often the combination of aleatoric, epistemic and sometimes distributional uncertainty for Out of distribution (OOD) methods. Based on the description of mainly epistemic uncertainty effects, it can be assumed that in most cases aleatoric uncertainty is not considered.

For a quantification of this assumption across 17 papers, I looked at the mentioned uncertainty papers on the GitHub page⁴ of the uncertainty baselines benchmark [120]. This benchmark aims to provide a unified testbed and evaluation for uncertainty methods. 15 out of 17 papers did not mention any aleatoric uncertainty but looked at epistemic uncertainty [67, 91, 110, 129, 135], calibration measure for model uncertainty [3, 68, 119, 123], OOD detection [101, 139, 187, 195, 196] or Bayesian learning [49]. Padhy et al. state that “[e]stimating the epistemic uncertainty is especially valuable”⁵ and “the field has dealt with three different paradigms of measuring predictive uncertainty, namely (1) in-distribution calibration of models measured on an i.i.d. test set, (2) robustness under dataset shift, and

⁴<https://github.com/google/uncertainty-baselines>

⁵quoted from [129] p.1

2.2. Other definitions

(3) *anomaly/out-of-distribution (OOD) detection*".⁶ Based on this summary, which was repeated in various versions or subsets in the mentioned papers, it can be taken as evidence for the claim that most authors mainly think about epistemic uncertainty when they talk about model uncertainty or predictive uncertainty.

The remaining two publications are [84, 176]. Kivlichan et al. agree that *"a common approach to evaluate a model's uncertainty quality is to measure its calibration performance"*.⁷ However, they also state that *"data uncertainty is inherent to the data generating process and is irreducible"*⁸ emphasizing also other sources of uncertainty are also relevant. The most diverse definition can be found by Tran et al. who describe their method Plex as a reliable method under many circumstances [176]. One of the three major circumstances is uncertainty, which is subdivided into calibration, selective prediction, open set recognition and label uncertainty. The first three topics are fully or partially described in the other 15 papers mentioned above. The last topic of label uncertainty describes the above defined label ambiguity, as they define *"[l]abel uncertainty can arise when, for example, human raters disagree about the label for an ambiguous input"*.⁹ It is important to note, that this small analysis of the papers mentioned in uncertainty baselines is not representative. It seems to be a good approximation of the diverse definitions about uncertainty in the literature and is consistent with our personal, less quantifiable perspective on the topic. Research such as [205] talks about this issue but does not call it aleatoric uncertainty as described in the next section on soft labels.

2.2.4 Soft Labels & Label Smoothing

Soft labels are used to represent the ambiguity in the GT distribution. This section discusses what the impact of soft labels is in the literature and contrasts it with the popular method of label smoothing.

The literature and my own research give several examples why one annotation is not enough to capture the ambiguity of the data [15, 38,

⁶quoted from [129] p.1

⁷quoted from [84] p.3

⁸quoted from [84] p.3

⁹quoted from [176] p.7

2. Data issues and their definitions

63, 152, 155, 177]. Thus, multiple annotations per image are needed. In many cases, an aggregation over multiple annotations is used as in our work [63, 152, 155]. However, other approaches have also been proposed. Wei et al. investigated the use of multiple separate labels instead of a majority vote aggregation and found that it could “*be more beneficial than label aggregation when the noise rates are high or the number of labelers is insufficient*”.¹⁰ Collins et al. [35] used multiple soft annotations instead of many hard annotations to estimate an aggregated soft-label as discussed in Section 2.2.1. This dramatically increases the annotation cost and introduces another uncertainty. Zhou et al. estimated the soft labels based on k-Means clustering for a poor quality image to avoid human re-annotation [205]. This approach has the drawback that the labels are not human verified and therefore may be incorrect or ill-defined, which is unacceptable in critical systems such as autonomous driving or healthcare. Bagherinezhad et al. showed that it can be beneficial to use soft labels in combination with hard labels for self-supervised image classification [9]. They acknowledge that many self-supervised learning methods rely on the assumption that the random crops of an image have the same semantic class as the original image which may be wrong for zoomed in regions without the original main object, such as a person holding a small ball. Moreover, they found that some cropped regions, for example of a “dough” and “butternut squash” share a similar texture and are even indistinguishable by humans anymore as shown in Figure 2.6.

A similar but distinct idea of using soft labels is called label smoothing. Muller et al. summarized label smoothing as the process of using a blended version of the hard-encoded GT distribution $P(L^x = k)$ for class k with a uniform distribution over all classes K with the blending factor α : $P(L^x = k)(1 - \alpha) + \alpha/K$ [116]. For $\alpha = 0$, there are no changes applied. For $\alpha = 1$, the uniform probability distribution of $1/K$ is preferred over all previous distribution information of $P(L^x = k)$. This approach is successfully applied in many DL approaches to improve the overall performance [89, 182], to mitigate annotation errors [107] or as a regularization [106]. The literature is divided on when or why label smoothing helps [116, 192]. What they have in common is that they seem to mitigate

¹⁰quoted from [194] p.12

2.2. Other definitions

some problem in the data. But this problem is modeled only over the whole dataset or in some cases per class [106] and not per image as in this dissertation. The connections between the success of label smoothing and the use of soft labels for ambiguity share many core ideas. In this work, only per-sample soft labels are investigated, as they are a generalization of the techniques discussed above.

2. Data issues and their definitions



(a) butternut squash



(b) cutout



(c) dough



(d) cutout

Figure 2.6. Examples of butternut squash and dough – Both look different from far away but share the same texture in a cutout region. Image sources: dough¹¹, butternut squash¹²

¹¹[https://www.cookipedia.co.uk/recipes_wiki/Pizza_dough_\(TM\)](https://www.cookipedia.co.uk/recipes_wiki/Pizza_dough_(TM))

¹²<https://www.flickr.com/photos/farmanac/5827322918>

Quantity & Quality of Data

This dissertation aims to enhance the data for image classification. However, the term “improving” may refer to either the quantity or quality of the data. Therefore, this section examines the latest developments in the areas of quality and quantity of labeled data. Each section provides a brief overview of the covered topics and research directions that go beyond the scope of this dissertation. The dissertation focuses on improving quality as the majority of research in semi-supervised learning pertains to quantity.

Quantity Quantity refers to the amount of labeled and/or unlabeled data available for the task. The amount of unlabeled data is often not under the control of the researcher, but is determined by the circumstances. Recent research looks at generative models as a data source [79] or even just learning from visual noise patterns [12] to fill the gap when not enough unlabeled data is available. Another approach could be domain adaptation [188, 189] from a larger known dataset to the target domain. This dissertation will focus on the amount of labeled data required.

The impact of the required labeled data can be adjusted by requiring less labeled data or by increasing the labeled data by making it more cost effective. Research in self-supervised, semi-supervised and unsupervised learning can reduce the required amount of labeled data by up to a factor of 100 [165, 174]. An extension of our survey [151] on this topic is given in Section 3.1.

Another common approach to make the annotation process more effective is to select the samples to be annotated using a sophisticated strategy. These approaches can be summarized as active learning [118, 140, 168]. Active learning can be described as a heuristic for selecting the most relevant images to annotate next. It can also be used to iteratively improve

3. Quantity & Quality of Data

test performance [86]. This concept is often complementary to other semi-supervised algorithms, since there is little overlap between improving a model and the data selection process. While active learning can also be used to improve data quality, for example, by diversifying the images used, it is often seen as an interactive approach to reducing the amount of labeled data. Tifrea et al. even point out that active learning based on entropy may not be suitable for settings with a low number of samples and/or a high mixture of classes [173]. In addition, active learning can most likely be used in combination with the data enhancement strategies investigated in this dissertation, which simply select samples at random. For all these reasons, this topic is beyond the scope of the rest of this work.

Another frequently used term for working with limited labeled data is weakly supervised learning. According to [206], weak supervision can stem from three reasons: incomplete, inexact, or inaccurate supervision. Incomplete supervision refers to the issue of having unlabeled data in addition to labeled data which has been discussed in greater detail above in relation to semi-supervised learning and active learning. Inexact supervision implies that the labels are insufficiently specified. In the case of images, inexact labels may consist of cropped image regions [30] or more general hierarchical class structures [167]. Given that inexact labels can be viewed as supplementary constraints, this dissertation will not delve into the details of inexact labels. Inaccurate supervision, which amounts to incorrect labeling, is regarded as a quality issue in this dissertation and will therefore be examined further below.

Quality In this dissertation, data quality refers to the quality of the associated label for labeled data. It does not consider image quality, such as resolution or artifacts, data selection or general issues like long-tailed distributions [4]. All of these affect the quality and thus the results, but the focus in this work is on the labels. Often humans are used to provide the label on the same modality that is shown to the neural network during training. Only in rare cases is an external GT available, such as the fracture of a bone in a retro-perspective [153] or from another domain [146]. Thus, the common case is that humans provide the annotations to generate the label. It is known that these annotations can vary between annotators (inter) or over time for the same annotator (intra), leading to inter- or

3.1. Semi-, Self- and Unsupervised Learning

intra-operator variability. This kind of variability/uncertainty/ambiguity can arise from multiple sources. Humans make mistakes, especially over time when they are less focused. The task definition may be ambiguous or poorly defined, leaving subjective interpretations in corner cases, such as what to do if two items are present but only one can be classified [98]. The task may be subjective by nature, such as rating emotions for different people [111]. The image quality is poor, e.g. blurred, which makes a simple task difficult and does not allow a clear distinction between classes [20, 133, 149]. As shown in Figure 2.2, there may be intermediate states that belong to a combination of classes.

A lot of research is done, for example, in the field of consensus and annotator training, to counteract this variability [2, 10, 132] by solving this problem on the human side. Chang et al. propose to combine different tasks such as classification, description, and revoting to identify and better label ambiguity [26]. This dissertation focuses only on proposal-driven annotation. Proposal-driven annotation means that annotators have access to a class proposal during the annotation process, which should guide their decision. A review of the literature is given in Section 3.2.

3.1 Semi-, Self- and Unsupervised Learning

This section is an extension of my survey [151] of the year 2021, and will closely follow the structure of that work. Thus, all abbreviations¹ are used in this dissertation without further explanation. Please refer to the initial survey in Section A.1.2 for comprehensive explanations. The structure of the survey is shown in Figure 3.1. It can be roughly explained as methods can use several common ideas and are grouped based on the amount of labeled data and the time step at which they are used. These groups are called training strategies and are described in more detail in Figure 3.2. The survey aimed to be as robust as possible for future developments in the field, but noted that *“we know that these ideas need to be extended in the future as new common ideas will arise, old ones will disappear, and focus will*

¹e.g. Cross-Entropy (CE), Cross-Entropy (used not as common supervised loss) (CE*), Clustering Loss (CL), Entropy Minimization (EM), Kullback-Leibler divergence (KL), Mutual Information (MI), Mean Squared Error (MSE), Mix up (MU), Overclustering (OC), Pseudo-Labeling (PL), Pretext Task (PT), Virtual Adversarial Training (VAT)

3. Quantity & Quality of Data

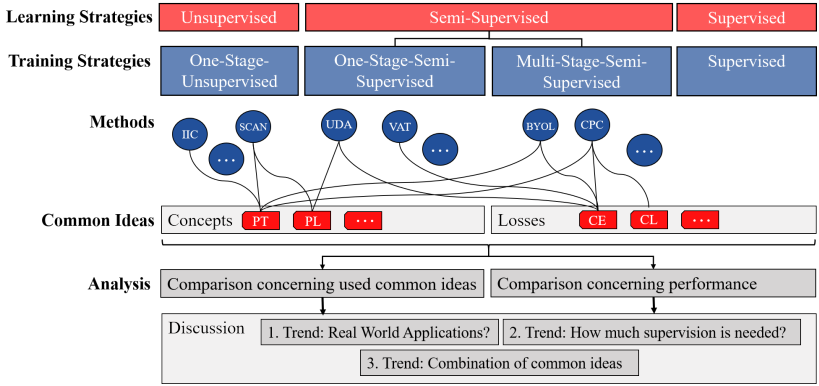


Figure 3.1. Overview of the structure of my survey [151] – The learning strategies unsupervised, semi-supervised, and supervised are commonly used in the literature. Since semi-supervised learning includes many methods, we have defined training strategies that subdivide semi-supervised learning. For details about the training strategies, see Figure 3.2. Each method belongs to a training strategy and uses several common ideas. A common idea can be a concept like a pretext task or a loss like cross entropy. Additional methods and common ideas are given in Section 3.1.1 and Section 3.1.2. The methods are compared to each other in Section 3.1.3 in terms of the common ideas they use and their performance with respect to the previously identified trends. Image and caption adapted from [151]

*shift to other ideas. In contrast to detailed taxonomies, these new ideas can easily be integrated as new tags”.*² Following the argument, new common ideas (Section 3.1.1) and methods (Section 3.1.2) with regard to the old elements from the original survey [151] are added in this section.

The added ideas and methods are sorted by name in descending order. The results tables from the survey are updated in Section 3.1.3 and it is checked if the originally defined trends are still valid, if they need to be updated or if completely new ones emerge.

3.1. Semi-, Self- and Unsupervised Learning










Learning Strategy	Σ	I	II	Training Strategy
Supervised			-	Supervised
Unsupervised			-	One-Stage-Unsupervised
Semi-Supervised			-	One-Stage-Semi-Supervised
Semi-Supervised				Multi-Stage-Semi-Supervised

Figure 3.2. Illustration of the different training strategies – Each row represents a different combination of data usage during the first and second stages of the training. The first column gives the name of the learning strategy commonly used in the literature for that use, while the last column gives the name of the training strategy used in this study. The second column represents the total amount of data used. The third and fourth columns represent the data used in the first or second stage. The blue and gray (half) circles represent the usage of the labeled data X_l and the unlabeled data X_u in each stage or in total. A minus sign means that no further stage is used. The dashed half-circle in the last row indicates that this dashed part of the data can be used, image and caption adapted from [151]

3.1.1 Extended common ideas

Masked Autoencoder (MAE)

He et al. proposed MAE, a self-supervised pre-training that is well suited for Vision Transformer (ViT) in [71]. The idea is to encode batches of images in an arbitrary feature space with a high masking rate, e.g. 75%. A decoder uses the mask tokens and the generated coding to reconstruct the full image without masking. This asymmetric design leads to a significant speedup in training, since the more complicated encoder only has to look at a small portion of the full image. In general, masking is done by random sampling, but other approaches have been discussed in this and other

²quoted from [151] p.6

3. Quantity & Quality of Data

works [57]. This type of pretext task found a wide adaptation in other methods such as [21, 57]. For this reason, it is described as its own general idea.

Vision Transformer (ViT)

Vision transformers are, as the name implies, transformers for Computer Vision (CV) on images. Transformers were originally proposed by Vaswani et al. for Natural Language Processing (NLP) [182]. They replaced commonly used recurrent neural networks such as LSTM [74] because they solved problems such as vanishing gradients and bottleneck information. Transformers accomplish this by applying multi attention heads to the entire sentence at each prediction step. Dosovitskiy et al. proposed how transformers which are designed to work with sequential data can be applied to a grid data such as an image [46]. The key idea is to split the image into several patches, flatten the input, assign positions and then treat the image as sequential data. This data is fed into a transformer encoder and an additional multi-layer perceptron is used to make the final class prediction. With enough data, ViT can perform state-of-the-art classifications. However, in general, several millions of images and huge amounts of computing power are required to train the transformer encoder. After this initial encoder training, training or fine-tuning the multi-layer perceptron is relatively easy and less costly. Unless otherwise noted, all other models discussed in this extension of [151] are based on Convolutional Neural Networks such as ResNet50 [69].

Vision Language Model (VLM)

Some recent successes in computer vision with few or no labels have been achieved using Vision Language Model (VLM) trained on millions of image-text pairs. These models show a high degree of generalizability to different tasks and scale with larger models and datasets. They are particularly useful for zero-shot or unsupervised learning because natural language can more easily describe the classes and the classification task. However, Liu et al. showed that this general knowledge may not be sufficient for specialized applications [100].

3.1. Semi-, Self- and Unsupervised Learning

3.1.2 Extended methods

Contrastive Language-Image Pretraining, CLIP

Radford et al. point out that common supervision is restricted to certain classes and that general image-text pairs allow for a broader scope during training [137]. Thus, they only train for the correct assignment of captions to images. For this purpose, they use a text and an image encoder to embed the images and texts in a feature space. CLIP is trained with a contrastive loss, which is a symmetric CE loss variant as in [179], on the pairwise cosine similarities between the image and text embeddings. CLIP can be used as a zero-shot transfer for image classification by using a text embedding of class descriptions such as “photo of {class}” and comparing it to the predicted image features with cosine similarity. For this reason, CLIP can be considered a one-stage unsupervised method, and with the addition of a linear classification head, a multistage semi-supervised method.

Common ideas: CE, CL, (MI),PT, ViT, VLM

DivideMix

Li et al. expand upon the MixMatch method [18] by incorporating Gaussian Mixture Models to divide the given data into labeled and unlabeled parts for two parallel models [96]. The primary objective is to ascertain during training which elements are noisy and therefore, their labels should be overlooked in the unlabeled data section. One major distinction compared to other semi-supervised approaches like MixMatch is the use of pre-trained models on ImageNet rather than randomly initialized networks. In general, this approach closely resembles MixMatch in terms of the used common ideas and is one-stage semi-supervised. *Common ideas: CE, (EM), MSE, MU, PL*

Self-distillation with no label (DINO)

Caron et al. investigated how self-supervised learning can be applied to vision transformers [24]. DINO uses a teacher-student approach where the teacher is an exponential moving average of the student. The difference

3. Quantity & Quality of Data

between the student and the teacher is minimized by CE. An important step to avoid mode collapse is the newly proposed centering, which encourages convergence to the uniform distribution. This effect can be balanced by the applied sharpening of the softmax output, which is an inherent minimization of entropy. They find that SSL works well with ViT, but also that the trained self-supervised attention heads can be used for semantic segmentation. If the ViT is trained supervised or if convolutional neural networks are used, this property is lost or the performance is degraded. The network is fine-tuned at the end of the desired target dataset and thus this method is a multi-stage semi-supervised method.

Common ideas: CE, (EM) PT, ViT

FreeMatch

Wang et al. argue that semi-supervised methods such as FixMatch [165] make suboptimal use of the unlabeled data because they use predefined thresholds [191]. FreeMatch uses an adaptive threshold on a global and local scale, which leads to better data utilization especially at the beginning of training. In the early stages, the model confidences are not sharp, and thus only adaptive thresholds can determine meaningful images for training. The global adaptive threshold is an exponential moving average of the confidences over the entire unlabeled dataset. The local adaptive threshold is similar but defined per class and thus allows to “*modulate the global threshold in a class-specific fashion to account for the intra-class diversity and the possible class adjacency*”.³ To stabilize the training, CE is used to keep the marginal distribution of all predictions in the desired range. Training is done in a single stage, making it a one-stage semi-supervised method.

Common ideas: CE, CE, PL*

Masked Autoencoders (MAE)

Masked Autoencoders are not only a common idea but also a method presented in [71]. It is a Multi-Stage-Semi-Supervised method, because

³quoted from [191] p.5

3.1. Semi-, Self- and Unsupervised Learning

after the pretext task where the loss is reduced with MSE, the ViT is fine-tuned on a target dataset.

Common ideas: CE, MAE, MSE, PT, ViT

Noisy Student

Xie et al. use the established student-teacher approach, but they alternate their roles [198]. First, a teacher is trained on the labeled data. A student is trained on the teacher's predictions, but with injected noise in the form of data augmentation and dropout, for example. This should ensure that the student generalizes better than the teacher and thus the student becomes the new teacher for the next training iteration. It is important to note that they used 300 million additional images as unsupervised data during this process, but is trained in one stage. The process might be iterative but is still considered one stage with regard to the definition in [151].

Common ideas: CE, PL

Masked Auto-Encoding approach to train an Omnivorous visual encoder (OmniMAE)

Girdhar et al. noticed that MAE are used for image and video vision tasks in different papers, but they are trained only for single tasks and not in unison [57]. They proposed to simplify the network architecture to learn representations that are suitable for multiple computer vision tasks such as image and video classification. They show that using image and video data together during the self-supervised MAE allows to increase the masking ratio to 90 or 95%, which further improves the training time. The ViT is finally fine-tuned to the target domain, which makes it a multistage semi-supervised method.

Common ideas: CE, MAE, MSE, PT, ViT

Retrieval-Augmented Customization (REACT)

Liu et al. found that high generalizability is achieved with Vision Language Model (VLM) by using more and more data for pretraining [100]. They suggest adjusting the data used during fine-tuning. The idea is that a human would also filter for relevant information before learning. REACT

3. Quantity & Quality of Data

can be used without any further labeled data by using the class names as queries to the VLM and using these images as labeled data for fine-tuning a classification head, thus it is a multi-stage semi-supervised method. It is debatable whether this form of supervision (text-image pairs) on the Web should be considered unlabeled data or weakly labeled data. This process could also be replaced or supplemented with image label pairs from the target dataset. They can achieve acceptable results for many datasets, but achieve inferior results for highly domain-specific datasets such as the cancer cell regonition benchmark. Liu et al. attribute this problem to the fact that the images retrieved for model adaptation are not similar enough to the target domain dataset, and thus the adaptation fails.

Common ideas: CE, CL, (MI), PT, PL, ViT, VLM

Representation Learning via Invariant Causal Mechanisms (RELICv2)

Tomasev et al. extend the previously introduced method RELIC [113] to real-world classification [174]. They argue that self-supervised learning may learn based on spurious features such as background or unrelated items, which is undesirable. RELIC uses instance classification via contrastive learning as in [27] and invariant prediction by penalizing with KL the difference between the predictions of two different augmentations of the same image. Tomasev et al. extend this by adding saliency masking and more and different sizes for the augmented views during training. The saliency masking is a form of pseudo-labeling and is used to randomly remove the background of each image with a small probability. This motivates the learning of more robust features. The network is then fine-tuned on a target dataset, making it a multi-stage semi-supervised method. They claim to be the first to consistently beat the supervised baseline, regardless of the ResNet architecture.

Semi-Supervised Vision Transformers (Semi-ViT)

Cai et al. propose a three-stage semi-supervised training protocol with ViT [21]. The first stage is self-supervised pre-training with MAE, the second is supervised fine-tuning on the labeled dataset, and the third stage is

3.1. Semi-, Self- and Unsupervised Learning

semi-supervised fine-tuning using additional unlabeled data. Overall, this makes it a Multi-Stage-Semi-Supervised method. The last stage is inspired by the FixMatch [165] training protocol, but allows an exponential moving update of the student to the teacher weights. They also propose Probabilistic Pseudo Mixup, which applies MixUp to labeled and unlabeled data. For unlabeled data, the pseudo-labels are used and the mixup ratio for blending the images depends on the confidence about these pseudo-labels. *Common ideas: CE, CE*, PL, ViT, MSE, PT, MAE, MU*

3.1.3 Extended comparison

The extended summary tables from the survey are given and discussed in this section. The extended comparison of the relationship between the added methods and the common ideas is given in Table 3.1. The reported results are given for single-stage methods in Table 3.2 and for multi-stage methods in Table 3.3. Table 3.4 reports additional results with even less supervision than the default values defined in [151].

The three previously identified trends are discussed in relation to these new findings in their respective subsections below.

Trend 1: Real World Applications?

The survey found that “*most methods are not scalable to high-resolution and complex image classification problems*”,⁴ but also that some methods like FixMatch were suitable for this task. This trend is clearly continuing and even increasing. Most papers no longer report results on simpler datasets like CIFAR-10 or STL-10, but only on the ImageNet dataset. For example, MAE can achieve about 87% ACC on Imagenet when a linear evaluation head is trained on the full dataset, and REACT can even achieve about 79% without using any labels from the official ImageNet dataset, which is better than FixMatch with 10% of the labels at about 72%.

⁴quoted from [151] p.19

3. Quantity & Quality of Data

Table 3.1. Overview of methods and common ideas used - On the left, the reviewed methods are sorted by training strategy. The top row lists the common ideas. The last column and some rows summarize the use of ideas per method or training strategy. *Legend:* (X) The idea is used only indirectly. Table adapted from [151]. Newly added methods and ideas are separated from the original data.

	CE	CE*	EM	CL	KL	MSE	MU	MU	OC	PT	PL	VAT	MAE	VIT	VLM	Overall Sum
One-Stage-Semi-Supervised																
Pseudo-Label [95]	X	X									X					3
π model [92]	X					X										2
Temporal Ensembling [92]	X					X										2
Mean Teacher [171]	X					X										2
VAT [114]	X											X				2
VAT + EntMin [114]	X		X									X				3
ICT [183]	X					X	X				X					4
fast-SWA [7]	X					X										2
MixMatch [18]	X		(X)			X	X				X					5
EnAET [190]	X		(X)		X	X	X			AET	X					7
UDA [197]	X		X		X						(X)					4
SPamCO [108]	X	X				X										4
ReMixMatch [17]	X	X	(X)				X	(X)		Rotation	X					7
FixMatch [165]	X	X	(X)								X					4
DivideMix [96]	X		(X)			X	X				X					3
FreeMatch [191]	X	X									X					3
NoisyStudent [198]	X										X					2
Sum	14+3	4+1	6+1	0	2	8+1	4+1	1	0	2	8+3	2	0	0	0	47+10
Multi-Stage-Semi-Supervised																
Exemplar [45]	X	X								Augmentation						3
Context [43]	X	X								Context						3
Jigsaw [124]	X	X								Jigsaw						3
DeepCluster [22]	X	X							X	Clustering	X					5
Rotation [56]	X	X								Rotation						3
CPC [72, 179]	X	(X)		X				(X)		CL						5
CMC [172]	X	(X)		X				(X)		CL						5
DIM [73]	X							X		MI						3
AMDIM [8]	X							X		MI						3
DMT [99]	X	X								Metric	X					4
IIC [80]	X									MI						4
S ⁴ [203]	X	X	X					X	X	Rotation	X	X				6
SimCLR [27]	X	(X)								CL						3
MoCo [70]	X			X						Metric						3
BYOL [62]	X					X				Bootstrap						3
FOC [149]	X	(X)						X	X	MI						5
SimCLRv2 [28]	X	(X)		X						CL	X					5
CLIP [137]	X			X				(X)		CL				X	X	6
DINO [24]	X		(X)							CE				X		4
MAE [71]	X					X				MAE			X	X		5
OminiMAE [57]	X					X				MAE			X	X		5
REACT [100]	X			X				(X)		CL	X			X	X	7
RelicV2 [174]	X	X		X	X					CL	X					6
Semi-ViT [21]	X	X				(X?)	X			MAE	X		X	X		8
Sum	17+7	11+2	1+1	4+3	0+1	1+3	0+1	6+2	3	17+7	4+3	1	0+3	0+6	0+2	65+41
One-Stage-Unsupervised																
DAC [25]											X					1
IMSAT [76]								X				X				2
IIC [80]								X	X	MI						3
FOC [149]								X	X	MI						3
SCAN [180]								X		CL	X					3
CLIP [137]				X				(X)		CL				X	X	5
REACT [100]				X				(X)		CL	X			X	X	6
Sum	0	0	0	0+2	0	0	0	3+2	3	3+2	2+1	1	0	0+2	0+2	12+11
Overall Sum	31+10	54+3	7+1	4+4	2+1	9+4	4+1	10+4	6	22+9	14+7	4	0+3	0+8	0+4	124+62

3.1. Semi-, Self- and Unsupervised Learning

Table 3.2. Overview of the reported accuracies for one-stage methods - The first column indicates the method used. For the supervised baseline, we used the best reported results that were considered as baselines in the referenced papers. The original paper is given in parentheses after the result. The architecture is given in the second column. The last four columns give the top-1 accuracy score in % for the respective dataset. *Legend:* [†] 100% of the labels are used instead of the default value defined in [151]. [‡] A multilayer perceptron is used for fine tuning instead of a fully connected layer. Notes on special architectures and scoring: ¹ Architecture includes shake-shake regularization. ² Network uses wider hidden layers. ³ Method uses ten random classes out of default 1000 classes. ⁴ Network predicts only 20 superclasses instead of the default 100 classes. ⁵ Inputs are pre-trained ImageNet features. ⁶ The method uses different copies of the net for each input. ⁷ The network uses selective kernels [97]. ⁸ Uses an additional 300 million unlabeled images. ⁹ Uses a vision language model to define the desired classes. Table adapted from [151].

	Architecture	Publication	CIFAR-10	CIFAR-100	STL-10	ILSVRC-2012	ILSVRC-2012 (Top-5)
Supervised (100% labels)	Best reported	-	98.01[190]	79.82[8]	68.7 [73]	85.7 [175]	97.6 [175]
One-Stage-Semi-Supervised							
Pseudo-Label [95]	ResNet50v2 [69]	2013					82.41 [203]
π model [92]	CONV-13	2017	87.64				
Temporal Ensembling [92]	CONV-13	2017	87.84				
Mean Teacher [171]	CONV-13	2017	87.69				
Mean Teacher [171]	Wide ResNet-28	2017	89.64				90.9[28]
VAT [114]	CONV-13	2018	88.64				
VAT [114]	ResNet50v2	2018					82.78 [203]
VAT + EntMin [114]	CONV-13	2018	89.45				
VAT + EntMin [114]	ResNet50v2	2018	86.41 [203]				83.3 [203]
ICT [183]	Wide ResNet-28	2019	92.34				
ICT [183]	CONV-13	2019	92.71				
fast-SWA [7]	CONV-13	2019	90.95	66.38			
fast-SWA [7]	ResNet-26 ¹	2019	93.72				
MixMatch [18]	Wide ResNet-28	2019	95.05	74.12	94.41		
EnAET [190]	Wide ResNet-28	2019	94.65	73.07	95.48		
UDA [197]	Wide ResNet-28	2019	94.7			68.66	88.52
SPamCo [108]	Wide ResNet-28	2020	92.95				
ReMixMatch [17]	Wide ResNet-28	2020	94.86	76.97[165]			
FixMatch [165]	Wide ResNet-28	2020	95.74	77.40			
FixMatch [165]	ResNet-50	2020				71.46	89.13
Noisy Student [198]	EfficientNet-L2 [169]	2020				88.4 ^{1,8}	98.7 ^{1,8}
One-Stage-Unsupervised							
DAC [25]	All-ConvNet	2017	52.18	23.75	46.99	52.72 ³	
IMSAT [76]	Autoencoder ⁵	2017	45.6	27.5	94.1		
IIC [80]	ResNet34	2019	61.7	25.7 ⁴	59.6		
FOC [149]	ResNet34	2020			60.45		
SCAN [180]	ResNet18	2020	88.3	50.7 ⁴	80.9		
CLIP [137]	ViT-L	2021				75.3 ⁹ [100]	
REACT [100]	ViT-L	2023				78.5 ⁹	

3. Quantity & Quality of Data

Table 3.3. Overview of the reported accuracies for multi-stage methods - The first column indicates the method used. For the supervised baseline, we used the best reported results that were considered as baselines in the referenced papers. The original paper is given in parentheses after the result. The architecture is given in the second column. The last four columns give the top-1 accuracy score in % for the respective dataset. *Legend:* [†] 100% of the labels are used instead of the default value defined in [151]. [‡] A multilayer perceptron is used for fine tuning instead of a fully connected layer. Notes on special architectures and scoring: ¹ Architecture includes shake-shake regularization. ² Network uses wider hidden layers. ³ Method uses ten random classes out of default 1000 classes. ⁴ Network predicts only 20 superclasses instead of the default 100 classes. ⁵ Inputs are pre-trained ImageNet features. ⁶ The method uses different copies of the net for each input. ⁷ The network uses selective kernels [97]. ⁸ Uses an additional 300 million unlabeled images. ⁹ Uses a vision language model to define the desired classes. Table adapted from [151].

	Architecture	Publication	CIFAR-10	CIFAR-100	STL-10	ILSVRC-2012	ILSVRC-2012 (Top-5)
Supervised (100% labels)	Best reported	-	98.01[190]	79.82[8]	68.7 [73]	85.7 [175]	97.6 [175]
Multi-Stage-Semi-Supervised							
Exemplar [45]	ResNet50	2015				46.0 [†] [85]	81.01 [203]
Context [43]	ResNet50	2015				51.4 [†] [85]	
Jigsaw [124]	AlexNet	2016				44.6 [†] [85]	
DeepCluster [22]	AlexNet	2018			73.4 [80]	41 [†]	
Rotation [56]	AlexNet	2018				55.4 [†] [85]	
Rotation [56]	ResNet50v2	2018					78.53 [203]
CPC [72]	ResNet-170	2020	77.45 [†] [73]		77.81 [†] [73]	61.0	84.88
CMC [172]	AlexNet	2019			86.88 [†]		
CMC [172]	ResNet-50 ⁶	2019				70.6	89.7*
DIM [73]	AlexNet	2019			72.57 [†]		
DIM [73]	GAN Discriminator	2019	75.21 ^{††}	49.74 ^{††}			
AMDIM [8]	ResNet18	2019	91.3 [†] / 93.6 ^{††}	70.2 [†] / 73.8 ^{††}	93.6 / 93.8 [†]	60.2 [†] / 60.9 ^{††}	
DMT [99]	Wide ResNet-28	2019	88.70				
IIC [80]	ResNet34	2019			85.76 [149] / 88.8 [†]		
S ⁴ L [203]	ResNet50v2 ²	2019				73.21	91.23*
SimCLR [27]	ResNet50v2 ²	2020				74.4 [28] / 76.5 [†]	92.6 / 93.2 [†]
MOCO [70]	ResNet50v2 ²	2020				68.6 [†]	
MOCO [70]	ResNet50v2	2020				60.6 [†] / 71.1 ^{††} [29]	
MOCO [70]	ViT-S	2020				72.7 [†] [24]	
BYOL [62]	ResNet200 ²	2020				77.7	93.7
BYOL [62]	ResNet200 ²	2020				79.6 [†]	94.8 [†]
BYOL [62]	ViT-S	2020				71.4 [†] [24]	
FOC [149]	ResNet34	2020			86.49		
SimCLRv2 [28]	ResNet-152 ^{2,7}	2020				80.9 [†]	95.5 [†]
DINO [24]	Resnet50	2021				75.3 [†]	
DINO [24]	ViT-B	2021				80.1 [†]	
CLIP [137]	ViT-L	2021				84.7[100]	
MAE [71]	ViT-B	2021				83.6 [†]	
MAE [71]	ViT-H	2021				86.9 [†]	
RELICv2 [174]	ResNet50v2 ²	2022				79.4 [†]	
RELICv2 [174]	ResNet200 ²	2022				80.6 [†]	
Semi-ViT [21]	ViT-H	2022				84.3	96.6
OmniMAE [57]	ViT-B	2022				83.0 [†]	
OmniMAE [57]	ViT-H	2022				86.5 [†]	
REACT [100]	ViT-B	2023				81.8 [†]	
REACT [100]	ViT-L	2023				85.1 ^{†9}	

3.1. Semi-, Self- and Unsupervised Learning

Table 3.4. Overview of reported accuracies with less labels - The first column shows the method used. The last seven columns report the top-1 accuracy in % for the given data set and number of labels. The number is given either as an absolute number or as a percentage. An empty entry represents the fact that no result was reported. [†] only 100 labels per class Table adapted from [151].

	CIFAR-10			STL-10		ILSVRC-2012		ILSVRC-2012 (Top-5)	
	4000	1000	250	5000	1000	10%	1%	10%	1%
One-Stage-Semi-Supervised									
Mean Teacher [171]	89.64	82.68	52.68						
ICT [183]	92.71	84.52	61.4 [18]						
MixMatch [18]	93.76	92.25	88.92	94.41	89.82				
EnAET [190]	94.65	93.05	92.4	95.48	91.96				
UDA [197]	95.12[165]		91.18[165]		92.34[165]	68.66		88.52	
ReMixMatch [17]	94.86	94.27	93.73		93.82				
FixMatch [165]	95.74		94.93		94.83	71.46	56.34 [†] [191]	89.13	78.20 [†] [191]
FreeMatch [191]	95.90	95.12	95.10		94.37		59.43 [†]		81.23 [†]
Multi-Stage-Semi-Supervised									
DMT [99]	88.70		80.3				58.6		
SimCLR [27]						74.4[28]	63.0[28]	92.6	85.8
BYOL [62]						77.7	71.2	93.7	89.5
SimCLRv2 [28]						80.9	76.6	95.5	93.4
Semi-ViT [21]						84.3	80.0	96.6	93.1
CLIP [137]						84.7	80.5		
React [100]						85.1	81.6		

The survey also noted that methods are often evaluated only on balanced and manually cleaned datasets. Based on the reported results, this trend continues, but some papers also report results on more difficult subsets of ImageNet [198] or even benchmarks with multiple datasets [100]. However, ImageNet still seems to be the most common evaluation dataset.

Trend 2: How much supervision is needed?

The second trend describes the shrinking gap between supervised training and SSL. Secondly, it states that unsupervised methods may not be as important as semi-supervised methods with very few samples. The gap is narrowing, as described in the previous paragraph, and sometimes it is already surpassed as described in the RelicV2 paper [174]. In particular, the use of Vision Language Model (VLM) drastically reduces the required amount of labeled data and allows some kind of unlabeled approaches with high performance. Depending on the interpretation, if defining the classes with natural language is equivalent to using no or few samples per class, the use of semi- or unsupervised methods is more promising.

3. Quantity & Quality of Data

Trend 3: Combination of common ideas

The survey revealed that some common ideas are not often used in combination with each other. These missing combinations represent opportunities for future research. The sample size of this survey is too small to draw conclusions about all the research done, but the top performing approaches do not seem to combine more different approaches. However, ideas such as MAE, CL, PL, ViT are used in most of the newer methods.

3.2 Proposal guided annotations

The usage of proposals, pseudo-labels, model suggestions or computer-generated entries as labels or as a prior for annotation is well documented in the literature under various names [41, 105, 130, 159]. Positive effects on the labeling speed and quality have been reported in small user studies (n=54) [42] and by us in a proof-of-concept evaluation [149, 154]. The literature describes different information that can be provided to the annotators besides the prediction and finds different effects [47, 141]. Dudely et al. show that the quality of the predictions is important and that they lead to improved results in deception detection [47]. In contrast to this, Zhang et al. find that for income prediction, only the model confidence has a significant positive impact compared to the prediction itself [204]. Even when the problem of ambiguity is recognized, proposals are often only used to automatically [131] or human-verified [16] correct error cases, rather than to estimate the underlying distribution.

The downside of using proposals is the *default effect* [78]. Annotators are more likely to accept the default option that is provided as proposal. We have shown that this effect leads to skewed distributions/introduction of a bias [154, 155] compared to annotations without proposals. In [155] we also analyze this bias and provide possible correction methods, which are described in detail in Section 5.4 and Section 6.3. Schulz et al. found no noteworthy bias in the segmentation and classification of epistemic activities in diagnostic reasoning texts, while reporting improved speed and quality [159].

Despite the potential downside of the default effect, proposals remain highly valuable due to their documented positive impact on annotation

3.2. Proposal guided annotations

speed and consistency. They provide additional information during the annotation process that can help to guide and enhance the process. While biases need to be addressed, proposals play an important role in error correction and estimation of underlying distributions.

Part II

Own methodical and data contributions

Overview about the methodology

To answer Research Question 2: “How can improved data quality for image classification be quantified?” (RQ2) and Research Question 3: “What are the implications of using proposals to guide the annotation process for image classification?” (RQ3), an overview of the methodology used is given in this chapter before it is described in detail in the next chapters. The necessary components of the provided methodology vary depending on the use case. The relevant distributions are introduced in Figure 4.1. They are described in detail and are followed by a description how these distributions can be converted into each other and lastly possible use cases are described as examples.

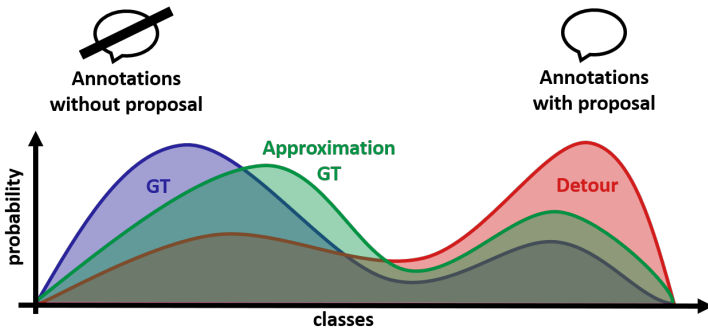


Figure 4.1. Illustration of relevant distributions for creating high-quality data – The figure illustrates the creation of the desired distribution (green), the approximation of the ground-truth, through the adoption of detour distribution (red) in order to be as close as possible to the Ground-Truth (GT) distribution (blue).

4. Overview about the methodology

GT distribution (blue) As discussed in Section 2.2.4, we need soft GT data for each image. However, there are only a few datasets [111, 133, 193] with soft labels as GT which allow to measure an improved probability distribution. For this reason, we proposed a data-centric benchmark (see Section 5.1) to allow a multi-domain evaluation on soft labels in [152]. Moreover, this benchmark proposes a two-phase evaluation scheme, which allows a better separation of the improvement of the data or the model. This gives us the blue distribution shown in Figure 4.1 and Figure 4.2.

Approximation of GT distribution (green) Ideally, another distribution like the green one in Figure 4.1 would be approximated first and then compared to the GT distribution (blue). However, this direct approximation requires multiple annotations per image to create high-quality soft labels. Thus, direct approximation of the green distribution is expensive even for smaller labeled datasets.

Detour distribution (red) Therefore, a slightly different distribution (shown in red) is approximated first, which can be achieved more cost-effectively. Such a distribution can be obtained by averaging over multiple annotations that have been annotated using a proposal. The advantage of using proposals is that they allow faster annotation and thus reduce costs. The methods for creating such proposals with Fuzzy Overclustering (FOC) and Data-Centric Classification & Clustering (DC3) are discussed in Section 6.1 and Section 6.2. They can introduce bias because people tend to take the default option (see the default effect in Section 3.2), which results in the shifted red distribution instead of the desired green distribution. It can be shown that using overclustering proposals directly is not optimal, and thus this dissertation proposes Overclustering Stochastic Proposal (OSP) to exploit the full potential of overclustering proposals (see Section 6.4).

How to create distributions from other distributions As mentioned above, a good prediction of the GT (blue) distribution is desirable. Therefore, transitions between the blue/green and red distributions must be computable. These transitions are called Simulated Proposal Acceptance (SPA)

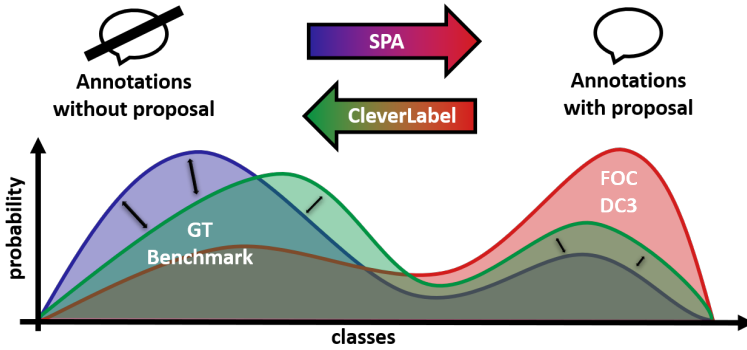


Figure 4.2. Illustration of proposed methodology to create high-quality data on the example of an idealized distribution across all classes for a single image – The different distributions are introduced in Figure 4.1. The white text indicates proposed methodology of this dissertation. The GT distribution is often not computed in public datasets, so the proposed benchmark gives the distribution for 10 different datasets. The transition to the red distribution is necessary because it would be too costly to estimate the green/blue distribution directly. The transition from the GT distribution (blue) to the red distribution created by annotations with proposals can be created with the proposed method Simulated Proposal Acceptance (SPA) for faster development. The proposals mentioned above can be generated by the proposed methods Fuzzy Overclustering (FOC) or Data-Centric Classification & Clustering (DC3). The correction of a possible introduced bias can be achieved by the proposed method Cost-effective LabEling using Validated proposal-guidEd annotations and Repaired LABELs (CleverLabel), which describes the transition from the red to the green distribution.

and Cost-effective LabEling using Validated proposal-guidEd annotations and Repaired LABELs (CleverLabel) (see Figure 4.2). SPA can be used to simulate the acceptance behaviour of humans with regard to proposals for a repeatable and low cost simulation during the development process. CleverLabel combines two distribution enhancement methods to revert or minimize the effect of the introduced bias in the red distribution and convert it to the desired green distribution. Both methods are described in detail in Section 5.4 and Section 6.3, respectively. For evaluation, this dissertation uses the Simulated Proposal Acceptance (SPA) and human

4. Overview about the methodology

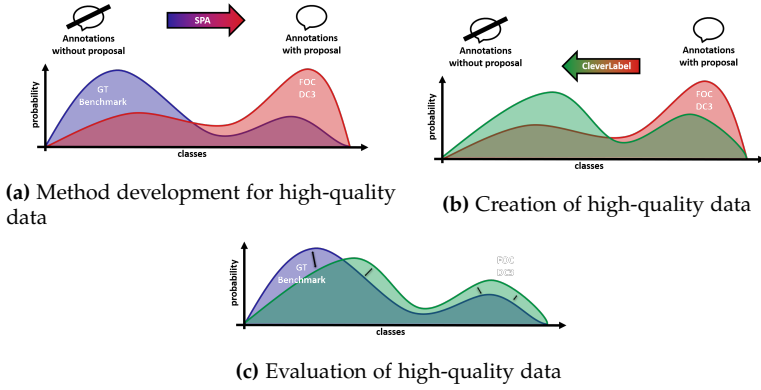


Figure 4.3. Different use cases for the provided methodology in this dissertation

annotations on real-world applications described in detail in Section 5.3 for the biased red distribution.

Metric A measure of the difference between the two distributions (blue, green) needs to be defined after the distribution based on proposals (red) has been adopted with CleverLabel as a better prediction (green) of the GT distribution (blue). The main metrics used for this comparison are described in Section 5.2.

Resulting strategy Given the results of the previous methodology, our Strategy for creating high-quality iMage Annotations with human Reliability and judgement enhancement (SMART) is described in Section 6.5. This strategy includes the process suggested above and possible alternatives based on a specific use case. Thus, the SMART answers Research Question 4: “How should ambiguous data be annotated based on the results of RQ1, RQ2 and RQ3?” (RQ4) and can be seen as the main result of this dissertation.

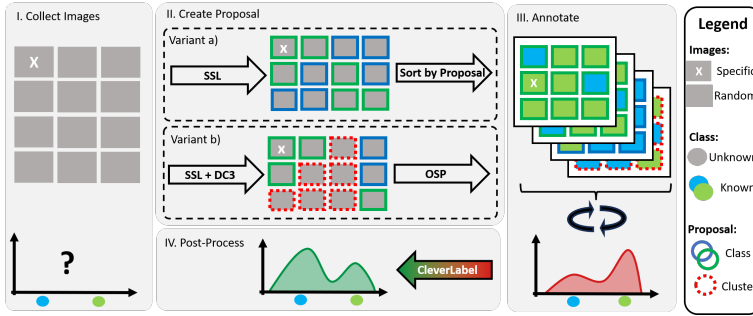


Figure 4.4. Workflow from unlabeled image data to high-quality label distributions – This workflow is founded on SMART (see Section 6.5 for more details). The process begins with a set of unlabeled images (represented by grey squares) with the distribution of the marked image (x) across the classes blue and green being unknown (see “I. Collect Images”). Step two, “II. Create Proposal”, generates proposals for each image using variant a or b, with and without Overclustering (OC), respectively. Variant a uses SSL to create high quality pseudo-labels for each image which can be used as proposals. For the next step, all proposals are grouped based on their respective proposed class. Variant b is analogous but additionally uses DC3 to group some images into clusters instead of just classes. The method OSP is necessary to create meaningful proposals from these clusters for every image in the next step. It is noteworthy that both options yield comparable results; Variant A is less complex, whereas Variant B yields higher downstream image classification performance. In step “III. Annotate” all images are annotated multiple times. For each iteration, the images are shown in a grid (see four example grids in figure). It is important that all proposals for each image in the current grid are identical. The annotator only needs to reannotate the anomalies of the proposed category/cluster. This setup reduces annotation time and increases consistency in annotations based on proposals. The average across all iterations can be summarized for the marked image (x) as the red distribution across the blue and green classes. The use of proposals may introduce bias towards the proposed class (see red distribution) as it is more likely to be accepted by an annotator than any other class. The CleverLabel method equalizes the distribution to minimize potential bias, resulting in the green distribution shown in the figure. This output can be utilized to train a neural network or compared to a GT distribution (not shown in the figure) to measure quality.

4. Overview about the methodology

Use Cases In Figure 4.3, three different use cases are described for which this dissertation methodology could be used.

The first use case, describes the research about how to create high-quality data. To collect such annotations, it is not practical to rely on human input for every iteration in the development process. SPA enables the creation of comparable annotations from the known distribution of GT to expedite the development process. This approximation facilitated the investigation for generating high-quality data in this dissertation.

The creation of high-quality data is explained in the second use case. The detour distribution is determined through human annotations or generated as in the first use case. The CleverLabel technique can be utilized to produce an enhanced label distribution, resulting in higher quality data. It is important to note that in general, objective evaluations of data quality cannot be made. This use case can be seen as the most important one because it has the broadest scope. Our overall strategy, as outlined in Section 6.5, is reflected in the detailed workflow presented in Figure 4.4.

The third use case serves to demonstrate the potential for evaluating data quality. Only if GT data is available can an evaluation of the quality be performed by comparing the created approximation (green) with the GT distribution.

The complete methodology of this dissertation, comprising of all use cases, is illustrated in Figure 4.2.

How to measure data quality?

Before discussing improvement methods for data quality in the next chapter (Chapter 6), it is necessary to define how to measure improved data quality. Therefore, the main datasets used in this dissertation and a corresponding benchmark are presented in Section 5.1. The metrics used to measure the differences are defined in Section 5.2.

To evaluate the influence of proposal systems on human annotations, it is important to either conduct human case studies (see Section 5.3) or to simulate their behavior (see Section 5.4). Both options are explored and their main results are summarized in their respective sections.

5.1 Benchmark

The proposed benchmark holds great significance for this dissertation because it outlines the primary evaluation setup, which differs from traditional model-centric approaches. While not every method is assessed using this benchmark (refer to Chapter 6), the cohesive evaluation presented in Chapter 7 is solely reliant on this setup. The creation of our benchmark was motivated by two factors: insufficient data and a data-centric approach. Both reasons are discussed below.

The number of datasets that provide multiple annotations per image to approximate $P(L^x = \cdot)$ is limited. Only CIFAR10H [133], CIFAR10N [193] and ImageNet ReaL [19] are known datasets with this desired property. However, CIFAR10N contains more images than CIFAR10H but fewer annotations per image and ImageNet ReaL contains 1000 classes of manually cleaned images which was not a common real-world classification problem. Prior to the benchmark, our previous models were evaluated on other real world datasets such as plankton [149], turkey [184, 186] or mice

5. How to measure data quality?

Table 5.1. Overview of the used datasets – # is an abbreviation for number. The class imbalance is given as the percentage of the smallest and largest class with regard to the complete dataset. The agreement is the percentage of annotations that agree with the majority vote. The scores ACC and \hat{ACC} are given for the supervised baseline across three test folds. The metrics are defined in Section 5.2. The access describes if the (raw) data is available openly, requires permission (restricted) or was not previously available (N/A) [before our benchmark]. In the last column, datasets with modifications to the original data are marked with X. A modification might be adding more annotations or crop images to a region of interest. Copied/Adapted from [152]

Name	# classes	Input size [px]	# Images	Class Imbalance [%]		Agreement [%] Mean \pm STD	# Annotations Mean \pm STD	ACC [%] Mean \pm STD	\hat{ACC} [%] Mean \pm STD	Access	Updated
				Smallest	Largest						
Benthic	10	112×112	4867	2.31	39.66	82.61 \pm 19.67	4.54 \pm 2.01	64.17 \pm 0.63	83.36 \pm 0.47	Restricted	X
CIFAR-10H	10	32×32	10000	9.88	10.16	95.44 \pm 8.91	51.10 \pm 1.54	90.75 \pm 0.39	95.72 \pm 0.12	Open	
Mice Bone	3	224×224	7240	14.75	70.48	85.06 \pm 17.52	15.30 \pm 21.90	61.88 \pm 9.44	78.39 \pm 1.95	Restricted	X
Pig	4	96×96	10237	7.82	41.23	65.32 \pm 19.50	7.26 \pm 2.29	35.97 \pm 3.61	64.77 \pm 0.79	N/A	X
Plankton	10	96×96	12280	4.16	30.37	93.26 \pm 13.60	24.38 \pm 44.17	89.89 \pm 0.82	92.41 \pm 0.41	Restricted	
Quality MRI	2	224×224	310	34.84	64.16	71.56 \pm 12.27	99.94 \pm 13.44	66.62 \pm 3.55	75.81 \pm 0.17	Restricted	X
Synthetic	6	224×224	15000	16.17	17.57	74.41 \pm 24.28	98.86 \pm 0.99	87.85 \pm 0.48	74.65 \pm 0.34	N/A	X
Treeversity#1	6	224×224	9489	9.98	30.67	88.60 \pm 16.13	14.78 \pm 7.06	79.50 \pm 1.53	89.20 \pm 0.31	Open	X
Treeversity#6	6	224×224	9826	8.77	31.26	66.53 \pm 19.48	35.45 \pm 11.47	56.71 \pm 4.89	68.88 \pm 0.72	Open	X
Turkey	3	192×192	8040	10.88	75.95	91.56 \pm 13.82	14.85 \pm 20.95	75.51 \pm 2.80	86.89 \pm 1.03	Restricted	X

bone fiber classification [148]. The benchmark includes CIFAR10H [133], the three datasets mentioned above and 6 new datasets from the areas of seafloor classification [94, 157], pig tail injury classification [20], MRI image quality classification [125, 166], synthetic color discrimination [149] and plant image classification¹. A complete overview of the main dataset characteristics is given in Table 5.1. All datasets are divided randomly into five stratified folds based on the class distribution. Three subsets per dataset are generated by rearranging which three folds are used for training, validation, and testing. This subset is referred to as a data set split. Every subset is evaluated, enabling cross-validation across different splits. While more than three splits are possible with five folds in theory, this would significantly increase the required computing resources. For more information, refer to [152] and Section A.1.5.

The second motivation for the benchmark was to improve the data, not just the model. As shown in Figure 5.1, methods are often evaluated directly on the same data on which they are trained. Thus, if the training data contains ambiguous labels that lower the prediction performance, it is difficult to distinguish between the effects of the model and the data. The

¹<https://arboretum.harvard.edu/research/data-resources/>

5.1. Benchmark

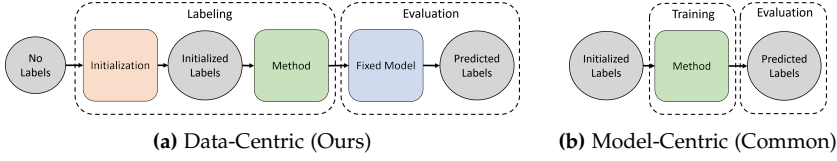


Figure 5.1. Comparison of our data-centric approach with the commonly used model-centric approach. The circles and arrows represent the label information available in addition to the corresponding images. The squares represent the methods that generate / modify these labels. There are two main differences between our approach and the common model-centric approach. Firstly, we also consider how the raw unlabeled data is initialized and thus how many annotations are needed. Secondly, we use a fixed model to evaluate the performance of benchmarked method. These differences lead to a greater separation of data quality and method performance on the final scores of the predicted labels. Cited/Copied from [152]

benchmark therefore consists of a two phase approach to better separate their effects. The structure can be described as follows:

“The main structure of the benchmark is divided into a Labeling and an Evaluation phase (see Figure 5.1a) which is comparable to established Teacher-Student-Approaches [92, 171]. Using this denotation, the benchmarked method will, as a teacher, improve labels during the first phase. These are then benchmarked in the second phase by analyzing their quality as training input to a student model. Be aware that we do not allow a knowledge transfer from the second phase to the first phase.

In detail, during the Labeling phase, we use samples from the distribution of the above-mentioned annotations to get different realistic label estimates as an initialization. The task of the benchmarked method is to improve these estimates for better performance of an other classification model in the second phase. In that phase (Evaluation), the obtained labels are used as input for training a fixed model and its performance is measured on a testing subset of the original data. In contrast to common model-centric deep learning approaches (see Figure 5.1b), we can vary the initialization for the same method and better separate its performance from the data improvement. The fixed model is used for the evaluation to facilitate distinguishing between performance gains due to improved input data and better

5. How to measure data quality?

learning of the method itself."² A full overview of the benchmark pipeline is also given as pseudo-code in Algorithm 1. For more details see the original paper [152] or the published source. code³

In the benchmark, 20 methods were evaluated as baseline comparisons. The default reference point (called Baseline in [152] and called GT(-Sampler) in this work) samples from the soft GT distribution ($P(L^x \cdot)$) and can be seen as an approximation of the performance of human annotators without the guidance of proposals for the classification task. The other 19 methods are as described in [152]: *"The supervised methods are Heteroscedastic [34], SNGP [101], and ELR+ [102]. The semi-supervised methods are Mean-Teacher [171], π -Model [92], FixMatch [165], DC3 [154], Pseudo-Label [95], and DivideMix [96]. The self-supervised methods are BYOL [62], MO-Cov2 [29], SimCLR [27], and SWAV [23]. [...] For DC3[154], we investigated the combinations with Mean-Teacher, π -Model, FixMatch, and Pseudo-Label. For Pseudo-Label, we used two different implementations (v1 and v2) and variants with or without pretraining and soft or hard labels as input."*⁴ For more details about the methods, please see the original benchmark [152] or our survey [151].

The main results of our benchmark paper are

- ▷ one annotation is not enough to capture the label ambiguity, while soft labels are a way to capture this ambiguity.
- ▷ improved data quality leads to improved performance on multiple metrics, but comes at a higher cost in terms of annotation time (see Section 5.2.4).
- ▷ a shift from optimizing only hard label classification performance to improving GT distribution estimation, for example by using soft labels, could potentially overcome issues such as overconfident models. This conclusion is based on the reported lower *ECE* scores with increased budget (*b*) (see Section 5.2).

²quoted from [152] p.2

³<https://github.com/Emprime/dcic>

⁴quoted from [152] p.6

Algorithm 1 Benchmark pipeline – For one dataset, the images are provided without labels in three subsets for training, validation, and testing, represented by X_{Train} , X_{Val} , and X_{Test} respectively. The oracle Ω can provide human-like annotations but at the cost of increasing the budget b . However, applying the oracle Ω on any images within the test set (X_{Test}) is prohibited. The variables $p_{labeled}$ and n_{annos} indicate the percentage of training and validation data to be annotated during initialization (init, orange block in Figure 5.1a) and the number of annotations per image, respectively. L_{Train} , L_{Val} , and L_{Test} represent the labels generated in the pipeline for training, validation, and testing, respectively. Each set may contain soft-labels corresponding to the image set (X). SSL algorithms can still use this data while supervised deep learning methods will ignore this image during training or validation. The function `method` serves as the benchmark method (green block in Figure 5.1a). This function may ignore, extend, enhance, or alter the provided labels. The metrics (`metrics`) are calculated solely on the test set (X_{test}) thus labels must be provided for every image $x \in X_{Test}$. Detailed descriptions of the calculated metrics are provided in Section 5.2. The function `train` describes a supervised training on the provided labeled images in X_{Train} and X_{Val} . This training can not be influenced by the user except from the provided L_{Train} , L_{Val} . This yields in a trained model Φ , which predicts (`predict`) soft-labels for all test images X_{Test} . The function `metrics` is used to calculate the final results of the benchmark (see details in Section 5.2).

Require: $X_{Train}, X_{Val}, X_{Test}, \Omega, p_{labeled}, n_{annos}$

▷ Labeling phase

$L_{Train}, L_{Val} \leftarrow \text{init}(X_{Train}, X_{Val}, X_{Test}, \Omega, p_{labeled}, n_{annos})$

▷ init orange block in Figure 5.1a

$L_{Train}, L_{Val}, L_{Test} \leftarrow \text{method}(X_{Train}, X_{Val}, X_{Test}, \Omega, L_{Train}, L_{Val}, L_{Test})$

▷ method is a user-defined function, green block in Figure 5.1a

$\hat{K}L, \hat{ACC}, \hat{F1} \leftarrow \text{metrics}(L_{Test})$

▷ details in Section 5.2

▷ Evaluation phase

$\Phi \leftarrow \text{train}(X_{Train}, X_{Val}, L_{Train}, L_{Val})$

▷ Blue block in Figure 5.1a

$L_{Test} \leftarrow \text{predict}(\Phi, X_{Test})$

$KL, ACC, F1 \leftarrow \text{metrics}(L_{Test})$

▷ details in Section 5.2

5. How to measure data quality?

5.2 Metrics

The main objective of this dissertation is to handle ambiguous data and thus a measure of improved data quality must be defined (see Research Question 2.2: “What metrics and algorithms are needed to quantify improved data quality?” (RQ2.2)). Based on the definition of the GT distribution ($P(L^x = \cdot)$), this objective can be interpreted as minimizing the difference between the network output $\Phi(x)$ and $P(L^x = \cdot)$ (both interpreted as probability distributions). Let K be the number of classes and $\Phi(x), P(L^x = \cdot) \in [0, 1]^K$.

5.2.1 Measuring distributional differences

The Kullback-Leibler divergence (KL) is an established metric to measure the difference between two probability distributions [90, 117] and is in our case given by

$$\begin{aligned} KL(P(L^x = \cdot), \Phi(x)) &= \sum_{k=1}^K P(L^x = k) \cdot \log\left(\frac{P(L^x = k)}{\Phi(x)_k}\right) \\ &= - \sum_{k=1}^K P(L^x = k) \cdot \log\left(\frac{\Phi(x)_k}{P(L^x = k)}\right). \end{aligned} \quad (5.1)$$

Closely related to this is the common loss function for image classification, the cross-entropy, which in our case is given by

$$\begin{aligned} CE(P(L^x = \cdot), \Phi(x)) &= - \sum_{k=1}^K P(L^x = k) \cdot \log(\Phi(x)_k) \\ &= - \sum_{k=1}^K P(L^x = k) \cdot \log(P(L^x = k)) \\ &\quad + KL(P(L^x = \cdot), \Phi(x)) \\ &= H(P(L^x = \cdot)) + KL(P(L^x = \cdot), \Phi(x)) \end{aligned} \quad (5.2)$$

where H is the entropy of $P(L^x = \cdot)$. CE is identical to KL up to the entropy H and the full proof is given in the supplement of [149]. The metric CE is not optimal due to the fact that the entropy is a fixed value

for a given soft GT distribution that varies between images. Thus, it would distort the results and make them less interpretable.

It is important to note that when the two-stage approach of the benchmark is used, the metric KL refers to the comparison of the predictions of the fixed model with the GT. The comparison of the results of the first stage with the GT is defined in Section 5.2.3. The results of the second stage are used instead of the first stage because this dissertations aims to assess data quality by evaluating the downstream classification performance of the training labels.

5.2.2 Associated metrics

Accuracy (ACC) is a common image classification metric that counts the correctly predicted class labels for a set of images X with N images and is defined by

$$\begin{aligned}
 ACC(X) &= \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{k=k'}, \\
 k &= \operatorname{argmax}_{i=1}^K P(L^x = k), \\
 k' &= \operatorname{argmax}_{i=1}^K \Phi(x)_k
 \end{aligned} \tag{5.3}$$

where $\mathbb{1}$ is the indicator function. In this dissertation, ACC does not refer to the average accuracy of the complete image set X , but to the macro average over all classes. The macro ACC averages the ACC per class to obtain interpretable metrics even for long-tailed distributions. The literature also suggests other classification metrics, such as the $F1$ -score. However, we found in [152] that the macro averages of $F1$ and ACC have a correlation of 0.99, so this dissertation will rely on ACC . The disadvantage of this measure is that it does not account for uncertainty.

Expected calibration error (ECE) is a metric for measuring the calibration of a model [65]. The motivation for ECE is that if a model predicts, say, a label with 90% confidence, it should be correct 90% of the time. The

5. How to measure data quality?

metric is not defined per image, but over a complete set of N images X by

$$\begin{aligned} ECE(X) &= \sum_{b=1}^{\hat{B}} \frac{n_b}{N} |ACC(B_b) - conf(B_b)| \\ conf(B_b) &= \frac{1}{n_b} \sum_{x \in B_b} \Phi(x)_k \end{aligned} \quad (5.4)$$

for \hat{B} bins $B_1, \dots, B_{\hat{B}}$ with $\dot{\bigcup}_{b=1}^{\hat{B}} B_b = X$; $n_b = |B_b|$ images per bin and acc and $conf$ the average accuracy and model confidence per bin, respectively. While ECE is often used in the literature, it does not directly measure the difference between distributions as RQ2.2 intends. Recent work even criticizes the ECE metric for having ill-defined properties, such as a bias toward dataset size, as pointed out in [64] which is not desirable.

The Cohen's Kappa Score (κ) [112] is a metric for the agreement, for example, between two annotators by

$$\kappa = \frac{p_0 - p_c}{1 - p_c}. \quad (5.5)$$

where p_0 is the agreement rate, e.g. ACC , and p_c the agreement rate with random assignment, which is $p_c = \frac{1}{K}$ for the described discrete K class problem.

5.2.3 Special benchmark metrics

In the benchmark (see Section 5.1), the metrics defined above, such as KL or ECE , are computed for the output of the first or second stage (see labeling and evaluation stage in Figure 5.1). To distinguish between them, the metrics of the first stage are called \hat{KL} , \hat{ACC} , $\hat{F1}$ and \hat{ECE} . They can be seen as the performance of a model in a common model-centric setup and define the input quality for the evaluation model of the second phase in the benchmark.

5.2.4 Costs

The cost of annotating data is basically the number of annotations required. Assuming a constant cost per annotation and that the total cost of anno-

5.3. Human user studies

Table 5.2. Overview of human studies for proposal evaluation

Goal	# Annotators	# images	Avg. annotations	# annotations	Source
Measure intra- and intervariability	16	24	19	456	[148]
Proof-of-concept consistency gain with overclustering	1	7,424	2	14,848	[149]
Proof-of-concept benefit proposal for ambiguous data	5	3,748	27	101,196	[154]
Verification of Annotation Strategy	6	3,721	106	394,426	[156]

tating all images in image set X once is one, this dissertation refers to the cumulative cost as budget (b). In general, b can be calculated by the *number of images* to annotate times *annotations per image*. For example, if 20% of all data X is annotated three times, the final budget is $b = 0.2 \cdot 3 = 0.6$. The number of annotated images (given as a percentage of the total size of X) is called initial supervision (*in. sup.*).

If two rounds of annotation are used, as in [155], the calculation of b is given by *in. sup.* + (*percentage of annotatins of X · number of annotations per image*) / S). A “Speedup (S) [...] can be expected due to using proposals [...] [as opposed to no guidance during the annotation process]. For this reason we include this parameter [...] with [e.g.] the values 1 (no speedup), 2.5 as in [154] or 10 as in [158]”.⁵ A common setup in [155] was to annotate 20% of the data initially once and then annotate the whole dataset 5 times with an expected S of 2.5, resulting in a budget of $b = 0.2 + (5 \cdot 1) / 2.5 = 2.2$.

5.3 Human user studies

An overview of the four published proof-of-concept human case studies in different papers [148, 149, 154, 156] is given in Table 5.2. The used methods will be explained later in Chapter 6. This section discusses higher level questions with regard to humans and thus the methodological details are not so important.

In [148] the intra- and intervariability of humans for semantic segmentation of collagen fiber orientation was investigated 15 different people and 5 results over time from the same person resulted in a total of 19 samples for 24 images each. The intra-person segmentation had a lower variability than the inter-person segmentation but still yielded only about

⁵quoted from [155] p.11

5. How to measure data quality?

Table 5.3. Consistency comparison on plankton dataset – The consistency is rated by experts over the complete data and a subset without the class no-fit. The score is given overall and as average per cluster with standard deviation. Copied and adapted from [149].

Method	all data		ignore class no-fit	
	overall	per cluster	overall	per cluster
FixMatch [165]	82.56	78.78 \pm 28.22	77.11	69.61 \pm 29.41
FOC (Ours)	87.80	79.66 \pm 18.88	86.31	86.41 \pm 13.68

78.29% macro accuracy. The most interesting part is that no method could outperform this result, while higher results in earlier validation could be verified as overfitting the data. The problem was that the validation and test data were also labeled by the person from the intra-person check and thus even better results could not be detected due the noise and ambiguity in the ground truth data. More details can be found in the original publication in [148].

In [149], the acceptance rate by a domain expert of the results proposed by FOC (see Section 6.1) and FixMatch[165] was measured. *“We can judge the consistency of each image within its cluster with the help of experts as a quality measure. An image is consistent if an expert views it as visually similar to the majority of the cluster. The consistency is calculated by dividing the number of consistent images by all images. The consistency over all classes or per class for FOC and FixMatch is given in [Table 5.3]”*.⁶ The domain expert found the generated clusters cleaner than the class separation generated by FixMatch. *“This means the clusters produced by FOC are more relevant in practice because there are fewer low-quality clusters which can not be used. Overall, this higher consistency can lead to faster and more reliable annotations.”*⁷

[154] used a combination of domain experts and hired workers to annotate the data. A total of 5 annotators were asked, including two domain experts. Their annotations were used to investigate the benefit of using proposal with class structure or our new combined classification and clustering over no proposals on four different datasets. *“We see a*

⁶quoted from [149] p.10

⁷quoted from [149] p.10

5.3. Human user studies

Table 5.4. Consistency comparison of generated labels from proposals – The first column describes the annotator selection and the used proposals. The Cohen’s Kappa Score (κ) measures the agreement of between the used repetitions and Time gives annotation time in minutes. Results which are within one percent or minute of the best result per dataset and annotator selection are marked bold, copied and adopted from [154]

	Plankton		Turkey		Mice Bone		CIFAR-10H	
	κ [%] \uparrow	Time [min] \downarrow	κ [%] \uparrow	Time [min] \downarrow	κ [%] \uparrow	Time [min] \downarrow	κ [%] \uparrow	Time [min] \downarrow
A1	73.00 \pm 1.51	51.09 \pm 2.36	88.08 \pm 3.43	14.56 \pm 0.84	71.35 \pm 2.56	13.94 \pm 2.25	92.70 \pm 1.69	40.58 \pm 1.93
A1 + SSL	85.00 \pm 2.52	12.69 \pm 3.37	85.63 \pm 3.66	10.70 \pm 0.44	72.00 \pm 2.87	6.59 \pm 1.65	94.85 \pm 0.91	14.33 \pm 1.48
A1 + DC3	90.29 \pm 1.41	11.32 \pm 1.43	91.95 \pm 1.22	11.57 \pm 0.64	81.36 \pm 2.17	6.74 \pm 1.05	94.70 \pm 0.52	14.65 \pm 0.60
A2	85.25 \pm 1.79	61.99 \pm 10.98	81.54 \pm 0.89	18.11 \pm 4.30	68.63 \pm 6.66	11.06 \pm 3.60	98.81 \pm 0.14	33.08 \pm 5.36
A2 + SSL	94.88 \pm 0.52	9.23 \pm 0.76	81.10 \pm 3.39	9.48 \pm 0.83	59.63 \pm 6.20	12.07 \pm 4.77	98.00 \pm 0.27	12.66 \pm 0.69
A2 + DC3	94.04 \pm 0.67	10.32 \pm 0.07	81.83 \pm 1.98	9.91 \pm 0.39	72.19 \pm 3.23	9.13 \pm 2.98	98.29 \pm 0.19	14.27 \pm 0.69
A3	84.74 \pm 1.02	21.54 \pm 1.54	78.27 \pm 1.08	19.35 \pm 1.16	56.27 \pm 4.03	10.15 \pm 2.12	93.22 \pm 1.01	21.96 \pm 1.10
A3 + SSL	88.59 \pm 0.84	9.02 \pm 0.20	88.44 \pm 1.74	13.24 \pm 0.32	72.32 \pm 0.61	8.02 \pm 1.23	92.37 \pm 1.78	9.79 \pm 0.52
A3 + DC3	88.57 \pm 0.62	7.76 \pm 0.27	91.94 \pm 1.04	14.05 \pm 0.51	72.77 \pm 2.74	9.56 \pm 1.71	94.81 \pm 0.96	9.50 \pm 0.74

general trend that the consistency improves and the annotation time decreases when proposals are used instead of None. Using DC3 proposals instead of SSL proposals, either leads to a similar or better consistency while the annotation time is often increased by one or two minutes. For this improvement, we credit the cleaner and more fine-grained outputs of the network. The additional [verification] of the clusters could lead to the slightly increased annotation time. [...] On average across all annotators and datasets, we achieve an improved consistency of 6.74%, a relative speed-up of 2.4 and a maximum speed-up of 4.5 with DC3 proposals in comparison to the baseline”⁸.

5.3.1 Evaluation study of annotation strategy

The main goal of this dissertation is to answer Research Question 4: “How should ambiguous data be annotated based on the results of RQ1, RQ2 and RQ3?” (RQ4) and Section 6.5 will present an answer to this question, which was described in detail in [156]. This subsection describes the user study used to validate the proposed strategy. The data used and the task were described in [156] as follows: “We use the Verse datasets for better reproducibility of our osteoporotic vertebral fracture classification. As in [104], we consider only at thoracolumbar vertebrae because osteoporotic fractures occur mainly in this region of the spine, as shown in Figure 5.2. For the classification of

⁸quoted from [154] p.12

5. How to measure data quality?

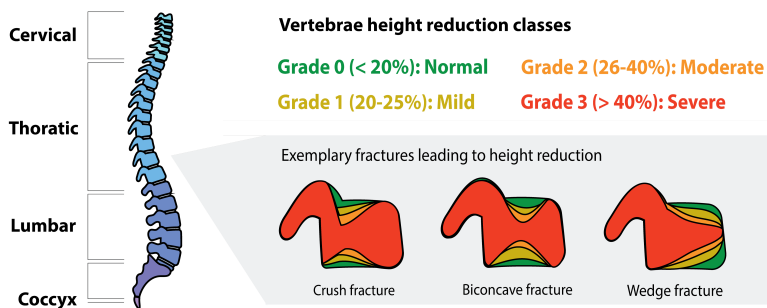


Figure 5.2. Illustration of a spine and definition of height reduction classes, copied from [156]

the fractures, we use an adaptation of the most commonly used score, the semi-quantitative Genant score [54]. As shown in Figure 5.2, the score defines four classes, which are mainly guided by the height reduction of the vertebrae compared to its neighbors or unfractured vertebrae. However, degenerative deformities (e.g. short vertebral height) of up to 20% are ignored for the osteoporotic fracture classification. Partially, these degenerative cases can not be assessed in the dataset because the resolution is lower than in normal radiographs. As shown in Table 5.1, this results in 3,761 individual vertebrae with a high class imbalance towards class zero. Overall, we have at least 1.1% samples per class and can apply common semi- or supervised techniques.”⁹

Two differences to the original work [104] are described in [156]. First, 2D projections of the 3D data were used because the Genant score was originally defined in 2D and it allows to use previous techniques [154]. Second, five hired students with non-medical background annotated the data because medical experts are in high demand and cannot annotate thousands of images multiple times. As the author, I worked with the students on some preliminary training annotations. For more details, see the original publication [156].

In Table 5.5 the original overview of human performance on the task is given. It can be seen that the annotations with proposals achieve a higher macro $F1$ score than without and that there is no significant difference

⁹quoted from [156] p.7

5.3. Human user studies

Table 5.5. Comparison of human performance on Verse2019 [104] – copied and adapted from [156]

	F1
2D - all	0.6380 ± 0.0323
2D - no proposal	0.6297 ± 0.0345
2D - with proposal	0.6462 ± 0.0302
3D - no proposal	0.6243 ± 0.0370

Table 5.6. Comparison of dataset specific variables – δ is the data specific offset when annotating with proposals described in [155]. S is the speedup difference between annotating with and without proposals. \hat{p}_c describes the percentage of data with at least 95% consistent annotations. Copied and adapted from [156]

name	δ	\hat{p}_c	# classes	largest class [%]	S
Verse [156]	0.1143	0.6833	4	90.11	1.1636
CIFAR10-H [154]	0.0973	0.7766	10	10.16	2.4665
MiceBone [154]	0.3636	0.4775	3	70.48	1.4471
Plankton [154]	0.5784	0.7368	10	30.37	4.4319
Turkey [154]	0.2164	0.6863	3	75.95	1.4877

between 2D and 3D annotators. An analysis of the dataset specific variables in Table 5.6 showed that the Verse data is comparable to previous [154] datasets. However, the measured speedup S was the smallest and it was hypothesized that “[t]his relationship is theoretically justified because a proposal is less important since the majority of images already belong to a class and thus define some kind of a proposal”.¹⁰

Key findings included that

- ▷ human annotation performance was higher with proposal as guidance.
- ▷ the annotations on 2D data with non-medical experts produce results comparable to the original annotations [104].

¹⁰quoted from [156] p.10

5. How to measure data quality?

Algorithm 2 Simulated Proposal Acceptance (SPA), i is the desired iteration step, a_i^x is the simulated annotation for image x in the i -th iteration, ρ_x is the proposed class, $P(L^x = \cdot)$ is the given GT distribution which is used to calculate the simulated annotation, copied and adapted from [155]

Require: Proposal ρ_x ; $a_i^x \in \{0\}^K$
 Calculate acceptance probability A
 $r \leftarrow \text{random}(0,1)$
if $r \leq A$ **then** \triangleright Accept proposal
 $a_{i,\rho_x}^x \leftarrow 1$
else \triangleright Sample from remaining classes
 $k \leftarrow \text{sampled from } P(L^x = k \mid \rho_x \neq k)$
 $a_{i,k}^x \leftarrow 1$
end if

\triangleright the learning effect of the annotators was measured over time and improved with proposals.

In Chapter 7, a comparison of the network performance based on the human annotations with different options for the proposal and correction methods is given.

5.4 Simulated Proposal Acceptance (SPA)

A major problem with human studies is that they are, by definition, dependent on humans. This is an issue because during the development process of new methods it is often not feasible to wait weeks or months to get feedback from real people on the new developments. For this reason, we proposed in [155] a simulation of human proposal acceptance (SPA), which calculates the biased distribution $P(L_b^x = \cdot)$.

Based on our previous user study in [154], we investigated how people annotate the data with and without proposals. Due to the default effect [78], people tend to accept the proposed class more often than the others. After this acceptance they mostly decided as if no proposal was given for the remaining classes. This led to two main conclusions, the acceptance probability A of a proposal ρ_x can be approximated for all images x as

5.4. Simulated Proposal Acceptance (SPA)

Table 5.7. Used offsets (δ) for proposal acceptance, copied from [155]

dataset	Benthic	CIFAR10H	MiceBone	Pig	Plankton
User Study	N/A	9.73%	36.36%	N/A	57.84%
Calculated	40.17%	0.00%	41.03%	25.72%	64.81%
dataset	QualityMRI	Synthetic	Treeversity#1	Treeversity#6	Turkey
User Study	N/A	N/A	N/A	N/A	21.64%
Calculated	0.00%	26.08%	26.08%	20.67%	14.17%

shown in Equation 5.6 and the simulation of the whole process can be modeled as shown in Algorithm 2.

$$A = \delta + (0.99 - \delta)P(L^x = \rho_x) \quad (5.6)$$

The acceptance probability A is roughly the offset GT probability $P(L^x = \rho_x)$ with a dataset specific offset (δ) up to 99%. The upper limit of 99% is used to allow for some noise even in highly certain situations, since there may be errors or conflicting ideas. In the original paper [155], we show how δ can be approximated based on multiple annotations of 20 images with a known GT probability $P(L^x = \rho_x)$. In Table 5.7, the computed values of δ with this approach and the real values from the user studies (where applicable) are shown for all datasets of the benchmark [152]. It is important to note that a lower offset δ means less bias is introduced into the annotations and $\delta = 0$ means no bias at all.

*“A visual comparison of the real and simulated results for all uncertainty bins can be seen in [Figure 5.3]. The main diagonal line contains the accepted proposals while the rest, especially the upper right corner are the rejected images. We see that the presented matrices are very similar, even in overlapping regions between accepted and rejected proposals as for uncertainty bin 0.41-0.6 of the proposed and annotated classes.”*¹¹ For more details on uncertainty bins or a more detailed comparison, see the original paper [155].

¹¹quoted from [155] p.9

5. How to measure data quality?

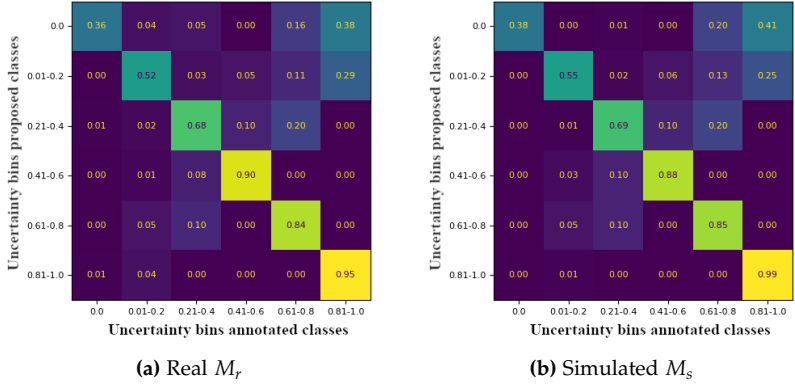


Figure 5.3. Visual comparison of the uncertainty bins for real vs. simulated proposal acceptance on the MiceBone dataset, normalized per row / per proposal uncertainty bin, the uncertainty bins can be interpreted as a histogram of the GT probability for the proposed and accepted class for each annotation, meaning for example that in real M_r , if a class is proposed with soft ground truth probability 0.21-0.4, in 0.10 of the cases a class with ground truth probability 0.41-0.6 is annotated. Therefore, some cells are 0 by default, copied and adapted from [155]

In general, the paper showed that SPA produces similar results to the original human study and thus can be used to approximate the human behavior. However, SPA is not a perfect representation of the reality and conclusions should be verified in larger real human studies as suggested above in Section 5.3.1. A unified presentation of the results is given in Chapter 7 and in the original paper [155].

Improving data quality

This chapter describes the methods developed as part of the research leading up to this dissertation. This answers Research Question 3: “What are the implications of using proposals to guide the annotation process for image classification?” (RQ3). An overview of the methods, their motivation and individual results is given in this section, while a unified evaluation is given in Chapter 7. It is important to note that the results presented in this chapter are individual and not comparable to each other due to the lack of a unified setup as described in Chapter 7. Details about differences and similarities are highlighted in each section, and the full information is reported in the corresponding original papers.

All methods presented here are motivated by the desire to improve label quality in less time during the annotation process. The motivation for why proposals are a good option for this goal, was described in Section 3.2. Over the years, we have developed three main algorithms that build on each other for this purpose. FOC [149] was the first algorithm and took the ideas of classification and clustering from [80] and extended them to real-world data. In particular, it defined the loss term CE^{-1} . DC3 [154] is a generalization of FOC to overcome problems such as unstable training and long training times. Moreover, it is method agnostic and thus generally applicable to most SSL algorithms and datasets. CleverLabel [155] solves the problem of the possible introduction of a bias due to the proposals used and the default effect [78].

In addition, the novel idea of Overclustering Stochastic Proposal (OSP) is introduced to solve the problem that more consistent annotations do not automatically translate into better data quality. Finally, all these results are unified in a **Strategy for creating high-quality iMAGE Annotations with human Reliability and judgement enhancementT (SMART)**. SMART can be

6. Improving data quality

considered as the most important method and methodological conclusion of the whole dissertation, because it answers Research Question 4: “How should ambiguous data be annotated based on the results of RQ1, RQ2 and RQ3?” (RQ4) and integrates all previous knowledge.

A unified evaluation of the individual methods and a verification of SMART based on the proposed human study in Section 5.3 is given in the next chapter (Chapter 7).

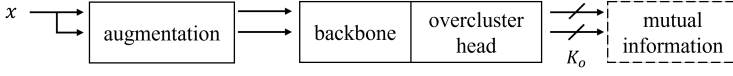
6.1 Fuzzy Overclustering (FOC)

FOC was developed to apply SSL to ambiguous plankton data, referred to as fuzzy data in the original paper. The work was inspired by the method Invariant Information Clustering (IIC) [80] which was state of the art in 2019 on the SSL benchmark STL-10 [32] and is graphically depicted in Figure 6.1a. Ji et al. proposed an unsupervised clustering loss based on the mutual information theory which incentivized that two augmented versions of the same images have a high mutual information in a clustering space [80]. Importantly, they used overclustering which means that they had more clusters K_o than ground-truth classes K . In the next step, the cluster heads are dropped and the retained backbone is fine-tuned on labeled data.

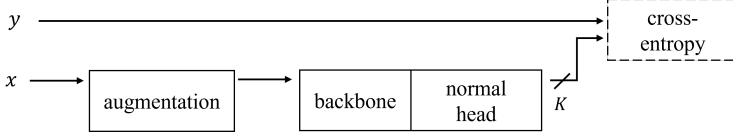
In contrast to IIC, FOC aimed to keep overclustering in combination with classes because it allowed a better representation of ambiguous data. Ambiguous data may not fit perfectly into only one class. A graphical representation of FOC is given in Figure 6.1b. The loss proposed by Ji et al. [80] could be applied to unlabeled data, while standard cross-entropy could be applied to labeled data. However, due to the larger number of clusters than classes, cross-entropy could not be applied to labeled data for clustering since no ground-truth class is known. The proposed loss CE^{-1} should ensure that dissimilar images are not in the same cluster and is described in Section 6.1.1. For labeled data, dissimilar images can be computed based on different class labels. It is important to note that the mutual information loss proposed by IIC gives slightly better results but easily leads to a collapse in IIC and FOC. Therefore, Ji et al. proposed several stabilization methods, such as repeating images several times in

6.1. Fuzzy Overclustering (FOC)

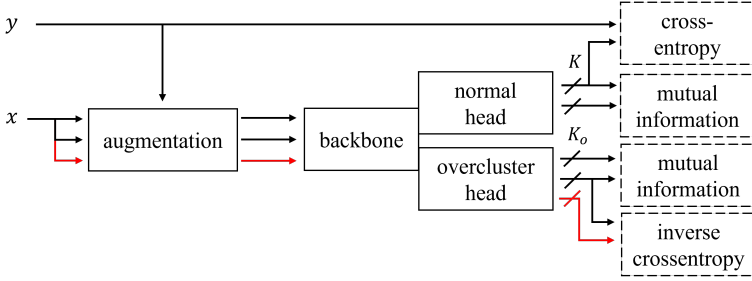
1. Pretext Task



2. Finetune



(a) IIC [80]



(b) FOC [149]

Figure 6.1. Method comparison of IIC[80] and FOC – The arrows represent the flow of the image/label information. The red arrow is a negative example, which is known not to be of the same class as x . The numbers at the outputs define how many output nodes were used (more or equal to the number of classes). The main similarities are that the mutual information loss proposed by Ji et al. [80] was used for the overclustering training and CE for the normal head. The main differences are that IIC splits the training into two stages and that our novel CE^{-1} loss was used for the overclustering training. More details are described in the original paper [149] and in the main text, figure adapted from [149]

a batch or averaging over multiple output heads [80]. To prevent mode collapse, FOC had to continue using these methods even though they drastically increased the run time up to several days. The results are shown in Table 6.1. *"We see that FOC reaches a performance of about 86% on [non-ambiguous] data but is not able to reach the performance of FixMatch.*

6. Improving data quality

Table 6.1. Comparison of state-of-the-art on certain and fuzzy data – The original paper or original author’s code was utilized to replicate the results, and the reported metrics differ depending on the dataset. The STL-10 dataset shows the metric as ACC, whereas the Plankton dataset uses the macro F1 score, both of which are calculated based on the validation dataset. The runtime required for a successful training session on the Plankton dataset is measured in hours on a single Nvidia RTX 2080 Ti GPU. For two-stage approaches, separate runtimes are provided for each stage. The best results are marked bold. Legend: [†] A MLP used for fine-tuning. [‡] Used only 1000 labels instead of 5000. * Unsupervised method.

Method	Network	Type of Data		Runtime
		Certain	Ambiguous	
SCAN* [180]	ResNet18	80.9	38.34	48 + 2
IIC [80]	ResNet34	85.76	66.63	38 + 4
IIC [†] [80]	ResNet34	88.8	69.92	38 + 4
Mean-Teacher [171]	Wide ResNet28	78.577 [‡]	74.51	10
Pi [92]	Wide ResNet28	73.77 [‡]	74.03	9
Pseudo-Label [95]	Wide ResNet28	72.01 [‡]	75.14	7
FixMatch [165]	Wide ResNet28	94.83[‡]	75.97	96
FOC - Light (Ours)	ResNet50	–	75.16	4
FOC (Ours)	ResNet50	86.98	77.60	16 + 42

FixMatch outperforms FOC by a clear margin of nearly 8% while using a fifth of the labels. This performance is expected as FOC does not focus like the others on classifying [non-ambiguous STL-10] but [ambiguous plankton] data. If we look at the less curated ambiguous Plankton dataset, we see that FOC outperforms all methods by a small margin. All previous methods focus on [non-ambiguous] and curated data and we see this leads to a huge performance degeneration if they are applied to [ambiguous] data. FixMatch reaches in both datasets the best performance except for our method FOC. We conclude that the overclustering from FOC is the key for handling [ambiguous] data because it allows more flexibility during training. Previous semi-supervised methods did not consider the issue of inter- and intraobserver variability and thus are worse than FOC in classifying [ambiguous] data. If we use FOC-Light without the loss and stabilization of [80] the F1-Score drops slightly to 75% but the used GPU hours can be decreased

6.1. Fuzzy Overclustering (FOC)

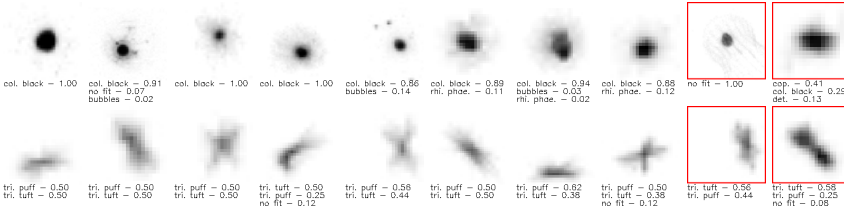


Figure 6.2. Qualitative comparison of clusters generated by FOC – The results in each row are from the same predicted cluster. The top three ambiguous labels based on the citizen scientists’ annotations are shown below the image. The last two items with the red box in each row show examples that do not match the majority of the cluster. Image and caption adapted from [149]

from 58 to 4 h.”¹ It is important to note that these results are not directly comparable to the results presented in Chapter 7 due to the fact that the ambiguous plankton dataset was generated slightly differently from later versions of the benchmark [152]. Some qualitative results are shown in Figure 6.2. “All images in a cluster are visually similar, even the probably wrongly assigned images (red box). For the images in the first row, the annotators are certain that the images belong to the same class. In the second row, annotators show a high uncertainty of assignment between the two variants of the same biological object. This illustrates the benefit of overclustering since visual similar items are in the same cluster even for uncertain [classes].”²

The analysis showed that

- ▷ the state-of-the-art SSL was better on manually cleaned benchmarks such as STL-10, but our method FOC was the best on ambiguous plankton data.
- ▷ the runtime was negatively affected by the mutual loss proposed in [80] by a factor of up to 20 while improving the results by only about 2%.
- ▷ overclustering was the key to handling ambiguous data, resulting in more consistent annotations based on these proposals.

The details can be found in [149].

¹quoted from [149] p.9+10

²quoted from [149] p.10

6. Improving data quality

6.1.1 Inverse cross-entropy (CE^{-1})

Cross-entropy is used in many supervised learning algorithms as the main loss function [22, 69, 165, 171, 202]. As noted above, a similarly well-performing loss is needed for clusters, while there are potentially more clusters than classes. While the ground-truth distribution $P(L_x = \cdot)$ may be known, no knowledge of the connection of classes to clusters can be expected. Thus, we reformulated cross-entropy which does not pull arbitrary probability distribution p, q together but pulls p and an inverse distribution of q together. This inverse distribution should ensure that dissimilar images have dissimilar distributions and was approximated by $1 - q$. The equation is given in Equation 6.1 for the overclustering case with K_o clusters and the arbitrary distributions p, q . In FOC and DC3 these distributions p, q are just the predicted distributions of a neural network over the K_o clusters. Note that $1 - q$ is no longer a probability distribution except in the binary case. However, it still works as desired in practical experiments.

$$CE^{-1}(p, q) = - \sum_{k=1}^{K_o} p(k) \cdot \ln(1 - q(k)) \quad (6.1)$$

6.2 Data-Centric Classification & Clustering (DC3)

DC3 can be seen as a generalization of FOC, which suffered from the inherited drawbacks of [80] as described above. The main idea that motivated DC3 is to apply overclustering to any common SSL algorithm. For this reason, DC3 is a modular extension of any SSL algorithm, as illustrated in Figure 6.3. Normally, a neural network Φ takes only one image x as input and outputs a classification as a soft distribution over the desired K classes with an output head of size K . DC3 adds an overclustering head $p_o(x)$ and an ambiguity estimation head $p_a(x)$. The overclustering head has more outputs than K classes and thus can produce an overclustering as in FOC. The ambiguity estimation is an output of size one that predicts for each image the probability that the given image is ambiguous or not.

6.2. Data-Centric Classification & Clustering (DC3)

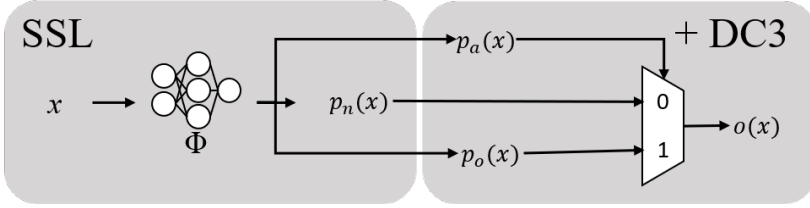


Figure 6.3. Our method DC3 and an extended arbitrary SSL method – The SSL algorithm passes an image x through the network Φ and outputs a classification $p_n(x)$. DC3 adds two additional outputs: an overclustering $p_o(x)$ and an ambiguity estimation $p_a(x)$. The ambiguity estimation $p_a(x)$ is used to determine whether to use the classification or the overclustering output. Since only a few labels are available for the classification output, most images have to be trained completely self-supervised on all outputs. Image and caption copied/adapted from [154].

Based on this estimation, DC3 can decide whether to use overclustering in the ambiguous case or to use classification.

The network is trained by a weighted sum of different loss functions. It can be roughly characterized by one loss for each of the three output heads. For the normal head $p_n(x)$, the SSL classification loss provided by the underlying vanilla method is applied. For the overclustering head $p_o(x)$, the loss CE^{-1} (see Section 6.1.1) is used for training. The negative example (from another class) is computed based on the given labels or based on pseudo-labels. The third loss is for the ambiguity estimation head $p_a(x)$.

It is important to note that during training at most one annotation per image is known and thus the ambiguity of an image cannot be estimated. The loss of the ambiguity head L_A can not be conditioned on the ambiguity of an image if this information is not known during training. If one has knowledge of a prior probability $p_A \in [0, 1]$, which is usually set to $p_A = 0.6$, one could calculate pseudo-labels $h(x)$ for each image. The probability p_A defines the expected proportion of ambiguous images in the dataset and thus in each mini-batch during training. *"The loss L_A is the binary cross-entropy between the pseudo-label $h(x)$ and $p_a(x)$. The usage of hot-encoded pseudo-labels forces the network to make more confident predictions."*

6. Improving data quality

Table 6.2. Performance across different methods and datasets for DC3 – The vanilla algorithm is highlighted in light grey. Better results in comparison to the vanilla algorithm are marked bold. The definition of the metrics are given in [154]. CE stands for supervised Cross-Entropy training. All values are given in %. H – Excluded due to hardware restrictions, copied and adopted from [154]

Methods	Plankton			Turkey			Mice Bone			CIFAR-10H			STL-10
	F1 ↑	d ↓	$(d-F1)$ ↓	F1 ↑	d ↓	$(d-F1)$ ↓	F1 ↑	d ↓	$(d-F1)$ ↓	F1 ↑	d ↓	$(d-F1)$ ↓	F1 ↑
CE	86.71	30.45	-56.26	83.84	42.98	-40.86	69.55	54.75	-14.80	67.71	55.80	-11.91	80.48
CE + DC3	78.24	23.41	-54.84	85.79	27.64	-58.14	93.88	36.58	57.30	78.27	54.52	-23.75	88.45
Mean-Teacher [171]	88.72	25.84	-62.88	81.82	45.12	-36.70	66.41	48.83	-17.58	73.53	46.93	-26.59	80.67
Mean-Teacher [171] + DC3	91.30	24.84	-66.46	86.45	33.92	-52.53	89.4	35.11	-54.73	85.13	52.44	-32.69	89.28
Pi-Model [92]	87.57	28.43	-39.14	82.11	39.46	-42.65	68.15	54.11	-14.04	71.53	49.13	-22.40	82.56
Pi-Model [92] + DC3	79.79	19.08	-60.71	87.43	23.33	-64.10	88.01	30.99	-57.02	83.05	43.40	-39.65	89.54
Pseudo-Label [95]	87.62	27.42	-60.20	82.37	44.88	-37.49	66.60	57.03	-9.57	69.70	53.30	-16.40	82.48
Pseudo-Label [95] + DC3	89.31	31.76	-57.55	83.44	35.04	-48.41	86.58	37.52	-49.06	83.74	51.32	-32.42	88.87
FixMatch [165]	85.81	30.29	-55.52	82.14	43.33	-38.81	H	H	H	78.09	41.99	-36.10	89.35
FixMatch [165] + DC3	87.20	31.28	-55.92	83.56	28.17	-55.39	H	H	H	83.09	49.49	-33.60	91.45

The formulation is given below [in Equation 6.2] with i as the index of the image x inside the given batch, when all images inside the batch are sorted in ascending order based on p_a .³

$$\begin{aligned}
 L_A(x) &= CE(h(x), p_a(x)) \\
 &= -(1 - h(x)) \cdot \ln(p_a(x = 0)) \\
 &\quad - h(x) \cdot \ln(p_a(x = 1)) \text{ with} \\
 h(x) &= \begin{cases} 1 & i \leq \text{batch size} \cdot p_A \\ 0 & \text{else} \end{cases}
 \end{aligned} \tag{6.2}$$

The original main results table from [154] is shown in Table 6.2. The metrics are defined in detail in the paper. In short, $F1$ is the macro $F1$ -score which is comparable to ACC and was evaluated only on the non-ambiguous data, if applicable. d is the inner distance of the clusters defined by the GT distribution and was evaluated only on the ambiguous cluster data, if applicable. $(d-F1)$ is the difference of both scores and was used to define a unified balanced score of both that could be used as a target for minimization. The datasets Plankton, Turkey, MiceBone and CIFAR-10H are comparable to the those reported datasets in [152] and Section 5.1. However, the data splits may differ or additional annotations may be utilized. For additional information on these variances, please refer to the

³quoted from [154] p. 6

6.2. Data-Centric Classification & Clustering (DC3)

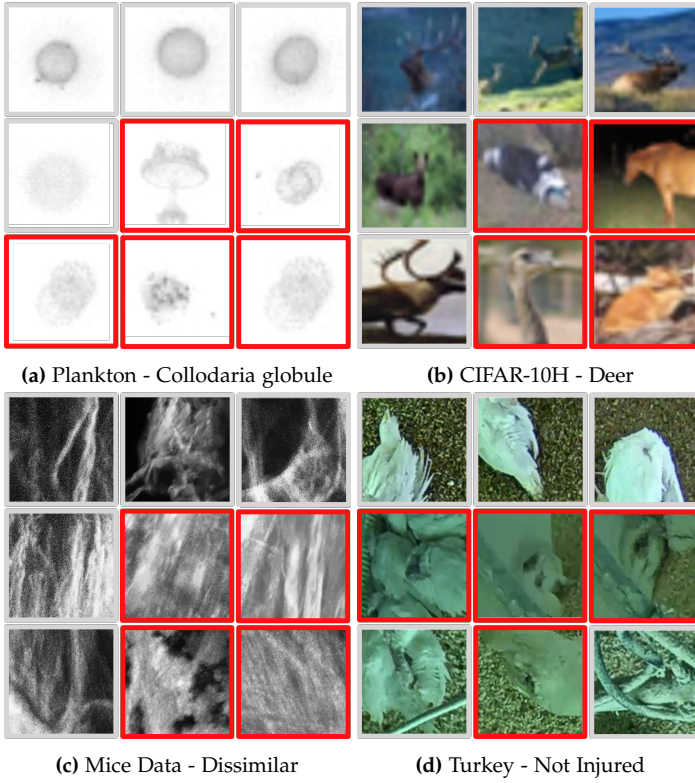


Figure 6.4. Qualitative results for selected classes over different confidences and ambiguity predictions – Incorrect classifications based on the normal head are highlighted in red. The ground-truth class is given in the caption. Graphics taken/adapted from [154]

appendix in the original paper. All evaluations are calculated on a separate validation subset. Regardless of the method or dataset, in most cases an improvement over the vanilla algorithms is shown when using the DC3 extension.

Some qualitative results for DC3 are shown in Figure 6.4. In the original paper, an image was called confident if it had a high predicted probability for the maximum class on the normal head $p_n(x)$ and a certain image if the

6. Improving data quality

ambiguity estimate $p_a(x)$ was low. “We show [9] randomly picked examples for selected classes across the datasets [...]. The images in each row have a similar value for $p_n(x)$ and $p_a(x)$. The first row presents highly confident predictions on certain predicted images and shows no errors in the given random picks. The middle row shows highly confident predictions on ambiguous predicted images. Some of these images are false and would lower the performance without the additional ambiguity prediction. The last row shows non-confident or uncertain ($0.4 < p_a(x) < 0.6$) predictions which are often wrong.”⁴ The conclusion was that ambiguity estimation leads to better insights of the model.

Key findings included that

- ▷ DC3 can be successfully applied to multiple methods and datasets.
- ▷ no hyperparameter tuning was required between training different methods and dataset.
- ▷ the combination of the individual loss terms ensured better training and avoided degeneration.
- ▷ ambiguity and confidence estimation provide a better interpretability than the confidence estimation alone.

The details can be found in [154]. A re-evaluation of the results is provided in Chapter 7.

6.3 Cost-effective labeling using validated proposals and repaired labels (CleverLabel)

As discussed in Section 3.2, the use of proposal can reduce the required annotation time while increasing the consistency. However, this is only possible because of the default effect [78] which means that people are more likely to accept the given option than without the proposal. This is generally desirable, but it also introduces a systematic bias as shown in [155].

In Section 5.4, it was discussed how the knowledge of how people annotate with proposal p_x for image x can be used to simulate with SPA

⁴quoted from [154] (supplementary) p.22+23

6.3. Cost-effective labeling using validated proposals and repaired labels (CleverLabel)

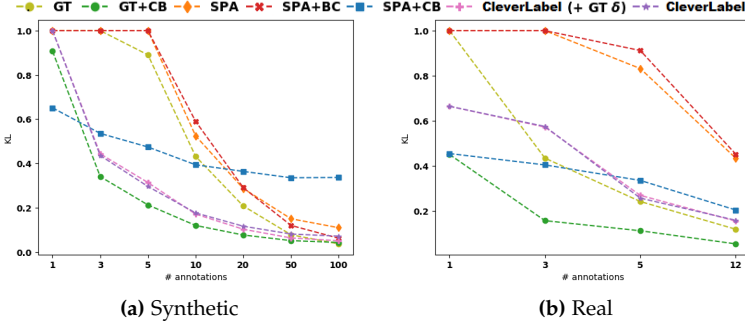


Figure 6.5. Evaluation of label improvement on synthetic data created with SPA and real user study across different amounts of annotations. Results are clamped for visualization to the range 0 to 1.

the biased distribution $P(L_b^x = \cdot)$ based on the GT distribution $P(L^x = \cdot)$. This knowledge can also be used to develop a Bias Correction (BC) and a heuristic for Class Blending (CB) to mitigate the bias as shown in [155]. The Bias Correction (BC) uses the same mathematical model as SPA, but reverses the introduced bias. The heuristic Class Blending (CB) blends the calculated distribution with a fixed class transition matrix c of the mixture of classes. For example, a cat and a dog are more likely to be confused than a cat and a bus. Both label improvement methods (BC,CB) are called together CleverLabel, even if SPA is not used to generate the biased distribution $P(L_b^x = \cdot)$. It is important to note that CB can be applied to any distribution and was used to enhance distributions without proposals. The details can be found in [155].

The original paper shows that CleverLabel can be used to improve the biased distribution $P(L_b^x = \cdot)$. However, the benefit of each parts depends on the number of annotations with proposal available. *"CB improves the results with fewer (10-) annotations [and] BC improves the results [with more] (20+)"*⁵ annotations. It could be confirmed that CleverLabel shows similar trends when applied to real annotations and synthetic ones created with SPA as shown in Figure 6.5. The reported results are the KL between the created and GT distribution on the complete dataset and *"are the median*

⁵quoted from [155] p.14

6. Improving data quality

Table 6.3. Comparison between normal and overclustering proposals (DC3) – All results are reported based on the user study executed in [154] and the distributions are enhanced with CleverLabel. The reported findings present the KL values between the generated distributions and the GT distributions across the entire dataset. The datasets utilized are the same as those in [152], but evaluations are conducted using a single-stage approach. Therefore, the reported KL metric is comparable to the KL metric presented in the benchmark study. The better results per dataset and number of annotation is marked bold. Datasets ordered with increasing δ . The expectation would be that DC3 proposals are better because they are yielding also more consistent annotations. This is not always the case which motivates the development of OSP.

Dataset	Proposal	Number of annotations			
		1	3	5	12
CIFAR10H	Normal	0.47	0.29	0.16	0.13
CIFAR10H	DC3	0.33	0.19	0.15	0.13
Turkey	Normal	0.77	0.60	0.27	0.08
Turkey	DC3	0.81	0.45	0.21	0.09
MiceBone	Normal	0.55	0.54	0.26	0.16
MiceBone	DC3	0.66	0.64	0.26	0.16
Plankton	Normal	0.90	0.97	0.59	0.40
Plankton	DC3	1.44	1.54	0.79	0.49

performance across different annotation offsets or datasets for the synthetic and real data, respectively.”⁶ The graphic also indicates “that label improvement is possible for synthetic and real data and that the combination of CB and BC with an offset of 0.1 is in most cases the strongest improvement”⁷. A major conclusion was that with a high speed-up S and the use of proposals, CleverLabel outperforms previous state-of-the-art methods and even sampling from the GT distribution directly in combination with CB. A detailed summary of the results is available in Chapter 7.

⁶quoted from [155] p. 10

⁷quoted from [155] p. 11

6.4. Overclustering Stochastic Proposal (OSP)

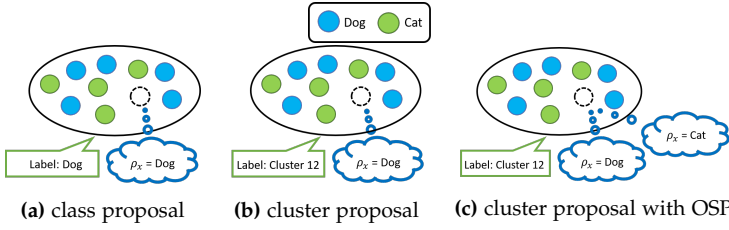


Figure 6.6. Illustration of Overclustering Stochastic Proposal (OSP) – The black bubbles illustrate the proposed classes or clusters. The green speech bubbles describe the estimated label, while the blue thought bubbles give the proposal ρ_x for the unknown image x (circle with dashed line). The blue and green dots represent data points of the class dog and cat. A class label (a) can be used directly to generate a proposal ρ_x . For a cluster (b), the class must be inferred for the proposal. A possible naive solution would be to use the majority class of the cluster. OSP uses stochastic proposals based on the distribution of classes in the cluster. For example in (c), OSP would use the class “dog” with probability $\frac{5}{9}$ and otherwise cat.

6.4 Overclustering Stochastic Proposal (OSP)

In Section 5.3, the benefits of using DC3 proposals for more consistent and faster annotations were discussed. However, as described in Chapter 4, this work aims to measure the difference between the predicted distribution and the GT distribution ($P(L_x = \cdot)$). It seems likely that more consistent annotations will lead to a better approximation of the desired distribution.

Comparing the resulting KL of normal and DC3 proposals from the human annotations in [154], an advantage of DC3 is often not visible. Even with the applied extension CleverLabel shown in Table 6.3, the use of DC3 proposals is inferior to using normal class predictions as proposals. This issue is plausible because greater consistency in the use of proposals means that the distribution is also more skewed toward that proposal. This justification is supported by the fact that the difference in proposals used increases with higher δ , and for the lowest δ the use of DC3 proposals is actually beneficial.

6. Improving data quality

This problem motivated the development of Overclustering Stochastic Proposal (OSP). If overclustering leads to faster and more consistent annotations, but clusters can be a mixture of classes, they should not be used as class proposals. To recognize the difference from class predictions, one should use probabilistic proposals based on the cluster consistency or class content. This effect is illustrated in Figure 6.6. An evaluation of this proposed method is given in Chapter 7.

6.5 Strategy for creating high-quality image annotations (SMART)

This section describes the annotation strategy proposed in [156]. The annotation strategy consists of aggregated guidelines from the literature and is illustrated in Figure 6.7. This strategy combines the result of all of my research and can be seen as the main result of this dissertation.

The **Strategy for creating high-quality iMage Annotations with human Reliability and judgement enhancemenT** (SMART) has 5 steps, which are briefly described below. For a full explanation, see the original work [156].

1. *Definition - What?* – The first phase focuses on defining the task and collecting the raw data to be annotated. In this step, it is important to define the basics of the data that will be used.
2. *Definition - Who?* – The second step answers the question of who will annotate the collected data and with what platform.
3. *Definition - How?* – This step answers the question of whether to annotate with or without proposals. The two main questions build on the results of [155].
4. *Annotation Process* – This step creates the annotations either with or without proposals. The number of annotations for each image depends on whether or not a consensus can be reached.
5. *Post-Processing* – The last step applies the proposed label improvement algorithms CleverLabel or CB from [155].

6.5. Strategy for creating high-quality image annotations (SMART)

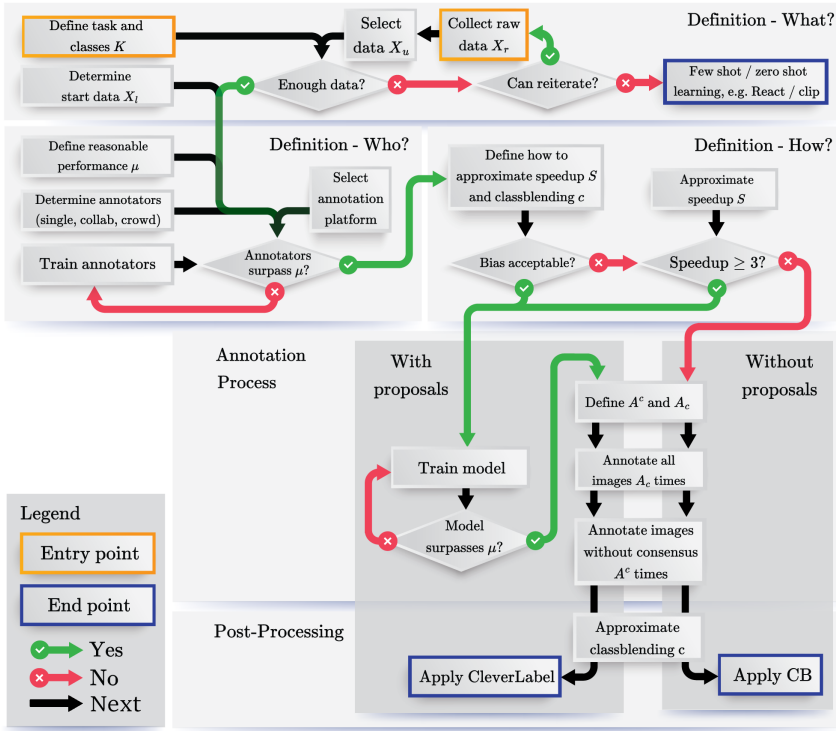


Figure 6.7. Flowchart with guidelines on how to annotate ambiguous data – The square boxes describe tasks to be performed, and the black arrows describe the next step. All square boxes pointing to a box must be completed before the box itself can be addressed. The diamonds pose questions that guide users through different paths based on yes/no questions. Best viewed in color. Copied and adapted from [156]

The strategy was verified on the Verse dataset described in Section 5.3.1 and [156]. The evaluation showed that the optimal solution would have been proposed regardless of the path chosen in Figure 6.7. A repeated and extended analysis of the main result including DC3 proposals is given in Chapter 7.

Part III

Analysis

Unified Evaluation

The evaluation chapter is divided into three sections. The goal is to present an overview of the published and new results in a unified structure that allows for a better comparison than before (see Chapter 6), where only individual methods were described. In Section 7.1, our proposed methods are compared to previous state-of-the-art. The analysis is based on up to 6450 experiments which allows a broad comparison across non-overclustering methods. Furthermore, we quantify the performance gain with respect to the budget. In Section 7.2, the benefit of using Overclustering (OC) and Overclustering Stochastic Proposal (OSP) in combination with our proposal methods is illustrated. In Section 7.3, the Strategy for creating high-quality iMage Annotations with human Reliability and judgement enhancementT (SMART) is experimentally verified on the proposed human study from Section 5.3.1.

7.1 Evaluation on benchmark

A uniform evaluation on the proposed benchmark [152] with respect to KL is shown in Table 7.1 and in more detail in Table 7.2. The GT method refers to sampling from the GT distribution as introduced in Section 5.1. This performance is an approximation of human performance on this task and can be improved with CB (see Section 6.3). The previous state-of-the-art methods (Best previous †) are independently selected as the best results over multiple methods [23, 27, 29, 34, 62, 92, 95, 96, 101, 102, 165, 171] using hard labels (see Section 5.1 and Section A.1.5 for more details about the methods). Pseudo-Labeling soft was introduced in [152] and uses Pseudo-Labels [95] but as soft labels as opposed to hard labels. DivideMix [96]+CleverLabel [155] is the main proposed method, which

7. Unified Evaluation

Table 7.1. Results overview over all datasets and different annotations for the proposed benchmark [152] – The results are given per method in each row and across different numbers of annotations (columns). The reported score is the *KL* score and is aggregated across the 10 different datasets with median or mean \pm mean of the standard deviations (STD). The best results per column except GT and ViT are bold and results better than GT+CB are italicized. A dagger (†) indicates the best previously reported result from [23, 27, 29, 34, 62, 92, 95, 96, 101, 102, 165, 171]. Median and mean individually selected over previous methods per aggregation and number of annotations.

Number annotations	1		3		10	
	Median	Mean \pm STD	Median	Mean \pm STD	Median	Mean \pm STD
GT	0.52	0.69 +- 0.08	0.34	0.37 +- 0.04	0.29	0.33 +- 0.04
GT+CB	0.44	0.60 +- 0.29	0.26	0.32 +- 0.03	0.24	0.30 +- 0.02
Best previous†	0.38	0.48 +- 0.04	0.45	0.54 +- 0.06	0.45	0.57 +- 0.10
Pseudo-Labeling soft	0.43	0.52 +- 0.07	0.37	0.39 +- 0.04	0.30	0.33 +- 0.02
DivideMix+CleverLabel	0.29	0.37 +- 0.02	0.28	0.33 +- 0.02	0.25	0.30 +- 0.02
DivideMix+CleverLabel (ViT)	N/A	N/A	0.22	0.23 +- 0.01	0.18	0.20 +- 0.01

uses the predictions of DivideMix as proposals. DivideMix was utilized because it was among the top-performing methods in [152] and employs pretrained weights from ImageNet [88], which facilitates better adaptation to previously unobserved domains. Human acceptance is simulated as in [155] with SPA and then enhanced with CleverLabel. The ViT version is the same method but uses a stronger network backbone during the second evaluation phase of the benchmark.

It is important to note that these tables only compare the number of annotations and not the budget (b) as introduced in Section 5.2.4. However, the number of annotations is the main linear factor in the calculation of b , and a comparison in terms of b is shown below. So the columns are comparable, but the comparison is slightly in favor of the proposal using techniques for one annotation because the initial setup cost is not considered, and the other way around for more annotations because the Speedup (S) is not considered.

Keeping these limitations in mind, the results reported in [152] can be confirmed, such as more annotations improve the *KL* score and soft labels make a crucial difference, as Pseudo-Labeling soft is better for at least 3 annotations than all other previously reported and evaluated methods. In

7.1. Evaluation on benchmark

Table 7.2. Unified overview about results on benchmark [152] per dataset – The results are given per method in each row and over different numbers of annotations (first row). The KL score is given for all datasets individually with an aggregation of the mean \pm mean of the standard deviations (STD) with respect to three independent repetitions. The best results per column except GT are bold and results better than GT+CB are italicized. Abbreviations: BP† \triangleq Best previously reported result from [23, 27, 29, 34, 62, 92, 95, 96, 101, 102, 165, 171], the results are independently chosen per dataset from all previous methods, PS \triangleq Pseudo-Labeling with soft labels, DM+CL \triangleq DivideMix + CleverLabel

# Annos	Benthic	CIFAR10H	MiceBone	Pig	Plankton	QualityMRI	Synthetic	Treeversity#1	Treeversity#6	Turkey
1	GT	1.17 \pm 0.04	0.41 \pm 0.02	0.55 \pm 0.06	0.75 \pm 0.05	0.34 \pm 0.02	1.73 \pm 0.48	0.08 \pm 0.01	0.49 \pm 0.04	1.02 \pm 0.03
	GT+CB	0.81 \pm 0.04	0.27 \pm 0.01	0.35 \pm 0.05	0.67 \pm 0.08	0.25 \pm 0.01	0.47 \pm 0.10	0.10 \pm 0.00	0.41 \pm 0.01	0.61 \pm 0.03
	BP†	0.70 \pm 0.01	0.28 \pm 0.02	0.29 \pm 0.01	0.56 \pm 0.01	0.24 \pm 0.03	0.12 \pm 0.02	0.05 \pm 0.00	0.41 \pm 0.01	0.42 \pm 0.01
	PS	1.00 \pm 0.08	0.41 \pm 0.02	0.40 \pm 0.08	0.70 \pm 0.06	0.32 \pm 0.06	0.62 \pm 0.16	0.10 \pm 0.01	0.46 \pm 0.01	0.83 \pm 0.13
	DM+CL	0.78 \pm 0.03	0.25 \pm 0.01	0.31 \pm 0.03	0.61 \pm 0.02	0.26 \pm 0.01	0.24 \pm 0.09	0.08 \pm 0.01	0.39 \pm 0.02	0.47 \pm 0.01
3	GT	0.81 \pm 0.04	0.30 \pm 0.02	0.27 \pm 0.03	0.56 \pm 0.02	0.23 \pm 0.02	0.39 \pm 0.19	0.05 \pm 0.00	0.37 \pm 0.01	0.50 \pm 0.02
	GT+CB	0.68 \pm 0.02	0.25 \pm 0.01	0.24 \pm 0.01	0.55 \pm 0.01	0.21 \pm 0.02	0.22 \pm 0.17	0.08 \pm 0.00	0.37 \pm 0.02	0.37 \pm 0.00
	BP†	0.68 \pm 0.01	0.28 \pm 0.01	0.30 \pm 0.04	0.65 \pm 0.08	0.27 \pm 0.02	0.24 \pm 0.12	0.13 \pm 0.01	0.44 \pm 0.01	0.51 \pm 0.02
	PS	0.76 \pm 0.06	0.41 \pm 0.02	0.29 \pm 0.03	0.63 \pm 0.03	0.33 \pm 0.08	0.21 \pm 0.10	0.07 \pm 0.00	0.41 \pm 0.02	0.47 \pm 0.02
	DM+CL	0.76 \pm 0.06	0.26 \pm 0.01	0.30 \pm 0.00	0.54 \pm 0.02	0.23 \pm 0.01	0.12 \pm 0.03	0.07 \pm 0.01	0.37 \pm 0.01	0.37 \pm 0.02
10	GT	0.75 \pm 0.03	0.26 \pm 0.02	0.23 \pm 0.02	0.54 \pm 0.04	0.20 \pm 0.02	0.31 \pm 0.14	0.04 \pm 0.00	0.36 \pm 0.02	0.33 \pm 0.01
	GT+CB	0.65 \pm 0.01	0.25 \pm 0.01	0.23 \pm 0.02	0.51 \pm 0.04	0.20 \pm 0.01	0.24 \pm 0.13	0.07 \pm 0.00	0.36 \pm 0.02	0.31 \pm 0.00
	BP†	0.75 \pm 0.05	0.28 \pm 0.02	0.31 \pm 0.05	0.57 \pm 0.01	0.27 \pm 0.03	0.16 \pm 0.03	0.20 \pm 0.09	0.45 \pm 0.01	0.60 \pm 0.09
	PS	0.70 \pm 0.02	0.43 \pm 0.02	0.22 \pm 0.02	0.61 \pm 0.07	0.25 \pm 0.01	0.12 \pm 0.04	0.06 \pm 0.01	0.37 \pm 0.02	0.35 \pm 0.01
	DM+CL	0.67 \pm 0.02	0.24 \pm 0.01	0.26 \pm 0.04	0.51 \pm 0.00	0.22 \pm 0.01	0.11 \pm 0.05	0.08 \pm 0.00	0.35 \pm 0.01	0.33 \pm 0.01

addition, DivideMix+Cleverlabel is the best of all evaluated methods as reported in [155]. For multiple annotations, the proposed baseline using GT information in combination with our proposed CB in [155] is still slightly better, if the possible Speedup (S) mentioned above is not considered. A more modern architecture such as Vision Transformer (ViT) allows further improvement with our method, highlighting that our proposed methods are combinable with future advances in deep learning.

Table 7.2 allows a comparison per dataset over different numbers of annotations. This comparison is even more unfair to DivideMix+CleverLabel because it selects the previous best method per dataset, not the best average over all datasets as in Table 7.1. Because of this unfair effect, DivideMix+CleverLabel is the best method for only 3 out of 10 datasets for an annotation, but is still better than GT+CB for 9 out of 10 datasets. Looking at more annotations (e.g. 10), we can see that DivideMix+CleverLabel is the best method 7 times out of 10, with a difference to the best method of only 0.04. A conclusion of [152] was that no method is the best over different modalities. However, DivideMix+CleverLabel is the best or one of the best methods over all datasets and number of annotations. Only

7. Unified Evaluation

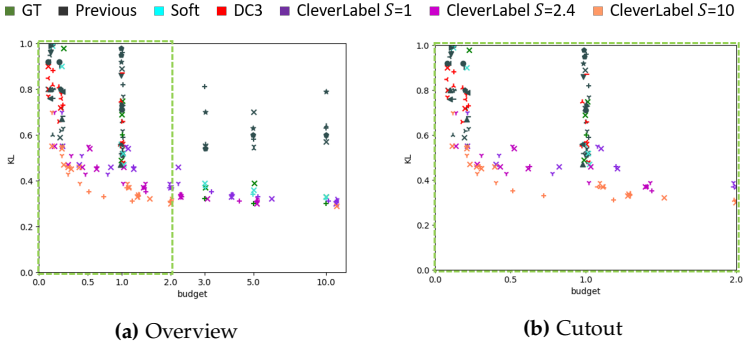


Figure 7.1. Overview of 6450 experiments taken from [152, 155] on the proposed benchmark [152] and averaged over datasets and repetitions – The x-axis represents the budget (b) and the y-axis the Kullback-Leibler divergence (KL). Each color represents a different algorithm group. “GT” describes methods that use the GT data and thus represent human performance. “Previous” describes previously reported methods using hard labels. “Soft” refers to methods that use soft labels. “DC3” describes methods using the method DC3 but not as proposals. “CleverLabel” are methods that use the method CleverLabel for correction, SPA for simulation and possibly DC3 proposals. S describes the expected Speedup. The markers represent different methods in each group. The left image shows the complete overview with a logarithmic budget scale. The right image (green striped border) depicts the budget up to 2.0 with a linear budget scale. The markers are randomized for better visualization with respect to the budget with a uniform distribution of $[-0.025, 0.025]$. Best viewed digitally and idealized version in Figure 7.2

the baseline GT+CB is consistently better than DivideMix+CleverLabel for multiple annotations and when no Speedup is considered.

The effect of Speedup (S) on budget (b) is illustrated in Figure 7.1 over 34 methods, 215 individual experiments across all 10 datasets and 3 repetitions defined in [152]. The evaluation uses the idealized Pareto front visualization from Figure 7.2 for visual interpretability. The organizing groups are described in the caption of Figure 7.1. The group “GT” methods, which can be seen as in [152] as human annotation performance, decrease approximately logarithmically with increasing budget (b). The “Soft” group performs similarly, but is slightly better at lower budgets, and vice versa at budgets roughly greater than one. “Previous” is the group of

7.1. Evaluation on benchmark

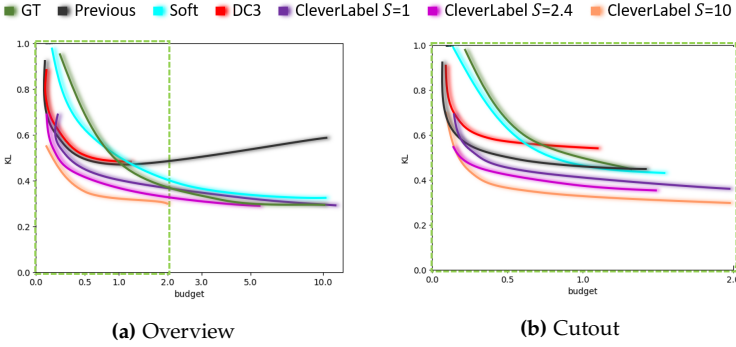


Figure 7.2. Idealized Pareto front visualizations of Figure 7.1 – Each color represents a different group of algorithms. Their meaning is described in the original figure. The idealization is created by manually inserting Pareto fronts for illustration purposes. Best viewed digitally

previously reported methods in [152] using hard labels. The performance is stronger than the groups mentioned above for a budget less than one, but for a budget larger than one the result seems to degenerate. This effect could come from more overconfident models or from fewer samples with a higher budget, resulting in a less refined Pareto front. The use of DC3 without taking advantage of the proposal potential seems to lead to worse results. See Section 7.2 for how to use overclustering proposal for better performance. The combination of CleverLabel and potentially DC3 proposals shows the best results overall. Even without an expected Speedup ($S=1$), CleverLabel outperforms all methods except “Previous” for a budget less than 0.2 and “GT” for a budget greater than 2. With a reported Speedup of 2.4 ([154]) or 10 ([158]), the results outperform all other methods. Most importantly, with the proposed method and the reported Speedup, improved performance can be achieved with a budget of less than one. “Cost is already a limiting factor when annotating data and thus only results with a better performance for a budget of less than one (which equals the current annotation [scheme] of every image once) can be expected to be applied in real world applications”.¹ This means that all methods which perform

¹quoted from [155] p.13

7. Unified Evaluation

better than the “GT” (green) group at budget one with a smaller budget are of high interest. The proposed groups utilizing CleverLabel stand out as the sole consistently present entities within this area, underscoring their importance in practical scenarios.

As mentioned above, the performance of a method with respect to KL seems to be logarithmic with increasing b . Based on the experiments from [152, 155] and their visualization in Figure 7.1, a logarithmic regression was performed for each method with the formula $kl = \log(b) \cdot \alpha + \beta$. On average, the regressed functions had a mean absolute error of 0.0370 with a standard deviation of 0.0430, with a minimum of 0.0006 and a maximum of 0.2068 across all 34 methods. The means with standard deviations for α and β are -0.0920 ± 0.1139 and 0.5032 ± 0.1921 , respectively. The conclusion is that KL is indeed logarithmic in b , but the exact values depend on the dataset and the method used.

7.2 Benefit of using Overclustering Stochastic Proposal

In this work, we found that more consistent annotations do not always lead to better downstream performance with respect to KL , as discussed in Section 6.4. This issue can also be verified on the proposed benchmark in Table 7.3 with an average of all comparable methods with and without DC3. This table shows that no significant difference can be measured between the vanilla SSL method and the DC3 extension. In this case, the predictions are used directly and not as proposals. Possible additions such as CleverLabel and OSP are not taken into account.

In Table 7.4 a comparison of Mean-Teacher where DC3 predictions are used as proposals with the enhancements CleverLabel and OSP is given for the metrics KL and \hat{KL} . The Mean-Teacher approach was employed due to its demonstrated ability to generalize well across semi-supervised methods and datasets, as evidenced by [154]. Other methods were not evaluated due to technical constraints. For both metrics, there is no measurable benefit of DC3 proposals without OSP. However, when OSP is applied, the results improve over all numbers of annotations, especially for the \hat{KL} metric. This different behavior for different metrics may be due to

7.2. Benefit of using Overclustering Stochastic Proposal

Table 7.3. Comparison on benchmark with and without DC3 averaged over all datasets and multiple methods from [152] – Results are aggregated with the median or mean \pm mean of the standard deviations (STD) over all datasets of the benchmark. It is important to note that the DC3 predictions were not used as proposals. There is no positive or negative effect of using DC3 in this comparison.

budget (b)	0.1		0.2		0.5		1.0	
	Median	Mean \pm STD	Median	Mean \pm STD	Median	Mean \pm STD	Median	Mean \pm STD
SSL	0.79	0.86 \pm 0.13	0.65	0.73 \pm 0.10	0.49	0.56 \pm 0.08	0.45	0.51 \pm 0.05
SSL+DC3	0.78	0.85 \pm 0.14	0.67	0.76 \pm 0.13	0.48	0.58 \pm 0.09	0.45	0.58 \pm 0.16

Table 7.4. Comparison of the effect of DC3, OSP and CleverLabel on Mean-Teacher (MT) [171] evaluated on the proposed benchmark [152] over different number of annotations – The first half represents the KL results and the second half the \bar{KL} results (see Section 5.2). The results are aggregated with the median or mean \pm mean of the standard deviations (STD) over all datasets of the benchmark. The best results per column are marked bold. A dagger (\dagger) means that probabilistic proposals are used for all images, not just for overclustered images.

Number annotations	1		3		10	
	Median	Mean \pm STD	Median	Mean \pm STD	Median	Mean \pm STD
MT+CleverLabel	0.34	0.37 \pm 0.04	0.28	0.34 \pm 0.03	0.27	0.32 \pm 0.04
MT+DC3+CleverLabel	0.34	0.37 \pm 0.03	0.29	0.34 \pm 0.03	0.26	0.31 \pm 0.02
MT+DC3+OSP+CleverLabel \dagger	0.39	0.39 \pm 0.04	0.27	0.33 \pm 0.02	0.25	0.31 \pm 0.02
MT+DC3+OSP+CleverLabel	0.29	0.37 \pm 0.03	0.27	0.34 \pm 0.02	0.24	0.31 \pm 0.02
MT+CleverLabel	0.81	0.81 \pm 0.04	0.38	0.35 \pm 0.01	0.18	0.16 \pm 0.01
MT+DC3+CleverLabel	0.88	0.84 \pm 0.03	0.41	0.37 \pm 0.01	0.18	0.17 \pm 0.01
MT+DC3+OSP+CleverLabel \dagger	1.05	0.95 \pm 0.03	0.40	0.37 \pm 0.01	0.18	0.17 \pm 0.01
MT+DC3+OSP+CleverLabel	0.88	0.81 \pm 0.05	0.35	0.34 \pm 0.02	0.16	0.15 \pm 0.01

saturation effects for the ResNet backbone as discussed in [155]. OSP uses probabilistic proposals for all overclustered images. The question is whether the idea of probabilistic proposals can be beneficially applied to normal class prediction. The results in Table 7.4 marked with \dagger indicate that this approach actually leads to inferior results. This work hypothesizes that this difference is due to the fact that overclustering can actually be a combination of different classes, while class predictions should only be of one class. Introducing the confidence of the network into the proposed class may introduce unnecessary and seemingly negative noise.

7. Unified Evaluation

Table 7.5. Comparison of *KL* results on the Verse data [104, 161] from [156] – The first method uses OSP, CB and BC as recommended above. The next three methods are ablations, each without one of these elements. The last three methods do not use proposals which means their annotation time is also slower in comparison than the previous four. Thus, one should not compare the first four and the last three with the same number of annotations, but with a higher number of annotations for DC3 proposals (first four), because the higher number can potentially be achieved in the same or less time. The results are given as the mean \pm standard deviation (STD). The best comparable results are in bold. The best overall results per column are in italics. Copied and adapted from [156]

method			number of annotations			
proposal	blending (CB)	correction (BC)	1	3	5	10
DC3+OSP	balanced	yes	0.4481 \pm 0.1957	0.2425 \pm 0.0741	0.1832 \pm 0.0095	0.1615 \pm 0.0129
DC3	balanced	yes	0.6804 \pm 0.4510	0.4582 \pm 0.3392	0.4150 \pm 0.1721	0.1553 \pm 0.0097
DC3+OSP	balanced	no	1.0309 \pm 0.1750	0.2704 \pm 0.0760	0.2105 \pm 0.0256	0.4546 \pm 0.4343
DC3+OSP	no	yes	1.7382 \pm 0.1553	0.2499 \pm 0.0374	0.2081 \pm 0.0171	0.2798 \pm 0.1325
no	only blends	no	0.2311 \pm 0.0224	0.2394 \pm 0.0930	0.1904 \pm 0.0209	0.1644 \pm 0.0062
no	balanced	no	0.8565 \pm 0.4466	0.1966 \pm 0.0245	0.1678 \pm 0.0224	0.1568 \pm 0.0165
no	no	no	0.5578 \pm 0.1259	0.2451 \pm 0.0127	0.5435 \pm 0.4285	0.1898 \pm 0.0380

7.3 Verification of annotation strategy

In Table 7.5 the original results of [156] are reported. It also shows the results without OSP. The best results can only be achieved with OSP, balanced blending (CB) and Bias Correction (BC), except for one outlier within 0.01 of the best result. The results without DC3 proposals achieve better results, especially for higher numbers of annotations. However, as mentioned above, the number of annotations is not exactly equal to budget (b). “For example [with] a speedup [S] [...] of 3, 5 annotations with proposal and 3 annotations without proposal, we have a lower annotation cost based on the calculations form [155] ($5 \cdot 0.33$ vs. 3) and have a lower *KL* score”.²

In [156], the impact of an introduced bias was acceptable and thus the annotations with proposals yield the best results. The original work analyzed all possible paths through the flowchart defined in Figure 6.7. “We can validate that the strategy leads to the best results even if a bias would not be acceptable. Based on the speedup of about 1.2, the recommendation would have been to not annotate with proposals and to use only balanced blending,

²quoted from [156] p.12

7.3. Verification of annotation strategy

*which is the best method for multiple annotations. Theoretically, if we had a high speedup, we could compare a higher number of annotations with proposals to fewer annotations without proposals.”*³ As described above, CleverLabel yields the best results for a Speedup of at least at the decision boundary of 3. The conclusion of this evaluation regarding SMART was “[b]ased on the results for all possible scenarios, we were able to verify the effectiveness of our strategy based on the literature for the given use case”.⁴

³quoted from [156] p.12

⁴quoted from [156] p.12

Discussion

This chapter is divided into two sections. The first section (see Section 8.1) discusses the answers to all the research questions from Section 1.3. The second section (see Section 8.2) discusses the limitations of this work and concludes future research topics based on these limitations.

8.1 Research question discussion

Reflecting on and summarizing the results presented in Chapter 7 and the individual descriptions of the methods in Chapter 6 and the human studies in Section 5.3, this dissertation comes to the following conclusions. These conclusions are structured by the corresponding research question from Section 1.3.

8.1.1 Research Question 1: “What are the characteristics of ambiguous labels in relation to other data quality issues?” (RQ1)

The definition of ambiguity for this dissertation is given in Section 2.1. It describes the inherent uncertainty in an image during the annotation process that leads to different results by the same or different people. Most importantly, this definition is distinct from noise and can be described as a form of heteroscedastic aleatoric uncertainty. In this dissertation, this ambiguity is captured by using soft labels. The idea of using soft labels, or rather a distribution over the classes, has already been introduced with label smoothing [116, 192]. However, while both techniques may have a similar effect, the soft labels of this dissertation are based on multiple

8. Discussion

annotations and are thus image dependent instead of the common label smoothing which uses a constant factor for all images.

8.1.2 Research Question 2: “How can improved data quality for image classification be quantified?” (RQ2)

As mentioned in Section 5.2 “[b]ased on the definition of the GT distribution ($P(L^x = \cdot)$), this objective can be interpreted as minimizing the difference between the network output $\Phi(x)$ and $P(L^x = \cdot)$ (both interpreted as probability distributions).” Thus, minimizing this difference is an improvement in data quality with respect to our GT distribution definition ($P(L^x = \cdot)$).

Research Question 2.1: “What data are required to quantify improved data quality?” (RQ2.1)

The target distribution $P(L^x = \cdot)$ must be known for each image, otherwise no quantization of the improvement is possible. Possible solutions to approximate $P(L^x = \cdot)$ are described in Section 2.1. The conclusion was that averaging multiple annotations per image is feasible for image classification. It is still expensive and therefore not common in published datasets. The benchmark described in Section 5.1 provides multiple datasets with the desired characteristics. For this reason, this benchmark was mainly used to analyze and verify the other presented methods and results.

Another aspect is that if proposals are to be used to answer RQ3, human studies are required. In Section 5.3 the human studies are outlined to verify or falsify assumptions with real people.

Research Question 2.2: “What metrics and algorithms are needed to quantify improved data quality?” (RQ2.2)

In Section 5.2 several metrics are discussed and the decision to use them is justified. The choice of metrics can have a significant impact on the results, as shown previously in [152]. We concluded in that paper that the focus on metrics such as ACC may be the reason for the well-known problem of overconfident models [60, 123].

8.1. Research question discussion

As noted above, human studies are needed to verify results based on human interactions with proposals. However, human studies are too expensive to be used for method development, so a simulation such as SPA is needed to allow rapid iteration during development (see Section 5.4). It is important to note that the development of SPA enabled our research into CleverLabel and ultimately SMART.

8.1.3 Research Question 3: “What are the implications of using proposals to guide the annotation process for image classification?” (RQ3)

Proposals or Pseudo-Labels are often used in modern machine learning to speed up or improve the annotation process as described in Section 3.2. The positive and negative effects are often not studied in detail. This dissertation finds that the default effect [78] can reduce annotation time and increase consistency, but introduces a bias that needs to be minimized to achieve overall improved data quality. This bias arises because annotators are more likely to annotate the proposed class, resulting in a skewed GT distribution $P(L^x = \cdot)$ towards the proposed class.

Research Question 3.1: “What positive effects can be achieved by using proposals to guide the annotation process for image classification?” (RQ3.1)

In Section 6.1 and Section 6.2, this dissertation presented two methods for generating proposals based on overclustering. The DC3 method is a model agnostic extension for most SSL models and has shown reliable performance on several datasets as shown in Section 6.2. In addition, the proposals based on overclustering have been shown to lead to higher consistency and annotation time due to the default effect. This section also shows the benefit of using proposals based on overclustering rather than just normal predictions. A conclusion of [63, 149, 154] is that overclustering allows a better separation of the data into smaller pieces that are less ambiguous and therefore can be annotated more consistently.

8. Discussion

Research Question 3.2: “How can negative effects be minimized when using proposals to guide the annotation process for image classification?” (RQ3.2)

The use of the default effect can also be seen as a drawback because it introduces a bias into the annotation process. Based on the understanding of the introduction of this bias (see SPA), two data enhancement methods (CB, BC) have been developed in Section 6.3. Together they are called CleverLabel. As shown in Chapter 7, they can improve the data quality when using proposals and thus minimize the introduced bias. In addition, it was shown that CB can also be used to enhance data quality when no proposals are used. The improvement of CleverLabel, depending on Speedup (S), can lead to an overall improved quality at a lower cost than one annotation per image.

8.1.4 Research Question 4: “How should ambiguous data be annotated based on the results of RQ1, RQ2 and RQ3?” (RQ4)

A unified strategy has been defined with SMART in Section 6.5. The evaluation in Section 7.3 and [156] showed that regardless of the decision made in SMART, the resulting path leads to the highest performance in a real-world dataset conditioned on the decisions made. The proposed strategy was developed based on our own and other research. It has been verified on another domain and thus it can be concluded that SMART provides the optimal annotation strategy for the defined use cases in [156]. The main questions that led to the development of SMART are answered below.

Research Question 4.1: “Is one annotation enough to capture the ambiguity in image classification tasks?” (RQ4.1)

No, one annotation is not enough, as shown in Chapter 7 and discussed in [152, 155]. The use of multiple annotations and thus a soft GT label leads to a better training with respect to the metrics KL , ECE and ACC . Even the selection of the best previously reported methods in Chapter 7 leads

8.2. Limitations & Future Work

often to inferior performance compared to using pseudo-labeling with soft labels. This conclusion is consistent with previous own conclusions in [63, 149] and has also been discussed by others [15, 38, 177]. In addition, the use of label smoothing is an established technique to improve the training of neural networks [106, 107, 116]. The ambiguity in the data could be a possible reason why and when it works. This is important because these questions are still open in the literature.

Research Question 4.2: “Are proposals based on classification or overclustering better for the annotation process?” (RQ4.2)

Overclustering has been reported to be more consistent compared to classifications as proposals in [149, 154] and Section 5.3. The annotation process is faster than without proposals [154]. It could be shown that more consistent results based on the overclustering proposal do not directly lead to improved data quality (see Section 6.4 and Chapter 7). However, when the proposal is varied based on the clustering content (see Section 6.4), the resulting distributions from the overclustering proposal outperform no proposals (see Section 7.2 and Section 7.3). The conclusion of this dissertation is that OSP is essential to translate the higher consistency of overclustering into higher quality image annotations.

Research Question 4.3: “How does an increased or decreased budget during the annotation process affect the resulting data quality?” (RQ4.3)

As noted in [152], performance improves with increasing b . It is important to note that this improvement is given for different neural networks and training protocols. The expected gain is approximately logarithmic in b (see Section 7.1) and depends on the given dataset, model and b . Saturation of performance is expected as b increases. Saturation can be avoided or reduced with more advanced models such as ViT.

8.2 Limitations & Future Work

All research in [149, 152, 154, 155, 156] uses various image domains, resolutions, class distributions and number of annotations, but the investigated

8. Discussion

datasets and research share common assumptions. The conclusions are expected to be generalizable to datasets and research not sharing these assumptions, but this verification is beyond the scope of this dissertation.

All my research involving annotations created by humans assume helpful and supportive annotators who annotate to the best of their ability. Malicious introduction of errors or intentionally varying performance of annotators is not considered nor prevented. Moreover, as stated in 6.5, it is essential to ensure a certain level of annotator quality to produce meaningful annotations. However, this dissertation does not investigate the impact of higher quality annotations, such as those from domain experts with more advanced skills, on the results. The datasets in [152] and [156] were annotated by a diverse group of annotators, ranging from unskilled workers to trained domain experts. Thus, it appears reasonable to suggest that higher quality annotations can also improve data quality further. However, the observed trends, methods, and strategies are applicable irrespective of the worker's skill level if a certain level of quality is maintained.

My research was limited to datasets with up to 15 000 images and up to 10 classes. The described effects and method should also work for larger datasets with million of images and thousands of classes. Based on the success of language models like GPT-4 [127, 128], a diminishing return can be expected when the dataset is large enough. The idea behind this is that while the label per image might still be wrong or inaccurate, there might be several hundred or thousand images with very similar content which were described differently. Thus, the average across multiple images is approximately the same as soft labeling a single image. For more classes, this dissertation expects a reverse trend. It is more difficult and time consuming to annotate thousands of fine-grained classes, which allows more ambiguity between classes [181]. Thus, a proposal system which allows to create consistent and fine-grained classifications will be helpful.

One issue to consider is the effect of long-tail distributions. While the used datasets can have a high class imbalance (up to nearly 90 times the smallest class), no distributions with few (< 10) labeled examples or even no labels during training were investigated. This could lead to inferior network proposals. These issues are beyond the scope of this work, but can most likely be combined with our results, since only the model training needs to be adapted to handle these rare training cases.

8.2. Limitations & Future Work

The interpretability of the network prediction has been improved by providing more meaningful confidence predictions. However, actual explanations for the computed decisions of the network as in [13, 162, 199] are not given. Due to the fact that the proposed methods improve the data quality, it should be possible to leverage this benefit for network interpretability in the future.

Overall, this topic has been analyzed with the fewest limitations possible. Research about extending and intensifying this research is left for future researchers.

Conclusion

This dissertation investigated the challenges posed by ambiguous data in the context of image classification. One of the fundamental issues in deep learning is the reliance on high-quality annotated data. However, due to the inconsistency and ambiguity of the annotators, the data itself becomes ambiguous. In line with previous research, this dissertation emphasizes the inadequacy of a single annotation per image to deal with this ambiguity. Recognizing the limitations of existing annotation approaches, this research compares the notion of ambiguity with other concepts in the literature, proposes a benchmark, a simulation, and metrics to enable further research, while describing methods to overcome these challenges. The significance of this work lies in its contribution to improving data quality with an efficient strategy to improve the data annotation process.

This dissertation is guided by the research questions (see Section 1.3). Starting with the definition of ambiguity itself, this work compares similarities and differences with other concepts reported in the literature (see RQ1 and Chapter 2). This highlights the need to determine how improved data quality can be measured (see RQ2 and Chapter 5). Based on this, the methods for improving data quality using proposals could be described (see RQ3 and Chapter 6). This research led to the definition of an annotation strategy and its verification (see RQ4, Section 6.5 and Chapter 7).

Chapter 5 describes how improved data quality can be measured. In this dissertation, improved data quality is defined as the minimization of the difference between the network output $\Phi(x)$ and the unknown GT distribution ($P(L^x = \cdot)$). This introduces the problem that the GT distribution is unknown, and a dataset with multiple annotations is required to approximate it. A multi-domain benchmark that meets this requirement is described in Section 5.1. The evaluation of human interaction with propos-

9. Conclusion

als requires human studies or a simulation of human interaction, which are described in Section 5.3 and Section 5.4.

The key idea for improving data quality is to use the proposals as a guide during the annotation process. With FOC and DC3, two methods have been introduced in Chapter 6 that provide overclustering proposals. While annotation proposals may inadvertently introduce biases, CleverLabel was proposed in Section 6.3 to reverse and minimize these biases. OSP describes how overclustering proposals can be used to improve not only data consistency but also data quality in Section 6.4. These proposed methods and their associated evaluations led to a unified Strategy for creating high-quality iMAGE Annotations with human Reliability and judgement enhancementT (SMART). This strategy describes several possible ways to optimally annotate the data for image classification based on a simple flowchart. The advantage of using this strategy is to improve data quality at the same or lower cost. This strategy combines all the aforementioned research and can be seen as the main result of this dissertation.

All proposed methods are evaluated in a unified setup on the proposed benchmark in Chapter 7. The results experimentally prove that the use of overclustering proposals generated by DC3 with OSP and enhanced with CleverLabel leads to superior results and can potentially even increase the quality while decreasing the cost. Furthermore, the analysis of SMART verified that all possible paths defined by the strategy are optimal given the specific circumstances that led to the path. Due to the fact that SMART was verified experimentally on a chosen classification problem that has not been evaluated before, it can be assumed that the strategy will also work for any other classification problem.

In conclusion, this dissertation has provided an investigation into the fundamental issues of data quality for ambiguous classification tasks. The results of this research have broad implications for the field of deep learning. By providing an effective annotation strategy for image classification tasks, one can improve the consistency and quality of annotations while reducing the time and cost associated with the annotation process. The improvements in annotation quality and efficiency offered by this dissertation have the potential to shape the way deep learning models are developed and deployed.

Part IV

Appendix

Own previous papers

In the following chapter, all original works of relevance for this work are presented in chronological order.

A.1 Long papers

A.1.1 2D and 3D Segmentation of uncertain local collagen fiber orientations in SHG microscopy

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-33676-9_26[insert DOI].

A. Own previous papers

2D and 3D Segmentation of uncertain local collagen fiber orientations in SHG microscopy

Lars Schmarje¹, Claudius Zelenka¹, Ulf Geisen², Claus-C. Glüer², and Reinhard Koch¹

¹ Multimedia Information Processing Group, Kiel University, Germany
{las,cze,rk}@informatik.uni-kiel.de

² Molecular Imaging North Competence Center, Kiel University, Germany
{ulf.geisen,glueer}@rad.uni-kiel.de

Abstract. Collagen fiber orientations in bones, visible with Second Harmonic Generation (SHG) microscopy, represent the inner structure and its alteration due to influences like cancer. While analyses of these orientations are valuable for medical research, it is not feasible to analyze the needed large amounts of local orientations manually. Since we have uncertain borders for these local orientations only rough regions can be segmented instead of a pixel-wise segmentation. We analyze the effect of these uncertain borders on human performance by a user study. Furthermore, we compare a variety of 2D and 3D methods such as classical approaches like Fourier analysis with state-of-the-art deep neural networks for the classification of local fiber orientations. We present a general way to use pretrained 2D weights in 3D neural networks, such as Inception-ResNet-3D a 3D extension of Inception-ResNet-v2. In a 10 fold cross-validation our two stage segmentation based on Inception-ResNet-3D and transferred 2D ImageNet weights achieves a human comparable accuracy.

Keywords: comparison 2D and 3D · weight transfer from 2D to 3D · osteogenesis imperfecta · second harmonic generation · uncertain borders · rough semantic segmentation

1 Introduction

In a variety of medical issues and research activities, computed tomography (CT) scans are used for bone examinations. However, most CT scans only have a resolution in the millimeter range. Special CT procedures allow resolutions of a few μm [4,17]. For this reason single collagen fiber bundles of about 2-3 μm [1] can not be detected well in CT scans. The structure and orientation of these bundles allow us to make conclusions about changes in the bone (e.g. by growth or disease) [1]. These characteristics of the inner bone structure are valuable for research in the fields of age determination, disease detection and cancer research.

Second harmonic generation (SHG) microscopy can visualize these structures of collagen due to its higher resolution. This methods allows us to generate large

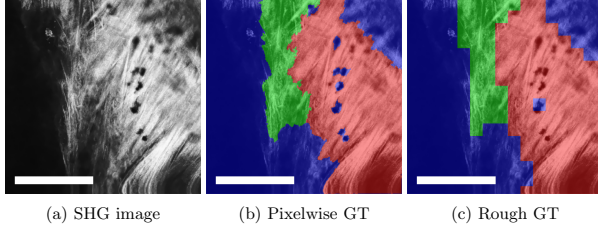


Fig. 1: Example of a SHG input image and the corresponding ground-truth - Note that while (b) is a pixel-wise ground-truth annotation we are in a setting with uncertain borders. Hence we should try to recreate a rough segmentation as shown in (c). The white scale bar depicts a size of 100 μm . Color code: Similar orientation - red, Dissimilar orientation - green, Not of interest - blue

amounts of dense 3D scans of collagen fibers in bones. It is time-consuming to create statistics of fiber bundles orientations or to mark regions of interest by experts. Moreover, manual annotations for large datasets are not feasible due to time constraints, budget and subjective biases. These biases are results of the uncertain borders in local fiber orientations. Therefore, an automatic analysis of fiber bundles orientation in large amounts of SHG data would benefit a variety of medical research activities. Large scale studies are not practical without automatic analysis.

The disease osteogenesis imperfecta (OI), also known as brittle bone disease, changes the orientation of fiber bundles in the bones of affected people and animals[1]. Hence SHG data of healthy and diseased mice is predestined for the evaluation of new methods.

Consequently, we developed different automatic algorithms for fiber bundle orientation analysis. We considered the orientation of fiber bundles in a local region as a classification problem. We focused ourselves on three classes: Bundles with similar (S) and dissimilar (D) orientations and everything else or not of interest (N) (e.g. noise, background). Figure 1 shows an example input and the corresponding ground-truth segmentation images.

The classification of a local region can also be defined as a rough semantic segmentation of the entire image. Since borders of local orientations are not well defined and can be described as highly uncertain and fluent, the goal is not to create pixel-wise segmentation. Therefore, we are interested in rough localization of these regions and their classification. Due to large regions which are not of interest (N) a large class imbalance in favor of this class exists and must be addressed. We analyzed the effect of uncertain borders on human performance in the task of rough fiber orientation segmentation by a user study.

A. Own previous papers

We investigated classical approaches like Fourier analysis and state-of-the-art methods like deep neural networks. Instead of reporting only the best results we present a complete overview and comparison for future research in rough semantic segmentation. Most of our neural networks use a state-of-the-art backbone and domain specific adaption. We aimed to change as little as possible in the original backbones to allow interchangeability with other backbones in the future. We show how a state-of-the-art 2D backbone can be used in 3D rough semantic segmentation. We call this network Inception-ResNet-3D. Especially we present a way to transfer pretrained 2D weights into the 3D case. The code is publicly available for reproducibility.¹

To sum up, the main contributions of our work are:

- We report a systematical comparison of algorithms for classification and segmentation in 2D and 3D with uncertain borders. We use a novel dense 3D SHG dataset with more than 4500 slices for method development and testing. Human performance to classify uncertain collagen fiber orientations on this dataset is also reported.
- We show a general way to convert weights from the 2D into 3D.
- Our two stage approach with Inception-ResNet-3D and transferred 2D weights achieves a performance comparable to humans for uncertain collagen fiber segmentation in a 10 fold cross-validation.

2 Related Work

Currently neural networks are state-of-the-art in the field of image data classification (e.g. ImageNet [18]). A variety of neural networks have emerged over the years [10,21,7,8,24,23]. These networks started with a simple architecture (e.g. VGG-16 [21]). They integrated new structure elements like residual [7] and inception blocks [24] as they were developed and proved their superior performance. This development led to an increase of the top1 accuracy on the ImageNet test set from 71.3% with VGG-16 to 80.3% with Inception-ResNet-v2. Parallel the depth and thereby the complexity increased from 23 to 572 layers².

Semantic segmentation gives a classification for every pixel in an image and is an extension of a classification problem. Shelhamer et al. [20] first proposed to use fully convolutional networks to solve semantic segmentation. U-Net [16] is a network for semantic segmentation which was designed for medical images. Often semantic segmentation networks consist of a down- and a upsampling part [20,16].

However, the current state-of-the-art approaches for image classification and semantic segmentation have two major drawbacks in the context of uncertain local fiber orientation classification. We have 3D data and a high uncertainty for the borders. Most research focuses on 2D data while Zhou et al. [26] showed

¹ <https://github.com/Emprime/uncertain-fiber-segmentation>

² Values are based on the reference implementation in Keras

that it is beneficial to use the 3D information for organ segmentation. Networks like PointNet [14] can classify 3D point clouds yet they do not consider dense 3D input as we have. The network 3D-U-Net [3] represents an expansion of U-Net to 3D data. It is typically used to segment 3D objects like organs [3]. This fixes the first drawback while the second one remains. Objects with uncertain borders like our fiber orientations are not well represented.

While 3D extensions of Inception-ResNet-v2 have been presented in [9,6] the usage of 2D pretraining is not so widely used. Parallel to our research Shan et al. proposed a 2D weight transfer strategy to 3D [19] which is most similar to ours (see subsection 4.1).

Collagen structures in SHG images have been analyzed in several publications [2,25,5,1,22,15,11]. They were analyzed in tissue [2] and bones [5,25]. Rao et al. [15] presented how Fourier analysis can be used to investigate the orientation of collagen fibers. The Fourier analysis was extended from small regions to the whole scan in [1,22,11]. The analysis classified small image parts as anisotropic, isotropic and dark. These classifications were used to calculate the distributions of classes over an image. In [22] these distributions were used to detect injured tendons. Moreover, Ambekar et al. [1] showed the change of distribution due to aging can be used to determine the age of pigs. Lau et al. [11] used the 3D information of SHG data and could show an increase in performance.

Nevertheless, their analysis is based on only few (<100) images. An analysis on larger amounts of data is not known. The data shown in the papers seems to be of overall of a good quality. Artifacts, noise and blurring and impact of performance was not reported.

Liang et al. [12] state to be the first to analyze SHG images with neural networks. They estimated the elastic properties of collagenous tissue. A classification or segmentation of fibers were not part of their investigation.

To our knowledge, we are the first who use neural networks to automatically classify and segment local collagen fiber orientation in large amounts of 2D or 3D data. In contrast to previous neural network literature we use 3D data and adapt our networks to uncertain borders. In comparison to earlier fiber analyses we utilize neural networks to process large amount of mixed quality data.

3 User study

While we knew that we operated in a context with uncertain borders we did not know how this would impact performance. Therefore, we investigated this issue by a random sample user study. Our goal was to examine how well humans can classify and segment local fiber orientations. We compared 15 different people with each other (interpersonal) and 5 results of the same person over time (intrapersonal).

A. Own previous papers

The participants were given two tasks. The first task was to chose one annotation out of 5 given example annotations for one image. This task was repeated for 10 different images. The second task was to create an annotation for 24 images.

For the first task we calculated the Pearson correlation coefficient between the annotation selections of all participants (interpersonal) or over time (intrapersonal). This leads to a mean absolute coefficient of 0.44 with a standard deviation of 0.26 for the intrapersonal comparison. The interpersonal comparison results in a mean absolute coefficient of 0.24 with a standard deviation of 0.2.

For the second task we calculated the accuracies of the created annotations with the ground-truth (see subsection 5.2 for the metric definition). The intrapersonal comparison reached a mean accuracy of 78.29% with a standard deviation of 2.40% over all 24 images. The interpersonal comparison resulted in a mean accuracy of 58.83% with a standard deviation of 7.44%.

All in all we see that it is more difficult for different people to select or create consistent annotations than for one person over time. However, even for a person over time the selection and creation is not perfect. We can state that humans achieve only about 78% accuracy consistency with themselves. If we train and evaluate a neural network on human created ground-truth with this consistency rate we can not expect that an algorithm performs significantly better.

4 Methods

All methods use the same datasets although the 2D methods ignore the inherent three-dimensional information. Therefore, 2D data will be referred to as scan slice or image and 3D data as scan. For all methods we investigated a variety of hyperparameters such as batchsize, backbones and loss variations. We will mention in the method description only important hyperparameter selections and specialties. For further details see the supplementary materials.

4.1 Weight transfer 2D to 3D

We want to utilize the pretrained ImageNet weights in our 3D Networks and thus we have to transfer the 2D kernel weights into 3D kernel weights. Technically this is a function $I : \mathbb{R}^{w \times h \times c} \rightarrow \mathbb{R}^{w \times h \times d \times c}$ that transforms a 3D (width, height, channels) matrix M_1 into a 4D (width, height, depth, channels) matrix M_2 with $w, h, d, c \in \mathbb{N}$.

We investigated two methods for the weight transformation. We denote the set $\{1, \dots, N\}$ by $[n]$. The first approach is to divide M_1 by d and stack them d times to create M_2 for a given depth $d \in \mathbb{N}$:

$$M_{2(i,j,k,l)} = M_{1(i,j,l)} / d \text{ for all } i \in [w], j \in [h], k \in [d], \text{ and } l \in [c]. \quad (1)$$

The second approach is to insert the 3D matrix into the 4D matrix and fill the

A.1. Long papers

6 L. Schmarje et al.

rest up with zeros. Shan et al. proposed in [19] a similar method. For given odd depth $d \in \mathbb{N}$ and center element $\hat{c} = \frac{(d-1)}{2} + 1$ this is defined as

$$M_{2(i,j,k,l)} = \begin{cases} M_{1(i,j,l)} & \text{for } k = \hat{c} \\ 0 & \text{for } k \in [d] \setminus \{\hat{c}\} \end{cases} \quad (2)$$

and all $i \in [w], j \in [h], k \in [d], m \in [c]$.

Henceforth, we refer to these transformations if we talk about 2D weights in a 3D context or transferred weights. See subsection 4.4 for further details on the selected transfer strategies. We use this weight transformations in the network Inception-Resnet-3D a 3D extension of Inception-ResNet-v2 [23]. In general the architecture is the same but with 3D layers as done before by [9,6]. In the case of asymmetric input data we introduce asymmetric strides in the downsamplings in the stem block. These asymmetric strides create symmetric input for deeper layers.

4.2 Weighted Focal Loss

As mentioned before we have to address the issue of class imbalance in our training data and chose to investigate different loss variations. Lin et al. defined the novel loss function Focal Loss in [13] which should automatically balance the contribution of classes in skewed cases. As mentioned in [13] the loss $L : [0, 1]^n \times \{0, 1\}^n \rightarrow [0, 1]$ can be extended to the non binary case as shown in Equation 3 with $n \in \mathbb{N}$ being the number of classes and γ the Focal Loss parameter. Furthermore, we integrated weights $w \in \mathbb{R}^n$ for every prediction $\hat{y} \in [0, 1]^n$ and ground-truth $y \in \{0, 1\}^n$. In our case $n = 3$ and the values of y and \hat{y} correspond to these three classes.

$$L(\hat{y}, y) = - \sum_{i \in [n]} w_i (1 - \hat{y}_i)^\gamma y_i \log(\hat{y}_i) \quad (3)$$

Keep in mind that Equation 3 is equal to cross entropy if $\gamma = 0$ and $w = \{1\}^3$.

4.3 2D methods

Fourier analysis As a reference we reimplemented a classification based on the Fourier analysis in small image patches [15]. We used thresholds like in [15] to discriminate different classes.

Classification A straight forward approach for local region classification is to create such regions by splitting the images into smaller image parts each with the same width and height. An equal class distribution was enforced on these image regions and a graphical representation of the input is given in Figure 2a. A rough segmentation for the image can be generated out of the classifications for all image parts.

A. Own previous papers

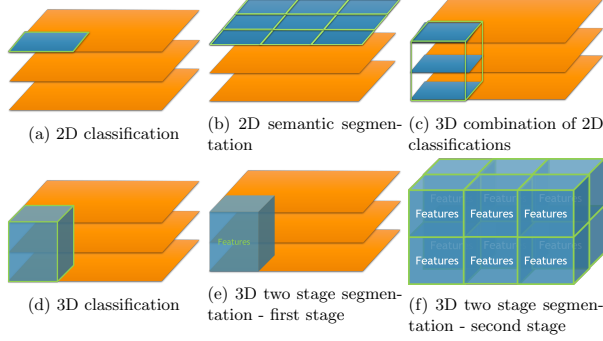


Fig. 2: Graphical representation of the used input and output for the different proposed methods - The orange slices represent the SHG images which form a scan together. The blue tiles or blocks are the inputs to the network while the green markings show the output format. For example (a) shows a 2D image part as an input and one value as output. While (f) takes features as a 3D matrix and outputs a 3D matrix.

During development we discovered that classification accuracy was higher if the image part size was larger. These larger sizes lead to a rougher segmentation and thus we used an ensemble and majority voting to combine the benefits of a smaller image part size and the higher accuracy of larger image parts. We used an Inception-ResNet-v2 [23] backbone with pretrained weights, $\gamma = 0$ and $w = \{1\}^3$.

Semantic segmentation Classification of image parts has two major drawbacks. It is time consuming since a lot of image parts have to be processed and a post processing step is needed to combine the classifications to a segmentation. A parallel rough semantic segmentation of an image can overcome both these problems. In contrast to other literature [20,16] we use only a downscaling part and not an upsampling part in our segmentation network. As described earlier we are not interested in a fine segmentation and can drop the upsampling because of this (see Figure 2b).

The segmentation networks differ from the original backbones mostly in the output. The architecture is shown in Figure 3a. We use an average pooling layer and a 1×1 convolutional layer instead of a global average pooling layer and a fully connected layer for multiple reasons. Firstly, we want to create a matrix as an output which consists of softmax outputs for every row and column. Secondly, we can incorporate the neighborhood information through the pooling layer.

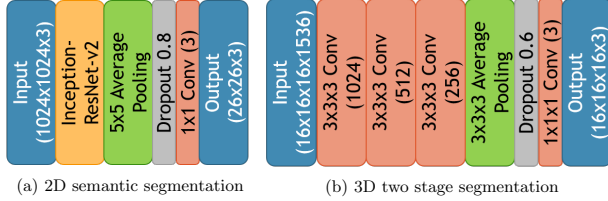


Fig. 3: Architecture of the 2D semantic segmentation network with backbone and the second stage network for the 3D two stage segmentation - Different layers have different colorings. The main part of Inception-ResNet-v2 up to the global average pooling layer is described as one block. The dimensions for in- and output and the number of feature maps for convolutional layers are given in brackets. The kernel size is given before the layer name.

Thirdly, we have to use a convolutional layer for the output because otherwise the number of parameters for the fully connected layer would have become unmanageable. In addition we can create a finer segmentation than the ensemble above (subsection 4.3) and also use neighborhood information due to average pooling layers. We used an Inception-ResNet-v2 [23] backbone with pretrained weights, $\gamma = 0$ and $w = (\frac{16}{41}, \frac{24}{41}, \frac{1}{41})$. The weights are needed to account for the class imbalance in the input data.

4.4 3D methods

Combination of 2D classifications This method is an extension of the 2D classification by combination. We use the 2D classifications and include 3D information by averaging over the classifications which are positioned next to each other. This simple aggregation yields a 3D classification but inherits the drawbacks of the 2D classification. A graphical representation is shown in Figure 2c.

Classification This method is an extension of the 2D classification in 3D. Instead of image parts (square) we use scan blocks (cuboid) for the classification. The graphical representation of the input and output is shown in Figure 2d. However, we do not use an ensemble to combine different 3D classifications. We used our proposed Inception-ResNet-3D with transferred 2D weights based on ImageNet, $\gamma = 2$, $w = \{1\}^3$ and used the transfer strategy based on Equation 1.

Two stage segmentation In the 2D case parallel segmentation of a complete image could utilize the neighborhood information for every entry in the output matrix. In order to combine the information of a whole scan for an output and still fit in the memory of one GPU we had to take a two stage approach. The idea

A. Own previous papers

is to extract features with a pretrained network and then combine these features in a second network to create a 3D matrix where every entry corresponds to the three classes (graphical representation see Figure 2e and Figure 2f).

We used Inception-ResNet-3D as an extraction network with transferred ImageNet weights with the transfer strategy based on Equation 1. Unlike in 3D classification we do not want one classification but the features as output. The second stage is a small network out of convolutional and average pooling layer to combine the 3D matrix of features to class predictions. The architecture is shown in Figure 3b and was inspired by Figure 3a.

5 Experimental Results

5.1 Dataset

We developed our methods on one dataset which was created by the MOIN CC¹. The dataset consists of 4736 SHG images from 35 scans of 6 mice where 3 mice had the disease OI and the others do not. The scans were taken on different parts of the legs and had a resolution of 1000 x 1000 px or 1024 x 1024 px while capturing 250 μm x 250 μm of the bone. We cannot downscale these images because we would lose the necessary resolution fine fiber structures. The depth of each scan was variable and ranged from 78 to 214 images while the distance in the bone between each image is 0.5 μm .

A main property of the data is the class imbalance. Roughly 2% of the data belongs to the class S (similar orientation) and 3% to the class D (dissimilar orientation). The remaining 95% belong to the class N (not of interest) and are, therefore, not interesting in medical research. The data was split into a training, validation and test set. Figure 4 displays three examples of used SHG images which represent the variety in the input data.

Moreover, investigations of selected background regions showed a high scanner noise. The average grey value (0-255) of the background should be zero but varied between 2.91 and 40.2. On average the registered grey values differ from the real values with a mean of 9.18 and a standard deviation of 8.79.

5.2 Evaluation Metrics

We use an adapted accuracy function to measure the performance of our results. In short we use the mean of accuracies per class as our accuracy measure. Our function *meanacc* for prediction $\hat{y} \in [n]^k$ and ground-truth $y \in [n]^k$ with $n \in \mathbb{N}$ number of classes and $k \in \mathbb{N}$ entries is defined as follows

$$\text{meanacc}(\hat{y}, y) = \frac{1}{n} \sum_{j=1}^n \frac{\sum_{i=1}^k \mathbb{1}_{\hat{y}_i=y_i, \hat{y}_i=j}}{\sum_{i=1}^k \mathbb{1}_{y_i=j}}. \quad (4)$$

¹ Molecular Imaging North Competence Center

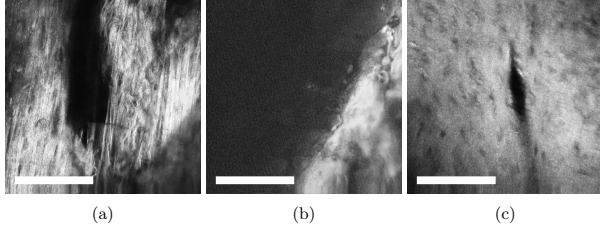


Fig. 4: Examples of used SHG images - (a) shows a desirable image. We have sharp and fine collagen structures while the noise ratio is low. (b) shows a noisy image where we can see collagen structures on the right hand side. Due to blurring and noise we can only detect rough shapes. (c) shows the macroscopic structure of the bone. We can detect the hole in the center but no single fibers or fiber bundles. The white scale bar depicts a size of 100 μm .

The function *meanacc* has the benefit of being stable against class imbalance and is the same as the normal accuracy in the case of class balance. It allows an estimation of performance in a single value without tuning weights. In this paper accuracy on our data always refers to *meanacc*.

5.3 Method comparison

We used a strict data separation during development to be able to compare methods. All hyperparameters were selected based on the validation set. The test set was only used during method comparison. In general we noticed two trends. Firstly, better network performance on ImageNet translates to improved accuracies in all our methods. This isn't noteworthy for tasks like 2D classification but for improvements in segmentation and feature extraction tasks it is. Secondly, pretrained weights ensure a good initialization and lead to greater accuracies. This is expected and reported for 2D classification tasks but the fact that pretraining can be interpolated to a 3D case and still ensures greater performance is significant.

Table 1 compares all presented methods with regard to run-time, resolution and accuracy. The Fourier analysis results in the worst performance even on the validation set. Due to this inferior performance and the long run-time we did not evaluate the method further. We believe that the high variability in the data can not be captured from such a simple approach. The method of 2D semantic segmentation has the fastest run-time and the finest resolution. The accuracy is with about 65% the best for all 2D methods. The methods 3D classification and 3D two stage segmentation score a higher accuracy but have a rougher resolution and a longer run-time. The best accuracy of 72% is achieved by 3D two stage segmentation.

A. Own previous papers

3D Segmentation of uncertain orientations of collagen fibers 11

Method	Run-time	Resolution	Accuracy
2D Fourier analysis	N/A*	16 x 16 x 1	55.00%*
2D classification	101 min	64 x 64 x 1	59.54%
3D combination of 2D classifications	101 min	64 x 64 x 16	59.68%
2D semantic segmentation	14 min	40 x 40 x 1	64.58%
3D classification	17 min	128 x 128 x 64	70.51%
3D two stage segmentation	72 min	64 x 64 x 16	72.18%

Table 1: Overview of the best results on the test set for each method - Run-time is the time it took to process the test set once which includes pre- and postprocessing. Resolution describes the number of pixels in the input that are mapped to one output value. Smaller resolutions result in finer segmentation but also in an accuracy drop for some methods and are, therefore, not reported here. The accuracy is reported on the test set. The best result of each column is marked bold. *Due to the inferior performance on the validation data and long run-time we did not evaluate the method Fourier analysis on the test set.

In general we see that it is beneficial to process as much data as possible simultaneously to achieve a high accuracy. This result can be explained due to the fact that simultaneous processing incorporates a larger neighborhood. Furthermore, we see that there is not one best method in all regards. Only a trade-off between different characteristics can be chosen.

It is remarkable that the features used in the second stage were extracted with transferred 2D ImageNet weights and still lead to the best results. Furthermore no adaption to domain specific weights is needed. In the context of a human consistency of about 78.29% in the user study (section 3) a result of 72% is remarkable.

5.4 Cross-validation

We did 10 fold cross-validation on the complete dataset to verify that the chosen random split in different sets introduced no bias and represent the real data distribution. The data was split 10 times into a training (50%), a validation (25%) and a test (25%) set. We split randomly but kept only splits where at least 2% of the data had the class S and another 2% the class D. We put scans from the same bone region into the same set.

We trained the method two stage segmentation on the training set, used the best weights on the validation set and evaluated on the test set. The mean accuracy over 10 runs is 75.79%. Figure 5 shows that the accuracies for all runs are in a margin of about 10% around the mean. Some runs achieve an accuracy above the expected accuracy based on the user study.

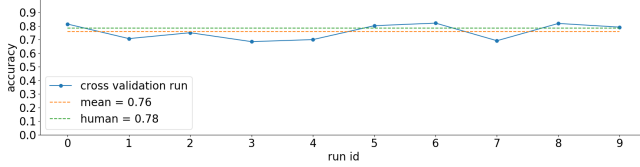


Fig. 5: Results of the cross-validation with mean and human performance based on the user study

6 Conclusion

We compared a variety of methods for rough semantic segmentation of collagen fiber orientation in 2D and 3D. As a dataset we used a novel collection of dense 3D SHG scans which is larger and more diverse as previously used datasets [1,11]. Our conducted user study implies that human can reach an average consistency of about 78.29% on the task of uncertain collagen fiber orientation segmentation. This results in a similar expected accuracy for trained algorithm due to the human annotated ground-truth. We showed how to use transformed 2D ImageNet weights in 3D networks in general and in Inception-ResNet-3D in particular. We proposed a two stage model that can simultaneously process large 3D inputs and use transformed 2D weights. This best method two stage segmentation achieves an average accuracy of 75.79% over 10 fold cross-validation. Based on the user study we can say that we created an algorithm with near human performance.

The presented user study led to great insights into possible performance of neural networks. It will be beneficial to repeat the user study at a larger scale. We are confident that two stage segmentation with transferred weights can be applied in different 3D classification and rough segmentation tasks. We will investigate these usages in the future. Furthermore, we will investigate how to create more objective ground-truth for example by leveraging pretrained features and reduced supervision.

A. Own previous papers

References

1. Ambekar, R., Chittenden, M., Jasiuk, I., Toussaint, K.C.: Quantitative second-harmonic generation microscopy for imaging porcine cortical bone: comparison to sem and its potential to investigate age-related changes. *Bone* **50** 3, 643–50 (2012)
2. Campagnola, P.J., Loew, L.M.: Second-harmonic imaging microscopy for visualizing biomolecular arrays in cells, tissues and organisms. *Nature Biotechnology* **21**, 1356–1360 (2003)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
4. Genant, H., Engelke, K., Prevrhal, S.: Advanced ct bone imaging in osteoporosis. *Rheumatology* **47**(suppl_4), iv9–iv16 (2008)
5. Genthial, R., Beaufort, E., Schanne-Klein, M.C., Peyrin, F., Farlay, D., Olivier, C., Bala, Y., Boivin, G., Vial, J.C., Débarre, D., et al.: Label-free imaging of bone multiscale porosity and interfaces using third-harmonic generation microscopy. *Scientific reports* **7**(1), 3419 (2017)
6. Hassani, B., Mahoor, M.H.: Facial expression recognition using enhanced deep 3d convolutional neural networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 2278–2288 (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)
8. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2261–2269 (2017)
9. Kang, G., Liu, K., Hou, B., Zhang, N.: 3d multi-view convolutional neural networks for lung nodule classification. *PloS one* **12**(11), e0188290 (2017)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2012)
11. Lau, T.Y., Ambekar, R., Toussaint, K.C.: Quantification of collagen fiber organization using three-dimensional fourier transform-second-harmonic generation imaging. *Optics express* **20** 19, 21821–32 (2012)
12. Liang, L., Liu, M., Sun, W.: A deep learning approach to estimate chemically-treated collagenous tissue nonlinear anisotropic stress-strain responses from microscopy images. *Acta biomaterialia* **63**, 227–235 (2017)
13. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2999–3007 (2017)
14. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 77–85 (2017)
15. Rao, R.A.R., Mehta, M.R., Toussaint, K.C.: Fourier transform-second-harmonic generation imaging of biological tissues. *Optics express* **17** 17, 14534–42 (2009)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
17. Rueckel, J., Stockmar, M.K., Pfeiffer, F., Herzen, J.: Spatial resolution characterization of a x-ray microct system. *Applied radiation and isotopes : including data,*

A.1. Long papers

14 L. Schmarje et al.

- instrumentation and methods for use in agriculture, industry and medicine **94**, 230–234 (2014)
18. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* **115**, 211–252 (2015)
19. Shan, H., Zhang, Y., Yang, Q., Kruger, U., Kalra, M.K., Sun, L., Cong, W., Wang, G.: 3-d convolutional encoder-decoder network for low-dose ct via transfer learning from a 2-d trained network. *IEEE Transactions on Medical Imaging* **37**, 1522–1534 (2018)
20. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3431–3440 (2015)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *CoRR* **abs/1409.1556** (2015)
22. Sivaguru, M., Durgam, S.S., Ambekar, R., Luedtke, D., Fried, G.A., Stewart, A.W., Toussaint, K.C.: Quantitative analysis of collagen fiber organization in injured tendons using fourier transform-second harmonic generation imaging. *Optics express* **18** **24**, 24983–93 (2010)
23. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Thirty-First AAAI Conference on Artificial Intelligence* (2017)
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1–9 (2015)
25. Thomas, B., McIntosh, D., Fildes, T., Smith, L., Hargrave, F., Islam, M., Thompson, T., Layfield, R., Scott, D., Shaw, B., et al.: Second-harmonic generation imaging of collagen in ancient bone. *Bone reports* **7**, 137–144 (2017)
26. Zhou, X., Yamada, K., Kojima, T., Takayama, R., Wang, S., Zhou, X., Hara, T., Fujita, H.: Performance evaluation of 2d and 3d deep learning approaches for automatic segmentation of multiple organs on ct images. In: *Medical Imaging 2018: Computer-Aided Diagnosis*. vol. 10575, p. 105752C. International Society for Optics and Photonics (2018)

A. Own previous papers

**A.1.2 A Survey on Semi-, Self- and Unsupervised Learning
for Image Classification**

Received April 19, 2021, accepted May 18, 2021, date of publication May 27, 2021, date of current version June 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3084358

A Survey on Semi-, Self- and Unsupervised Learning for Image Classification

LARS SCHMARJE[✉], MONTY SANTAROSSA[✉], SIMON-MARTIN SCHRÖDER[✉],
AND REINHARD KOCH[✉], (Member, IEEE)

Multimedia Information Processing Group, Kiel University, 24118 Kiel, Germany

Corresponding author: Lars Schmarje (las@informatik.uni-kiel.de)

This work was supported by Land Schleswig-Holstein through the Open Access Publikationsfonds Funding Program.

ABSTRACT While deep learning strategies achieve outstanding results in computer vision tasks, one issue remains: The current strategies rely heavily on a huge amount of labeled data. In many real-world problems, it is not feasible to create such an amount of labeled training data. Therefore, it is common to incorporate unlabeled data into the training process to reach equal results with fewer labels. Due to a lot of concurrent research, it is difficult to keep track of recent developments. In this survey, we provide an overview of often used ideas and methods in image classification with fewer labels. We compare 34 methods in detail based on their performance and their commonly used ideas rather than a fine-grained taxonomy. In our analysis, we identify three major trends that lead to future research opportunities. 1. State-of-the-art methods are scalable to real-world applications in theory but issues like class imbalance, robustness, or fuzzy labels are not considered. 2. The degree of supervision which is needed to achieve comparable results to the usage of all labels is decreasing and therefore methods need to be extended to settings with a variable number of classes. 3. All methods share some common ideas but we identify clusters of methods that do not share many ideas. We show that combining ideas from different clusters can lead to better performance.

INDEX TERMS Semi-supervised, self-supervised, unsupervised, image classification, deep learning, survey.

I. INTRODUCTION

Deep learning strategies achieve outstanding successes in computer vision tasks. They reach the best performance in a diverse range of tasks such as image classification [1]–[3], object detection [4], [5] or semantic segmentation [6], [7].

The quality of a deep neural network is strongly influenced by the number of labeled/supervised images [8]. ImageNet [1] is a huge labeled dataset with over one million images which allows the training of networks with impressive performance. Recent research shows that even larger datasets than ImageNet can improve these results [9]. However, in many real-world applications it is not possible to create labeled datasets with millions of images. A common strategy for dealing with this problem is transfer learning. This strategy improves results even on small and specialized datasets like medical imaging [10]. This might be a practical workaround for some applications but the fundamental issue

remains: Unlike humans, supervised learning needs enormous amounts of labeled data.

For a given problem we often have access to a large dataset of unlabeled data. How this unsupervised data could be used for neural networks has been of research interest for many years [11]. Xie *et al.* were among the first in 2016 to investigate unsupervised deep learning image clustering strategies to leverage this data [12]. Since then, the usage of unlabeled data has been researched in numerous ways and has created research fields like unsupervised, semi-supervised, self-supervised, weakly-supervised, or metric learning [13]. Generally speaking, unsupervised learning uses no labeled data, semi-supervised learning uses unlabeled and labeled while self-supervised learning generates labeled data on its own. Other research directions are even more different because weakly-supervised learning uses only partial information about the label and metric learning aims at learning a good distance metric. The idea that unifies these approaches is that using unlabeled data is beneficial during the training process (see Figure 1 for an illustration). It either makes the

The associate editor coordinating the review of this manuscript and approving it for publication was Varuna De Silva[✉].

A. Own previous papers

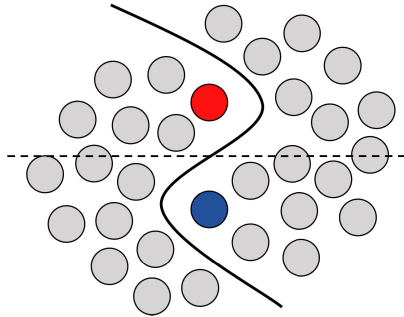


FIGURE 1. This image illustrates and simplifies the benefit of using unlabeled data during deep learning training. The red and dark blue circles represent labeled data points of different classes. The light grey circles represent unlabeled data points. If we have only a small number of labeled samples available we can only make assumptions (dotted line) over the underlying true distribution (solid line). This true distribution can only be determined if we also consider the unlabeled data points and clarify the decision boundary.

training with fewer labels more robust or in some rare cases even surpasses the supervised cases [14].

Due to this benefit, many researchers and companies work in the field of semi-, self-, and unsupervised learning. The main goal is to close the gap between semi-supervised and supervised learning or even surpass these results. Considering presented methods like [15], [16] we believe that research is at the breaking point of achieving this goal. Hence, there is a lot of research ongoing in this field. This survey provides an overview to keep track of the major and recent developments in semi-, self-, and unsupervised learning.

Most investigated research topics share a variety of common ideas while differing in goal, application contexts, and implementation details. This survey gives an overview of this wide range of research topics. The focus of this survey is on describing the similarities and differences between the methods.

Whereas we look at a broad range of learning strategies, we compare these methods only based on the image classification task. The addressed audience of this survey consists of deep learning researchers or interested people with comparable preliminary knowledge who want to keep track of recent developments in the field of semi-, self- and unsupervised learning.

A. RELATED WORK

In this subsection, we give a quick overview of previous works and reference topics we will not address further to maintain the focus of this survey.

The research of semi- and unsupervised techniques in computer vision has a long history. A variety of research, surveys, and books has been published on this topic [17]–[21].

Unsupervised cluster algorithms were researched before the breakthrough of deep learning and are still widely used [22]. There are already extensive surveys that describe unsupervised and semi-supervised strategies without deep learning [18], [23]. We will focus only on techniques including deep neural networks.

Many newer surveys focus only on self-, semi- or unsupervised learning [19], [20], [24]. Min *et al.* wrote an overview of unsupervised deep learning strategies [24]. They presented the beginning in this field of research from a network architecture perspective. The authors looked at a broad range of architectures. We focus on only one architecture which Min *et al.* refer to as “Clustering deep neural network (CDNN)-based deep clustering” [24]. Even though the work was published in 2018, it already misses the recent and major developments in deep learning of the last years. We look at these more recent developments and show the connections to other research fields that Min *et al.* did not include.

Van Engelen and Hoos give a broad overview of general and recent semi-supervised methods [20]. They cover some recent developments but deep learning strategies such as [14], [25]–[28] are not covered. Furthermore, the authors do not explicitly compare the presented methods based on their structure or performance.

Jing and Tian concentrated their survey on recent developments in self-supervised learning [19]. Like us, the authors provide a performance comparison and a taxonomy. Their taxonomy distinguishes between different kinds of pretext tasks. We look at pretext tasks as one common idea and compare the methods based on these underlying ideas. Jing and Tian look at different tasks apart from classification but do not include semi- and unsupervised methods without a pretext task.

Qi and Luo are one of the few who look at self-, semi- and unsupervised learning in one survey [29]. However, they look at the different learning strategies separately and give comparisons only inside the respective learning strategy. We show that bridging these gaps leads to new insights, improved performance, and future research approaches.

Some surveys focus not on the general overviews about semi-, self-, and unsupervised learning but special details. In their survey, Cheplygina *et al.* present a variety of methods in the context of medical image analysis [30]. They include deep learning and older machine learning approaches but look at different strategies from a medical perspective. Mey and Loog focused on the underlying theoretical assumptions in semi-supervised learning [31]. We keep our survey limited to general image classification tasks and focus on their practical application.

In this survey, we will focus on deep learning approaches for image classification. We will investigate the different learning strategies with a spotlight on loss functions. We concentrate on recent methods because older one are already adequately addressed in previous literature [17]–[21]. Keeping the above-mentioned limitations in mind, the topic of self-, semi-, and unsupervised learning still includes a broad

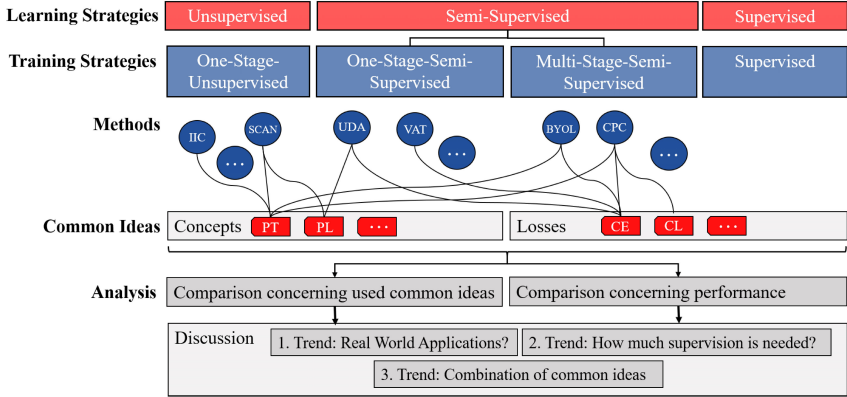


FIGURE 2. Overview of the structure of this survey – The learning strategies unsupervised, semi-supervised and supervised are commonly used in the literature. Because semi-supervised learning is incorporating many methods we defined training strategies which subdivides semi-supervised learning. For details about the training and learning strategies (including self-supervised learning) see subsection II-A. Each method belongs to one training strategy and uses several common ideas. A common idea can be a concept such as a pretext task or a loss such as cross-entropy. The definition of methods and common ideas is given in section II. Details about the common ideas are defined in subsection II-B. All methods in this survey are shortly described and categorized in section III. The methods are compared with each other based on this information concerning their used common ideas and their performance in subsection IV-C. The results of the comparisons and three resulting trends are discussed in subsection IV-D.

range of research fields. We have to exclude some related topics from this survey to keep the focus of this work for example because other research have a different aim or are evaluated on different datasets. Therefore, topics like metric learning [13] and meta learning such as [32] will be excluded. More specific networks like general adversarial networks [33] and graph networks such as [34] will be excluded. Also, other applications like pose estimation [35] and segmentation [36] or other image sources like videos or sketches [37] are excluded. Topics like few-shot or zero-shot learning methods such as [38] are excluded in this survey. However, we will see in subsection IV-D that topics like few-shot learning and semi-supervised can learn from each other in the future like in [39].

B. OUTLINE

The rest of the paper is structured in the following way. We define and explain the terms which are used in this survey such as method, training strategy and common idea in section II. A visual representation of the terms and their dependencies can be seen before the analysis part in Figure 2. All methods are presented with a short description, their training strategy and common idea in section III. In section IV, we compare the methods based their used ideas and their performance across four common image classification datasets. This section also includes a description of the datasets and evaluation metrics. Finally, we discuss the results of the comparisons in subsection IV-D and identify three

trends and research opportunities. In Figure 2, a complete overview of the structure of this survey can be seen.

II. UNDERLYING CONCEPTS

Throughout this survey, we use the terms training strategy, common idea, and method in a specific meaning. The *training strategy* is the general type/approach for using the unsupervised data during training. The training strategies are similar to the terms semi-supervised, self-supervised, or unsupervised learning but provide a definition for corner cases that the other terms do not. We will explain the differences and similarities in detail in subsection II-A. The papers we discuss in detail in this survey propose different elements like an algorithm, a general idea, or an extension of previous work. To be consistent in this survey, we call the main algorithm, idea, or extension in each paper a *method*. All methods are briefly described in section III. A method follows a training strategy and is based on several *common ideas*. We use the term common idea, or in short idea, for concepts and approaches that are shared between different methods. We roughly sort the methods based on their training strategy but compare them in detail based on the used common ideas. See subsection II-B for further information about common ideas.

In the rest of this chapter, we will use a shared definition for the following variables. For an arbitrary set of images X we define X_l and X_u with $X = X_l \cup X_u$ as the labeled and unlabeled images, respectively. For an image $x \in X_l$ the corresponding label is defined as $z_x \in Z$. An image

A. Own previous papers

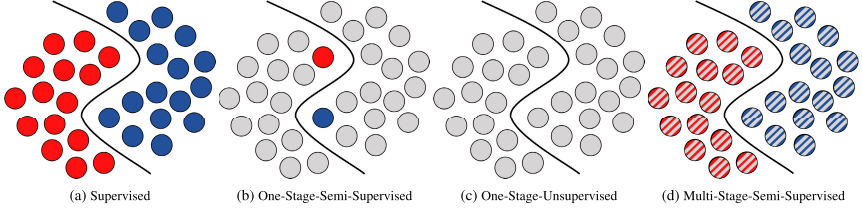


FIGURE 3. Illustrations of supervised learning (a) and the three presented reduced training strategies (b-d) - The red and dark blue circles represent labeled data points of different classes. The light grey circles represent unlabeled data points. The black lines define the underlying decision boundaries between the classes. The striped circles represent data points that do not use the label information in the first stage and can access this information in a second stage. For more details on stages and the different learning strategies see subsection II-A.

$x \in X_u$ has no label otherwise it would belong to X_l . For the distinction between X_u and X_l , only the usage of the label information during training is important. For example, an image $x \in X$ might have a label that can be used during evaluation but as long as the label is not used during training we define $x \in X_u$. The learning strategy LS_X for a dataset X is either unsupervised ($X = X_u$), supervised ($X = X_l$) or semi-supervised ($X_l \cap X_u \neq \emptyset$). During different phases of the training, different image datasets X_1, X_2, \dots, X_n with $n \in \mathbb{N}$ could be used. Two consecutive datasets X_i and X_{i+1} with $i \leq n$ and $i \in \mathbb{N}$ are different as long as different images ($X_i \neq X_{i+1}$) or different labels ($X_{l_i} \neq X_{l_{i+1}}$) are used. The learning strategy LS_i up to the dataset X_i during the training is calculated based on $X_u = \bigcup_{j=1}^i X_{u_j}$ and $X_l = \bigcup_{j=1}^i X_{l_j}$. Consecutive phases of the training are grouped into *stages*. The stage changes during consecutive datasets X_i and X_{i+1} iff the learning strategy is different ($LS_{X_i} \neq LS_{X_{i+1}}$) and the overall learning strategy changes ($LS_i \neq LS_{i+1}$). Due to this definition, only two stages can occur during training and the seven possible combinations are visualized in Figure 4. For more details see subsection II-A. Let C be the number of classes for the labels Z . For a given neural network f and input $x \in X$ the output of the neural network is $f(x)$. For the below-defined formulations, f is an arbitrary network with arbitrary weights and parameters.

A. TRAINING STRATEGIES

Terms like semi-supervised, self-supervised, and unsupervised learning are often used in literature but have overlapping definitions for certain methods. We will summarize the general understanding and definition of these terms and highlight borderline cases that are difficult to classify. Due to these borderline cases, we will define a new taxonomy based on the stages during training for a precise distinction of the methods. In subsection IV-C, we will see that this taxonomy leads to a clear clustering of the methods regarding the common ideas which further justifies this taxonomy. A visual comparison between the learning-strategies semi-supervised and unsupervised learning and the training strategies can be found in Figure 4.

Unsupervised learning describes the training without any labels. However, the goal can be a clustering (e.g. [14], [27]) or good representation (e.g. [25], [40]) of the data. Some methods combine several unsupervised steps to achieve firstly a good representation and then a clustering (e.g. [41]). In most cases, this unsupervised training is achieved by generating its own labels, and therefore the methods are called self-supervised. A counterexample for an unsupervised method without self-supervision would be k-means [22]. Often, self-supervision is achieved on a pretext task on the same or a different dataset and then the pretrained network is fine-tuned on a downstream task [19]. Many methods that follow this paradigm say their method is a form of representation learning [25], [40], [42]–[44]. In this survey, we focus on image classification, and therefore most self-supervised or representation learning methods need to fine-tune on labeled data. The combination of pretraining and fine-tuning can neither be called unsupervised nor self-supervised as external labeled information are used. Semi-supervised learning describes methods that use labeled and unlabeled data. However, semi-supervised methods like [16], [26], [45]–[49] use the labeled and unlabeled data from the beginning in comparison to representation learning methods like [25], [40], [42]–[44] which use them in different stages of their training. Some methods combine ideas from self-supervised learning, semi-supervised learning and unsupervised learning [15], [27] and are even more difficult to classify.

From the above explanation, we see that most methods are either unsupervised or semi-supervised in the context of image classification. The usage of labeled and unlabeled data in semi-supervised methods varies and a clear distinction in the common taxonomy is not obvious. Nevertheless, we need to structure the methods in some way to keep an overview, allow comparisons and acknowledge the difference of research foci. We decided against providing a fine-grained taxonomy as in previous literature [29] because we believe future research will come up with new combinations that were not thought of before. We separate the methods only based on a rough distinction when the labeled or unlabeled data is used during the training. For detailed comparisons, we distinct the

methods based on their common ideas that are defined above and described in detail in subsection II-B. We call all semi-, self-, and unsupervised (learning) strategies together *reduced supervised* (learning) strategies.

We defined *stages* above (see section II) as the different phases/time intervals during training when the different learning strategies supervised ($X = X_l$), unsupervised ($X = X_u$) or semi-supervised ($X_u \cap X_l \neq \emptyset$) are used. For example, a method that uses a self-supervised pretraining on X_u and then fine-tunes on the same images with labels has two stages. A method that uses different algorithms, losses, or datasets during the training but only uses unsupervised data X_u has one stage (e.g. [41]). A method which uses X_u and X_l during the complete training has one stage (e.g. [26]). Based on the definition of stages during training, we classify reduced supervised methods into the training strategies: One-Stage-Semi-Supervised, One-Stage-Unsupervised, and Multi-Stage-Semi-Supervised. An overview of the stage combinations and the corresponding training strategy is given in Figure 4. As we concentrate on reduced supervised learning in this survey, we will not discuss any methods which are completely supervised.

Due to the above definition of stages a fifth combination of data usage between the stages exists. This combination would use only labeled data in the first stage and unlabeled data in the second stage. In the rest of the survey, we will exclude this training strategy for the following reasons. The case that a stage of complete supervision is followed by a stage of partial or no supervision is an unusual training strategy. Due to this unusual usage, we only know of weight initialization followed by other reduced supervised training steps where this combination could occur. We see the initialization of a network with pretrained weights from a supervised training on a different dataset (e.g. Imagenet [1]) as an architectural decision. It is not part of the reduced supervised training process because it is used mainly as a more sophisticated weight initialization. If we exclude weight initialization for this reason, we know of no method which belongs to this stage.

In the following paragraphs, we will describe all other training strategies in detail and they are illustrated in Figure 3.

1) SUPERVISED LEARNING

Supervised learning is the most common strategy in image classification with deep neural networks. These methods only use labeled data X_l and its corresponding labels Z . The goal is to minimize a loss function between the output of the network $f(x)$ and the expected label $z_x \in Z$ for all $x \in X_l$.

2) ONE-STAGE-SEMI-SUPERVISED TRAINING

All methods which follow the one-stage-semi-supervised training strategy are trained in one stage with the usage of X_l , X_u , and Z . The main difference to all supervised learning strategies is the usage of the additional unlabeled data X_u . A common way to integrate the unlabeled data is to add one or more unsupervised losses to the supervised loss.

Learning Strategy	Σ	I	II	Training Strategy
Supervised	X_l	X_l	-	Supervised
Unsupervised	X_u	X_u	-	One-Stage-Unsupervised
Semi-Supervised	$X_l \cup X_u$	$X_l \cup X_u$	-	One-Stage-Semi-Supervised
Semi-Supervised	$X_l \cup X_u$	X_u	$X_l \cup X_u$	Multi-Stage-Semi-Supervised

FIGURE 4. Illustration of the different training strategies – Each row stands for a different combination of data usage during the first and second stage (defined in section II). The first column states the common learning strategy name in the literature for this usage whereas the last column states the training strategy name used in this survey. The second column represents the used data overall. The third and fourth column represent the used data in stage one or two. The blue and grey (half-) circles represent the usage of the labeled data X_l and the unlabeled data X_u respectively in each stage or overall. A minus means that no further stage is used. The dashed half circle in the last row represents that this dashed part of the data can be used.

3) ONE-STAGE-UNSUPERVISED TRAINING

All methods which follow the one-stage-unsupervised training strategy are trained in one stage with the usage of only the unlabeled samples X_u . Therefore, many authors in this training strategy call their method unsupervised. A variety of loss functions exist for unsupervised learning [12], [14], [50]. In most cases, the problem is rephrased in such a way that all inputs for the loss can be generated, e.g. reconstruction loss in autoencoders [12]. Due to this self-supervision, some call also these methods self-supervised. We want to point out one major difference to many self-supervised methods following the multi-stage-semi-supervised training strategy below. One-Stage-Unsupervised methods give image classifications without any further usage of labeled data.

4) MULTI-STAGE-SEMI-SUPERVISED TRAINING

All methods which follow the multi-stage-semi-supervised training strategy are trained in two stages with the usage of X_u in the first stage and X_l and maybe X_u in the second stage. Many methods that are called self-supervised by their authors fall into this strategy. Commonly a pretext task is used to learn representations on unlabeled data X_u . In the second stage, these representations are fine-tuned to image classification on X_l . An important difference to a one-stage method is that these methods return useable classifications only after an additional training stage.

B. COMMON IDEAS

Different common ideas are used to train models in semi-, self-, and unsupervised learning. In this section, we present a selection of these ideas that are used across multiple methods in the literature.

It is important to notice that our usage of common ideas is fuzzy and incomplete by definition. A common idea should not be an identical implementation or approximation but the underlying motivation. This fuzziness is needed for two

A. Own previous papers

reasons. Firstly, a comparison would not be possible due to so many small differences in the exact implementations. Secondly, they allow us to abstract some core elements of a method and therefore similarities can be detected. Also, not all details, concepts, and motivations are captured by common ideas. We will limit ourselves to the common ideas described below since we believe they are enough to characterize all recent methods. At the same time, we know that these ideas need to be extended in the future as new common ideas will arise, old ones will disappear, and focus will shift to other ideas. In contrast to detailed taxonomies, these new ideas can easily be integrated as new tags.

We sorted the ideas in alphabetical order and distinguish loss functions and general concepts. Since ideas might reference each other, you may have to jump to the corresponding entry if you would like to know more.

LOSS FUNCTIONS

CROSS-ENTROPY (CE)

A common loss function for image classification is cross-entropy [51]. It is commonly used to measure the difference between $f(x)$ and the corresponding label z_x for a given $x \in X$. The loss is defined in Equation 1 and the goal is to minimize the difference.

$$\begin{aligned} CE(z_x, f(x)) &= -\sum_{c=1}^C P(c|z_x) \log(P(c|f(x))) \\ &= -\sum_{c=1}^C P(c|z_x) \log(P(c|z_x)) \\ &\quad - \sum_{c=1}^C P(c|z_x) \log\left(\frac{P(c|f(x))}{P(c|z_x)}\right) \\ &= H(P(\cdot|z_x)) \\ &\quad + KL(P(\cdot|z_x) \parallel P(\cdot|f(x))) \end{aligned} \quad (1)$$

P is a probability distribution over all classes and is approximated with the (softmax-)output of the neural network $f(x)$ or the given label z_x . H is the entropy of a probability distribution and KL is the Kullback-Leibler divergence. It is important to note that cross-entropy is the sum of entropy over z_x and a Kullback-Leibler divergence between $f(x)$ and z_x . In general, the entropy $H(P(\cdot|z_x))$ is zero due to the one-hot encoded label z_x .

The loss function CE could also be used with a different probability distribution than P based on the ground-truth label. These distributions could be for example be based on Pseudo-Labels or other targets in a self-supervised pretext task. We abbreviate the used common idea with CE^* if not the ground-truth labels are used to highlight this specialty.

CONTRASTIVE LOSS (CL)

A contrastive loss tries to distinguish positive and negative pairs. The positive pair could be different views of the same image and the negative pairs could be all other pairwise combinations in a batch [25]. Hadsell *et al.* proposed to

learn representations based on contrasting [53]. In recent years, the idea has been extended by self-supervised visual representation learning methods [25], [54]–[57]. Examples of contrastive loss functions are NT-Xent [25] and InfoNCE [55] and both are based on Cross-Entropy. The loss NT-Xent is computed across all positive pairs (x_i, x_j) in a fixed subset of X with N elements e.g. a batch during training. The definition of the loss for a positive pair is given in Equation 2. The similarity sim between the outputs is measured with a normalized dot product, τ is a temperature parameter and the batch consists of N image pairs.

$$l_{x_i, x_j} = -\log \frac{\exp(\text{sim}(f(x_i), f(x_j))/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(\text{sim}(f(x_i), f(x_k))/\tau)} \quad (2)$$

Chen and Li generalize the loss NT-Xent into a broader family of loss functions with an alignment and a distribution part [58]. The alignment part encourages representations of positive pairs to be similar whereas the distribution part “encourages representations to match a prior distribution” [58]. The loss InfoNCE is motivated like other contrastive losses by maximizing the agreement / mutual information between different views. Van der Oord *et al.* showed that InfoNCE is a lower bound for the mutual information between the views [55]. More details and different bounds for other losses can be found in [59]. However, Tschannen *et al.* show evidence that these lower bounds might not be the main reason for the successes of these methods [60]. Due to this fact, we count losses like InfoNCE as a mixture of the common ideas contrastive loss and mutual information.

ENTROPY MINIMIZATION (EM)

Grandvalet and Bengio noticed that the distributions of predictions in semi-supervised learning tend to be distributed over many or all classes instead of being sharp for one or few classes [61]. They proposed to sharpen the output predictions or in other words to force the network to make more confident predictions by minimizing entropy [61]. They minimized the entropy $H(P(\cdot|f(x)))$ for a probability distribution $(P(\cdot|f(x)))$ based on a certain neural output $f(x)$ and an image $x \in X$. This minimization leads to sharper / more confident predictions. If this loss is used as the only loss the network/predictions would degenerate to a trivial minimization.

KULLBACK-LEIBLER DIVERGENCE (KL)

The Kullback-Leibler divergence is also commonly used in image classification since it can be interpreted as a part of cross-entropy. In general, KL measures the difference between two given distributions [62] and is therefore often used to define an auxiliary loss between the output $f(x)$ for an image $x \in X$ and a given secondary discrete probability distribution Q over the classes C . The definition is given in Equation 3. The second distribution could be another network output distribution, a prior known distribution, or a ground-truth distribution depending on the goal of

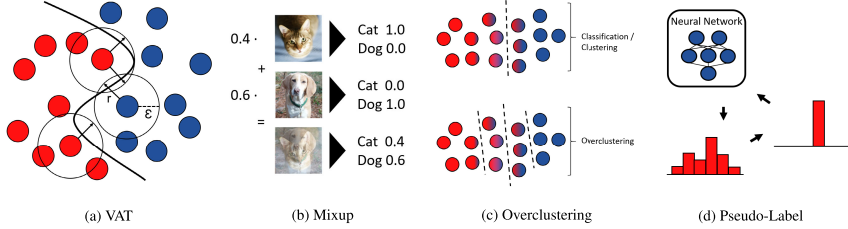


FIGURE 5. Illustration of four selected common ideas – (a) The blue and red circles represent two different classes. The line is the decision boundary between these classes. The ϵ spheres around the circles define the area of possible transformations. The arrows represent the adversarial change vector r which pushes the decision boundary away from any data point. (b) The images of a cat and a dog are combined with a parametrized blending. The labels are also combined with the same parameterization. The shown images are taken from the dataset STL-10 [52] (c) Each circle represents a data point and the coloring of the circle the ground-truth label. In this example, the images in the middle have fuzzy ground-truth labels. Classification can only draw one arbitrary decision boundary (dashed line) in the datapoints whereas overclustering can create multiple subregions. This method could also be applied to outliers rather than fuzzy labels. (d) This loop represents one version of Pseudo-Labeling. A neural network predicts an output distribution. This distribution is cast into a hard Pseudo-Label which is then used for further training the neural network.

the minimization.

$$KL(Q || P(\cdot|f(x))) = - \sum_{c=1}^C Q(c) \log \left(\frac{P(c|f(x))}{Q(c)} \right) \quad (3)$$

MEAN SQUARED ERROR (MSE)

MSE measures the Euclidean distance between two vectors e.g. two neural network outputs $f(x), f(y)$ for the images $x, y \in X$. In contrast to the loss CE or KL, MSE is not a probability measure and therefore the vectors can be in an arbitrary Euclidean feature space (see Equation 4). The minimization of the MSE will pull the two vectors or as in the example the network outputs together. Similar to the minimization of entropy, this would lead to a degeneration of the network if this loss is used as the only loss on the network outputs.

$$MSE(f(x), f(y)) = \|f(x) - f(y)\|_2^2 \quad (4)$$

MUTUAL INFORMATION (MI)

MI is defined for two probability distributions P, Q as the Kullback Leiber (KL) divergence between the joint distribution and the marginal distributions [63]. In many reduced supervised methods, the goal is to maximize the mutual information between the distributions. These distributions could be based on the input, the output, or an intermediate step of a neural network. In most cases, the conditional distribution between P and Q and therefore the joint distribution is not known. For example, we could use the outputs of a neural network $f(x), f(y)$ for two augmented views x, y of the same image as the distributions P, Q . In general, the distributions could be dependent as x, y could be identical or very similar and the distributions could be independent if x, y they are crops of distinct classes e.g. the background sky and the foreground object. Therefore, the mutual information needs to be approximated. The used approximation varies depending

on the method and the definition of the distributions P, Q . For further theoretical insights and several approximations see [59], [64].

We show the definition of the mutual information between two network outputs $f(x), f(y)$ for images $x, y \in X$ as an example in Equation 5. This equation also shows an alternative representation of mutual information: the separation in entropy $H(P(\cdot|f(x)))$ and conditional entropy $H(P(\cdot|f(x)) | P(\cdot|f(y)))$. Ji *et al.* argue that this representation illustrates the benefits of using MI over CE in unsupervised cases [14]. A degeneration is avoided because MI balances the effects of maximizing the entropy with a uniform distribution for $P(\cdot|f(x))$ and minimizing the conditional entropy by equalizing $P(\cdot|f(x))$ and $P(\cdot|f(y))$. Both cases lead to a degeneration of the neural network on their own.

$$\begin{aligned} I(P(\cdot|f(x)), P(\cdot|f(y))) &= KL(P(\cdot|f(x), f(y)) || P(\cdot|f(x) * P(\cdot|f(y)))) \\ &= \sum_{c=1}^C P(c, c'|f(x), f(y)) \\ &\quad \log \left(\frac{P(c, c'|f(x), f(y))}{P(c|f(x) * P(c'|f(y)))} \right) \\ &= H(P(\cdot|f(x)) + H(P(\cdot|f(x)) | P(\cdot|f(y))) \end{aligned} \quad (5)$$

VIRTUAL ADVERSARIAL TRAINING (VAT)

VAT [65] tries to make predictions invariant to small transformations by minimizing the distance between an image and a transformed version of the image. Miyato *et al.* showed how a transformation can be chosen and approximated in an adversarial way. This adversarial transformation maximizes the distance between an image and a transformed version of it over all possible transformations. The loss is defined in Equation 6 with an image $x \in X$ and the output of a given

A. Own previous papers

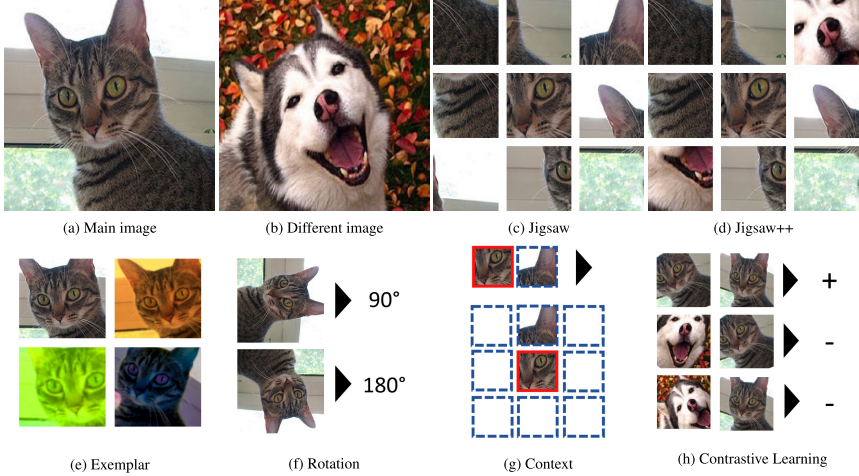


FIGURE 6. Illustrations of 8 selected pretext tasks – (a) Example image for the pretext task (b) Negative/different example image in the dataset or batch (c) The jigsaw pretext task consists of solving a simple jigsaw puzzle generated from the main image. (d) jigsaw++ augments the jigsaw puzzle by adding in parts of a different image. (e) In the exemplar pretext task, the distributions of a weakly augmented image (upper right corner) and several strongly augmented images should be aligned. (f) An image is rotated around a fixed set of rotations e.g. 0, 90, 180, and 270 degrees. The network should predict the rotation which has been applied. (g) A central patch and an adjacent patch from the same image are given. The task is to predict one of the 8 possible relative positions of the second patch to the first one. In the example, the correct answer is upper center. (h) The network receives a list of pairs and should predict the positive pairs. In this example, a positive pair consists of augmented views from the same image. Some illustrations are inspired by [40], [42], [44].

neural network $f(x)$.

$$\begin{aligned} \text{VAT}(f(x)) &= D(P(\cdot|f(x)), P(\cdot|f(x + r_{adv}))) \\ r_{adv} &= \underset{r: ||r|| \leq \epsilon}{\operatorname{argmax}} D(P(\cdot|f(x)), P(\cdot|f(x + r))) \end{aligned} \quad (6)$$

P is the probability distribution over the outputs of the neural network and D is a non-negative function that measures the distance. As illustrated in Figure 5a r is a vector and ϵ the maximum length of this vector. Two examples of used distance measures are cross-entropy [65] and Kullback-Leiber divergence [15].

CONCEPTS

MIXUP (MU)

Mixup creates convex combinations of images by blending them into each other. An illustration of the concept is given in Figure 5b. The prediction of the convex combination of the corresponding labels turned out to be beneficial because the network needs to create consistent predictions for intermediate interpolations of the image. This approach has been beneficial for supervised learning in general [66] and is therefore also used in several semi-supervised learning algorithms [26], [45], [46].

OVERCLUSTERING (OC)

Normally, if we have k classes in the supervised case we also use k clusters in the unsupervised case. Research showed that it can be beneficial to use more clusters than actual classes k exist [14], [27], [67]. We call this idea *overclustering*. Overclustering can be beneficial in semi-supervised or unsupervised cases due to the effect that neural networks can decide 'on their own' how to split the data. This separation can be helpful in noisy/fuzzy data or with intermediate classes that were sorted into adjacent classes randomly [27]. An illustration of this idea is presented in Figure 5c

PRETEXT TASK (PT)

A pretext task is a broad-ranged description of self-supervised training a neural network on a different task than the target task. This task can be for example predicting the rotation of an image [40], solving a jigsaw puzzle [43], using a contrastive loss [25], [55] or maximizing mutual information [14], [27]. An overview of most pretext task in this survey is given in Figure 6 and a complete overview is given in Table 1. In most cases the self-supervised, pretext task is used to learn representations which can then be fine-tuned for image classification [25], [40], [42]–[44],

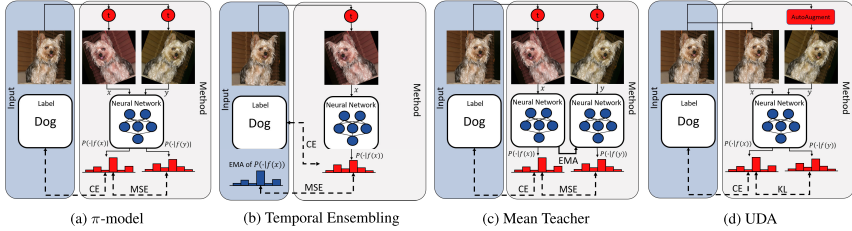


FIGURE 7. Illustration of four selected one-stage-semi-supervised methods – The used method is given below each image. The input including label information is given in the blue box on the left side. On the right side, an illustration of the method is provided. In general, the process is organized from top to bottom. At first, the input images are preprocessed by none or two different random transformations t . Special augmentation techniques like Autoaugment [69] are represented by a red box. The following neural network uses these preprocessed images (x, y) as input. The calculation of the loss $P(\cdot|f(x))$ is different for each method but shares common parts. All methods use the cross-entropy (CE) between label and predicted distribution $P(\cdot|f(x))$ on labeled examples. Details about the methods can be found in the corresponding entry in section III whereas abbreviations for common methods are defined in subsection II-B. EMA stands for the exponential moving average.

[55], [68]. In a semi-supervised context, some methods use this pretext task to define an additional loss during training [45].

PSEUDO-LABELS (PL)

A simple approach for estimating labels of unknown data is using Pseudo-Labels [47]. Lee proposed to classify unseen data with a neural network and use the predictions as labels. This process is illustrated in Figure 5d. What sounds at first like a self-fulfilling assumption works reasonably well in real-world image classification tasks. It is important to notice that the network needs additional information to prevent total random predictions. This additional information could be some known labels or a weight initialization of other supervised data or unsupervised on a pretext task. Several modern methods are based on the same core idea of creating labels by predicting them on their own [46], [48].

III. METHODS

This section shortly summarizes all methods in the survey in roughly chronological order and separated by their training strategy. Each summary states the used common ideas, explains their usage, and highlights special cases. The abbreviations for the common ideas are defined in subsection II-B. We include a large number of recent methods but we do not claim this list to be complete.

A. ONE-STAGE-SEMI-SUPERVISED

PSEUDO-LABELS

Pseudo-Labels [47] describes a common idea in deep learning and a learning method on its own. For the description of the common idea see above in subsection II-B. In contrast to many other semi-supervised methods, Pseudo-Labels does not use a combination of an unsupervised and a supervised loss. The Pseudo-Labels approach uses the predictions of a

neural network as labels for unknown data as described in the common idea. Therefore, the labeled and unlabeled data are used in parallel to minimize the CE loss. *Common ideas:* CE, CE^* , PL

π -MODEL AND TEMPORAL ENSEMBLING

Laine & Aila present two similar learning methods with the names π -model and Temporal Ensembling [49]. Both methods use a combination of the supervised CE loss and the unsupervised consistency loss MSE. The first input for the consistency loss in both cases is the output of their network from a randomly augmented input image. The second input is different for each method. In the π -model an augmentation of the same image is used. In Temporal Ensembling an exponential moving average of previous predictions is evaluated. Laine & Aila show that Temporal Ensembling is up to two times faster and more stable in comparison to the π -model [49]. Illustrations of these methods are given in Figure 7. *Common ideas:* CE, MSE

MEAN TEACHER

With Mean Teacher Tarvainen & Valpola present a student-teacher-approach for semi-supervised learning [48]. They develop their approach based on the π -model and Temporal Ensembling [49]. Therefore, they also use MSE as a consistency loss between two predictions but create these predictions differently. They argue that Temporal Ensembling incorporates new information too slowly into predictions. The reason for this is that the exponential moving average (EMA) is only updated once per epoch. Therefore, they propose to use a teacher based on the average weights of a student in each update step. Tarvainen & Valpola show for their model that the KL-divergence is an inferior consistency loss than MSE. An illustration of this method is given in Figure 7. *Common ideas:* CE, MSE

A. Own previous papers

VIRTUAL ADVERSARIAL TRAINING (VAT)

VAT [65] is not just the name for a common idea but it is also a one-stage-semi-supervised method. Miyato *et al.* use a combination of VAT on unlabeled data and CE on labeled data [65]. They showed that the adversarial transformation leads to a lower error on image classification than random transformations. Furthermore, they showed that adding EntMin [61] to the loss increased accuracy even more. *Common ideas: CE, (EM), VAT*

INTERPOLATION CONSISTENCY TRAINING (ICT)

ICT [70] uses linear interpolations of unlabeled data points to regularize the consistency between images. Verma *et al.* use a combination of the supervised loss CE and the unsupervised loss MSE. The unsupervised loss is measured between the prediction of the interpolation of two images and the interpolation of their Pseudo-Labels. The interpolation is generated with the mixup [66] algorithm from two unlabeled data points. For these unlabeled data points, the Pseudo-Labels are predicted by a Mean Teacher [48] network. *Common ideas: CE, MSE, MU, PL*

FAST-STOCHASTIC WEIGHT AVERAGING (FAST-SWA)

In contrast to other semi-supervised methods, Athiwaratun *et al.* do not change the loss but the optimization algorithm [71]. They analyzed the learning process based on ideas and concepts of SWA [72], π -model [49] and Mean Teacher [48]. Athiwaratun *et al.* show that averaging and cycling learning rates are beneficial in semi-supervised learning by stabilizing the training. They call their improved version of SWA fast-SWA due to faster convergence and lower performance variance [71]. The architecture and loss is either copied from π -model [49] or Mean Teacher [48]. *Common ideas: CE, MSE*

MixMatch

MixMatch [46] uses a combination of a supervised and an unsupervised loss. Berthelot *et al.* use CE as the supervised loss and MSE between predictions and generated Pseudo-Labels as their unsupervised loss. These Pseudo-Labels are created from previous predictions of augmented images. They propose a novel sharpening method over multiple predictions to improve the quality of the Pseudo-Labels. This sharpening also enforces implicitly a minimization of the entropy on the unlabeled data. Furthermore, they extend the algorithm mixup [66] to semi-supervised learning by incorporating the generated labels. *Common ideas: CE, (EM), MSE, MU, PL*

ENSEMBLE AutoEncoding TRANSFORMATION (EnAET)

EnAET [73] combines the self-supervised pretext task AutoEncoding Transformations [74] with MixMatch [46]. Wang *et al.* apply spatial transformations, such as translations and rotations, and non-spatial transformations, such as color distortions, on input images in the pretext task. The

transformations are then estimated with the original and augmented image given. This is a difference to other pretext tasks where the estimation is often based on the augmented image only [40]. The loss is used together with the loss of MixMatch and is extended with the Kullback Leiber divergence between the predictions of the original and the augmented image. *Common ideas: CE, (EM), KL, MSE, MU, PL, PT*

UNSUPERVISED DATA AUGMENTATION (UDA)

Xie *et al.* present with UDA a semi-supervised learning algorithm that concentrates on the usage of state-of-the-art augmentation [16]. They use a supervised and an unsupervised loss. The supervised loss is CE whereas the unsupervised loss is the Kullback Leiber divergence between output predictions. These output predictions are based on an image and an augmented version of this image. For image classification, they propose to use the augmentation scheme generated by AutoAugment [69] in combination with Cutout [75]. AutoAugment uses reinforcement learning to create useful augmentations automatically. Cutout is an augmentation scheme where randomly selected regions of the image are masked out. Xie *et al.* show that this combined augmentation method achieves higher performance in comparison to previous methods on their own like Cutout, Cropping, or Flipping. In addition to the different augmentation, they propose to use a variety of other regularization methods. They proposed Training Signal Annealing which restricts the influence of labeled examples during the training process to prevent overfitting. They use EntMin [61] and a kind of Pseudo-Labeling [47]. We use the term kind of Pseudo-Labeling because they do not use the predictions as labels but they use them to filter unsupervised data for outliers. An illustration of this method is given in Figure 7. *Common ideas: CE, EM, KL, (PL)*

SELF-PACED MULTI-VIEW CO-TRAINING (SpamCo)

Ma *et al.* propose a general framework for co-training across multiple views [76]. In the context of image classification, different neural networks can be used as different views. The main idea of the co-training between different views is similar to using Pseudo-Labels. The main differences in SpamCo are that the Pseudo-Labels are not used for all samples and they influence each other across views. Each unlabeled image has a weight value for each view. Based on an age parameter, more unlabeled images are considered in each iteration. At first only confident Pseudo-Labels are used and over time also less confident ones are allowed. The proposed hard or soft co-regularizers also influence the weighting of the unlabeled images. The regularizers encourage to select unlabeled images for training across views. Without this regularization the training would degenerate to an independent training of the different views/models. CE is used as loss on the labels and Pseudo-Labels with additional L_2 regularization. Ma *et al.* show further applications including text classification and object detection. *Common ideas: CE, CE*, MSE, PL*

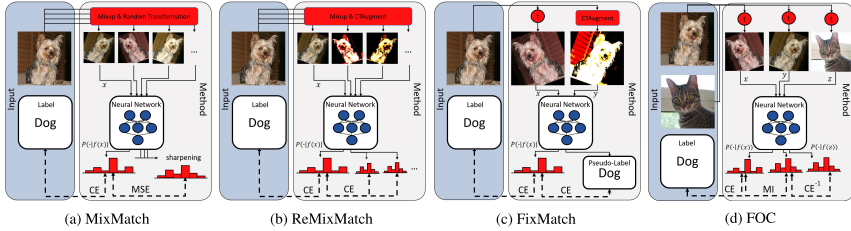


FIGURE 8. Illustration of four selected methods – The used method is given below each image. The input including label information is given in the blue box on the left side. On the right side, an illustration of the method is provided. For FOC the second stage is represented. In general, the process is organized from top to bottom. At first, the input images are preprocessed by none or two different random transformations f . Special augmentation techniques like CTaugment [45] are represented by a red box. The following neural network uses these preprocessed images (e.g. x, y) as input. The calculation of the loss (dotted line) is different for each method but shares common parts. All methods use the cross-entropy (CE) between label and predicted distribution $P(\cdot|f(x))$ on labeled examples. Details about the methods can be found in the corresponding entry in section III whereas abbreviations for common methods are defined in subsection II-B.

ReMixMatch

ReMixMatch [45] is an extension of MixMatch with distribution alignment and augmentation anchoring. Berthelot *et al.* motivate the distribution alignment with an analysis of mutual information. They use entropy minimization via “sharpening” but they do not use any prediction equalization like in mutual information. They argue that an equal distribution is also not desirable since the distribution of the unlabeled data could be skewed. Therefore, they align the predictions of the unlabeled data with a marginal class distribution over the seen examples. Berthelot *et al.* exchange the augmentation scheme of MixMatch with augmentation anchoring. Instead of averaging the prediction over different slight augmentations of an image they only use stronger augmentations as regularization. All augmented predictions of an image are encouraged to result in the same distribution with CE instead of MSE. Furthermore, a self-supervised loss based on the rotation pretext task [40] was added. *Common ideas:* CE, CE* (EM), (MI), MU, PL, PT

FixMatch

FixMatch [26] is building on the ideas of ReMixMatch but is dropping several ideas to make the framework more simple while achieving a better performance. FixMatch is using the cross-entropy loss on the supervised and the unsupervised data. For each image in the unlabeled data, one weakly- and one strongly-augmented version is created. The Pseudo-Label of the weakly-augmented version is used if a confidence threshold is surpassed by the network. If a Pseudo-Label is calculated the network output of the strongly-augmented version is compared with this hard label via cross-entropy which implicitly encourages low-entropy predictions on the unlabeled data [26]. Sohn *et al.* do not use ideas like Mixup, VAT, or distribution alignment but they state that they can be used and provide ablations for some of these extensions. *Common ideas:* CE, CE*, (EM), PL

B. MULTI-STAGE-SEMI-SUPERVISED

EXEMPLAR

Dosovitskiy *et al.* proposed a self-supervised pretext task with additional fine-tuning [68]. They randomly sample patches from different images and augment these patches heavily. Augmentations can be for example rotations, translations, color changes, or contrast adjustments. The classification task is to map all augmented versions of a patch to the correct original patch using cross-entropy loss. *Common ideas:* CE, CE*, PT

CONTEXT

Doersch *et al.* propose to use context prediction as a pretext task for visual representation learning [42]. A central patch and an adjacent patch from an image are used as input. The task is to predict one of the 8 possible relative positions of the second patch to the first one using cross-entropy loss. An illustration of the pretext task is given in Figure 6. Doersch *et al.* argue that this task becomes easier if you recognize the content of these patches. The authors fine-tune their representations for other tasks and show their superiority in comparison to the random initialization. Aside from fine-tuning, Doersch *et al.* show how their method could be used for Visual Data Mining. *Common ideas:* CE, CE*, PT

JIGSAW

Noroozi and Favaro propose to solve Jigsaw puzzles as a pretext task [43]. The idea is that a network has to understand the concept of a presented object to solve the puzzle using the classification loss cross-entropy. They prevent simple solutions that only look at edges or corners by including small random margins between the puzzle patches. They fine-tune on supervised data for image classification tasks. Noroozi *et al.* extended the Jigsaw task by adding image parts of a different image [44]. They call the extension Jigsaw++.

A. Own previous papers

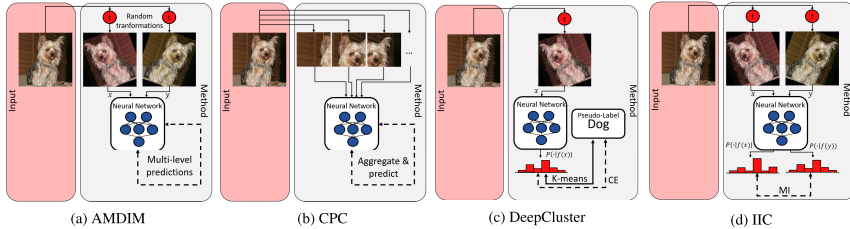


FIGURE 9. Illustration of four selected multi-stage-semi-supervised methods – The used method is given below each image. The input is given in the red box on the left side. On the right side, an illustration of the method is provided. The fine-tuning part is excluded and only the first stage/pretext task is represented. In general, the process is organized from top to bottom. At first, the input images are either preprocessed by one or two random transformations t or are split up. The following neural network uses these preprocessed images (x, y) as input. The calculation of the loss (dotted line) is different for each method. AMDIM and CPC use internal elements of the network to calculate the loss. DeepCluster and IIC use the predicted output distributions $P(-|f(x))$, $P(-|f(y))$ to calculate a loss. Details about the methods can be found in the corresponding entry in section III whereas abbreviations for common methods are defined in subsection II-B.

Examples for a Jigsaw or Jigsaw++ puzzle are given in Figure 6. *Common ideas: CE, CE*, PT*

DeepCluster

DeepCluster [67] is a self-supervised method that generates labels by k-means clustering. Caron *et al.* iterate between clustering of predicted labels to generate Pseudo-Labels and training with cross-entropy on these labels. They show that it is beneficial to use overclustering in the pretext task. After the pretext task, they fine-tune the network on all labels. An illustration of this method is given in Figure 9. *Common ideas: CE, OC, PL, PT*

ROTATION

Gidaris *et al.* use a pretext task based on image rotation prediction [40]. They propose to randomly rotate the input image by 0, 90, 180, or 270 degrees and let the network predict the chosen rotation degree. They train the network with cross-entropy on this classification task. In their work, they also evaluate different numbers of rotations but four rotations score the best result. For image classification, they fine-tune on labeled data. *Common ideas: CE, CE*, PT*

CONTRASTIVE PREDICTIVE CODING (CPC)

CPC [55], [56] is a self-supervised method that predicts representations of local image regions based on previous image regions. The authors determine the quality of these predictions with a contrastive loss which identifies the correct prediction out of randomly sampled negative ones. They call their loss InfoNCE which is cross-entropy for the prediction of positive examples [55]. Van den Oord *et al.* showed that minimizing InfoNCE maximizes the lower bound for MI between the previous image regions and the predicted image region [55]. An illustration of this method is given in Figure 9. The representations of the pretext task are then fine-tuned. *Common ideas: CE, (CE*), CL, (MI), PT*

CONSTRASTIVE MULTIVIEW CODING (CMC)

CMC [54] generalizes CPC [55] to an arbitrary collection of views. Tian *et al.* try to learn an embedding that is different for contrastive samples and equal for similar images. Like Oord *et al.* they train their network by identifying the correct prediction out of multiple negative ones [55]. However, Tian *et al.* take different views of the same image such as color channels, depth, and segmentation as similar images. For common image classification datasets like STL-10, they use patch-based similarity. After this pretext task, the representations are fine-tuned to the desired dataset. *Common ideas: CE, (CE*), CL, (MI), PT*

DEEP InfoMax (DIM)

DIM [77] maximizes the MI between local input regions and output representations. Hjelm *et al.* show that maximizing over local input regions rather than the complete image is beneficial for image classification. Also, they use a discriminator to match the output representations to a given prior distribution. In the end, they fine-tune the network with an additional small fully-connected neural network. *Common ideas: CE, MI, PT*

AUGMENTED MULTISCALE DEEP InfoMax (AMDIM)

AMDIM [78] maximizes the MI between inputs and outputs of a network. It is an extension of the method DIM [77]. DIM usually maximizes MI between local regions of an image and a representation of the image. AMDIM extends the idea of DIM in several ways. Firstly, the authors sample the local regions and representations from different augmentations of the same source image. Secondly, they maximize MI between multiple scales of the local region and the representation. They use a more powerful encoder and define mixture-based representations to achieve higher accuracies. Bachman *et al.* fine-tune the representations on labeled data to measure their quality. An illustration of this method is given in Figure 9. *Common ideas: CE, MI, PT*

DEEP METRIC TRANSFER (DMT)

DMT [79] learns a metric as a pretext task and then propagates labels onto unlabeled data with this metric. Liu *et al.* use self-supervised image colorization [80] or unsupervised instance discrimination [81] to calculate a metric. In the second stage, they propagate labels to unlabeled data with spectral clustering and then fine-tune the network with the new Pseudo-Labels. Additionally, they show that their approach is complementary to previous methods. If they use the most confident Pseudo-Labels for methods such as Mean Teacher [48] or VAT [65], they can improve the accuracy with very few labels by about 30%. *Common ideas: CE, CE*, PL, PT*

INVARIANT INFORMATION CLUSTERING (IIC)

IIC [14] maximizes the MI between augmented views of an image. The idea is that images should belong to the same class regardless of the augmentation. The augmentation has to be a transformation to which the neural network should be invariant. The authors do not maximize directly over the output distributions but over the class distribution which is approximated for every batch. Ji *et al.* use auxiliary overclustering on a different output head to increase their performance in the unsupervised case. This idea allows the network to learn subclasses and handle noisy data. Ji *et al.* use Sobel filtered images as input instead of the original RGB images. Additionally, they show how to extend IIC to image segmentation. Up to this point, the method is completely unsupervised. To be comparable to other semi-supervised methods they fine-tune their models on a subset of available labels. An illustration of this method is given in Figure 9. The first unsupervised stage can be seen as a self-supervised pretext task. In contrast to other pretext tasks, this task already predicts representations which can be seen as classifications. *Common ideas: CE, MI, OC, PT*

SELF-SUPERVISED SEMI-SUPERVISED LEARNING (S⁴L)

S⁴L [15] is, as the name suggests, a combination of self-supervised and semi-supervised methods. Zhai *et al.* split the loss into a supervised and an unsupervised part. The supervised loss is CE whereas the unsupervised loss is based on the self-supervised techniques using rotation and exemplar prediction [40], [68]. The authors show that their method performs better than other self-supervised and semi-supervised techniques [40], [47], [61], [65], [68]. In their *Mix Of All Models* (MOAM) they combine self-supervised rotation prediction, VAT, entropy minimization, Pseudo-Labels, and fine-tuning into a single model with multiple training steps. Since we discuss the results of their MOAM we identify S⁴L as a multi-stage-semi-supervised method. *Common ideas: CE, CE*, EM, PL, PT, VAT*

SIMPLE FRAMEWORK FOR CONTRASTIVE LEARNING OF VISUAL REPRESENTATION (SimCLR)

SimCLR [25] maximizes the agreement between two different augmentations of the same image. The method is similar

to CPC [55] and IIC [14]. In comparison to CPC Chen *et al.* do not use the different inner representations. Contrary to IIC they use normalized temperature-scaled cross-entropy (NT-Xent) as their loss. Based on the cosine similarity of the predictions, NT-Xent measures whether positive pairs are similar and negative pairs are dissimilar. Augmented versions of the same image are treated as positive pairs and pairs with any other image as negative pair. The system is trained with large batch sizes of up to 8192 instead of a memory bank to create enough negative examples. *Common ideas: CE, (CE*), CL, PT*

FUZZY OVERCLUSTERING (FOC)

Fuzzy Overclustering [27] is an extension of IIC [14]. FOC focuses on using overclustering to subdivide fuzzy labels in real-world datasets. Therefore, it unifies the used data and losses proposed by IIC between the different stages and extends it with new ideas such as the novel loss Inverse Cross-Entropy (CE^{-1}). This loss is inspired by Cross-Entropy but can be used on the overclustering results of the network where no ground truth labels are known. FOC is not achieving state-of-the-art results on a common image classification dataset. However, on a real-world plankton dataset with fuzzy labels, it surpasses FixMatch and shows that 5-10% more consistent predictions can be achieved. Like IIC, FOC can be viewed as a multi-stage-semi-supervised and an one-stage-unsupervised method. In general, FOC is trained in one unsupervised and one semi-supervised stage and can be seen as a multi-stage-semi-supervised method. Like IIC, it produces classifications already in the unsupervised stage and can therefore also be seen as an one-stage-unsupervised method. *Common ideas: CE, (CE*) MI, OC, PT*

MOMENTUM CONTRAST (MoCo)

He *et al.* propose to use a momentum encoder for contrastive learning [82]. In other methods [25], [55]–[57], the negative examples for the contrastive loss are sampled from the same mini-batch as the positive pair. A large batch size is needed to ensure a great variety of negative examples. He *et al.* sample their negative examples from a queue encoded by another network whose weights are updated with an exponential moving average of the main network. They solve the pretext task proposed by [81] with negative examples samples from their queue and fine-tune in a second stage on labeled data. Chen *et al.* provide further ablations and baseline for the MoCo Framework e.g. by using a MLP head for fine-tuning [83]. *Common ideas: CE, CL, PT*

BOOTSTRAP YOUR OWN LATENT (BYOL)

Grill *et al.* use an online and a target network. In the proposed pretext task, the online network predicts the image representation of the target network for an image [28]. The difference between the predictions is measured with MSE. Normally, this approach would lead to a degeneration of the network as a constant prediction over all images would also

A. Own previous papers

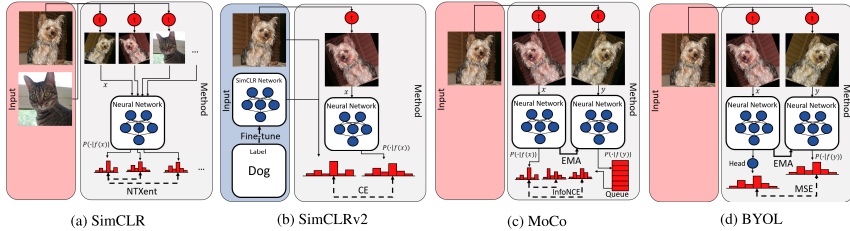


FIGURE 10. Illustration of four selected multi-stage-semi-supervised methods – The used method is given below each image. The input is given in the red (not using labels) or blue (using labels) box on the left side. On the right side, an illustration of the method is provided. The fine-tuning part is excluded and only the first stage/pretext task is represented. For SimCLRv2 the second stage or distillation step is illustrated. In general, the process is organized from top to bottom. At first, the input images are either preprocessed by one or two random transformations t or are split up. The following neural network uses these preprocessed images (x, y) as input. Details about the methods can be found in the corresponding entry in section III whereas abbreviations for common methods are defined in subsection II-B. EMA stands for the exponential moving average.

achieve the goal. In contrastive learning, this degeneration is avoided by selecting a positive pair of examples from multiple negative ones [25], [55]–[57], [82], [83]. By using a slow-moving average of the weights between the online and target network, Grill *et al.* show empirically that the degeneration to a constant prediction can be avoided. This approach has the positive effect that BYOL performance is depending less on hyperparameters like augmentation and batch size [28]. In a follow-up work, Richmond *et al.* show that BYOL even works when no batch normalization which might have introduced kind of a contrastive learning effect in the batches is used [84]. *Common ideas:* MSE, PT

SIMPLE FRAMEWORK FOR CONTRASTIVE LEARNING OF VISUAL REPRESENTATION (SimCLRv2)

Chen *et al.* extend the framework SimCLR by using larger and deeper networks and by incorporating the memory mechanism from MoCo [57]. Moreover, they propose to use this framework in three steps. The first is training a contrastive learning pretext task with a deep neural network and the SimCLRv2 method. The second step is fine-tuning this large network with a small amount of labeled data. The third step is self-training or distillation. The large pretrained network is used to predict Pseudo-Labels on the complete (unlabeled) data. These (soft) Pseudo-Labels are then used to train a smaller neural network with CE. The distillation step could be also performed on the same network as in the pretext task. Chen *et al.* show that even this self-distillation leads to performance improvements [57]. *Common ideas:* CE, (CE*), CL, PL, PT

C. ONE-STAGE-UNSUPERVISED

DEEP ADAPTIVE IMAGE CLUSTERING (DAC)

DAC [50] reformulates unsupervised clustering as a pairwise classification. Similar to the idea of Pseudo-Labels Chang *et al.* predict clusters and use these to retrain the network. The twist is that they calculate the cosine dis-

tance between all cluster predictions. This distance is used to determine whether the input images are similar or dissimilar with a given certainty. The network is then trained with binary CE on these certain similar and dissimilar input images. One can interpret these similarities and dissimilarities as Pseudo-Labels for the similarity classification task. During the training process, they lower the needed certainty to include more images. As input Chang *et al.* use a combination of RGB and extracted HOG features. *Common ideas:* PL

INFORMATION MAXIMIZING SELF-AUGMENTED TRAINING (IMSAT)

IMSAT [85] maximizes MI between the input and output of the model. As a consistency regularization Hu *et al.* use CE between an image prediction and an augmented image prediction. They show that the best augmentation of the prediction can be calculated with VAT [65]. The maximization of MI directly on the image input leads to a problem. For datasets like CIFAR-10, CIFAR-100 [86] and STL-10 [52] the color information is too dominant in comparison to the actual content or shape. As a workaround, Hu *et al.* use the features generated by a pretrained CNN on ImageNet [1] as input. *Common ideas:* MI, VAT

INVARIANT INFORMATION CLUSTERING (IIC)

IIC [14] is described above as a multi-stage-semi-supervised method. In comparison to other presented methods, IIC creates usable classifications without fine-tuning the model on labeled data. The reason for this is that the pretext task is constructed in such a way that label predictions can be extracted directly from the model. This leads to the conclusion that IIC can also be interpreted as an unsupervised learning method. *Common ideas:* MI, OC

FUZZY OVERCLUSTERING (FOC)

FOC [27] is described above as a multi-stage-semi-supervised method. Like IIC, FOC can also be seen as an one-



FIGURE 11. Examples of four random cats in the different datasets to illustrate the difference in quality.

stage-unsupervised method because the first stage yields cluster predictions. *Common ideas: MI, OC*

SEMANTIC CLUSTERING BY ADOPTING NEAREST NEIGHBORS (SCAN)

Gansbeke *et al.* calculate clustering assignments building on self-supervised pretext task by mining the nearest neighbors and using self-labeling. They propose to use SimCLR [25] as a pretext task but show that other pretext tasks [40], [81] could also be used for this step. For each sample, the k nearest neighbors are selected in the gained feature space. The novel semantic clustering loss encourages these samples to be in the same cluster. Gansbeke *et al.* noticed that the wrong nearest neighbors have a lower confidence and propose to create Pseudo-Labels on only confident examples for further fine-tuning. They also show that Overclustering can be successfully used if the number of clusters is not known before. *Common ideas: OC, PL, PT*

IV. ANALYSIS

In this chapter, we will analyze which common ideas are shared or differ between methods. We will compare the performance of all methods with each other on common deep learning datasets.

A. DATASETS

In this survey, we compare the presented methods on a variety of datasets. We selected four datasets that were used in multiple papers to allow a fair comparison. An overview of example images is given in Figure 11.

CIFAR-10 AND CIFAR-100

are large datasets of tiny color images with size 32×32 [86]. Both datasets contain 60,000 images belonging to 10 or 100 classes respectively. The 100 classes in CIFAR-100 can be combined into 20 superclasses. Both sets provide 50,000 training examples and 10,000 validation examples

(image + label). The presented results are only trained with 4,000 labels for CIFAR-10 and 10,000 labels for CIFAR-100 to represent a semi-supervised case. If a method uses all labels this is marked independently.

STL-10

is dataset designed for unsupervised and semi-supervised learning [52]. The dataset is inspired by CIFAR-10 [86] but provides fewer labels. It only consists of 5,000 training labels and 8,000 validation labels. However, 100,000 unlabeled example images are also provided. These unlabeled examples belong to the training classes and some different classes. The images are 96×96 color images and were acquired in combination with their labels from ImageNet [1].

ILSVRC-2012

is a subset of ImageNet [1]. The training set consists of 1.2 million images whereas the validation and the test set include 150,000 images. These images belong to 1000 object categories. Due to this large number of categories, it is common to report Top-5 and Top-1 accuracy. Top-1 accuracy is the classical accuracy where one prediction is compared to one ground-truth label. Top-5 accuracy checks if a ground truth label is in a set of at most five predictions. For further details on accuracy see subsection IV-B. The presented results are only trained with 10% of labels to represent a semi-supervised case. If a method uses all labels this is marked independently.

B. EVALUATION METRICS

We compare the performance of all methods based on their classification score. This score is defined differently for unsupervised and all other settings. We follow standard protocol and use the classification accuracy in most cases. For unsupervised learning, we use cluster accuracy because we need to handle the missing labels during the training. We need to find the best one-to-one permutations (σ) from the network

A. Own previous papers

cluster predictions to the ground-truth classes. For N images $x_1, \dots, x_N \in X_I$ with labels z_{x_i} and predictions $f(x_i) \in \mathbb{R}^C$ the accuracy is defined in Equation 7 whereas the cluster accuracy is defined in Equation 8.

$$ACC(x_1, \dots, x_N) = \frac{\sum_{i=1}^N \mathbb{1}_{z_{x_i} = \arg\max_{1 \leq j \leq C} f(x_i)}}{N} \quad (7)$$

$$ACC(x_1, \dots, x_N) = \max_{\sigma} \frac{\sum_{i=1}^N \mathbb{1}_{z_{x_i} = \sigma(\arg\max_{1 \leq j \leq C} f(x_i))}}{N} \quad (8)$$

C. COMPARISON OF METHODS

In this subsection, we will compare the methods concerning their used common ideas and performance. We will summarize the presented results and discuss the underlying trends in the next subsection.

COMPARISON CONCERNING USED COMMON IDEAS

In Table 1 we present all methods and their used common ideas. Following our definition of common ideas in subsection II-B, we evaluate only ideas that were used frequently in different papers. Special details such as the different optimizer for fast-SWA or the used approximation for MI are excluded. Please see section III for further details.

One might expect that common ideas are used equally between methods and training strategies. We rather see a tendency that common ideas differ between training strategies. We will step through all common ideas based on the significance of differentiating the training strategies.

A major separation between the training strategies can be based on CE and pretext tasks. All one-stage-semi-supervised methods use a cross-entropy loss during training whereas only two use additional losses based on pretext tasks. All multi-stage-semi-supervised methods use a pretext task and use CE for fine-tuning. All one-stage-semi-supervised methods use no CE and often use a pretext task. Due to our definition of the training strategies this grouping is expected.

However, further clusters of the common ideas are visible. We notice that some common ideas are (almost) solely used by one of the two semi-supervised strategies. These common ideas are EM, KL, MSE, and MU for one-stage-semi-supervised methods and CL, MI, and OC for multi-stage-semi-supervised methods. We hypothesize that this shared and different usage of ideas exists due to the different usage of unlabeled data. For example, one-stage-semi-supervised methods use the unlabeled and labeled data in the same stage and therefore might need to regularize the training with MSE.

If we compare multi-stage-semi-supervised and one-stage-unsupervised training we notice that MI, OC, and PT are often used in both. All three of them are not often used with one-stage-semi-supervised training as stated above. We hypothesize that this similarity arises because most multi-stage-semi-supervised methods have an unsupervised stage followed by a supervised stage. For the method IIC the

authors even proposed to fine-tune the unsupervised method to surpass purely supervised results. CE*, PL, and VAT are used in several different methods. Due to their simple and complementary idea, they can be used in a variety of different methods. UDA for example uses PL to filter the unlabeled data for useful images. CE* seems to be more often used by multi-stage-semi-supervised methods. The parentheses in Table 1 indicate that they often also motivate another idea like CE⁻¹ [27] or the CL loss [25], [55]. All in all, we see that the defined training strategies share common ideas inside each strategy and differ in the usage of ideas between them. We conclude that the definition of the training strategies is not only logical but is also supported by their usage of common ideas.

COMPARISON CONCERNING PERFORMANCE

We compare the performance of the different methods based on their respective reported results or cross-references in other papers. For better comparability, we would have liked to recreate every method in a unified setup but this was not feasible. Whereas using reported values might be the only possible approach, it leads to drawbacks in the analysis.

Kolesnikov *et al.* showed that changes in the architecture can lead to significant performance boost or drops [89]. They state that 'neither [...] the ranking of architectures [is] consistent across different methods, nor is the ranking of methods consistent across architectures' [89]. Most methods try to achieve comparability with previous ones by a similar setup but over time small differences still aggregate and lead to a variety of used architectures. Some methods use only early convolutional networks such as AlexNet [1] but others use more modern architectures like Wide ResNet-Architecture [90] or Shake-Shake-Regularization [91].

Oliver *et al.* proposed guidelines to ensure more comparable evaluations in semi-supervised learning [92]. They showed that not following these guidelines may lead to changes in the performance [92]. Whereas some methods try to follow these guidelines, we cannot guarantee that all methods do so. This impacts comparability further. Considering the above-mentioned limitations, we do not focus on small differences but look for general trends and specialties instead.

Table 2 shows the collected results for all presented methods. We also provide results for the respective supervised baselines reported by the authors. To keep fair comparability we did not add state-of-the-art baselines with more complex architectures. Table 3 shows the results for even fewer labels as normally defined in subsection IV-A.

In general, the used architectures become more complex and the accuracies rise over time. This behavior is expected as new results are often improvements of earlier works. The changes in architecture may have led to these improvements. However, many papers include ablation studies and comparisons to only supervised methods to show the impact of their method. We believe that a combination of more

A.1. Long papers

TABLE 1. Overview of the methods and their used common ideas – On the left-hand side, the reviewed methods from section III are sorted by the training strategy. The top row lists the common ideas. Details about the ideas and their abbreviations are given in subsection II-B. The last column and some rows sum up the usage of ideas per method or training strategy. *Legend:* (X) The idea is only used indirectly. The individual explanations are given in section III.

	CE	CE*	EM	CL	KL	MSE	MU	MI	OC	PT	PL	VAT	Overall Sum
One-Stage-Semi-Supervised													
Pseudo-Labels [47]	X	X									X		3
π model [49]	X					X							2
Temporal Ensembling [49]	X					X							2
Mean Teacher [48]	X					X							2
VAT [65]	X											X	2
VAT + EntMin [65]	X		X									X	3
ICT [70]	X					X	X				X		4
fast-SWA [71]	X					X							2
MixMatch [46]	X		(X)			X	X				X		5
EnAET [73]	X		(X)		X	X	X			AET	X		7
UDA [16]	X	X	X		X						(X)		4
SPamCO [76]	X	X				X					X		4
ReMixMatch [45]	X	X	(X)				X	(X)		Rotation	X		7
FixMatch [26]	X	X	(X)								X		4
Sum	14	4	6	0	2	8	4	1	0	2	8	2	47
Multi-Stage-Semi-Supervised													
Exemplar [68]	X	X								Augmentation			3
Context [42]	X	X								Context			3
Jigsaw [43]	X	X								Jigsaw			3
DeepCluster [67]	X	X							X	Clustering	X		5
Rotation [40]	X	X								Rotation			3
CPC [55], [56]	X	(X)		X				(X)		CL			5
CMC [54]	X	(X)		X				(X)		CL			5
DIM [77]	X							X		MI			3
AMDIM [78]	X							X		MI			3
DMT [79]	X	X				X				Metric	X		5
IIC [14]	X							X	X	MI			4
S ³ L [15]	X	X	X							Rotation	X	X	6
SimCLR [25]	X	(X)								CL			3
MoCo [82]	X			X						Metric			3
BYOL [28]	X					X				Bootstrap			3
FOC [27]	X	(X)						X	X	MI			5
SimCLRv2 [57]	X	(X)		X						CL	X		5
Sum	17	11	1	5	0	1	0	6	3	17	4	1	66
One-Stage-Unsupervised													
DAC [50]											X		1
IMSAT [85]								X			X		2
IIC [14]								X	X	MI			3
FOC [27]								X	X	MI			3
SCAN [41]									X	CL	X		3
Sum	0	0	0	0	0	0	0	3	3	3	2	1	12
Overall Sum	31	54	7	5	2	9	4	10	6	22	14	4	125

modern architecture and more advanced methods lead to improvements.

For the CIFAR-10 dataset, almost all multi- or one-stage-semi-supervised methods reach about or over 90% accuracy. The best methods MixMatch and FixMatch reach an accuracy of more than 95% and are roughly three percent worse than the fully supervised baseline. For the CIFAR-100 dataset, fewer results are reported. FixMatch is with about 77% on this dataset the best method in comparison to the fully supervised baseline of about 80%. Newer methods also provide results for 1000 or even 250 labels instead of 4000 labels. Especially

EnAET, ReMixMatch, and FixMatch stick out since they achieve only 1-2% worse results with 250 labels instead of with 4000 labels.

For the STL-10 dataset, most methods report a better result than the supervised baseline. These results are possible due to the unlabeled part of the dataset. The unlabeled data can only be utilized by semi-, self-, or unsupervised methods. EnAET achieves the best results with more than 95%. FixMatch reports an accuracy of nearly 95% with only 1000 labels. This is more than most methods achieve with 5000 labels.

A. Own previous papers

TABLE 2. Overview of the reported accuracies – The first column states the used method. For the supervised baseline, we used the best-reported results which were considered as baselines in the referenced papers. The original paper is given in brackets after the score. The architecture is given in the second column. The last four columns report the Top-1 accuracy score in % for the respective dataset (See subsection IV-B for further details). If the results are not reported in the original paper, the reference is given after the result. A blank entry represents the fact that no result was reported. Be aware that different architectures and frameworks are used which might impact the results. Please see subsection IV-C for a detailed explanation. Legend: ¹ 100% of the labels are used instead of the default value defined in subsection IV-A. ² Multilayer perceptron is used for fine-tuning instead of one fully connected layer. Remarks on special architectures and evaluations: ³ Architecture includes Shake-Shake regularization. ⁴ Network uses wider hidden layers. ⁵ Method uses ten random classes out of the default 1000 classes. ⁶ Network only predicts 20 superclasses instead of the default 100 classes. ⁷ Inputs are pretrained ImageNet features. ⁸ Method uses different copies of the network for each input. ⁹ The network uses selective kernels [87].

	Architecture	Publication	CIFAR-10	CIFAR-100	STL-10	ILSVRC-2012	ILSVRC-2012 (Top-5)
Supervised (100% labels)	Best reported	-	98.01 [73]	79.82 [78]	68.7 [77]	85.7 [88]	97.6 [88]
One-Stage-Semi-Supervised							
Pseudo-Label [47]	ResNet50v2 [2]	2013					82.41 [15]
π model [49]	CONV-13	2017	87.64				
Temporal Ensembling [49]	CONV-13	2017	87.84				
Mean Teacher [48]	CONV-13	2017	87.69				
Mean Teacher [48]	Wide ResNet-28	2017	89.64				90.9 [57]
VAT [65]	CONV-13	2018	88.64				
VAT [65]	ResNet50v2	2018					82.78 [15]
VAT + EntMin [65]	CONV-13	2018	89.45				
VAT + EntMin [65]	ResNet50v2	2018	86.41 [15]				83.3 [15]
ICT [70]	Wide ResNet-28	2019	92.34				
ICT [70]	CONV-13	2019	92.71				
fast-SWA [71]	CONV-13	2019	90.95	66.38			
fast-SWA [71]	ResNet-26 ¹	2019	93.72				
MixMatch [46]	Wide ResNet-28	2019	95.05	74.12	94.41		
EnAET [73]	Wide ResNet-28	2019	94.65	73.07	95.48		
UDA [16]	Wide ResNet-28	2019	94.7			68.66	88.52
SPamCo [76]	Wide ResNet-28	2020	92.95				
ReMixMatch [45]	Wide ResNet-28	2020	94.86	76.97 [26]			
FixMatch [26]	Wide ResNet-28	2020	95.74	77.40			
FixMatch [26]	ResNet-50	2020				71.46	89.13
Multi-Stage-Semi-Supervised							
Exemplar [68]	ResNet50	2015				46.0 ¹ [89]	81.01 [15]
Context [42]	ResNet50	2015				51.4 ¹ [89]	
Jigsaw [43]	AlexNet	2016				44.6 ¹ [89]	
DeepCluster [67]	AlexNet	2018			73.4 [14]	41 ¹	
Rotation [40]	AlexNet	2018				55.4 ¹ [89]	
Rotation [40]	ResNet50v2	2018					78.53 [15]
CPC [56]	ResNet-170	2020	77.45 ¹ [77]		77.81 ¹ [77]	61.0	84.88
CMC [54]	AlexNet	2019			86.88 ²		
CMC [54]	ResNet-50 ⁶	2019				70.6	89.7*
DIM [77]	AlexNet	2019			72.57 ¹		
DIM [77]	GAN Discriminator	2019	75.21 ^{1,2}	49.74 ^{1,3}			
AMDIM [78]	ResNet18	2019	91.3 ¹ / 93.6 ^{1,4}	70.2 ¹ / 73.8 ^{1,4}	93.6 / 93.8 ¹	60.2 ¹ / 60.9 ^{1,4}	
DMT [79]	Wide ResNet-28	2019	88.70				
IIC [14]	ResNet34	2019			85.76 [27] / 88.8 ¹		
S ² L [15]	ResNet50v2 ²	2019				73.2 ¹	91.23*
SimCLR [25]	ResNet50v2 ²	2020				74.4 [57] / 76.5 ¹	92.6 / 93.2 ¹
MOCO [82]	ResNet50 ²	2020				68.6	
MOCO [82]	ResNet50	2020				60.6 ¹ / 71.1 ^{1,3} [83]	
BYOL [28]	ResNet200 ²	2020				77.7	93.7
FOC [27]	ResNet34	2020			86.49		
SimCLRv2 [57]	ResNet-152 ^{2,7}	2020				80.9 ¹	95.5 ¹
One-Stage-Unsupervised							
DAC [50]	All-ConvNet	2017	52.18	23.75	46.99	52.72 ³	
IMSAT [85]	Autoencoder ⁵	2017	45.6	27.5	94.1		
IIC [14]	ResNet34	2019	61.7	25.7 ⁴	59.6		
FOC [27]	ResNet34	2020			60.45		
SCAN [41]	ResNet18	2020	88.3	50.7 ⁴	80.9		

The ILSVRC-2012 dataset is the most difficult dataset based on the reported Top-1 accuracies. Most methods only achieve a Top-1 accuracy which is roughly 20% worse than the reported supervised baseline with around 86%. Only the methods SimCLR, BYOL, and SimCLRv2 achieve an accuracy that is less than 10% worse than the baseline. SimCLRv2 achieves the best accuracy with a Top-1 accuracy of 80.9% and a Top-5 accuracy of around 96%. For fewer labels also SimCLR, BYOL and SimCLRv2 achieve the best results.

The unsupervised methods are separated from the supervised baseline by a clear margin of up to 10%. SCAN achieves the best results in comparison to the other methods as it builds on the strong pretext task of SimCLR. This also illustrates the reason for including the unsupervised method in a comparison with semi-supervised methods. Unsupervised methods do not use labeled examples and therefore are expected to be worse. However, the data show that the gap of 10% is not large and that unsupervised methods can benefit from ideas of self-supervised learning. Some paper report results

TABLE 3. Overview of the reported accuracies with fewer labels - The first column states the used method. The last seven columns report the Top-1 accuracy score in % for the respective dataset and amount of labels. The number is either given as an absolute number or in percent. A blank entry represents the fact that no result was reported.

	CIFAR-10			STL-10		ILSVRC-2012		ILSVRC-2012 (Top-5)	
	4000	1000	250	5000	1000	10%	1%	10%	1%
One-Stage-Semi-Supervised									
Mean Teacher [48]	89.64	82.68	52.68						
ICT [70]	92.71	84.52	61.4 [46]						
MixMatch [46]	93.76	92.25	88.92	94.41	89.82				
EnAET [73]	94.65	93.05	92.4	95.48	91.96				
UDA [16]	95.12 [26]		91.18 [26]		92.34 [26]	68.66		88.52	
ReMixMatch [45]	94.86	94.27	93.73		93.82				
FixMatch [26]	95.74		94.93		94.83	71.46		89.13	
Multi-Stage-Semi-Supervised									
DMT [79]	88.70		80.3			58.6			
SimCLR [25]						74.4 [57]	63.0 [57]	92.6	85.8
BYOL [28]						77.7	71.2	93.7	89.5
SimCLRv2 [57]						80.9	76.6	95.5	93.4

for even fewer labels as shown in Table 3 which closes the gap to unsupervised learning further. IMSAT reports an accuracy of about 94% on STL-10. Since IMSAT uses pretrained ImageNet features, a superset of STL-10, the results are not directly comparable.

D. DISCUSSION

In this subsection, we discuss the presented results of the previous subsection. We divide our discussion into three major trends that we identified. All these trends lead to possible future research opportunities.

1) TREND: REAL WORLD APPLICATIONS?

Previous methods were not scalable to real-world images and applications and used workarounds e.g. extracted features [85] to process real-world images. Many methods can report a result of over 90% on CIFAR-10, a simple low-resolution dataset. Only five methods can achieve a Top-5 accuracy of over 90% on ILSVRC-2012, a high-resolution dataset. We conclude that most methods are not scalable to high-resolution and complex image classification problems. However, the best-reported methods like FixMatch and SimCLRv2 seem to have surpassed the point of only scientific usage and could be applied to real-world classification tasks.

This conclusion applies to real-world image classification tasks with balanced and clearly separated classes. This conclusion also implicates which real-world issues need to be solved in future research. Class imbalance [93], [94] or noisy labels [27], [95] are not treated by the presented methods. Datasets with also few unlabeled data points are not considered. We see that good performance on well-structured datasets does not always transfer completely to real-world datasets [27]. We assume that these issues arise due to assumptions that do not hold on real-world datasets like a clear distinction between datapoints [27] and non-robust hyperparameters like augmentations and batch size [28]. Future research has to address these issues so that reduced

supervised learning methods can be applied to any real-world datasets.

2) TREND: HOW MUCH SUPERVISION IS NEEDED?

We see that the gap between reduced supervised and supervised methods is shrinking. For CIFAR-10, CIFAR-100 and ILSVRC-2012 we have a gap of less than 5% left between total supervised and reduced supervised learning. For STL-10 the reduced supervised methods even surpass the total supervised case by about 20% due to the additional set of unlabeled data. We conclude that reduced supervised learning reaches comparable results while using only roughly 10% of the labels.

In general, we considered a reduction from 100% to 10% of all labels. However, we see that methods like FixMatch and SimCLRv2 achieve comparable results with even fewer labels such as the usage of 1% of all labels. For ILSVRC-2012 this is equivalent to about 13 images per class. FixMatch even achieves a median accuracy of around 65% for one label per class for the CIFAR-10 dataset [26].

The trend that results improve overtime is expected. But the results indicate that we are near the point where semi-supervised learning needs very few to almost no labels per class (e.g. 10 labels for CIFAR10). In practice, the labeling cost for unsupervised and semi-supervised will almost be the same for common classification datasets. Unsupervised methods would need to bridge the performance gap on these classification datasets to be useful anymore. It is questionable if an unsupervised method can achieve this because it would need to guess what a human wants to have classified even when competing features are available.

We already see that on datasets like ImageNet additional data such as JFT-300M is used to further improve the supervised training [96]–[98]. These large amounts of data can only be collected without any or weak labels as the collection process has to be automated. It will be interesting to investigate if the discussed methods in this survey can also scale to such datasets while using only few labels per class.

A. Own previous papers

We conclude that on datasets with few and a fixed number of classes semi-supervised methods will be more important than unsupervised methods. However, if we have a lot of classes or new classes should be detected like in few- or zero-shot learning [38], [94], [99], [100] unsupervised methods will still have a lower labeling cost and be of high importance. This means future research has to investigate how the semi-supervised ideas can be transferred to unsupervised methods as in [14], [41] and to settings with many, an unknown or rising amount of classes like in [39], [96].

3) TREND: COMBINATION OF COMMON IDEAS

In the comparison, we identified that few common ideas are shared by one-stage-semi-supervised and multi-stage-semi-supervised methods.

We believe there is only a little overlap between these methods due to the different aims of the respective authors. Many multi-stage-semi-supervised papers focus on creating good representations. They fine-tune their results only to be comparable. One-stage-semi-supervised papers aim for the best accuracy scores with as few labels as possible.

If we look at methods like SimCLRv2, EnAET, ReMix-Match, or S⁴L we see that it can be beneficial to combine different ideas and mindsets. These methods used a broad range of ideas and also ideas uncommon for their respective training strategy. S⁴L calls their combined approach even “Mix of all models” [15] and SimCLRv2 states that “Self-Supervised Methods are Strong Semi-Supervised Learners” [57].

We assume that this combination is one reason for their superior performance. This assumption is supported by the included comparisons in the original papers. For example, S⁴L showed the impact of each method separately as well as the combination of all [15].

Methods like Fixmatch illustrate that it does not need a lot of common ideas to achieve state-of-the-art performance but rather that the selection of the correct ideas and combining them in a meaningful is important. We identified that some common ideas are not often combined and that the combination of a broad range and unusual ideas can be beneficial. We believe that the combination of the different common idea is a promising future research field because many reasonable combinations are yet not explored.

V. CONCLUSION

In this paper, we provided an overview of semi-, self-, and unsupervised methods. We analyzed their difference, similarities, and combinations based on 34 different methods. This analysis led to the identification of several trends and possible research fields.

We based our analysis on the definition of the different training strategies and common ideas in these strategies. We showed how the methods work in general, which ideas they use and provide a simple classification. Despite the difficult comparison of the methods’ performances due to different architectures and implementations, we identified three major trends.

Results of over 90% Top-5 accuracy on ILSVRC-2012 with only 10% of the labels indicate that semi-supervised methods could be applied to real-world problems. However, issues like class imbalance and noisy or fuzzy labels are not considered. More robust methods need to be researched before semi-supervised learning can be applied to real-world issues.

The performance gap between supervised and semi- or self-supervised methods is closing and the number of labels to get comparable results to fully supervised learning is decreasing. In the future, the unsupervised methods will have almost no labeling cost benefit in comparison to the semi-supervised methods due to these developments. We conclude that in combination with the fact that semi-supervised methods have the benefit of using labels as guidance unsupervised methods will lose importance. However, for a large number of classes or an increasing number of classes the ideas of unsupervised are still of high importance and ideas from semi-supervised and self-supervised learning need to be transferred to this setting.

We concluded that one-stage-semi-supervised and multi-stage-semi-supervised training mainly use a different set of common ideas. Both strategies use a combination of different ideas but there are few overlaps in these techniques. We identified the trend that a combination of different techniques is beneficial to the overall performance. In combination with the small overlap between the ideas, we identified possible future research opportunities.

ACKNOWLEDGMENT

This work was supported by Land Schleswig-Holstein through the Open Access Publikationsfonds Funding Program.

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 60, no. 6, New York, NY, USA: Association for Computing Machinery, 2012, pp. 1097–1105.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [3] J. Bringer, S. Dippel, R. Koch, and C. Veit, “‘Tailception’: Using neural networks for assessing tail lesions on pictures of pig carcasses,” *Animal*, vol. 13, no. 5, pp. 1030–1036, 2019.
- [4] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, *arXiv:1804.02767*. [Online]. Available: <http://arxiv.org/abs/1804.02767>
- [5] S. Clausen, C. Zelenka, T. Schwede, and R. Koch, “Parcel tracking by detection in large camera networks,” in *Pattern Recognition*, T. Brox, A. Bruhn, and M. Fritz, Eds. Cham, Switzerland: Springer, 2019, pp. 89–104.
- [6] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
- [8] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.
- [9] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Barambe, and L. van der Maaten, “Exploring the limits of weakly supervised pretraining,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 181–196.

A.1. Long papers

- [10] L. Schmarje, C. Zelenka, U. Geisen, C.-C. Glüer, and R. Koch, "2D and 3D segmentation of uncertain local collagen fiber orientations in SHG microscopy," in *Proc. DAGM German Conf. Pattern Recognit.*, in Lecture Notes in Computer Science, vol. 11824, 2019, pp. 374–386.
- [11] G. E. Hinton and T. J. Sejnowski, *Unsupervised Learning: Foundations of Neural Computation*. Cambridge, MA, USA: MIT Press, 1999.
- [12] J. Xie, R. B. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 1, 2016, pp. 740–749.
- [13] M. Kaya and H. S. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, Aug. 2019.
- [14] X. Ji, A. Vedaldi, and J. Henriques, "Invariant information clustering for unsupervised image classification and segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9865–9874.
- [15] L. Beyer, X. Zhai, A. Oliver, and A. Kolesnikov, "S4L: Self-supervised semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1476–1485.
- [16] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 1–20.
- [17] O. Chapelle, B. Schölkopf, and A. Zien, "Semi-supervised learning," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, p. 542, Mar. 2009.
- [18] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.
- [19] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 4, 2020, doi: 10.1109/TPAMI.2020.2992393.
- [20] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, Feb. 2020.
- [21] G. Ciocca, C. Cusano, S. Santini, and R. Schettini, "On the use of supervised features for unsupervised image categorization: An evaluation," *Comput. Vis. Image Understand.*, vol. 122, pp. 155–171, May 2014.
- [22] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, Oakland, CA, USA, vol. 1, 1967, pp. 281–297.
- [23] X. Zhu, "Semi-supervised learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep., 2008, vol. 2. [Online]. Available: <https://minds.wisconsin.edu/handle/1793/60444>
- [24] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A survey of clustering with deep learning: From the perspective of neural architecture," *IEEE Access*, vol. 6, pp. 39501–39514, 2018.
- [25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
- [26] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 1–21.
- [27] L. Schmarje, J. Bringer, M. Santarossa, S.-M. Schröder, R. Kiko, and R. Koch, "Beyond cats and dogs: Semi-supervised classification of fuzzy labels with overclustering," 2020, *arXiv:2012.01768*. [Online]. Available: <http://arxiv.org/abs/2012.01768>
- [28] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 1–35.
- [29] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," 2019, *arXiv:1903.11260*. [Online]. Available: <http://arxiv.org/abs/1903.11260>
- [30] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019.
- [31] A. Mey and M. Loog, "Improvability through semi-supervised learning: A survey of theoretical results," 2019, pp. 1–28, *arXiv:1908.09574*. [Online]. Available: <http://arxiv.org/abs/1908.09574>
- [32] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 3, Mar. 2017, pp. 1856–1868.
- [33] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Jun. 2014, pp. 2672–2680.
- [34] L. Liu, T. Zhou, G. Long, J. Jiang, L. Yao, and C. Zhang, "Prototype propagation networks (PPN) for weakly-supervised few-shot learning on category graph," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3015–3022.
- [35] N. Ukita and Y. Uematsu, "Semi- and weakly-supervised human pose estimation," *Comput. Vis. Image Understand.*, vol. 170, pp. 67–78, May 2018.
- [36] D. Mahapatra, "Combining multiple expert annotations using semi-supervised learning and graph cuts for medical image segmentation," *Comput. Vis. Image Understand.*, vol. 151, pp. 114–123, Oct. 2016.
- [37] P. Xu, Z. Song, Q. Yin, Y.-Z. Song, and L. Wang, "Deep self-supervised representation learning for free-hand sketch," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1503–1513, Apr. 2021.
- [38] L. Liu, T. Zhou, G. Long, J. Jiang, X. Dong, and C. Zhang, "Isometric propagation network for generalized zero-shot learning," in *Proc. Int. Conf. Learn. Represent.*, Feb. 2021, pp. 1–13.
- [39] Z. Yu, L. Chen, Z. Cheng, and J. Luo, "TransMatch: A transfer-learning scheme for semi-supervised few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12856–12864.
- [40] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16.
- [41] W. Van Gansbeke, S. Vandenheide, S. Georgoulis, M. Proesmans, and L. VanGool, "SCAN: Learning to classify images without labels," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 268–285.
- [42] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.
- [43] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.
- [44] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9359–9367.
- [45] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "ReMixMatch: Semi-supervised learning with distribution alignment and augmentation anchoring," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–13.
- [46] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5050–5060.
- [47] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Learn. Represent. (ICML)*, vol. 3, 2013, p. 2.
- [48] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–16.
- [49] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.
- [50] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, "Deep adaptive image clustering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5880–5888.
- [51] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [52] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 215–223.
- [53] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.
- [54] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.*, 2019, pp. 776–794.
- [55] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*. [Online]. Available: <http://arxiv.org/abs/1807.03748>
- [56] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. van den Oord, "Data-efficient image recognition with contrastive predictive coding," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 4182–4192.
- [57] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 1–18.
- [58] T. Chen and L. Li, "Intriguing properties of contrastive losses," 2020, *arXiv:2011.02803*. [Online]. Available: <http://arxiv.org/abs/2011.02803>

A. Own previous papers

L. Schmarje et al.: Survey on Semi-, Self- and Unsupervised Learning for Image Classification

IEEE Access

- [59] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5171–5180.
- [60] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, "On mutual information maximization for representation learning," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–16.
- [61] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 529–536.
- [62] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
- [63] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.
- [64] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and D. R. Hjelm, "Mutual information neural estimation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 531–540.
- [65] S.-S. Learning, T. Miyato, S.-I. Maeda, M. Koyama, S. Ishii, and M. Koyama, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.
- [66] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.
- [67] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 132–149.
- [68] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1734–1747, Sep. 2016.
- [69] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.
- [70] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1–23.
- [71] B. Athiwaratkun, M. Finzi, P. Izmailov, A. G. Wilson, D. Podoprikin, T. Garipov, D. Vetrov, and A. G. Wilson, "There are many consistent explanations of unlabeled data: Why you should average," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–22.
- [72] P. Izmailov, D. Podoprikin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," in *Proc. Conf. Uncertainty Artif. Intell.*, 2018, pp. 1–12.
- [73] X. Wang, D. Kihara, J. Luo, and G.-J. Qi, "EnAET: A self-trained framework for semi-supervised and supervised learning with ensemble transformations," 2019, *arXiv:1911.09265*. [Online]. Available: <http://arxiv.org/abs/1911.09265>
- [74] L. Zhang, G.-J. Qi, L. Wang, and J. Luo, "AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2542–2550.
- [75] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*. [Online]. Available: <http://arxiv.org/abs/1708.04552>
- [76] F. Ma, D. Meng, X. Dong, and Y. Yang, "Self-paced multi-view co-training," *J. Mach. Learn. Res.*, vol. 21, no. 57, pp. 1–38, 2020.
- [77] R. D. Hjelm, A. Fedorov, S. Lavioie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–24.
- [78] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 15509–15519.
- [79] B. Liu, Z. Wu, H. Hu, and S. Lin, "Deep metric transfer for label propagation with limited annotated data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.
- [80] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.
- [81] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.
- [82] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.
- [83] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*. [Online]. Available: <http://arxiv.org/abs/2003.04297>
- [84] P. H. Richemond, J.-B. Grill, F. Althé, C. Tallec, F. Strub, A. Brock, S. Smith, S. De, R. Pascanu, B. Piot, and M. Valko, "BYOL works even without batch statistics," 2020, *arXiv:2010.10241*. [Online]. Available: <http://arxiv.org/abs/2010.10241>
- [85] W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama, "Learning discrete representations via information maximizing self-augmented training," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1558–1567.
- [86] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," M.S. thesis, Univ. Toronto, Toronto, ON, Canada, 2009. [Online]. Available: <https://www.cs.toronto.edu/~kriz/>
- [87] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.
- [88] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy: FixEfficientNet," 2020, *arXiv:2003.08237*. [Online]. Available: <http://arxiv.org/abs/2003.08237>
- [89] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1920–1929.
- [90] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 87.1–87.12.
- [91] X. Gastaldi, "Shake-shake regularization," 2017, *arXiv:1705.07485*. [Online]. Available: <http://arxiv.org/abs/1705.07485>
- [92] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. J. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 3235–3246.
- [93] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
- [94] S.-M. Schröder, R. Kiko, and R. Koch, "MorphoCluster: Efficient annotation of plankton images by clustering," *Sensors*, vol. 20, no. 11, p. 3060, May 2020.
- [95] Q. Li, X. Peng, L. Cao, W. Du, H. Xing, Y. Qiao, and Q. Peng, "Product image recognition with guidance learning and noisy supervision," *Comput. Vis. Image Understand.*, vol. 196, Jul. 2020, Art. no. 102963.
- [96] H. Pham, Z. Dai, Q. Xie, M.-T. Luong, and Q. V. Le, "Meta pseudo labels," 2020, *arXiv:2003.10580*. [Online]. Available: <http://arxiv.org/abs/2003.10580>
- [97] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (BiT): General visual representation learning," in *Proc. Eur. Conf. Comput. Vis.*, Lecture Notes in Computer Science, 2020, pp. 491–507.
- [98] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10684–10695.
- [99] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, Jul. 2020.
- [100] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–37, 2019.



LARS SCHMARJE received the B.S. and M.S. degrees in computer science from Kiel University, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree in computer science.

In 2016, he worked as a Project Planner for a hybrid-e-mail-infrastructure with the direkt gruppe GmBH, Hamburg. From 2017 to 2018, he worked with Vater Solution GmbH, Kiel, as an IT Security Engineer. Since 2019, he has been a Research Assistant with the Multimedia Information Processing Group, Kiel University. He is also a part of a project that builds an autonomous racing car. His current research interest includes semi-supervised learning with a focus on fuzzy labels.

A.1. Long papers



MONTY SANTAROSSA received the B.S. and M.S. degrees in computer science from Kiel University, in 2017 and 2019, respectively, where he is currently pursuing the Ph.D. degree in computer science.

In 2019, in connection with his master thesis, he spent six months at the Daimler Research and Development, researching learned environment perception for autonomous cars. His current research interests at Kiel University include image-based diagnosis and prognosis of eye diseases, and multi-modal image understanding tasks, including deep-learning-based image registration, classification, and segmentation.

Mr. Santarossa won the Prof. Dr. Werner Petersen Preis der Technik 2019 in the Category Best Master Thesis.



SIMON-MARTIN SCHRÖDER received the B.Sc. and M.Sc. degrees in computer science from Kiel University, Kiel, Germany, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree in computer science.

His primary area of expertise is deep learning and image recognition and he is working on the recognition of plankton images. His research interests include supervised and unsupervised deep learning and the application of machine learning methods in natural sciences.



REINHARD KOCH (Member, IEEE) received the Diploma and Ph.D. degrees in electrical engineering from the University of Hannover, Germany, in 1985 and 1996, respectively.

After postdoctoral research at KU Leuven, Belgium, he joined the Department of Computer Science, Kiel University, Germany, as a Professor, in 1999, where he is currently the Director of the Department of Computer Science and the Vice Dean of the Faculty of Engineering. He is the

author or coauthor of over 250 peer-reviewed publications. His research interests include 3D computer vision and object tracking to multi-view analysis, light-field processing, computer graphics, augmented reality applications, and deep learning approaches to scene understanding and applications. He received numerous awards, including the Olympus Award for Pattern Recognition, in 1997, and the David Marr Price at ICCV 1998. He also serves as the President for the German Association for Pattern Recognition (DAGM) and a German Delegate for the Governing Board of the International Association for Pattern Recognition (IAPR).

...

A. Own previous papers

A.1.3 Fuzzy Overclustering: Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy



Article

Fuzzy Overclustering: Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy

Lars Schmarje ^{1,*} , Johannes Brünger ¹ , Monty Santarossa ¹ , Simon-Martin Schröder ¹ , Rainer Kiko ² and Reinhard Koch ¹

¹ Multimedia Information Processing Group, Kiel University, 24118 Kiel, Germany; jbr@informatik.uni-kiel.de (J.B.); msa@informatik.uni-kiel.de (M.S.); sms@informatik.uni-kiel.de (S.-M.S.); rk@informatik.uni-kiel.de (R.K.)

² Laboratoire d'Océanographie de Villefranche, Sorbonne Université, 06230 Villefranche-sur-Mer, France; rainer.kiko@obs-vlfr.fr

* Correspondence: las@informatik.uni-kiel.de

Abstract: Deep learning has been successfully applied to many classification problems including underwater challenges. However, a long-standing issue with deep learning is the need for large and consistently labeled datasets. Although current approaches in semi-supervised learning can decrease the required amount of annotated data by a factor of 10 or even more, this line of research still uses distinct classes. For underwater classification, and uncurated real-world datasets in general, clean class boundaries can often not be given due to a limited information content in the images and transitional stages of the depicted objects. This leads to different experts having different opinions and thus producing fuzzy labels which could also be considered ambiguous or divergent. We propose a novel framework for handling semi-supervised classifications of such fuzzy labels. It is based on the idea of overclustering to detect substructures in these fuzzy labels. We propose a novel loss to improve the overclustering capability of our framework and show the benefit of overclustering for fuzzy labels. We show that our framework is superior to previous state-of-the-art semi-supervised methods when applied to real-world plankton data with fuzzy labels. Moreover, we acquire 5 to 10% more consistent predictions of substructures.

Keywords: semi-supervised; fuzzy; deep learning; noisy; real-world; plankton; marine



Citation: Schmarje, L.; Brünger, J.; Santarossa, M.; Schröder, S.-M.; Kiko, R.; Koch, R. Fuzzy Overclustering: Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy. *Sensors* **2021**, *21*, 6661. <https://doi.org/10.3390/s21196661>

Academic Editor: Hyoungsik Nam

Received: 24 August 2021

Accepted: 2 October 2021

Published: 7 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Over the past years, we have seen the successful application of deep learning to many underwater computer vision problems [1–4]. Automatic analysis of underwater data allows us to monitor ecological changes by evaluating large amounts of for example plankton data [5,6]. While it is relatively easy to create a lot of underwater image data, its analysis is time-consuming and thus expensive because the annotation requires trained taxonomists. The possible reasons for this issue include the huge amounts of data, the high imbalance between classes and the variability of annotations [7].

In underwater classification, domain experts often differ in their annotations [7–9]. This issue arises due to the following reasons: Firstly, automatically captured underwater images often have a lower quality than images taken manually by humans. This difference in quality arises for example due to the underwater lighting conditions and no manual corrections to e.g. insufficient sharpness or not centering the target inside the focus. For example the analysis of benthic images can suffer from these issues [8,9]. Even in the best scenario, a single image generally does not contain most of the information needed for a clear identification (e.g., three-dimensional configuration, minute morphological details, fluorescence). Secondly, intermediate stages actually exist between classes [10]. For example, in Figure 1 we show two different physical appearances (puff & tuft richodermium, while the dataset also contains intermediate stages between these two classes.

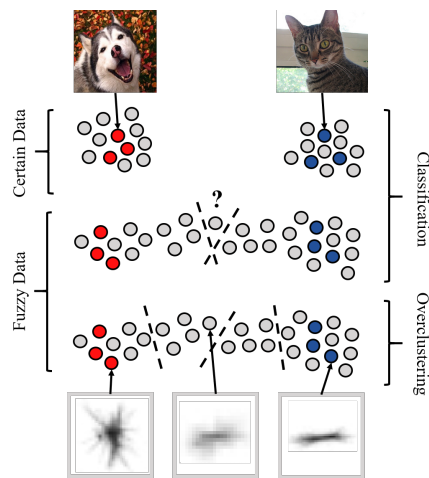


Figure 1. Illustration of fuzzy data and overclustering—The grey dots represent unlabeled data and the colored dots labeled data from different classes. The dashed lines represent decision boundaries. For **certain** data, a clear separation of the different classes with one decision boundary is possible and both classes contain the same amount of data (**top**). For **fuzzy** data determining a decision boundary is difficult because of intermediate datapoints between the classes (**middle**). These fuzzy datapoints can often not be easily sorted into one consistent class between annotators. If you overcluster the data, you get smaller but more consistent substructures in the fuzzy data (**bottom**). The images illustrate possible examples for certain data (cat & dog) and fuzzy plankton data (trichodesmium puff and tuft). The center plankton image was considered to be trichodesmium puff or tuft by around half of the annotators each. The left and right plankton image were consistently annotated.

This issue of different annotations is also known as *intra-* and *inter-observer* variability [11] and is common in many biological and medical application fields [8,9,12–17]. Even in a curated dataset [1], we quote Tarling et al. who state “there will very likely be inaccuracies, bias, and even inconsistencies in the labeling which will have affected the training capacity of the model and lead to discrepancies between predictions and ground truths” [18]. When aggregating multiple annotations per image, we call the resulting label fuzzy if we have different annotations between experts (non-zero variance), and *certain* if all annotations agree with each other. The mathematical formulation of a fuzzy label would be a unknown soft probability distribution l for k classes. The distribution $l \in (0, 1)^k$ can only be approximated with a high cost e.g., by averaging over multiple annotations.

Semi- and Self-Supervised Learning are promising approaches to decrease the needed amount of annotated data by a factor of 10 or even more [19–21]. These approaches leverage unlabeled data in addition to the normal labeled data to improve the training. A common strategy is to define a pretext task like image rotation prediction [22] or mutual information maximization [23] for pretraining. A broad overview of current trends, ideas and methods in semi-, self- and unsupervised learning is available in [24]. However, this research mainly focuses on established curated classification datasets such as STL-10 [25]. In these datasets, a clear distinction between classes such as cats and dogs are given. The hard partitioning of intermediate morphologies is not appropriate and does not allow the identification of substructures. We show that state-of-the-art semi-supervised algorithms are not well suited to handle fuzzy labels. These algorithms expect only *certain* labels as shown in the upper part of Figure 1. If we apply previous semi-supervised algorithms to fuzzy data which include fuzzy images, these algorithms arbitrarily assign undecidable images to one class (middle part of Figure 1).

Noisy labels are a common data quality issue and are discussed in the literature [11,26,27]. The fuzziness of labels is known as a special case of label noise that exist “due to subjectiveness of the task for human experts or the lack of experience in annotator[s]” [26]. In contrast to us, most methods [28–30] and literature surveys [11,26,27] interpret fuzzy labels as corrupted labels. We argue that fuzzy labels are valid signals derived from ambiguous images and that it is important to discover the substructures for real-world data handling [12–17].

Geng proposed to learn the label distribution to handle fuzzy data [31] and the idea was extended to the application of real-world images [32]. However, these methods are not semi-supervised and therefore depend on large labeled datasets. A variety of methods was proposed to handle fuzzy data in a semi-supervised learning approach [33–35]. These methods use lower-dimensional features spaces in contrast to images as input. Liu et al. proposed to use independent predictions of multiple networks as pseudo-labels for the estimation of the label distribution for photo shot-type classification [36]. We argue that the true label distribution is difficult to approximate and thus difficult to evaluate. We do not learn the label distribution but use clustering to identify substructures.

We propose *Fuzzy Overclustering* (FOC) which separates the fuzzy data into a larger number of visual homogeneous clusters (lower part, Figure 1) which can then be annotated very efficiently [10]. We will show on a Plankton dataset that state-of-the-art semi-supervised algorithms perform worse on fuzzy data in comparison to our method FOC which explicitly considers fuzzy images. Moreover, we will show that this leads to 5 to 10% more self-consistent predictions of plankton data.

One main idea is to rephrase the handling of fuzzy labels as a semi-supervised learning problem by using a small set of certain images and a large number of fuzzy images that are treated as unlabeled data. This approach allows us to use the idea of overclustering from semi-supervised literature [23,37] and apply it to fuzzy data. The difference to previous work is that we use overclustering not only to improve classification accuracy on the labeled data but improve the clustering and therefore the identification of substructures of fuzzy data. We show that overclustering allows us to cluster the fuzzy images in a more meaningful way by finding substructures and therefore allowing experts to analyze fuzzy images more consistently in the future.

We show the benefits of our method mainly on a plankton dataset which highlights the benefit for underwater classification. However, the issue of fuzzy labels is neither limited to plankton data nor to underwater classification. On a synthetic dataset, we show a proof-of-concept for the generalizability of our model to other datasets.

Our key contributions are:

- We identify an issue of semi-supervised algorithms that they do not work well with fuzzy labels. However, such fuzzy labels occur regularly in underwater image classification e.g. due to high natural variation of depicted objects which leads to a high inter- and intraobserver variability.
- We propose a novel framework for handling fuzzy labels with a semi-supervised approach. This framework uses overclustering to find substructures in fuzzy data and outperforms common state-of-the-art semi-supervised methods like FixMatch [38] on fuzzy plankton data.
- We propose a novel loss, *Inverse Cross-entropy* (CE^{-1}), which improves the overclustering quality in semi-supervised learning.
- We achieve 5 to 10% more self-consistent predictions on fuzzy plankton data.

2. Method

Our framework *Fuzzy Overclustering* (FOC) aims at creating an overclustering for fuzzy labels by using an auxiliary classification and not the other way round like previous literature [23,37]. In this section, we describe our framework in general and explain important parts in detail in the following subsections. We use the following notation for the given semi-supervised classification task. Our training data consists of the two subsets

X_l and X_u . X_l is a labeled image dataset with images $x \in X_l$ and corresponding labels y . X_u is an unlabeled image dataset, i.e., there is/exists no label for images $x \in X_u$.

We generate three inputs x_1, x_2, x_3 based on one image $x \in X_l \cup X_u$ depending on the availability of the corresponding label y . If y is not available, the images x_1 and x_2 are augmented views of x and x_3 is an augmented version of a random image $x' \in X_l \cup X_u$. If y is available, x_1 is an augmented view of x , x_2 is a supervised augmentation (see Section 2.3) and x_3 an inverse example. For the inverse example, we choose an image $x' \in X_l$ with a different label y' ($y' \neq y$). We use an augmented version of this image as third input $x_3 = g_3(x')$ with augmentation g_3 . We constraint the ratio from unlabeled to labeled data to a fixed ratio r to improve the run time of the model (see Section 2.4). The inputs are processed by a neural network Φ which is composed of a backbone like ResNet50 [39] and linear output prediction layers. Following [23], we call this linear predictors *heads* and use them either as normal or overclustering heads. As output we use the soft-max classifications of these normal and overclustering heads. If k_{GT} is the number of ground-truth classes a normal head outputs a probability for each of the k_{GT} classes. The overclustering head has k output nodes with $k > k_{GT}$ and give probabilities for more clusters than ground-truth classes (overclustering). Both type of heads are therefore fully connected layers with softmax activation but of different output size. We can average the training over multiple independent heads per type as shown in [23]. We use the notation Φ_{n_i} or Φ_{o_i} for the i -th normal or overclustering head respectively. An overview about the general pseudo code of FOC including the loss calculation is given in Algorithm 1.

For both heads the loss is different but can be written as the weighted sum of an unsupervised and a supervised loss as follows:

$$\mathcal{L} = \lambda_s \mathcal{L}_s + \lambda_u \mathcal{L}_u \quad (1)$$

\mathcal{L}_s is cross-entropy (\mathcal{L}_{CE}) for the normal head and our novel CE^{-1} loss ($\mathcal{L}_{CE^{-1}}$) for the overclustering head (see Section 2.1). For both heads \mathcal{L}_u is the mutual information loss \mathcal{L}_{MI} (see Section 2.2). An illustration of the complete pipeline is given in Figure 2. We initialize our backbones with pretrained weights and can therefore directly use RGB images as input. For further implementation details see Section 3.2.

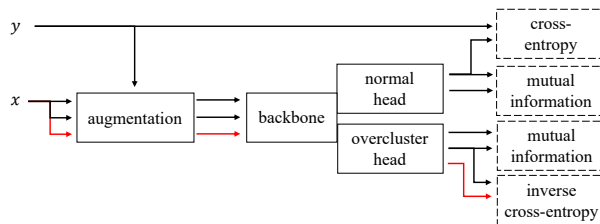


Figure 2. Overview of our framework FOC for semi-supervised classification—The input image is x and the corresponding label is y . The arrows indicate the usage of image or label information. Parallel arrows represent the independent copy of the information. The usage of the label for the augmentations is described in Section 2.3. The red arrow stands for an inverse example image x' with a different label than y . The output of the normal and the overclustering head have different dimensionalities. The normal head has as many outputs as ground-truth classes exist (k_{GT}) while the overclustering head has k outputs with $k > k_{GT}$. The dashed boxes on the right side show the used loss functions. More information about the losses inverse cross-entropy and mutual information can be found in Sections 2.1 and 2.2 respectively.

If we use FOC with $\lambda_s = 0$ and without supervised augmentations our model is comparable to the pretext task of Invariant Information Clustering (IIC) [23]. We can use this configuration as a warm-up to pretrain the weights. During the evaluation, we

will refer to using the pretext task for IIC and the warm-up of FOC synonymously. Our framework FOC can also be used to perform standard unsupervised clustering. The details about unsupervised clustering and a comparison to previous literature is given in the supplementary.

Algorithm 1: Pseudocode for our method Fuzzy Overclustering

Data: Batch of images of size b from labeled image data X_l and unlabeled image data X_u

Result: calculate loss value for one given batch for a network Φ with n normal and overclustering heads

L : matrix of size $b \times 2n$;

/ iterate over batch */*

for $i \leftarrow 0$ **to** b **do**

$x \leftarrow i$ -th image in batch;

if label y for image x_i available **then**

$x_1 \leftarrow g_1(x)$ with random augmentation g_1 ;

/ Supervised augmentation defined in Section 2.3 */*

$x_2 \leftarrow g_2(x)$ with supervised augmentation g_2 ;

/ Inverse example defined in Section 2 */*

$x_3 \leftarrow g_3(x')$ with random augmentation g_3 and inverse example x' ;

else

$x_1 \leftarrow g_1(x)$ with random augmentation g_1 ;

$x_2 \leftarrow g_2(x)$ with random augmentation g_2 ;

$x_3 \leftarrow g_3(x')$ with random augmentation g_3 and random image x' ;

end

/ iterate over heads */*

for $j \leftarrow 0$ **to** n **do**

 calculate forward pass for outputs Φ_{n_j} and Φ_{o_j} ;

/ CE loss for normal head */*

$L[i,j] \leftarrow \mathcal{L}_{CE}(\Phi_{n_j}(x_i), l_i)$ with l_i ;

/ CE^{-1} loss for overclustering head */*

$L[i,j+n] \leftarrow \mathcal{L}_{CE^{-1}}(x_1, x_2, x_3)$ with Equation (2);

end

end

/ calculate loss */*

$\mathcal{L}_s \leftarrow$ average supervised loss across heads and batch from L ;

$\mathcal{L}_u \leftarrow$ unsupervised MI loss across batch with Equations (3) and (4);

$\mathcal{L} \leftarrow \lambda_s \mathcal{L}_s + \lambda_u \mathcal{L}_u$;

2.1. Inverse Cross-Entropy (CE^{-1})

Inverse Cross-Entropy is a novel supervised loss for an overclustering head and one of the key contributions of this work. The loss is needed to use the label information for an overclustering head. For normal heads, we can use cross-entropy (CE) to penalize the divergence between our prediction and the label. We can not use CE directly for the overclustering heads since we have more clusters than labels and no predefined mapping between the two. However, we know that the inputs x_1/x_2 and x_3 should not belong to the same cluster. Therefore, our goal with CE^{-1} is to define a loss that pushes their output distributions (e.g., $\Phi(x_1)$ and $\Phi(x_3)$) apart from each other.

Let us assume we could define a distribution that $\Phi(x_3)$ should not be. In short, an inverse distribution $\Phi(x_3)^{-1}$. If we had such a distribution we could use CE to penalize the divergence for example between $\Phi(x_1)$ and $\Phi(x_3)^{-1}$.

One possible and easy solution for an inverse distribution is $\Phi(x_3)^{-1} = 1 - \Phi(x_3)$. For a binary classification problem, $\Phi(x_3)^{-1}$ can even be interpreted as a probability distribution again. This is not the case for a multi-class classification problem. We could use

a function like softmax to cast $\Phi(x_3)^{-1}$ into a probability distribution but decided against it for three reasons. Firstly, we would penalize correct behavior. For example in a three class problem with $\Phi_1(x_1) = 0.5 = \Phi_2(x_1)$ and $\Phi_3(x_3) = 1$ we only get $CE(\Phi(x_1), \Phi(x_3)^{-1}) = 0$ if $\Phi(x_3)^{-1}$ is not a probability distribution. Otherwise either $\Phi_1(x_3)^{-1}$ or $\Phi_2(x_3)^{-1}$ have to be real smaller than 1. Secondly, we are still minimizing the entropy of $\Phi(x_1)$ which leads to more confident predictions in semi-supervised learning [19,20,40–43]. The proof is given in the supplementary. Thirdly, it is easier and in practice, it is not needed. For the input $i = (x_1, x_2, x_3)$, we define the cross-entropy inverse loss $\mathcal{L}_{CE^{-1}}$ as shown in Equation (2).

$$\begin{aligned}\mathcal{L}_{CE^{-1}}(i) &= 0.5 \cdot CE^{-1}(\Phi(x_1), \Phi(x_3)) \\ &\quad + 0.5 \cdot CE^{-1}(\Phi(x_2), \Phi(x_3)), \text{ with} \\ CE^{-1}(p, q) &= - \sum_{c=1}^k p(c) \cdot \ln(1 - q(c)).\end{aligned}\quad (2)$$

2.2. Mutual Information (MI)

For the unlabeled data, we use the loss proposed by Ji et al. because it is calculated directly on the output clusters [23]. Therefore similar images are pulled to the same clusters while CE^{-1} pushes different images apart. For this purpose, we want to maximize the mutual information between two output predictions $\Phi(x_1), \Phi(x_2)$ with x_1, x_2 images which should belong to the same cluster and $\Phi: X \rightarrow [0, 1]^k$ a neural network with k output dimensions. We can interpret $\Phi(x)$ as the distribution of a discrete random variable z given by $P(z = c|x) = \Phi_c(x)$ for $c \in \{1, \dots, k\}$ with $\Phi_c(x)$ the c -th output of the neural network. With z, z' such random variables we need the joint probability distribution for $P_{cc'} = P(z = c, z' = c')$ for the calculation of the mutual information $I(z, z')$. Ji et al. propose to approximate the matrix P with the entry $P_{cc'}$ at row c and column c' by averaging over the multiplied output distributions in a batch of size n [23]. Symmetry of P is enforced as shown in Equation (3).

$$P = \frac{Q + Q^T}{2} \text{ with } Q = \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \cdot \Phi(x'_i)^T \quad (3)$$

We can maximize our objective $I(z, z')$ with the marginals $P_c = P_{c'} = P(z = c)$ given as sums over the rows or columns as shown in Equation (4).

$$I(z, z') = \sum_{c=1}^k \sum_{c'=1}^k P_{cc'} \cdot \ln \frac{P_{cc'}}{P_c \cdot P_{c'}} \quad (4)$$

2.3. Supervised Augmentations

In the unsupervised pretraining, we use the same image x to create the two inputs $x_1 = g_1(x)$ and $x_2 = g_2(x)$ based on the augmentations g_1 and g_2 . Otherwise, without supervision, it is difficult to determine similar images. However, if we have the label y for x we can use a secondary image $x' \in X_I$ with the same label to mock an ideal image transformation to which the network should be invariant. In this case we can create $x_2 = g_2(x')$ based on the different image. We call this *supervised augmentation*.

2.4. Restricted Unsupervised Data

Unlabeled data has a small impact on the results but drastically increases the runtime in most cases. The increased runtime is caused by the facts that we often have much more unlabeled data than labeled data and that a neural network runtime is normally linear in the number of samples it needs to process. However, unlabeled data is essential for our proposed framework and we can not just leave it out. We propose to restrict the unlabeled data to a fixed upper-bound ratio r in every batch and therefore the unlabeled data per epoch. Detailed examples and experiments are given in the supplementary. It is important

to notice that we restrict only the unlabeled data per batch/epoch. While for one epoch the network will not process all unlabeled data, over time all unlabeled data will be seen by the network. We argue that the impact on training time negatively outweighs the small benefit gained from all unlabeled data per epoch.

3. Experiments

We conducted our experiments mainly on a real-world plankton dataset. We used the common image classification dataset STL-10 as a comparison with only certain labels and a synthetic dataset for a proof-of-concept for the generalizability to other datasets. We compare ourselves to previous work and make several ablations. Additional results like unsupervised clustering, more detailed ablations and further details are given in the supplementary material.

3.1. Datasets

While the issue of fuzzy labels is present in multiple datasets [12–17], they are not well suited for evaluations. If we want to quantify the performance on fuzzy labels, we need a dataset with very good fuzzy ground-truth. This can only be achieved with a high cost e.g. by multiple annotations and thus is often not feasible. For all used datasets, we ensure that the labeled training data only consists of certain images and that the fuzzy images are used as unlabeled data. If we include fuzzy labels in the labeled data which is used as guidance during training, this will lead to worse performance as illustrated in the ablations (Table 3).

3.1.1. Plankton

The plankton dataset contains diverse grey-level images of marine planktonic organisms. The images were captured with an Underwater Vision Profiler 5 [44] and are hosted on EcoTaxa [45]. In the citizen science project PlanktonID (<https://planktonid.geomar.de/en> (accessed on 6 October 2021)), each sample was classified multiple times by citizen scientists. The data for the PlanktonID project is a subset of the data available on EcoTaxa [45]. It was resorted to contain a more balanced representation of the available classes. The dataset consists of 12,280 images in originally 26 classes. We merged minor and similar classes so that we ended up with 10 classes. The class no-fit represents a mixture of left-over classes. The merging was necessary because some classes had too few images for current state-of-the-art semi-supervised approaches. After this process, a class imbalance is still present with the smallest class containing about 4.16% and the largest class 30.37% of all samples. We use the mean over all annotations as the fuzzy label. The citizen scientists agree on most images completely. We call these images and their label *certain*. However, about 30% of the data has at least one disagreeing annotation. We call these images and their label *fuzzy* and use the most likely class as ground-truth if we need a hard label for evaluation. The fuzzy labeled images are used only as unlabeled data. More details about the mapping process, the number of used samples and graphical illustrations are given in the supplementary.

3.1.2. STL-10

STL-10 is a common semi-supervised image classification dataset [25] and a subset of ImageNet [46]. It consists of 5000 training samples and 8000 validation samples depicting everyday objects. Additionally, 100,000 unlabeled images are provided that may belong to the same or different classes than the training images. In contrast to the plankton and synthetic dataset, no labels are provided for the unlabeled data and no fuzzy datapoints exist. We use this dataset only to illustrate the difference in the performance of FOC to previous semi-supervised methods.

3.1.3. Synthetic Circles and Ellipses (SYN-CE)

This dataset is a mixture of circles and ellipses (bubbles) on a black background with different colors. The 6 ground-truth classes are blue, red and green circles or ellipses. An image is defined as certain if the hue of the color is 0 (red), 120 (green) or 240 (blue) and the main axis ratio of the bubble is 1 (circle) or 2 (ellipse). Every other datapoint is considered fuzzy and the ground-truth label l is calculated as the product of the interpolation of the color p_c and the geometry p_g distribution. More details are in the supplementary. The dataset consists of 1800 certain and 1000 fuzzy labeled images for train, validation and unlabeled data split. We look at three subsets: *Ideal*, *Real* and *Fuzzy*. The *Ideal* subset uses the maximal class of the fuzzy label l as a ground-truth class and represents the ideal case that we certainly know the most likely label to each image. For the *Real* subset, the ground-truth classes are randomly picked with the distribution of the fuzzy label l and represent the real or common case. For example due to only one annotation, the percentage that the label corresponds to the actual most likely class is linear to the fuzzy label. The *Fuzzy* subset only uses certain labeled images as training data and represent a cleaned training dataset. We will show that this handling of fuzzy labels leads to a higher classification performance in comparison to the *Real* dataset in Section 3.5.1. The *Ideal* and the *Real* subset can be evaluated on the unlabeled data of the *Fuzzy* subset with some overlap in the images.

3.2. Implementation Details

As a backbone for our framework, we used either a ResNet34 variant [23] or a standard ResNet50v2 [39]. The heads are single fully connected layers with a softmax activation function. Following [23], we use five randomly initialized copies for each type of head and repeat images per batch three times for more stable training. We alternated between training the different types of heads. The inputs are either sobel-filtered images or color images for pretrained networks. For the ResNet34 backbone, we use CIFAR20 (20 superclasses in CIFAR-100 [47]) weights and for the ResNet50v2 backbone ImageNet [46] weights. We use in general $\lambda_s = 1 = \lambda_u$ and an unlabeled data restriction of $r = 0.5$. We call our Framework FOC-Light if we use $\lambda_u = 0$ and no warm-up. This means we do not use the loss introduced by [23] and therefore also do not have to use their stabilization methods like repetitions. During the pretext task or warm-up and the main training, we train the framework with Adam and an initial learning rate of 1×10^{-4} for 500 epochs. When switching from the pretext task to fine-tuning, we train only the heads for 100 epochs with a learning rate of 1×10^{-3} before switching to the lower learning rate of 1×10^{-4} . The number of outputs for the overclustering head should be about 5 to 10 times the number of classes. The exact number is not crucial because it is only an upper bound for the framework. We use 70 for STL-10 and 60 for the plankton dataset. We selected all hyperparameters heuristically based on the STL-10 dataset and did not change them for the plankton dataset. We used the recommended hyperparameters by the original authors for the previous methods. We compared with the following methods Semantic Clustering by Adopting Nearest neighbors (SCAN) [48], Information Invariant Clustering (IIC) [23], Mean-Teacher [49], Pi(-Model) [29], Pseudo-label [50] and FixMatch [38]. More detailed descriptions are given in the supplementary.

3.3. Metrics

The evaluation protocols vary slightly depending on the used output and dataset. The used data splits training, validation and unlabeled are defined above in Section 3.1.

On STL-10, we calculate accuracy of the validation data. Accuracy is the portion of true positive and true negatives from the complete dataset.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

TP, TN, FP and FN are the true positive, true negative, false positive and false negative respectively. We calculate these values per class and then sum the up before calculating

the accuracy (micro averaging). For the overclustering head, we need to find a mapping between the output clusters and the given classes. We calculate this mapping based on the majority class in each cluster on the training data as in [23].

On the fuzzy plankton and synthetic datasets, we evaluate the macro-f1 score on the unlabeled data. We calculate the macro F1-Score i.e., the average of the F1-scores per class due to the skewed class distribution.

$$\text{F1-Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

(6)

Mind that a micro averaged F1-Score would be in our case the same as the above defined accuracy. We use the unlabeled data as evaluation dataset because the fuzzy images, in which we are interested, are only included in the unlabeled data split by definition. The mapping for the overclustering head is calculated based on the unlabeled data split because we expect human experts to be involved in this process for the identification of substructures. The best unlabeled results of the fuzzy Plankton and Synthetic dataset are reported based on the validation metrics.

If not stated otherwise, we report the maximum score for the overclustering and the normal head and the average and standard deviation over 3 independent repetitions.

3.4. Results

3.4.1. State-of-the-Art Comparison

We compare the state-of-the-art methods on certain and fuzzy data in Table 1.

Table 1. Comparison of state-of-the-art on certain and fuzzy data—We use STL-10 as a certain dataset and the Plankton data as a fuzzy dataset. We report the Accuracy for STL-10 and the F1-Score for the Plankton data due to class imbalance. It is important to notice that STL-10 is a curated dataset while the Plankton dataset still contains the fuzzy images. For more details about the metrics see Section 3.3. The results of previous methods are reported in the original paper or the original authors code was used to replicate the results. The best results are marked bold. Legend: † A MLP used for fine-tuning. ‡ Used only 1000 labels instead of 5000. * Unsupervised method.

Method	Network	Type of Data	
		Certain	Fuzzy
SCAN * [48]	ResNet18	76.80 ± 1.10	37.64 ± 3.56
IIC [23]	ResNet34	85.76 ± 1.36	65.47 ± 1.86
IIC † [23]	ResNet34	88.8	66.81 ± 1.85
Mean-Teacher [49]	Wide ResNet28	78.577 ± 2.39 ‡	72.85 ± 0.46
Pi [29]	Wide ResNet28	73.77 ± 0.82 ‡	74.34 ± 0.58
Pseudo-label [50]	Wide ResNet28	72.01 ± 0.83 ‡	75.04 ± 0.52
FixMatch [38]	Wide ResNet28	94.83 ± 0.63 ‡	76.28 ± 0.27
FOC-Light (Ours)	ResNet50	–	72.79 ± 2.99
FOC (Ours)	ResNet50	86.12 ± 1.22	76.79 ± 1.18

We see that FOC reaches a performance of about 86% on certain data but is not able to reach the performance of FixMatch. FixMatch outperforms FOC by a clear margin of nearly 8% while using a fifth of the labels. This performance is expected as FOC does not focus like the others on classifying certain but fuzzy data. If we look at the less curated fuzzy Plankton dataset, we see that FOC outperforms all all methods by a small margin. All previous methods focus on certain and curated data and we see this leads to a huge performance degeneration if they are applied to fuzzy data. FixMatch reaches in both datasets the best performance except for our method FOC. We conclude that the overclustering from FOC is the key for handling fuzzy data because it allows more flexibility during training. Previous semi-supervised methods did not consider the issue of inter- and intraobserver variability and thus are worse than FOC in classifying fuzzy data.

If we use FOC-Light without the loss and stabilization of [23] the F1-Score drops slightly to 75% but the used GPU hours can be decreased from 58 to 4 h. We conclude that the overclustering head is more suitable for handling fuzzy real-world data as we assumed at the beginning. Moreover, we see that the combination of cross-entropy and our novel loss CE^{-1} can also successfully train an overclustering head.

3.4.2. Consistency

Up to this point, we analyzed classification metrics based on the 10 ground-truth classes but the quality of substructures was not evaluated. We can judge the consistency of each image within its cluster with the help of experts as a quality measure. An image is consistent if an expert views it as visually similar to the majority of the cluster. The consistency is calculated by dividing the number of consistent images by all images. The consistency over all classes or per class for FOC and FixMatch is given in Table 2 and raw numbers are provided in the supplementary. We provide a comparison based on all data and without the no-fit class because this class contains a mixture of different plankton entities. Visual similarity is therefore difficult to judge because it can only be defined by not being similar the other nine classes. Based on the F1-Score, FixMatch and FOC perform similarly but if we look at the consistency we see that FOC is more than 5% more consistent than FixMatch. If we exclude the class no-fit from the analysis, FOC reaches a consistency of around 86% in comparison to 77% from FixMatch. For both sets, our method FOC reaches a higher average consistency per cluster and lower standard deviation. This means the clusters produced by FOC are more relevant in practice because there are fewer low-quality clusters which can not be used. Overall, this higher consistency can lead to faster and more reliable annotations.

Table 2. Consistency comparison on plankton dataset—The consistency is rated by experts over the complete data and a subset without the class no-fit. The score is given overall as as average per cluster with standard deviation and is described in Section 3.4.2. The best results are marked bold.

Method	All Data		Ignore Class No-Fit	
	Overall	Per Cluster	Overall	Per Cluster
FixMatch [38]	82.56	78.78 ± 28.22	77.11	69.61 ± 29.41
FOC (Ours)	87.80	79.66 ± 18.88	86.31	86.41 ± 13.68

3.4.3. Qualitative Results

We illustrate some qualitative results of FOC in Figure 3. All images in a cluster are visually similar, even the probably wrongly assigned images (red box). For the images in the first row, the annotators are certain that the images belong to the same class. In the second row, annotators show a high uncertainty of assignment between the two variants of the same biological object. This illustrates the benefit of overclustering since visual similar items are in the same cluster even for uncertain annotations. In a consensus process for the second row, experts could decide if the cluster should be the puff, tuft or a new borderline class. Moreover, this clustering could be beneficial for monitoring the current imaging process. We provide more randomly selected results in the supplementary.

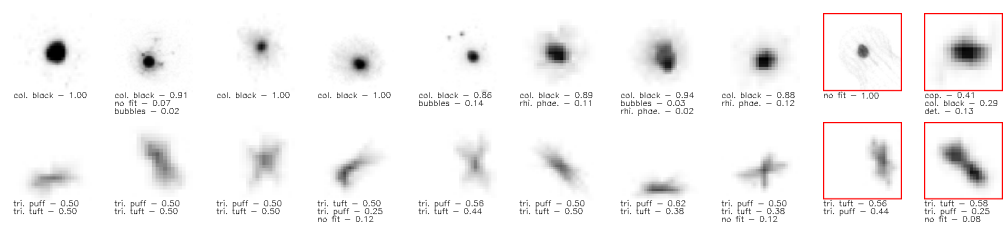


Figure 3. Qualitative results for unlabeled data—The results in each row are from the same predicted cluster. The three most important fuzzy labels based on the citizen scientists’ annotations are given below the image. The last two items with the red box in each row show examples not matching the majority of the cluster.

3.5. Ablation Studies
3.5.1. SYN-CE

We compare our framework with some previous methods on the three subsets of SYN-CE in Table 3. All semi-supervised method reach almost a F1-Score of 100% on the unlabeled fuzzy data for the subset *Ideal*. In real-world data, it is unlikely that we have the real fuzzy ground-truth labels. It is more likely that we have uncertain/wrong labels for training and validation or no labels at all for fuzzy data like in the subsets *Real* or *Fuzzy*. In both cases, we see that our method reaches a superior performance with up to 10% higher F1-Score. While FOC-Light is only slightly better in comparison to the other semi-supervised methods on the *Real* subset it is comparable to the complete framework on the *Fuzzy* dataset. This is one indication that CE^{-1} is one of the key components for successfully training the overclustering heads. We see the F1-Score on the *Fuzzy* subset is around 10% higher than on the *Real* subset. We conclude that these results supports our idea of separating certain and fuzzy data during training because we do not need to potentially falsely approximate the real fuzzy ground-truth label like in the *Real* subset.

Table 3. Comparison to state-of-the-art on SYN-CE datasets—Each column represent a subset of the dataset SYN-CE. The results are F1-Scores which were calculated on the unlabeled data which include the fuzzy labels. All results within a one percent margin of the best result are marked bold.

Method	Ideal	Real	Fuzzy
Mean-Teacher [49]	97.11 ± 0.78	73.23 ± 2.49	66.57 ± 16.27
Pi [29]	98.44 ± 0.28	72.74 ± 2.43	77.69 ± 5.02
Pseudo-label [50]	98.17 ± 0.30	75.70 ± 1.98	89.48 ± 1.94
FixMatch [38]	98.32 ± 0.01	71.81 ± 1.06	93.82 ± 1.83
FOC-Light (Ours)	97.46 ± 4.39	78.77 ± 7.83	94.29 ± 0.87
FOC (Ours)	97.72 ± 4.52	83.86 ± 4.21	94.15 ± 0.29

3.5.2. Loss & Network

In Table 4 multiple ablations for STL-10 and the plankton dataset are given. The scores are averaged across the different output heads of our framework. Based on these tables, we illustrate the impact of the warm-up, the initialization and the usage of the MI and CE^{-1} loss for our framework. The normal accuracy can be improved by about 10% when using the unsupervised warm-up on the STL-10 dataset. On the plankton dataset, the impact is less but tends to give better results of some percent. Warm-up in combination with the MI loss leads to a performance which is not more than 10% worse than the full setup for all ablations except for one. For this exception, CE^{-1} is needed to stabilize the overclustering performance due to the poor initialization with CIFAR-20 weights. We attribute this worse performance to the initialization and not the different backbone because on STL-10 the CIFAR-20 initializations of the ResNet34 backbone outperform the ImageNet

weights of the ResNet50v2 backbone. We believe the positive effects of ImageNet weights for its subset STL-10 and the better network are negated by the different loss.

IIC is similar to FOC with warm-up and no additional losses but we train also train an overclustering head for handling fuzzy data. Taking this into consideration, we achieve an 8 to 11% better F1-Score than IIC. A special case is FOC-light which does only use the CE^{-1} loss and therefore no stabilization method proposed in [23]. This decreases gpu memory usage and runtime and results in a total decrease of the GPU hours from 58 to 4 h. Overall, our novel loss CE^{-1} improves the overclustering performance regardless of the dataset and the weight initialization by 10% on STL-10 and up to 7% on the plankton dataset. We see that CE^{-1} is a key component for training an overclustering head and can even be trained without the stabilization of the warm-up and the MI loss.

Table 4. Ablation study—The second to fourth column indicates if a warm-up, the MI loss or our CE^{-1} loss were used respectively. The fifth column indicates if CIFAR-20 (C), ImageNet (I) or no (–) weights were used. Sobel filtered images are used as input for no weights. The Top1 and Top3 results are marked bold respectively. * Original authors code. † A MLP used for fine-tuning.

Method	Warm	MI	CE ^{−1}	Weight	Accuracy	
					Overcluster	Normal
FOC		X		–	70.92 ± 2.42	76.39 ± 0.05
IIC * [23]	X			–		85.76
FOC	X	X		–	73.88 ± 0.21	82.01 ± 5.31
FOC	X	X	X	–	82.59 ± 0.06	86.49 ± 0.01
FOC	X	X	X	C	84.36 ± 0.64	78.59 ± 7.40
FOC	X	X	X	I	83.57 ± 0.10	85.21 ± 0.03
(a) STL-10						
Method	Warm	MI	CE ^{−1}	Weight	F1-Score	
					Overcluster	Normal
IIC [23]	X			–	–	66.63
IIC † [23]	X			–	–	69.92
FOC				C	31.45 ± 6.02	39.35 ± 1.30
FOC		X		C	29.82 ± 2.98	60.65 ± 0.02
FOC		X	X	C	70.11 ± 1.99	64.10 ± 0.13
FOC	X			C	23.95 ± 2.63	58.71 ± 2.07
FOC	X	X		C	69.36 ± 0.05	56.59 ± 0.04
FOC	X	X	X	C	70.68 ± 0.10	58.09 ± 0.03
FOC				I	29.88 ± 2.75	54.92 ± 0.03
FOC-Light			X	I	74.93 ± 0.22	73.64 ± 0.06
FOC		X		I	72.70 ± 0.36	64.78 ± 0.04
FOC		X	X	I	73.93 ± 0.29	64.84 ± 0.03
FOC	X			I	73.93 ± 0.29	64.84 ± 0.03
FOC	X	X		I	69.64 ± 1.04	66.56 ± 0.08
FOC	X	X	X	I	74.01 ± 3.17	65.17 ± 0.18
(b) plankton dataset						

4. Conclusions

In this paper, we take the first steps to address real-world underwater issues with semi-supervised learning. Our presented novel framework FOC can handle fuzzy labels via overclustering. We showed that overclustering can achieve better results than previous state-of-the-art semi-supervised methods on fuzzy plankton data. The additional overclustering output is a key difference to previous work to achieve this superior performance.

While on certain data FOC is not state-of-the-art by a clear margin of over 10%, it slightly outperforms all other methods on the fuzzy plankton data. These beneficial effects have to be verified on other fuzzy datasets and with more semi-supervised algorithms in the future. Due to better performance of FOC on fuzzy data, we expect a similar outcome. We illustrated the visual similarity on qualitative results from these predictions and results in 5 to 10% more consistent predictions. We showed that CE^{-1} is the key component for training the overclustering head.

Supplementary Materials: The following are available at <https://www.mdpi.com/article/10.3390/s21196661/s1>, The details about unsupervised clustering and a comparison to previous literature.

Author Contributions: Conceptualization, L.S., J.B., M.S., S.-M.S., R.K. (Rainer Kiko) and R.K. (Reinhard Koch); methodology, L.S., J.B., M.S. and S.-M.S.; software, L.S.; validation, L.S.; formal analysis, L.S.; investigation, L.S., J.B., M.S., S.-M.S. and R.K. (Rainer Kiko); resources, L.S. and R.K. (Rainer Kiko); data curation, L.S., S.-M.S. and R.K. (Rainer Kiko); writing—original draft preparation, L.S., J.B., M.S., S.-M.S., R.K. (Rainer Kiko) and R.K. (Reinhard Koch); writing—review and editing, L.S., J.B., M.S., S.-M.S., R.K. (Rainer Kiko) and R.K. (Reinhard Koch); visualization, L.S., J.B., M.S., S.-M.S., R.K. (Rainer Kiko) and R.K. (Reinhard Koch); supervision, R.K. (Reinhard Koch); project administration, Not applicable; funding acquisition, Not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge funding of L. Schmarje by the ARTEMIS project (Grant number 01EC1908E) funded by the Federal Ministry of Education and Research (BMBF, Germany). We acknowledge funding of M. Santarossa by the KI-SIGS project (Grant number FKZ 01MK20012E) funded by the Federal Ministry for Economic Affairs and Energy (BMWi, Germany). S-M Schöder was supported by the “CUSCO—Coastal Upwelling System in a Changing Ocean” project (Grant number 03F0813) funded by the Federal Ministry of Education and Research (Germany). R Kiko also acknowledges support via a “Make Our Planet Great Again” grant of the French National Research Agency within the “Programme d’Investissements d’Avenir”; reference “ANR-19-MPGA-0012”. Funds to conduct the PlanktonID project were granted to R Kiko and R Koch (CP1733) by the Cluster of Excellence 80 “Future Ocean” within the framework of the Excellence Initiative by the Deutsche Forschungsgemeinschaft (DFG) on behalf of the German federal and state governments. This work was supported by Land Schleswig-Holstein through the Open Access Publikationsfonds Funding Program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The used STL-10 dataset is introduced in [25]. The raw image plankton data is hosted on EcoTaxa [45] and the annotations were created in the project PlanktonID <https://planktonid.geomar.de/de> (accessed on 6 October 2021). The annotations can be requested from the original data owners. The source code is available at <https://github.com/Emprime/FuzzyOverclustering> (accessed on 6 October 2021). The used data is available at <https://doi.org/10.5281/zenodo.5550918> (accessed on 6 October 2021).

Acknowledgments: We thank our colleagues, especially Claudius Zelenka, for their helpful feedback and recommendations on improving the paper. Moreover, we are grateful for all citizen scientist which participated in PlanktonID and the team of PlanktonID for providing us with their data. We thank Xu Ji, Ting Chen, Kihyuk Sohn and Wouter Van Gansbeke for answering our questions regarding their respective work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Saleh, A.; Laradji, I.H.; Konovalov, D.A.; Bradley, M.; Vazquez, D.; Sheaves, M. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Sci. Rep.* **2020**, *10*, 14671. [[CrossRef](#)] [[PubMed](#)]
2. Gómez-Ríos, A.; Tabik, S.; Luengo, J.; Shihavuddin, A.S.M.; Krawczyk, B.; Herrera, F. Towards highly accurate coral texture images classification using deep convolutional neural networks and data augmentation. *Expert Syst. Appl.* **2019**, *118*, 315–328. [[CrossRef](#)]
3. Thum, G.W.; Tang, S.H.; Ahmad, S.A.; Alrifayy, M. Toward a highly accurate classification of underwater ¹⁶¹ images via deep convolutional neural network. *J. Mar. Sci. Eng.* **2020**, *8*, 924. [[CrossRef](#)]

4. Knausgård, K.M.; Wiklund, A.; Sordalen, T.K.; Halvorsen, K.T.; Kleiven, A.R.; Jiao, L.; Goodwin, M. Temperate fish detection and classification: A deep learning based approach. *Appl. Intell.* **2021**. [\[CrossRef\]](#)
5. Lombard, F.; Boss, E.; Waite, A.M.; Uitz, J.; Stemmann, L.; Sosik, H.M.; Schulz, J.; Romagnan, J.B.; Picheral, M.; Pearlman, J.; et al. Globally consistent quantitative observations of planktonic ecosystems. *Front. Mar. Sci.* **2019**, *6*, 196. [\[CrossRef\]](#)
6. Giering, S.L.C.; Cavan, E.L.; Basedow, S.L.; Briggs, N.; Burd, A.B.; Darroch, L.J.; Guidi, L.; Irisson, J.O.; Iversen, M.H.; Kiko, R.; et al. Sinking Organic Particles in the Ocean—Flux Estimates From in situ Optical Devices. *Front. Mar. Sci.* **2020**, *6*, 834. [\[CrossRef\]](#)
7. Addison, P.F.E.; Collins, D.J.; Trebilco, R.; Howe, S.; Bax, N.; Hedge, P.; Jones, G.; Miloslavich, P.; Roelfsema, C.; Sams, M.; et al. A new wave of marine evidence-based management: Emerging challenges and solutions to transform monitoring, evaluating, and reporting. *ICES J. Mar. Sci.* **2018**, *75*, 941–952. [\[CrossRef\]](#)
8. Durden, J.M.; Bett, B.J.; Schoening, T.; Morris, K.J.; Nattkemper, T.W.; Ruhl, H.A. Comparison of image annotation data generated by multiple investigators for benthic ecology. *Mar. Ecol. Prog. Ser.* **2016**, *552*, 61–70. [\[CrossRef\]](#)
9. Schoening, T.; Bergmann, M.; Ontrup, J.; Taylor, J.; Dannheim, J.; Gutt, J.; Purser, A.; Nattkemper, T.W. Semi-automated image analysis for the assessment of megafaunal densities at the Arctic deep-sea observatory HAUSGARTEN. *PLoS ONE* **2012**, *7*, e38179. [\[CrossRef\]](#)
10. Schröder, S.M.; Kiko, R.; Koch, R. MorphoCluster: Efficient Annotation of Plankton images by Clustering. *Sensors* **2020**, *20*, 3060. [\[CrossRef\]](#)
11. Karimi, D.; Dou, H.; Warfield, S.K.; Gholipour, A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med. Image Anal.* **2020**, *65*, 101759. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Brünner, J.; Dippel, S.; Koch, R.; Veit, C. ‘Tailception’: Using neural networks for assessing tail lesions on pictures of pig carcasses. *Animal* **2019**, *13*, 1030–1036. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Schmarje, L.; Zelenka, C.; Geisen, U.; Glüer, C.C.; Koch, R. 2D and 3D Segmentation of Uncertain Local Collagen Fiber Orientations in SHG Microscopy. In *DAGM German Conference of Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11824 LNCS, pp. 374–386. [\[CrossRef\]](#)
14. De Fauw, J.; Ledsam, J.R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; Blackwell, S.; Askham, H.; Glorot, X.; O’Donoghue, B.; Visentin, D.; et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **2018**, *24*, 1342–1350. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Karimi, D.; Nir, G.; Fazli, L.; Black, P.C.; Goldenberg, L.; Salcudean, S.E. Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images—Role of Multiscale Decision Aggregation and Data Augmentation. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 1413–1426. [\[CrossRef\]](#)
16. dos Reis, F.J.C.; Lynn, S.; Ali, H.R.; Eccles, D.; Hanby, A.; Provenzano, E.; Caldas, C.; Howat, W.J.; McDuffus, L.A.; Liu, B.; et al. Crowdsourcing the general public for large scale molecular pathology studies in cancer. *EBioMedicine* **2015**, *2*, 681–689. [\[CrossRef\]](#)
17. Culverhouse, P.; Williams, R.; Reguera, B.; Herry, V.; González-Gil, S. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Mar. Ecol. Prog. Ser.* **2003**, *247*, 17–25. [\[CrossRef\]](#)
18. Tarling, P.; Cantor, M.; Clapés, A.; Escalera, S. Deep learning with self-supervision and uncertainty regularization to count fish in underwater images. *arXiv* **2021**, arXiv:2104.14964.
19. Berthelot, D.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Sohn, K.; Zhang, H.; Raffel, C. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. *arXiv* **2019**, arXiv:1911.09785.
20. Zhai, X.; Oliver, A.; Kolesnikov, A.; Beyer, L. S4L: Self-Supervised Semi-Supervised Learning. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1476–1485.
21. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv* **2020**, arXiv:2002.05709.
22. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised Representation Learning by Predicting Image Rotations. *arXiv* **2018**, arXiv:1803.07728.
23. Ji, X.; Henriques, J.F.; Vedaldi, A.; Ji, X.; Henriques, J.F.; Vedaldi, A. Invariant information clustering for unsupervised image classification and segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9865–9874.
24. Schmarje, L.; Santarossa, M.; Schroder, S.M.; Koch, R. A Survey on Semi-, Self-and Unsupervised Learning for Image Classification. *IEEE Access* **2021**, *9*, 82146–82168. [\[CrossRef\]](#)
25. Coates, A.; Ng, A.; Lee, H. An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA 11–13 April 2011; pp. 215–223.
26. Algan, G.; Ulusoy, I. Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey. *Knowl.-Based Syst.* **2021**, *215*, 106771. [\[CrossRef\]](#)
27. Song, H.; Kim, M.; Park, D.; Lee, J. Learning from Noisy Labels with Deep Neural Networks: A Survey. *arXiv* **2020**, arXiv:1406.2080.
28. Nguyen, D.T.; Mummadi, C.K.; Ngo, T.P.N.; Nguyen, T.H.P.; Beggel, L.; Brox, T. SELF: Learning to Filter Noisy Labels with Self-Ensembling. *arXiv* **2019**, arXiv:1910.01842.
29. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. *arXiv* **2016**, arXiv:1610.02242.
30. Li, J.; Socher, D.; Hoi, S.C.H. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. *arXiv* **2020**, arXiv:2002.07394.
31. Geng, X. Label distribution learning. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1734–1748. [\[CrossRef\]](#)

32. Gao, B.B.; Xing, C.; Xie, C.W.; Wu, J.; Geng, X. Deep Label Distribution Learning With Label Ambiguity. *IEEE Trans. Image Process.* **2017**, *26*, 2825–2838. [[CrossRef](#)] [[PubMed](#)]
33. Liu, J.; Ma, Y.; Qu, F.; Zang, D. Semi-supervised Fuzzy Min–Max Neural Network for Data Classification. *Neural Process. Lett.* **2020**, *51*, 1445–1464. [[CrossRef](#)]
34. Kowsari, K.; Bari, N.; Vichr, R.; Goodarzi, F.A. FSL-BM: Fuzzy Supervised Learning with Binary Meta-Feature for Classification. In *Future of Information and Communication Conference*; Springer: Cham, Switzerland, 2018; pp. 655–670.
35. El-Zahhar, M.M.; El-Gayar, N.F. A semi-supervised learning approach for soft labeled data. In Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications, Cairo, Egypt, 29 November–1 December 2010; pp. 1136–1141.
36. Liu, Y.; Liang, X.; Tong, S.; Kumada, T. Photo Shot-Type Disambiguation by Multi-Classifer Semi-Supervised Learning. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2466–2470.
37. Caron, M.; Bojanowski, P.; Joulin, A.; Douze, M. Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 132–149.
38. Sohn, K.; Berthelot, D.; Li, C.L.; Zhang, Z.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Zhang, H.; Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv* **2020**, arXiv:2001.07685.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, the Netherlands, 8–16 October 2016; pp. 630–645.
40. Grandvalet, Y.; Bengio, Y. Semi-supervised learning by entropy minimization. *Adv. Neural Inf. Process. Syst.* **2005**, *367*, 529–536.
41. Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.T.; Le, Q.V. Unsupervised Data Augmentation for Consistency Training. *arXiv* **2019**, arXiv:1904.12848.
42. Miyato, T.; Maeda, S.I.; Koyama, M.; Ishii, S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1979–1993.
43. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. Mixmatch: A holistic approach to semi-supervised learning. *arXiv* **2019**, arXiv:1905.02249.
44. Picheral, M.; Guidi, L.; Stemann, L.; Karl, D.M.; Iddoud, G.; Gorsky, G. The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnol. Oceanogr. Methods* **2010**, *8*, 462–473. [[CrossRef](#)]
45. Picheral, M.; Colin, S.; Irisson, J.O. EcoTaxa, a Tool for the Taxonomic Classification of Images. 2017. Available online: <https://ecotaxa.obs-vlfr.fr/> (accessed on 6 October 2021).
46. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *60*, 1097–1105. [[CrossRef](#)]
47. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. Technical Report. 2009. Available online: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf> (accessed on 6 October 2021).
48. Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; Van Gool, L. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 268–285.
49. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv* **2017**, arXiv:1703.01780.
50. Lee, D.H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning*; ICML: Atlanta, GA, USA, 2013; Volume 3, p. 2.

A. Own previous papers

A.1.4 A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering

A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering

Lars Schmarje¹[0000-0002-6945-5957], Monty Santarossa¹[0000-0002-4159-1367],
Simon-Martin Schröder¹[0000-0002-6603-9907], Claudius
Zelenka¹[0000-0002-9902-2212], Rainer Kiko²[0000-0002-7851-9107], Jenny
Stracke³[0000-0002-9986-9720], Nina Volkmann⁴[0000-0003-2870-9954], and
Reinhard Koch¹[0000-0003-4398-1569]

¹ MIP, Computer Science, Kiel University, Germany
{las,msa,sms,czw,rk}@informatik.uni-kiel.de

² LOV, Sorbonne Université, France rainer.kiko@obs-vlfr.fr

³ ITW, University of Bonn jenny.stracke@itw.uni-bonn.de

⁴ WING & ITTN, University of Veterinary Medicine Hannover
nina.volkmann@tiho-hannover.de

Abstract. Consistently high data quality is essential for the development of novel loss functions and architectures in the field of deep learning. The existence of such data and labels is usually presumed, while acquiring high-quality datasets is still a major issue in many cases. Subjective annotations by annotators often lead to ambiguous labels in real-world datasets. We propose a data-centric approach to relabel such ambiguous labels instead of implementing the handling of this issue in a neural network. A hard classification is by definition not enough to capture the real-world ambiguity of the data. Therefore, we propose our method "Data-Centric Classification & Clustering (DC3)" which combines semi-supervised classification and clustering. It automatically estimates the ambiguity of an image and performs a classification or clustering depending on that ambiguity. DC3 is general in nature so that it can be used in addition to many Semi-Supervised Learning (SSL) algorithms. On average, our approach yields a 7.6% better F1-Score for classifications and a 7.9% lower inner distance of clusters across multiple evaluated SSL algorithms and datasets. Most importantly, we give a proof-of-concept that the classifications and clusterings from DC3 are beneficial as proposals for the manual refinement of such ambiguous labels. Overall, a combination of SSL with our method DC3 can lead to better handling of ambiguous labels during the annotation process.⁵

Keywords: Data-Centric, Clustering, Ambiguous Labels

A. Own previous papers

2 L. Schmarje et al.

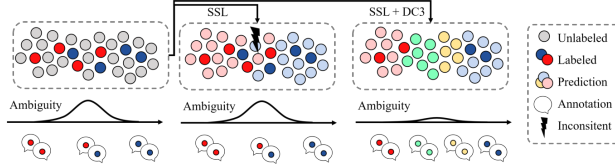


Fig. 1: Benefit of data-centric classification and clustering (DC3) over Semi-Supervised Learning (SSL) - Real-world datasets often suffer from intra- or inter-observer variability (IIV) during the annotation and thus no clear separation of classes is given as in common benchmark datasets. Images with a high variability between the annotations therefore have a ambiguous label. These ambiguous labels can perturb SSL approaches (see lightning bolt) and result in inconsistent predictions. Our method DC3 can be used in combination with SSL to identify ambiguous images automatically and cluster them, while classifying the rest as usual. Therefore, we avoid label ambiguity during training and generate cluster proposals which can be used to create more consistent labels.

1 Introduction

In recent years, deep learning has been successfully applied to many computer vision problems [21, 49, 11, 42, 40, 15]. The availability of large high-quality datasets was a main reason for this success, as this enabled machine learning to incorporate a wide variety of real world patterns [30]. Many novel loss functions and architectures have been proposed including options to handle imperfect data [51, 58]. This model-centric view mostly tries to deal with issues like label bias [35], label noise [26] or ambiguous labels [17] instead of improving the dataset during the annotation process. Following recent data-centric literature [4, 43, 45], we therefore investigate in this paper an approach to improve the dataset during the annotation process.

Specifically, we study the impact of ambiguous labels due to *intra- or inter-observer variability* (IIV). Such variability may arise from variability / inconsistency of annotations over time or between annotators. This issue is common when annotating data [39, 45, 26, 43, 47, 24, 44, 37, 5, 46, 16, 25, 14, 19]. The literature names different possible reasons for this variability such as low resolution [39], bad quality [22, 47], subjective interpretations of classes [25, 37] or mistakes [26, 33].

We assume that this variability can be modeled for each image with an unknown soft probability distribution $l \in [0, 1]^k$ for a classification problem with k classes. Many previous methods use a hard label instead of a soft label for training and therefore can not model this issue by definition. We call a label and its corresponding image *certain* if all annotators would agree on the

⁵ Source code is available at <https://github.com/Emprime/dc3>

classification ($l \in \{0, 1\}$) and *ambiguous* if they would disagree ($l \in (0, 1)$). In other words, ambiguous images are likely to have different annotations due to IIV while certain images do not. It is problematic that the unknown distribution l can only be estimated with expensive operations such as the acquisition of multiple annotations. Real-world example images with certain and ambiguous labels are given in Figure 3 and detailed definitions are given in subsection 2.1.

The goal of this paper is to introduce a method which provides predictions which are beneficial for improving ambiguous labels via relabeling in a downstream task. The quality of ambiguous labels and thus the performance of trained models [4, 60] can easily be improved with more annotations. However, more annotations are associated with a higher cost in the form of human working hours. Semi-Supervised Learning (SSL) can reduce these costs because it has shown great potential in reducing the amount of required labeled data to 10% or even 1% while maintaining classification performance [49, 27, 11, 61, 8]. SSL can even boost performance further [59, 40] on already large labeled datasets like ImageNet [30].

Therefore, we propose Data-Centric Classification & Clustering (DC3) which can be used in combination with many SSL algorithms to perform a combined semi-supervised classification and clustering. It simultaneously distinguishes between ambiguous and certain images, classifies the certain images and clusters visually similar ambiguous images. A graphical summary is provided in Figure 1. We will show that this approach leads to better classifications and more compact clusters across multiple semi-supervised algorithms and non-curated datasets. Furthermore, we give a proof-of-concept that these improvements lead to a greater consistency of labels based on proposals from DC3.

The key contributions of this paper are: (1) DC3 allows an SSL algorithm to predict on average a 7.6% better F1-Score for classifications and a 7.9% lower inner distance of clustering across multiple algorithms and non-curated datasets. The hyperparameters of DC3 are fixed across all algorithms and datasets which illustrates the general applicability of our method. (2) We give a proof-of-concept that these improved predictions can be used to create labels on average 2.4-fold faster and 6.74% more consistent, in comparison to the non-extended algorithms and a consensus process. This leads to higher quality data for further evaluation or model training. (3) DC3 can be used in combination with many SSL algorithms without a noticeable trade-off in terms of run-time or memory consumption, which should enable many further applications.

1.1 Related Work

Our method is mainly related to Data-Centric Machine Learning, Semi-Supervised Learning and Classification & Clustering.

Data-Centric Machine Learning aims at improving the data quality rather than improving the model alone [45, 36]. The data issues like imperfect, ambiguous or erroneous labels [4, 52, 60, 14, 24, 13, 1] are often handled in a model-centric approach by detecting errors or making the models more robust [50, 9, 26, 2]. We

A. Own previous papers

4 L. Schmarje et al.

want to use the predictions of our model to improve the annotation process and therefore prevent or minimize the quality issues before they need to be handled in particular.

Semi-Supervised Learning [10] is mainly developed on curated benchmark datasets [30, 12, 29] where the issue of IIV is not considered. In contrast to other SSL research [11, 61, 8, 18, 49], we are not evaluating on these curated benchmarks but work with new real-world datasets for two reasons. Firstly, curated datasets do not suffer so much from IIV because they were already cleaned. Recent research indicates that even these datasets suffer from errors in the labels which negatively impact the performance [39, 4]. Secondly, if we want to evaluate the IIV issue, we need an approximation of the variability of the label for each image e.g. in the form of multiple annotations per image. However, this information is not provided for current state-of-the-art benchmarks except for datasets like [39, 4].

Classification&Clustering was investigated in detail [41, 38, 6, 7, 54]. However, classical low dimensional approaches are difficult to extend to real-world images [41, 38, 6], and many deep-learning methods use the clustering only as a proxy task before the actual classification [55, 23, 43] or iterate between classifications and clusters [7, 54]. The work by Smieja et al. is a rare example where classification and clustering results are generated in parallel in each training step [48]. However, we want to automatically decide which data should be classified or clustered due to their underlying ambiguity.

2 Method

Our method Data-Centric Classification & Clustering (DC3) is not an individual method but an extension for SSL algorithms such as [3, 49, 53, 32, 31]. Any image classification model can be combined with DC3 as long as it is compatible with the definition of an arbitrary SSL algorithm below.

2.1 Definitions

We assume that every image $x \in X$ has an unknown soft probability distribution $l \in [0, 1]^k$ for a classification problem with k classes. This assumption is based on two main reasons. Firstly, inconsistent annotations exist due to subjective opinions from the annotators, e.g. the grading of an illness [25]. A hard label $l \in \{0, 1\}^k$ could not model such a difference over the complete annotator population. Secondly, if we consider biological processes, images of intermediate transition stages between two classes, such as the degeneration of a living underwater organism to dead biomass exist [43].

An image and its corresponding label l are *ambiguous* if $i, j \in \{1, \dots, k\}$ exist with $i \neq j, l_i > 0$ and $l_j > 0$. Otherwise the image and its label are *certain*. The ambiguity of a label is $1 - \max_{i \in \{1, \dots, k\}} l_i$. An image might be ambiguous because

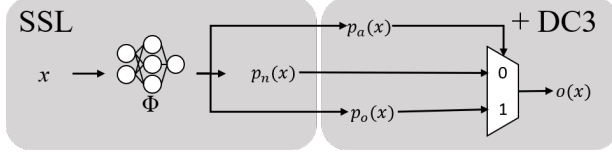


Fig. 2: Our method DC3 and an extended arbitrary SSL method – The SSL algorithm passes an image x through the network Φ and outputs a classification $p_n(x)$. We add two additional outputs: an overclustering $p_o(x)$ and a ambiguity estimation $p_a(x)$. The ambiguity estimation $p_a(x)$ is used to determine if the classification or the overclustering output is used for our method DC3. Only some labels are available for the classification output and therefore most images have to be trained completely self-supervised on all outputs.

it is actually an intermediate or uncertain combination of different classes as stated above. For this reason, ambiguous images are not just wrongly assigned images.

An SSL algorithm uses a labeled dataset X_l and an unlabeled dataset X_u for the training of a neural network Φ with $X = X_l \cup X_u$. For all images $x \in X_l$ a hard label l is available while no label information is available for $x \in X_u$. The output $p_n(x) := \Phi(x)$ is a probability distribution over the k classes.

2.2 DC3

Our method DC3 extends an arbitrary SSL algorithm. The SSL algorithm passes an image x through the network Φ and predicts a classification $p_n(x) \in [0, 1]^k$. DC3 calculates two additional outputs without a noticeable impact on training time or memory consumption: a clustering assignment $p_o(x) \in [0, 1]^{k'}$ with $k' > k$ and an ambiguity estimation $p_a(x) \in [0, 1]$. The cluster assignment partitions visually similar images in more clusters than classes exist (overclustering with $k' > k$). The ambiguity estimation is used to determine if a classification ($p_n(x)$) or an (over)clustering ($p_o(x)$) should be used as the final output. The image is predicted as certain and the classification is used if $p_a(x) < 0.5$. Otherwise, the image is estimated as ambiguous and the clustering is used as output.

A key difference to previous literature [23, 43, 7] is that we do not create an additional or only a clustering of all samples. We create SSL classifications for certain images while ambiguous images are clustered without prescribed knowledge. Moreover, it is not feasible to determine ambiguous images before this classification/clustering and thus we have no ground-truth for this decision as well. These conditions led us to formulate three goals for our development: 1. The underlying SSL training must be possible and not negatively impacted while computing an additional overclustering. 2. A degeneration to one or random cluster assignments has to be avoided as no ground-truth is available for the

A. Own previous papers

6 L. Schmarje et al.

clustering. 3. A balance between certain and ambiguous images is needed as the same argument (no-ground truth) applies to the ambiguity estimation $p_a(x)$ For this purpose, the network is trained by minimizing the following loss function which benefits from SSL but avoids the described degenerations.

$$L(x) = L_{SSL}(x) \cdot [1 - p_a(x)] + \lambda_{CE^{-1}} L_{CE^{-1}}(x) \cdot [1 - p_a(x)] \\ + \lambda_a L_A(x) + \lambda_s L_S(x) \cdot p_a(x) \quad (1)$$

The first three loss terms correspond to the outputs $p_n(x)$, $p_o(x)$ and $p_a(x)$ and the three goals described above, respectively. The last term (L_S) is optional and stabilizes the training. The λ values are weights to balance the impact of each term. The first loss L_{SSL} is the loss calculated by the original SSL algorithm and is only scaled with $[1 - p_a(x)]$ to prevent the original SSL training on images the network predicts as ambiguous.

The second loss $L_{CE^{-1}}$ incentives visually homogeneous clusters of the images by pushing images from different classes into different clusters. This loss is needed to prevent a degeneration of the clustering. A similar loss was used in [43] but could only be trained on labeled data, with pretrained networks and several inefficient stabilizing methods like repeating every sample 3-5 times per batch. We generalized the formula for two input images x, x' of the same mini-batch which should not be of the same class:

$$CE^{-1}(p_o(x), p_o(x')) = - \sum_{c=1}^k p_o(x)_c \cdot \ln(1 - p_o(x')_c). \quad (2)$$

For the selection of x, x' , we use either the ground-truth label l of x if it is available or the Pseudo-Label based on the network prediction $p_n(x)$. The loss is also scaled with $[1 - p_a(x)]$ because it uses an estimate of the class for an image which could be wrong / ill-suited for ambiguous images.

The third loss L_A allows the ambiguity estimation. As stated above, the underlying distribution l is unknown and thus we do not know during training if x is ambiguous or certain. However, we can expect to know or be given a prior probability $p_A \in [0, 1]$ of the expected percentage of ambiguous images in the total dataset. We set p_A to a fixed value which balances certain and ambiguous images and the details are given in subsection 3.3. Based on this probability, we can estimate a Pseudo-Label of the ambiguity of each image in a batch during training. The loss L_A is the binary cross-entropy between the Pseudo-Label $h(x)$ and $p_a(x)$. The usage of hot-encoded Pseudo-Labels forces the network to make more confident predictions. The formulation is given below with i as the index of the image x inside the given batch, when all images inside the batch are sorted in ascending order based on p_a .

$$L_A(x) = CE(h(x), p_a(x)) \\ = -(1 - h(x)) \cdot \ln(p_a(x = 0)) \\ - h(x) \cdot \ln(p_a(x = 1)) \text{ with} \quad (3) \\ h(x) = \begin{cases} 1 & i \leq \text{batch size} \cdot p_A \\ 0 & \text{else} \end{cases}$$

The fourth term L_S is the cross-entropy (CE) between $p_o(x)$ and $p_o(x')$ for two differently augmented versions x, x' of the same image. This loss is scaled with $p_a(x)$ and incentives that augmented versions of the same ambiguous image are in the same output cluster. We use CE because it indirectly minimizes also the entropy of $p_o(x)$ which leads to sharper predictions. Many SSL algorithms already use a differently augmented version x' of x as secondary input [3, 49, 53, 31, 23] which allows an easy computation. Otherwise, the fourth term is not calculated and treated as zero.

It is important to note that only the proposed combination of the individual parts leads to a successful training of all desired outputs. We show in section 4 that the combined clustering and classification (CC) based on $p_a(x)$ and the loss L_{CE-1} are the two essential parts to DC3.

3 Experiments

3.1 Datasets

That our method can be applied to many SSL algorithms across different real-world ambiguous datasets without major changes is a major advantage. While many datasets [39, 26, 43, 47, 24, 37, 5, 16, 25, 14] suffer from annotation variability, we do not know the unknown underlying distribution l to evaluate the ambiguity or any related metrics. We can approximate l with the average over multiple annotations from humans. An annotation is the hard coded guess $a = (a_1, \dots, a_k) \in \{0, 1\}^k$ of a class for an image from a human with exactly one $i' \in \{1, \dots, k\} : a_{i'} = 1$ and for all $j \in \{1, \dots, k\} \setminus \{i'\} : a_j = 0$. We assume that the approximation \hat{l} as the average of n annotations is identical to the unknown distribution l for $n \rightarrow \infty$. This leaves the issue that we need multiple annotations per image for a dataset with ambiguous labels which are often not available. However, all datasets summarized in Table 1 have multiple annotations and thus allow the approximation of \hat{l} . Nine visual examples for all datasets are given in Figure 3 and the datasets are shortly introduced below.

The *Plankton* dataset was introduced in [43]. The dataset contains 10 plankton classes and has multiple labels per image due to the help of citizen scientists. In contrast to [43], we include ambiguous images in the training and validation set and do not enforce a class balance which results in a slightly different data split as shown in Figure 3. Moreover, we processed the data by recentering the images and removing artifacts like scale bars.

The *Turkey* dataset was used in [56, 57]. The dataset contains cropped images of potential injuries of the birds which were separately annotated by three experts as not injured or injured.

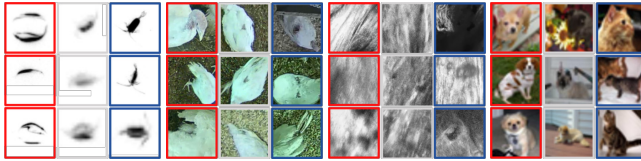
The *Mice Bone* dataset is based on raw data which was published in [47]. The raw data are 3D scans from collagen fibers in mice bones. The three proposed classes are similar as well as dissimilar collagen fiber orientations and not relevant regions due to noise or background. We used the given segmentations to cut image regions from the original 2D image slices which mainly consist of one

A. Own previous papers

8 L. Schmarje et al.

Table 1: Overview of the used datasets – # is an abbreviation for number. The class imbalance is given as the percentage of the smallest and largest class with regard to the complete dataset. \hat{p}_A is the expected prior ambiguity probability of the dataset. n is the average of annotations per image.

Name	# classes	Input size [px]	# Images			Class Imbalance [%]		\hat{p}_A [%]	n
			Train	Val	Unlabeled	Smallest	Largest		
Plankton [43]	10	96x96	1964	2456	7860	4.16	30.37	44	24
Turkey [7]	2	96x96	1299	1542	5199	9.66	90.33	22	3
Mice Bone [47]	3	224x224	277	169	278	10.81	63.98	65	3
CIFAR-10H [39]	10	32x32	1600	2000	6400	9.88	10.16	32	51



(a) Plankton [43] (b) Turkey [56] (c) Mice Bone [47] (d) CIFAR-10H [39]

Fig. 3: Example images for the ambiguous real-world datasets – All datasets have certain images (red & blue) and ambiguous images between these classes (grey). The classes are Bubble & Copepod, Not Injured & Injured, Similar & Dissimilar Orientations and Dog & Cat respectively.

class. We generated ambiguous GT labels on 10% of the generated images by averaging over three classifications from an expert.

The *CIFAR-10H* [39] dataset provides multiple annotations for the test set of *CIFAR-10*[29]. This dataset is interesting because it illustrates that even the hard labels from benchmark datasets like *CIFAR-10* are based on soft labels due to IIV.

As stated above the approximation of \hat{l} is only possible with multiple annotations per image. For the *STL-10* dataset [12], only one annotation / label per image is given. We still include some results of this dataset to illustrate the performance on previous benchmarks.

For all datasets, we split our images X into a labeled X_l and an unlabeled X_u training set. We keep additional images as a validation subset. On X_l , we use for each image a random hard label sampled from the corresponding \hat{l} . This simulates the noisy approximation of the true ground truth label l . On X_u , we can only use the image information and not any label information. The validation data is used to compare the trained networks and to detect issues like overfitting.

3.2 Metrics

We want to measure the quality of classification and clusters over the certain and ambiguous data respectively which we assume are better proposals in the annotation or evaluation process. Based on this reasoning, we decided to use the weighted F1-Score on certain data and the mean inner distance on ambiguous data. The ambiguity is determined by the network output p_a . We define the metrics in detail below and give in subsection 3.5 a proof-of-concept for the higher consistency of labels based on proposals selected by the defined metrics. Common metrics like accuracy are not used as the class imbalance of several of our datasets would lead to misleading results.

During training we do not enforce a balance between ambiguous and certain predictions to keep the required prior knowledge minimal. This can lead to uninformative metrics and therefore we call a training *degenerated* if no more than 10% of the validation data are either predicted as ambiguous or certain. We use the *weighted F1-Score* on certain images, based on the number of images per class to avoid instability due to classes with no or very few certain (predicted) images. For the ambiguous images, we use the mean inner euclidean distance (d) to the centroid on the soft / ambiguous Ground-Truth (GT) labels. The metric d is based on the soft GT and thus also minimal for classifications of the majority class which allows an evaluation also on classified data. The equation for a set of clusters of images X is given in Equation 4 with sets $C \in X$ as clusters and the corresponding approximated soft label distribution \hat{l}_x for each image $x \in C$. The centroid per cluster is given as μ_C .

$$\begin{aligned} d(X) &:= \frac{1}{|X|} \sum_{C \in X} \frac{1}{|C|} \sum_{x \in C} \|\hat{l}_x - \mu_C\|_2 \text{ with} \\ \mu_C &:= \frac{1}{|C|} \sum_{x \in C} \hat{l}_x \end{aligned} \tag{4}$$

We use the vanilla (unchanged) SSL algorithms as baseline experiments. For these experiments and some ablation experiments, we have no ambiguity prediction $p_a(x)$. In these cases, we assume all images to be certain and use $p_n(x)$ as output. We often noticed that the classification improved while the clustering degenerated and the other way round. Therefore, we determine the best performance considering the difference (d -F1) between distance and F1-Score (smaller is better). It is important to note that this balancing is arbitrary, but we give a proof of concept that the proposals calculated by these metrics lead to more consistent annotations which justifies their definition. In general, we have 3 runs per setup but we exclude results that degenerate as described above. We report the best of these runs based on the (d -F1)-score over all non-degenerated runs. All scores are calculated on the validation data which is in general about 20% of all the data (see details in Table 1).

A. Own previous papers

10 L. Schmarje et al.

3.3 Implementation Details

All methods use the same code base and share major hyperparameters which is crucial for valid comparisons [28]. We use the prior ambiguity $p_A = 0.6$ and loss weights $\lambda_{CE^{-1}} = 10$, $\lambda_f = 0.1$ and $\lambda_s = 0.1$ across all experiments. It is important to note that we do not use the actual prior probability of ambiguous images p_A as given in Table 1 because the probability is unknown or would require multiple annotations per image. We use a constant approximation across all datasets and show in section 4 that this approximation is comparable or even better than p_A . This parameter is essential for balancing the certain and ambiguous images. The batch size was 64 for all datasets except for the mice bone dataset with a batch size of 8. The additional losses L_A and L_S are only applied on the unlabeled data while $L_{CE^{-1}}$ is also calculated on the labeled data. These hyperparameters were determined heuristically on the Plankton dataset with Mean-Teacher and show strong results across different methods and datasets as shown in subsection 3.4. Most likely these parameters are not optimal for an individual combination of a method and a dataset but they show the general applicability across methods and datasets. We want to show that DC3 can be applied successfully to other datasets without hyperparameter optimization and thus did not investigate all combinations in detail. Nevertheless, we refer to the supplementary for more detailed insights about individual hyperparameters and the complete pseudo code for the loss calculation.

3.4 Evaluation

The comparison between different SSL algorithms and their extension with DC3 is given in Table 2. The best results were selected as described in subsection 3.2. The complete results and additional plots are given in the supplementary. We see that DC3 improves the classification and clustering performance across the majority of classes and methods by 5 to 10%. ($d-F1$) is improved by up to 40% for 16 out of 19 method-dataset-combinations. On average, we achieve a 7.6% higher F1-Score for certain classifications and a 7.9% lower inner distance for clusterings of ambiguous images if we look at all non excluded method-dataset-combinations. Even on STL-10 (without the possibility to evaluate ambiguous labels) DC3 creates up to 9% better classifications. Overall, we see the most benefit on the Mice Bone and Turkey dataset which we attribute to the worse initial approximation of \bar{l} . The different vanilla algorithms achieve quite similar results for each dataset. Only FixMatch achieves a more than 5% better F1-Score on the curated STL-10 and CIFAR-10H dataset. In general, we see that DC3 can be beneficially applied to a variety of datasets and methods and predicts better classifications and more compact clusters.

Additional results about the impact of ambiguous data, the unlabeled data ratio and the interpretability can be found in the supplementary.

Table 2: Performance across different methods and datasets – The vanilla algorithm is highlighted in light grey. Better results in comparison to the vanilla algorithm are marked bold. The definition of the metrics are given in subsection 3.2. CE stands for supervised Cross-Entropy training. All values are given in %. Reasons for exclusion: H - Hardware Restrictions

Methods	Plankton			Turkey			Mice Bone			CIFAR-10H			STL-10
	F1 ↑	d ↓	(d-F1) ↓	F1 ↑	d ↓	(d-F1) ↓	F1 ↑	d ↓	(d-F1) ↓	F1 ↑	d ↓	(d-F1) ↓	F1 ↑
CE	86.71	30.45	-56.26	83.84	42.98	-40.86	69.55	54.75	-14.80	67.71	55.80	-11.91	80.48
CE + DC3	78.24	23.41	-54.84	85.79	27.64	-58.14	93.88	36.58	57.30	78.27	54.52	-23.75	88.45
Mean-Teacher [53]	88.72	25.84	-62.88	81.82	45.12	-36.70	66.41	48.83	-17.58	73.53	46.93	-26.59	80.67
Mean-Teacher [53] + DC3	91.30	24.84	-66.46	86.45	33.92	-52.53	89.4	35.11	-54.73	85.13	52.44	-32.69	89.28
Pt-Model [31]	87.57	28.43	-59.14	82.11	39.46	-42.65	68.15	54.11	-14.04	71.53	49.13	-22.40	82.56
Pt-Model [31] + DC3	79.79	19.08	-60.71	87.43	23.33	-64.10	88.01	30.99	-57.02	83.05	43.40	-39.65	89.54
Pseudo-Label [32]	87.62	27.42	-60.20	82.37	44.88	-37.49	66.60	57.03	-9.57	69.70	53.30	-16.40	82.48
Pseudo-Label [32] + DC3	89.31	31.76	-57.55	83.44	35.04	-48.41	86.58	37.52	-49.06	83.74	51.32	-32.42	88.87
FixMatch [49]	85.81	30.29	-55.52	82.14	43.33	-38.81	H	H	H	78.09	41.99	-36.10	89.35
FixMatch [49] + DC3	87.20	31.28	-55.92	83.56	28.17	-55.39	H	H	H	83.09	49.49	-33.60	91.45

Table 3: Consistency comparison of generated labels from proposals – The first column describes the annotator selection and the used proposals. The Cohen's kappa coefficient κ measures the agreement of between the used repetitions and Time gives annotation time in minutes. Results which are within one percent or minute of the best result per dataset and annotator selection are marked bold.

	Plankton				Turkey				Mice Bone				CIFAR-10H			
	κ [%] ↑	Time [min] ↓	κ [%] ↑	Time [min] ↓	κ [%] ↑	Time [min] ↓	κ [%] ↑	Time [min] ↓	κ [%] ↑	Time [min] ↓	κ [%] ↑	Time [min] ↓	κ [%] ↑	Time [min] ↓	κ [%] ↑	Time [min] ↓
A1	73.00 ± 1.51	51.09 ± 2.36	88.08 ± 3.43	14.56 ± 0.84	71.35 ± 2.56	13.94 ± 2.25	92.70 ± 1.69	40.58 ± 1.93								
A1 + SSL	85.00 ± 2.52	12.69 ± 3.37	85.63 ± 3.66	10.70 ± 0.44	72.00 ± 2.87	6.59 ± 1.65	94.85 ± 0.91	14.33 ± 1.48								
A1 + DC3	90.29 ± 1.41	11.32 ± 1.43	91.95 ± 1.22	11.57 ± 0.64	81.36 ± 2.17	6.74 ± 1.05	94.70 ± 0.52	14.65 ± 0.60								
A2	85.25 ± 1.79	61.99 ± 10.98	81.54 ± 0.89	18.11 ± 4.30	68.63 ± 6.66	11.06 ± 3.60	98.81 ± 0.14	33.08 ± 5.36								
A2 + SSL	94.88 ± 0.52	9.23 ± 0.76	81.10 ± 3.39	9.48 ± 0.83	59.63 ± 6.20	12.07 ± 4.77	98.00 ± 0.27	12.66 ± 0.69								
A2 + DC3	94.04 ± 0.67	10.32 ± 0.07	81.83 ± 1.98	9.91 ± 0.39	72.19 ± 3.23	9.13 ± 2.98	98.29 ± 0.19	14.27 ± 0.69								
A3	84.74 ± 1.02	21.54 ± 1.54	78.27 ± 1.08	19.35 ± 1.16	56.27 ± 4.03	10.15 ± 2.12	93.22 ± 1.01	21.96 ± 1.10								
A3 + SSL	88.59 ± 0.84	9.02 ± 0.20	88.44 ± 1.74	13.24 ± 0.32	72.32 ± 0.61	8.02 ± 1.23	92.37 ± 1.78	9.79 ± 0.52								
A3 + DC3	88.57 ± 0.62	7.76 ± 0.27	91.94 ± 1.04	14.05 ± 0.51	72.77 ± 2.74	9.56 ± 1.71	94.81 ± 0.96	9.50 ± 0.74								

3.5 Proof-of-concept improved data quality

We show above that DC3 can lead to better classifications and clusters than SSL alone. In accordance with previous literature [45, 43], we give a proof-of-concept in Table 3 that the annotation process can be improved with cluster-based proposals. As an SSL algorithm we used Mean-Teacher and for the datasets Plankton, Turkey and CIFAR-10H we used a random subsample of 10% for the evaluation. We conducted experiments with a pool of 6 annotators which consisted of domain experts and inexperienced hired workers which were paid a fixed wage per hour. We assigned 3 annotators from the pool per dataset. This means that annotator named e.g. A1 might be a different person between datasets in Table 3. We compare the annotations over time from each annotator. We investigated three different proposals for the annotation. The baseline is not using any proposals, the second is using the SSL predictions (classification) and the third is using the DC3 predictions (classification + clusters). For each cluster, a rough description was given as guidance during the annotation. After a training

A. Own previous papers

12 L. Schmarje et al.

Table 4: Ablation results averaged over different methods – The vanilla algorithms / baselines are highlighted in light grey. Each lower row extends this baseline individually with CE^{-1} [43], Clustering & Classification (CC) or both (DC3). CC can be interpreted as DC3 without CE^{-1} . The prior ambiguity estimate p_A is given in brackets if applicable. Results that improve over the baseline are marked in bold. The metrics are defined in subsection 3.2. The column ‘Ambiguous’ gives the percentage of predicted ambiguous data and the last column gives the number of non-degenerated runs over which we averaged

	F		d		(d-F)		Ambiguous		# Runs				
	best	mean \pm std	best	mean \pm std	best	mean \pm std	best	mean \pm std					
CIFAR-10H													
Baseline	0.7809	0.7153 \pm 0.0359	0.4199	0.5027 \pm 0.0469	-0.3611	-0.2126 \pm 0.0827	-	-	15				
+ CE ⁻¹	0.7383	0.7191	0.0164	0.4929	0.0243	-0.2091	-0.2262	0.4044	-	12			
+ CC ($p_A = 0.6$)	0.8565	0.7471	0.1246	0.8657	0.8768	0.0129	0.0092	0.1297	0.1374	0.6145	0.5923	\pm 0.0322	12
+ DC3 ($p_A = 0.32$)	0.6656	0.6970	0.0469	0.2155	0.3684	0.1227	-0.4501	-0.3286	0.0836	0.2910	0.3115	\pm 0.0140	12
+ DC3 ($p_A = 0.6$)	0.8305	0.7457	0.1097	0.4340	0.4741	0.0584	-0.3965	-0.2716	0.0928	0.6125	0.5860	\pm 0.0290	15
Plankton													
Baseline	0.8872	0.8652	\pm 0.0212	0.2584	0.2915	\pm 0.0240	-0.6287	-0.5737	\pm 0.0444	-	-	-	15
+ CE ⁻¹	0.8896	0.8803	0.0060	0.2540	0.2690	0.0098	-0.6356	-0.6113	\pm 0.0154	-	-	-	12
+ CC ($p_A = 0.6$)	0.8919	0.9128	\pm 0.0427	0.4085	0.7702	\pm 0.1630	-0.4833	-0.1426	\pm 0.1375	0.6242	0.5927	\pm 0.0127	12
+ DC3 ($p_A = 0.44$)	0.8625	0.9049	\pm 0.0340	0.2192	0.3269	\pm 0.0526	-0.6433	-0.5780	\pm 0.0305	0.4365	0.4451	\pm 0.0204	11
+ DC3 ($p_A = 0.6$)	0.9130	0.8768	\pm 0.0640	0.2484	0.3004	\pm 0.0750	-0.6646	-0.5764	\pm 0.0416	0.6164	0.5893	\pm 0.0202	14
Turkey													
Baseline	0.8211	0.8213	\pm 0.00469	0.3946	0.4428	\pm 0.0209	-0.4265	-0.3786	\pm 0.0230	-	-	-	15
+ CE ⁻¹	0.7998	0.7998	\pm 0.0000	0.3338	0.3338	\pm 0.0000	-0.4660	-0.4660	\pm 0.0000	-	-	-	12
+ CC ($p_A = 0.6$)	0.8527	0.8264	\pm 0.0469	0.3400	0.3435	\pm 0.0408	-0.5127	-0.4829	\pm 0.0128	0.5837	0.5646	\pm 0.0427	12
+ DC3 ($p_A = 0.22$)	0.7998	0.7998	\pm 0.0000	0.1675	0.2252	\pm 0.0646	-0.6322	-0.5746	\pm 0.0646	0.5090	0.3674	\pm 0.2054	4
+ DC3 ($p_A = 0.6$)	0.8743	0.8432	\pm 0.0350	0.2333	0.3270	\pm 0.0692	-0.6410	-0.5162	\pm 0.0643	0.8093	0.6387	\pm 0.2354	12

phase for the inexperienced annotators, we averaged across three repetitions for every annotator, proposal and dataset combination.

We see a general trend that the consistency improves and the annotation time decreases when proposals are used instead of None. Using DC3 proposals instead of SSL proposals, either leads to a similar or better consistency while the annotation time is often increased by one or two minutes. For this improvement, we credit the cleaner and more fine-grained outputs of the network. The additional verifications of the clusters could lead to the slightly increased annotation time. The individual benefits vary between the datasets and annotators. For example, the gains on the curated CIFAR-10H dataset are lower than on the uncurated Mice Bone dataset. On average across all annotators and datasets, we achieve an improved consistency of 6.74%, a relative speed-up of 2.4 and a maximum speed-up of 4.5 with DC3 proposals in comparison to the baseline.

4 Discussion

Ablation Study We pooled the runs between all methods to evaluate the impact of the individual components of our method DC3 and show the results in Table 4. The method FixMatch and the Mice Bone dataset are excluded from this ablation due to the up to 12 times higher required GPU hours and degenerated runs as before. Across the datasets, we see the best $(d-F1)$ -scores are achieved

by DC3. The impact of the components varies between the datasets. We see that CE^{-1} positively impacts the clustering results which confirms the benefit of using CE^{-1} for overclustering [43]. CC often reaches a better F1-Score than the baseline and even surpasses DC3 sometimes. However, the inner distance (d) may increase as well. We conclude that CC and CE^{-1} on their own can lead to improvements but only the combination of both parts results in a stable algorithm across datasets and methods. Additionally, we see that the number of not degenerated runs is highest with the combination of CE^{-1} and CC. If we use an realistic amount of ambiguity \hat{p}_A in each dataset as p_A , we see that in general the F1-Score decreases and d -score improves. We attribute this difference to the lower prior ambiguity p_A because DC3 tries to predict more certain than ambiguous images. This leads to a lower inner distance but also includes more difficult images in the classification of the certain data. We believe this parameter is essential for balancing the improvements in the F1- and d -score for a specified usecase. We chose a p_A of 0.6 because we wanted to weight certain and ambiguous images almost equally but ensure very certain /fewer classifications.

Qualitative Analysis with t-SNE We investigated some t-SNE [34] visualizations in Figure 4. Comparing the predicted (DC3) classes and ambiguity with the ground truth (GT), we see more wrong classifications on ambiguous images. DC3 outputs higher ambiguity than expected due to the higher value of p_A , but the predicted ambiguous clusters are often located nearby of ambiguous regions in the GT. Additionally, the clusters in (c) partition the feature space in smaller regions which can be more easily verified by humans as shown in subsection 3.5. Overall, we see a better representation of the ambiguous feature space.

Limitations We showed that DC3 generalizes to different SSL algorithms and datasets without hyperparameter changes. However, the datasets only consist of up to several thousand images. Due to the required multiple annotations per image for the evaluation it is difficult to obtain datasets with millions of images. We focused on improving the classification and clustering and gave a proof-of-concept for the increased consistency of relabeled data. Due to the required human labor during the relabeling step, we could not investigate the consistency across more datasets and algorithms or investigate the usage of the improved data. We proposed to improve the annotation process based on human-validated network predictions. This could introduce a not-desired bias into the data. This might lead to a negative impact for humans or a group of humans for certain usecases but we believe a small bias can be accepted in most applications because it is human controlled and systematically.

5 Conclusion

In real-world datasets, we often encounter ambiguous labels, due to intra- or interobserver variability, but also as intermediate classes might exist. We propose our method DC3 which is an extension to many SSL algorithms and allows to classify images with certain labels and cluster ambiguous ones. DC3

A. Own previous papers

14 L. Schmarje et al.

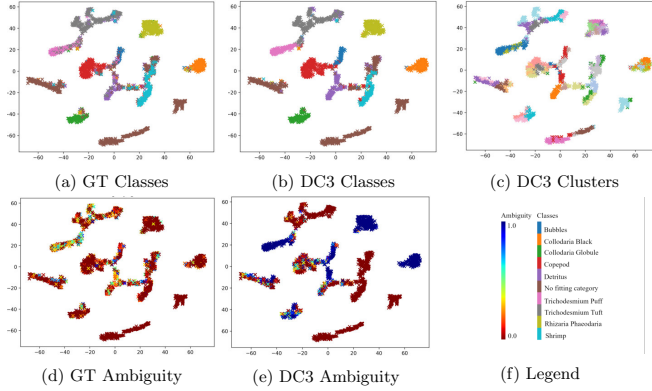


Fig. 4: t-SNE plots for Plankton dataset with Mean-Teacher – The same color was used 2–3 times for different clusters to ensure distinct colors.

also automatically determines which image should be treated as certain or ambiguous only based on a given prior probability p_A . On average, we achieve an increased F1-Score of 7.6% and a lower inner distance of clusters of 7.9% over all method-dataset-combinations. We give a proof-of-concept that these improved predictions can be used beneficially as proposals to create more consistent annotations. On average, we achieve an improved consistency of 6.74% and a relative speed-up of 2.4 when using DC3 proposals instead of no proposals. Therefore, SSL algorithms with DC3 are better suited to handle real-world datasets including ambiguous labeled images either by an improved classification / clustering or as a proposal during the annotation process with more insight.

Acknowledgements We acknowledge funding of LS by the ARTEMIS project (grant no. 01EC1908E) funded by the Federal Ministry of Education and Research (BMBF), Germany. SMS was funded by BMBF projects CUSCO (grant no. 03F0813D) and MOSAiC (grant no. 03F0917B). RKi was supported via a “Make Our Planet Great Again” grant of the French National Research Agency within the “Programme d’Investissements d’Avenir”; reference “ANR-19-MPGA-0012”. Funding for PlanktonID project were granted to RKi and RKo (CP1733) by the Cluster of Excellence 80 “Future Ocean” within the Excellence Initiative by the Deutsche Forschungsgemeinschaft on behalf of the German federal and state governments. Turkey data set was collected in the project “RedAlert – detection of pecking injuries in turkeys using neural networks” which was supported by the “Animal Welfare Innovation Award” of the “Initiative Tierwohl”.

References

1. Addison, P.F.E.E., Collins, D.J., Trebilco, R., Howe, S., Bax, N., Hedge, P., Jones, G., Miloslavich, P., Roelfsema, C., Sams, M., Stuart-Smith, R.D., Scanes, P., Von Baumgarten, P., McQuatters-Gollop, A.: A new wave of marine evidence-based management: Emerging challenges and solutions to transform monitoring, evaluating, and reporting. *ICES Journal of Marine Science* **75**(3), 941–952 (2018). <https://doi.org/10.1093/icesjms/fsx216>
2. Algan, G., Ulusoy, I.: Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey. *Knowledge-Based Systems* (2020). <https://doi.org/10.1016/j.knosys.2021.106771>
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: *Advances in Neural Information Processing Systems*. pp. 5050–5060 (2019)
4. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., van den Oord, A.: Are we done with ImageNet? arXiv preprint arXiv:2006.07159 (2020)
5. Brünger, J., Dippel, S., Koch, R., Veit, C.: ‘Tailception’: using neural networks for assessing tail lesions on pictures of pig carcasses. *Animal* **13**(5), 1030–1036 (2019). <https://doi.org/10.1017/S1751731118003038>
6. Cai, W., Chen, S., Zhang, D.: A simultaneous learning framework for clustering and classification. *Pattern Recognition* **42**(7), 1248–1259 (2009). <https://doi.org/10.1016/j.patcog.2008.11.029>
7. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 132–149 (2018)
8. Caron, M., Goyal, P., Misra, I., Bojanowski, P., Mairal, J., Joulin, A.: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (2020)
9. Cevikalp, H., Benligiray, B., Gerek, O.N.: Semi-supervised robust deep neural networks for multi-label image classification. *Pattern Recognition* **100**, 107164 (2020). <https://doi.org/https://doi.org/10.1016/j.patcog.2019.107164>
10. Chapelle, O., Scholkopf, B., Zien, A., Schölkopf, B., Zien, A.: Semi-supervised learning. *IEEE Transactions on Neural Networks* **20**(3), 542 (2006)
11. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big Self-Supervised Models are Strong Semi-Supervised Learners. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)* (2020)
12. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. pp. 215–223 (2011)
13. Crawford, K., Paglen, T.: Excavating AI: The Politics of Images in Machine Learning Training Sets. *AI and Society* pp. 1–12. <https://doi.org/10.1007/s00146-021-01162-8>
14. Culverhouse, P., Williams, R., Reguera, B., Herry, V., González-Gil, S.: Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series* **247**, 17–25 (2003). <https://doi.org/10.3354/meps247017>
15. Damm, T., Schmarje, L., Koser, N., Reinhold, S., Yilmaz, E., Krekieleh, N., Lui, L.Y., Cummings, S.R., Koch, R., Glueer, C.C.: Artificial intelligence-driven hip fracture prediction based on pelvic radiographs exceeds performance of DXA: the “Study of Osteoporotic Fractures” (SOF). *Journal of Bone and Mineral Research* **37**, 193–193 (2021)

A. Own previous papers

16 L. Schmarje et al.

16. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., Others, O'Donoghue, B., Visentin, D., Others: Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* **24**(9), 1342–1350 (2018)
17. Gao, B.B., Xing, C., Xie, C.W., Wu, J., Geng, X.: Deep Label Distribution Learning With Label Ambiguity. *IEEE Transactions on Image Processing* **26**(6), 2825–2838 (2017)
18. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Dorsch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised Learning. *Advances in Neural Information Processing Systems* 33 pre-proceedings (NeurIPS 2020) (2020)
19. Grossmann, V., Schmarje, L., Koch, R.: Beyond Hard Labels: Investigating data label distributions. *arXiv preprint arXiv:2207.06224* (2022)
20. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: *International Conference on Machine Learning*. pp. 1321–1330. PMLR (2017)
21. He, K., Gkioxari, G., Dollár, P., Girshick, R., R-cnn, M., Doll, P., Girshick, R.: Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2961–2969 (2017)
22. Jenckel, M., Parkala, S.S., Bukhari, S.S., Dengel, A.: Impact of Training LSTM-RNN with Fuzzy Ground Truth. In: *ICPRAM* (2018)
23. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 9865–9874. No. Iic (2019)
24. Jungo, A., Meier, R., Ermis, E., Blatti-Moreno, M., Herrmann, E., Wiest, R., Reyes, M.: On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: *Medical Image Computing and Computer Assisted Interventions, MICCAI*. pp. 682–690. Springer (2018)
25. Karimi, D., Nir, G., Fazli, L., Black, P.C., Goldenberg, L., Salcudean, S.E.: Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images—Role of Multiscale Decision Aggregation and Data Augmentation. *IEEE Journal of Biomedical and Health Informatics* **24**(5), 1413–1426 (2020). <https://doi.org/10.1109/JBHI.2019.2944643>
26. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Medical Image Analysis* **65** (2020)
27. Kim, B., Choo, J., Kwon, Y.D., Joe, S., Min, S., Gwon, Y.: SelfMatch: Combining Contrastive Self-Supervision and Consistency for Semi-Supervised Learning (NeurIPS) (2021)
28. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. pp. 1920–1929 (2019)
29. Krizhevsky, A., Hinton, G., Others: Learning multiple layers of features from tiny images. *Tech. rep.* (2009)
30. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. vol. 60, pp. 1097–1105. Association for Computing Machinery (2012). <https://doi.org/10.1145/3065386>
31. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: *International Conference on Learning Representations* (2017)

32. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 2 (2013)
33. Li, J., Socher, R., Hoi, S.C.H.: DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In: International Conference on Learning Representations. pp. 1–14 (2020)
34. der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *Journal of machine learning research* **9**(11) (2008)
35. Menon, A.K., Rawat, A.S., Yu, F., Jayasumana, S., Kumar, S., Reddi, S., Jitkrittum, W.: Disentangling sampling and labeling bias for learning in large-output spaces. In: International Conference on Machine Learning (2021)
36. Motamed, M., Sakharaykh, N., Kaldewey, T.: A Data-Centric Approach for Training Deep Neural Networks with Less Data. *NeurIPS 2021 Data-centric AI workshop* (2021)
37. Ooms, E.A., Zonderland, H.M., Eijkemans, M.J.C., Kriege, M., Mahdavian Delavary, B., Burger, C.W., Ansink, A.C.: Mammography: Interobserver variability in breast density assessment. *The Breast* **16**(6), 568–576 (2007). <https://doi.org/10.1016/j.breast.2007.04.007>
38. Peikari, M., Salama, S., Nofech-mozes, S., Martel, A.L.: A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification. *Scientific Reports* (April), 1–13 (2018). <https://doi.org/10.1038/s41598-018-24876-0>
39. Peterson, J., Battleday, R., Griffiths, T., Russakovsky, O.: Human uncertainty makes classification more robust. *Proceedings of the IEEE International Conference on Computer Vision* **2019-Octob**, 9616–9625 (2019). <https://doi.org/10.1109/ICCV.2019.00971>
40. Pham, H., Dai, Z., Xie, Q., Luong, M.T., Le, Q.V.: Meta Pseudo Labels (2020)
41. Qian, Q., Chen, S., Cai, W.: Simultaneous clustering and classification over cluster structure representation. *Pattern Recognition* **45**(6), 2227–2236 (2012). <https://doi.org/10.1016/j.patcog.2011.11.027>
42. Santarossa, M., Kilic, A., von der Burchard, C., Schmarje, L., Zelenka, C., Reinhold, S., Koch, R., Roeder, J.: MedRegNet: unsupervised multimodal retinal-image registration with GANs and ranking loss. In: *Medical Imaging 2022: Image Processing*. vol. 12032, pp. 321–333. SPIE (2022)
43. Schmarje, L., Brünge, J., Santarossa, M., Schröder, S.M., Kiko, R., Koch, R.: Fuzzy Overclustering: Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy. *Sensors* **21**(19), 6661 (2021). <https://doi.org/10.3390/s21196661>
44. Schmarje, L., Grossmann, V., Zelenka, C., Dippel, S., Kiko, R., Oszust, M., Pastell, M., Stracke, J., Valros, A., Volkmann, N., Koch, R.: Is one annotation enough? A data-centric image classification benchmark for noisy and ambiguous label estimation. *arXiv preprint arXiv:2207.06214* (2022)
45. Schmarje, L., Koch, R.: Life is not black and white - Combining Semi-Supervised Learning with fuzzy labels. *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen"*, (2021)
46. Schmarje, L., Liao, Y.H., Koch, R.: A Data-Centric Image Classification Benchmark. *NeurIPS 2021 Data-centric AI workshop* (2021)
47. Schmarje, L., Zelenka, C., Geisen, U., Glüer, C.C., Koch, R.: 2D and 3D Segmentation of uncertain local collagen fiber orientations in SHG microscopy. *DAGM German Conference of Pattern Recognition* **11824 LNCS**(November), 374–386 (2019). https://doi.org/10.1007/978-3-030-33676-9_26

A. Own previous papers

18 L. Schmarje et al.

48. Śmieja, M., Struski, L., Figueiredo, M.A.T.: A Classification-Based Approach to Semi-Supervised Clustering with Pairwise Constraints (2020)
49. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)* (2020)
50. Song, H., Kim, M., Park, D., Lee, J.G., Shin, Y., Lee, J.G.: Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* pp. 1–19 (2022). <https://doi.org/10.1109/TNNLS.2022.3152527>
51. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis* **63**, 101693 (2020). <https://doi.org/https://doi.org/10.1016/j.media.2020.101693>
52. Tarling, P., Cantor, M., Clapés, A., Escalera, S.: Deep learning with self-supervision and uncertainty regularization to count fish in underwater images pp. 1–22 (2021)
53. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: *ICLR* (2017)
54. Tian, Y., Henaff, O.J., van den Oord, A.: Divide and Contrast: Self-supervised Learning from Uncurated Data (2021)
55. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L.: Scan: Learning to classify images without labels. In: *Proceedings of the European Conference on Computer Vision*. pp. 268–285 (2020)
56. Volkmann, N., Brünger, J., Stracke, J., Zelenka, C., Koch, R., Kemper, N., Spindler, B.: Learn to train: Improving training data for a neural network to detect pecking injuries in turkeys. *Animals* 2021 **11**, 1–13 (2021). <https://doi.org/10.3390/ani11092655>
57. Volkmann, N., Zelenka, C., Devaraju, A.M., Brünger, J., Stracke, J., Spindler, B., Kemper, N., Koch, R.: Keypoint Detection for Injury Identification during Turkey Husbandry Using Neural Networks. *Sensors* **22**(14), 5188 (2022). <https://doi.org/10.3390/s22145188>
58. Wei, Y., Feng, J., Liang, X., Cheng, M.m.: Object Region Mining with Adversarial Erasing : A Simple Classification to Object Region Mining with Adversarial. *CVPR (March)*, 1568–1576 (2017)
59. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V., Luong, M.T., Le, Q.V., Hovy, E., Le, Q.V.: Self-Training With Noisy Student Improves ImageNet Classification. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10684–10695. *IEEE* (2020). <https://doi.org/10.1109/CVPR42600.2020.01070>
60. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-Labeling ImageNet: From Single to Multi-Labels, From Global to Localized Labels. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2340–2350 (2021)
61. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow Twins: Self-Supervised Learning via Redundancy Reduction (2021)

A.1.5 Is one annotation enough? - A data-centric image classification benchmark for noisy and ambiguous label estimation

Is one annotation enough?

A data-centric image classification benchmark for noisy and ambiguous label estimation

Lars Schmarje^{1*} Vasco Grossmann¹ Claudius Zelenka¹ Sabine Dippel² Rainer Kiko³
Mariusz Oszust⁴ Matti Pastell⁶ Jenny Stracke⁵ Anna Valros⁷ Nina Volkmann⁷
Reinhard Koch¹

¹MIP, Kiel University ²ITT, Friedrich-Loeffler-Institut ³LOV, Sorbonne Université

⁴Rzeszow University of Technology ⁵ITW, University Bonn ⁶University of Helsinki

⁷Luke, Natural Resources Institute Finland ⁸WING, University of Veterinary Medicine Hannover

Abstract

High-quality data is necessary for modern machine learning. However, the acquisition of such data is difficult due to noisy and ambiguous annotations of humans. The aggregation of such annotations to determine the label of an image leads to a lower data quality. We propose a data-centric image classification benchmark with ten real-world datasets and multiple annotations per image to allow researchers to investigate and quantify the impact of such data quality issues. With the benchmark we can study the impact of annotation costs and (semi-)supervised methods on the data quality for image classification by applying a novel methodology to a range of different algorithms and diverse datasets. Our benchmark uses a two-phase approach via a data label improvement method in the first phase and a fixed evaluation model in the second phase. Thereby, we give a measure for the relation between the input labeling effort and the performance of (semi-)supervised algorithms to enable a deeper insight into how labels should be created for effective model training. Across thousands of experiments, we show that one annotation is not enough and that the inclusion of multiple annotations allows for a better approximation of the real underlying class distribution. We identify that hard labels can not capture the ambiguity of the data and this might lead to the common issue of overconfident models. Based on the presented datasets, benchmarked methods, and analysis, we create multiple research opportunities for the future directed at the improvement of label noise estimation approaches, data annotation schemes, realistic (semi-)supervised learning, or more reliable image collection. ²

1 Introduction

High-quality data is the fuel of modern machine learning and almost all models improve with higher quality data [8, 80, 50]. Therefore, such data are a key component for developing future techniques. The acquisition of a large amount of data is considered particularly challenging due to the participation of humans in the process. Their mistakes or subjective interpretations of annotation tasks can lead to *noisy* or *ambiguous* labels, respectively [54, 16, 61, 28, 52, 9, 18, 29]. Consequently, the labels suffer from heteroscedastic aleatoric uncertainty which means that the data contains inherent noise, which is class- or even sample-dependent and negatively affects the quality [14].

In Figure 1, we present the impact of this uncertainty on the class "cat" in the CIFAR-10 dataset [32]. While all images have the same ground truth label in CIFAR-10, humans created agreeing annotations

*Corresponding Author, las@informatik.uni-kiel.de

²The source code is available at <https://github.com/Emprime/dcic>.

The datasets are available at <https://doi.org/10.5281/zenodo.7152309>.

A.1. Long papers

only with varying rates from four to 100 percent [54]. This means that individual annotations can be expected to be noisy as they diverge from the majority opinion. Furthermore, a majority vote across multiple annotations can not capture the ambiguity between different images. In some extreme cases (red borders), we even see a disagreeing majority vote across all annotators from the expected ground truth class. We raise the question if all images should be treated equally if human annotations show such varying certainties. Taking a *data-centric* perspective [47, 62, 45, 25], we investigate the data in contrast to only the model for answering this question. Specifically, we propose a data-centric image classification (DCIC) benchmark that indirectly measures a method’s ability to identify noisy and ambiguous labels and correct them. DCIC consists of ten real-world datasets of different domains (see Figure 2) and multiple human annotations for each image. The benchmark focuses on a data-centric view of the image classification problem by separating the data quality improvement and the classification performance into two tasks.

The main structure of the benchmark is divided into a *Labeling* and an *Evaluation* phase (see Figure 3a) which is comparable to established Teacher-Student-Approaches [70, 36]. Using this denotation, the benchmarked method will, as a teacher, improve labels during the first phase. These are then benchmarked in the second phase by analyzing their quality as training input to a student model. Be aware that we do not allow a knowledge transfer from the second phase to the first phase.

In detail, during the Labeling phase, we use samples from the distribution of the above-mentioned annotations to get different realistic label estimates as an *initialization*. The task of the benchmarked method is to improve these estimates for better performance of an other classification model in the second phase. In that phase (Evaluation), the obtained labels are used as input for training a fixed model and its performance is measured on a testing subset of the original data. In contrast to common model-centric deep learning approaches (see Figure 3b), we can vary the initialization for the same method and better separate its performance from the data improvement. The fixed model is used for the evaluation to facilitate distinguishing between performance gains due to improved input data and better learning of the method itself.



Figure 1: Are all images showing a cat? – Based on their ground truth labels from CIFAR-10 [32] they should all be cats. However, we give the agreement rate with the class cat from [54] in the lower right corner and see a wide range from four to 100%. Based on a majority vote, the last images (red border) would have not to be labeled as a cat but as dog, frog, dog, and deer, respectively. Based on these observations, we answer in our paper the question of whether all images should be treated equally as cats or if we should use multiple annotations and the resulting soft labels to capture this intrinsic noise and ambiguity.

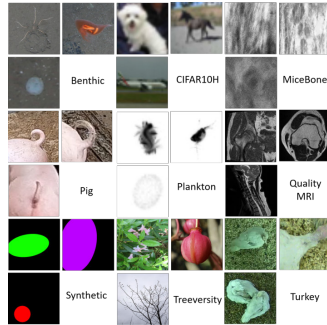


Figure 2: Three example images for all datasets. Details about the statistics of the dataset are given in Table 1.

Our benchmark is not only useful to evaluate existing methods, but will support research into algorithms for realistic datasets. Especially, it can bridge the research between semi-supervised learning and noise estimation based on realistic ambiguous noise patterns. We provide multiple algorithms as baselines and support the integration of more algorithms by common dataloaders for the two most popular deep learning frameworks: Tensorflow [1] and Pytorch [53]. We analyzed thousands of combinations of baseline methods, different initializations, and datasets. The obtained results confirm that the improvement of data quality leads to performance gains. Additionally, we investigated factors that influence the data quality and identified trends that lead to better learning of the underlying distribution.

A. Own previous papers

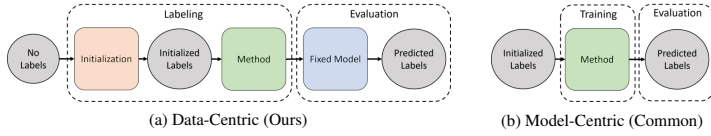


Figure 3: Comparison of our data-centric approach with the commonly used model-centric approach. The circles and arrows represent the available label information in addition to the corresponding images. The squares represent the methods which generate / change these label. There are two main differences between our and the common model-centric approach. Firstly, we also look at how the raw unlabeled data is initialized and thus how many annotations are required. Secondly, we use a fixed model to evaluate the output of the benchmarked method. These differences lead to a greater separation of data quality and method performance on the final scores on the predicted labels.

Our key contributions are: (1) We collected and created ten real-world image classification datasets with multiple annotations per image. These annotations allow a realistic simulation of noise patterns and will be helpful for future research in machine learning on real world data sets. (2) We provide a multi-domain benchmark based on these datasets for noisy and ambiguous label estimation. We implemented 20 methods as comparison. The benchmark also covers the topic of cost and bridges research between semi-supervised learning and ambiguous label estimation. (3) We show that one annotation per image is not enough because model performance improves as more labels are given for each input. We identify that the current focus on hard labels for classifications is ill-suited to learn the underlying ground truth distribution. A change in data preprocessing especially in annotation protocols could mitigate this and lead to less overconfident models.

1.1 Related Work

Human annotations of one image can differ due to complex reasons. Next to mere individual errors, cognitive sciences have shown that human judgement under uncertainty is driven by a subjective bias and the context of the annotation process [66]. As labeling relies on human perception, data quality problems, including issues with noisy and ambiguous labels, have been broadly discussed in the literature [68, 4, 54, 77, 3]. While voting strategies have proven to be robust tools to remove outlying annotation errors in a single label scenario, they also eliminate subjective disagreement, even in cases with more than just one valid interpretation [56]. The impact of this information loss has been discussed in numerous studies, indicating limitations in capturing ground truth by just one label [71, 17, 22, 6]. We empirically support these arguments across 9 datasets and 20 methods.

As label noise can severely degrade the classification performance [50], learning with flawed training data has become a substantial field of research in which numerous strategies have been proposed: while sample selection methods separate clean and noisy data by evaluating small-loss or disagreement [78, 79], correction methods aim at relabeling wrongly assigned labels by either learning class prototypes [24] or by pseudo-labeling strategies that utilize confident predictions [38]. Multiple methods have been proposed, but are often evaluated on synthetic noise [44, 40]. However, Wei et al. showed that synthetic noise is different from real noise by humans, which limits the generality of findings [77]. Gao et al. proposed synthetic annotators with individual labeling behavior instead of random noise to reduce uncertainty in predictions [20]. We go beyond this by using human annotations to reproduce realistic noise pattern and we do not only look at annotation errors but also at ambiguous annotations from subjective interpretations.

Datasets like CIFAR-10H [54] and CIFAR-10N [77] address the problem of realistic noise by providing multiple annotations per image for example of the original CIFAR-10 dataset. By doing so, both publications demonstrate an improved performance and a higher robustness, while also claiming that further research in dealing with human noise is still inevitable. The utilization of soft label distributions instead of hard one-hot label encodings enables the detailed representation of subjective disagreement and improves the generalization with ambiguous datasets [54, 5, 34]. In our benchmark, we extend this idea to eight diverse datasets apart from CIFAR-10 for a broader evaluation.

If we want to use multiple annotations per image, we need to consider the cost of such annotations to make it feasible in a project. Current research such as semi-supervised learning [12, 67, 64, 63] could

be used to analyze only a portion of the data with multiple annotations. However, approaches which combine noisy labels with semi-supervised learning have not been extended to real-world image classification tasks or do not consider the possibility that one labels is not enough to capture the ambiguity of subjectivity [42, 81, 76]. We provide with our benchmark the datasets and infrastructure to bridge the research between semi-supervised learning and noise estimation.

Prediction uncertainty can be attributed to unexplainable noise in the given training or test dataset (aleatoric uncertainty) or a wrong model inference (epistemic uncertainty) and it can be difficult to approximate and differentiate between them [58, 2, 72, 30]. Several real-world noisy datasets have been utilized as a foundation for classification benchmarks [41, 39, 54], Song et al. provide a current survey on datasets and methods [68]. Moreover, most robust methods are evaluated based on the test set accuracy [40, 68, 77, 59, 82]. However, even a small change in the structure or parameters of a method can directly impact its performance, limiting the comparability [31]. Other fields, such as Bayesian Neural Networks, address this issue by comparing results to statistical simulations, for example [27]. A recent benchmark [49] tries to overcome this issue by providing a baseline for noisy labels as a form of uncertainty estimation [14]. However, this benchmark relies on synthetic noise or noisy datasets without knowledge about the underlying ground truth distributions [41]. We use a data-centric approach to minimize the impact of implementation detail differences and measure the impact of the data indirectly during the Labeling phase by evaluating on a fixed model.

2 Benchmark

Our benchmark is divided into two major phases: *Labeling* and *Evaluation*. In alignment with the Data-Centric Idea [47], we separate the improvement of the data (Labeling) from the improvement of the models (Evaluation). The benchmark can be utilized to analyze a variety of research questions, but we focus on evaluating methods that estimate noisy or ambiguous labels.

We use the terms *noisy* and *ambiguous* throughout this work synonymously because we often do not differentiate between their cause during the annotation process. As mentioned above, these causes are errors or mistakes of human annotators which can be recovered for noisy label or noise. Subjective interpretations, imprecise task descriptions or poor image quality lead to ambiguous labels.

In general, we have an image dataset X with k known classes and use human annotations to approximate the image labels. Each image $x \in X$ has an often unknown soft ground truth label $\hat{l}_x \in [0, 1]^k$. Therefore, we use N hard human annotations $a_i \in \{0, 1\}^k$ with $i \in 1, \dots, N$ as estimates of \hat{l}_x . We assume that an average of annotations ($l_x = \sum_{i=0}^N \frac{a_i}{N}$) is an approximation of this target label \hat{l}_x as in [64]. Based on this definition, an annotation a_i or hard label sampled from the distribution l_x are in general of lower quality because they can not capture the aleatoric uncertainty of the soft label \hat{l}_x . We split the data equally and randomly in five *folds* and ensure a similar class distribution between the folds as best as possible. For one run, we use three folds as training (X_T) and one fold each as validation (X_V) and test (X_E) data, respectively. We call such an assignment of folds to the training, validation and test data *slice*.

Labeling The Labeling phase consists of two steps. In the first step an initialization is used to get label estimates and in the second step, the benchmarked method Θ aims to improve these labels. As initialization, we acquire $m \in \mathbb{N}$ annotations for $n \in [0, 100]$ percent of the training and validation images.

We call the total number of required annotations *budget* $b = m \cdot n$ and report it as proportions of training and validation images ($|X_T \cup X_V|$). In general, a classification task gets easier with more annotations or a higher budget. Be aware that the same initialization results in the the same budget while the same budget can achieved by different initializations.

The used initialization schemes per method are defined later in this section. We chose fixed initialization schemes for better comparability between the methods. How these labels are improved by the method Θ is not restricted. However, annotations aside from the given initialization are not allowed to be used. Since we measure the quality by training a different fixed network in the next phase, a good label would be presumably as close as possible to \hat{l}_x .

A. Own previous papers

Table 1: Overview of the used datasets – # is an abbreviation for number. The class imbalance is given as the percentage of the smallest and largest class with regard to the complete dataset. The agreement is the percentage of annotations that agree with the majority vote. The scores ACC and \hat{ACC} are given for the supervised baseline across three test folds. The access describes if the (raw) data is available openly, requires permission (restricted) or was not previously available (N/A). In the last column, datasets with modifications to the original data are marked with X. A modification might be adding more annotations or crop images to a region of interest.

Name	# classes	Input size [px]	# Images	Class Imbalance [%]		Agreement [%]	# Annotations	ACC [%]		\hat{ACC} [%]	Access	Updated
				Smallest	Largest			Mean \pm STD	Mean \pm STD			
Benthic	10	112×112	4867	2.31	39.66	82.61 \pm 19.67	4.54 \pm 2.01	64.17 \pm 0.63	83.36 \pm 0.47	Restricted	X	
CIFAR-10H	10	32×32	10000	9.88	10.16	95.44 \pm 8.91	51.10 \pm 1.54	90.75 \pm 0.39	95.72 \pm 0.12	Open	X	
Mice Bone	3	224×224	7240	14.75	70.48	85.06 \pm 17.52	15.30 \pm 21.90	61.88 \pm 9.44	78.39 \pm 1.95	Restricted	X	
Pig	4	96×96	10237	7.82	41.23	65.32 \pm 19.50	7.26 \pm 2.29	35.97 \pm 3.61	64.77 \pm 0.79	N/A	X	
Plankton	10	96×96	12280	4.16	30.37	93.26 \pm 13.60	24.38 \pm 44.17	89.89 \pm 0.82	92.41 \pm 0.41	Restricted	X	
Quality MRI	2	224×224	310	34.84	64.16	71.56 \pm 12.27	99.94 \pm 13.44	66.62 \pm 3.55	75.81 \pm 0.17	Restricted	X	
Synthetic	6	224×224	15000	16.17	17.57	74.41 \pm 24.28	98.86 \pm 0.99	87.85 \pm 0.48	74.65 \pm 0.34	N/A	X	
Treesvity#1	6	224×224	9489	9.98	30.67	88.60 \pm 16.13	14.78 \pm 7.06	79.50 \pm 1.53	89.20 \pm 0.31	Open	X	
Treesvity#6	6	224×224	9826	8.77	31.26	66.53 \pm 19.48	35.45 \pm 11.47	56.71 \pm 4.89	68.88 \pm 0.72	Open	X	
Turkey	3	192×192	8040	10.88	75.95	91.56 \pm 13.82	14.85 \pm 20.95	75.51 \pm 2.80	86.89 \pm 1.03	Restricted	X	

Evaluation In the Evaluation phase, the model and its hyperparameters are fixed to measure only the impact of the provided labels ($\Theta(x)$). The training of this fixed model Φ is calculated on the provided $\Theta(x)$ with $x \in X_T$. The best network parameters during training are selected based on a minimal divergence between $\Phi(x)$ and $\Theta(x)$ with $x \in X_V$. The generalization is then tested by measuring the difference between $\Phi(x)$ and l_x for $x \in X_E$.

Metrics Kullback-Leibler divergence (KL) [35] between $\Phi(x)$ and l_x for $x \in X_E$ has been used as our main metric since it is an established method to measures the difference between two distributions [48]. We averaged in a 3-fold cross-validation per dataset for a high reproducibility. We used the three slices defined by $X_{V_i} = \{f_{i+1}\}$, $X_{E_i} = \{f_{i+2}\}$ and the rest as training ($X_{T_i} = \{f_i, f_{((i+2)\%5)+1}, f_{((i+3)\%5)+1}, \dots\}$ with $\%$ for modulo) for the folds f_1, \dots, f_5 with $i \in 1, 2, 3$ as the index of the slices. While KL directly allows to measure the desired distribution divergence, we provide additional metrics as comparisons. We evaluate the accuracy (ACC) and F1-Score ($F1$) between $\Phi(x)$ and l_x for $x \in X_E$ per class and report the mean across the classes, which is commonly called the macro value and allows evaluation even in the presence of class imbalance. We used the most likely class based on the evaluated distributions for these metrics. We analyze the calibration of the models by reporting the Expected Calibration Error (ECE) [23]. As reference, we provide all of these metrics also on the difference between the proposed label before the second training $\Theta(x)$ and the expected ground-truth l_x for $x \in X_E$. The metrics are noted as \hat{ACC} , $F1$ and \hat{ECE} . We report the Cohen’s Kappa Score (κ) [46] as the measurement of the consistency of $\Theta(x)$ between the folds because more consistent labels result in higher model performance.

Datasets We include ten real-world classification datasets in our benchmark. Since we need multiple annotations per image for the evaluation of the quality of labels and this information is often not available in existing datasets or insufficient for our benchmark, we collected, adopted, or extended annotations of the following datasets. Their details are shortly described below, while their properties and exemplary images are shown in Table 1 and Figure 2, respectively. As presented, the datasets vary across all reported properties, giving an opportunity to comprehensively evaluate considered methods. More challenging datasets are characterized by a high-class imbalance, a low average agreement, or a low number of annotations per image. Detailed reports about the collection process and remaining dataset specifics are given in the supplementary.

1. *Benthic* depicts images from the seafloor and consists of underwater flora and fauna. We used annotations from [65, 37] but filtered for at least three annotations per object and cropped the main image to this object. We combined classes with too few images in agreement with domain experts. 2. *CIFAR-10H* is a variant of CIFAR-10 [32] introduced in [54]. Peterson et al. analyzed the underlying class distribution like us by reannotating the CIFAR-10 test set. In contrast to other variants like CIFAR-10N [77], this dataset provides more annotations per image. 3. *MiceBone* consists of Second-Harmonic-Generation images of collagen fibers in mice [60]. The raw images were preprocessed as described in [64]. Since there is a need for multiple annotations per image, we hired workers to increase their number by a factor of five. 4. *Pig* consists cropped tail images from European farms. The annotations were collected by hired workers with high domain knowledge. The goal is

A.1. Long papers

Table 2: Overview of used methods grouped into supervised, semi-supervised and self-supervised. The second to fifth column describe if the method uses unlabeled data, makes noise estimation, what pretraining is the used input of the initialized dataset are hard or soft labels, respectively. The initialization schemes columns describe which schemes were evaluated for individual methods. The average runtime of the labeling phase is given in the last column.

Name	Unlabeled Data	Noise Estimation	Pretraining	Labels	Initialization Schemes				Avg. Runtime [h]
					SL	SL+	SSL	SSL+	
Baseline				Soft	X	X	X	X	0.00
Heteroscedastic (Het) [15]		X		Hard	X	X	X	X	0.50
SNGP [43]		X		Hard	X	X	X	X	0.29
ELR+ [44]		X	ImageNet	Hard	X	X	X	X	0.09
Mean-Teacher (Mean) [70]	X			Hard	X		X		1.08
Mean-Teacher (Mean+DC3) [64]	X			Hard	X		X		1.20
π -Model (π) [36]	X			Hard	X		X		1.03
π -Model (π +DC3) [64]	X			Hard	X		X		1.15
FixMatch [67]	X			Hard	X		X		4.53
FixMatch+DC3 [64]	X			Hard	X		X		4.01
Pseudo-Label (Pseudo v1) [38]	X			Hard	X		X		1.10
Pseudo-Label (Pseudo v1 +DC3) [64]	X			Hard	X		X		1.40
Pseudo-Label (Pseudo v2 hard) [38]	X		ImageNet	Hard	X	X	X	X	0.16
Pseudo-Label (Pseudo v2 soft) [38]	X		ImageNet	Soft	X	X	X	X	0.12
Pseudo-Label (Pseudo v2 not) [38]	X		ImageNet	Soft	X	X	X	X	0.12
DivideMix [40]	X	X	ImageNet	Hard	X	X	X	X	1.39
BYOL [21]	X		Self-Supervised	Hard	X		X		2.59
MOCOv2 [13]	X		Self-Supervised	Hard	X		X		7.94
SimCLR [11]	X		Self-Supervised	Hard	X		X		5.89
SWAV [10]	X		Self-Supervised	Hard	X		X		4.17

the classification of the injury degree of the tail. 5. *Plankton* is a collection of underwater plankton images with multiple annotations from citizen scientists [61]. We use the preprocessing described in [64]. 6. *QualityMRI* consists of human magnetic resonance images (MRI) with a varying quality and multiple subjective quality ratings gathered in tests with radiologists. It was introduced and evaluated in [51, 69]. 7. *Synthetic* dataset was generated for the purpose of this study. It consists of images that contain one blue, red, or green circle or ellipse on a black background. To create ambiguous images, we added color and axis interpolations of these classes. 8+9. *TreeVersity* is a publicly available crowdsourced dataset of plant images from the Arnold Arboretum of Harvard University³. In the crowdsourcing project, the images were tagged with a given set of labels. We used a simplified version with six classes where we combined classes with too few images. Only images with at least three tags were used. Tags are not the same as class labels, therefore, we provide two subsets of TreeVersity. In TreeVersity#1, we filtered for exactly one given tag of the six possible ones per user which is similar to a classification. In TreeVersity#6, we filtered for a maximum of six different tags which means we did not apply any restrictions. 10. *Turkey* is a dataset with images of turkeys and their injuries [74, 75]. We used the preprocessing described in [64] and extended the original annotations, increasing their number by a factor of five with hired workers.

Methods We compare a variety of recent supervised, self-supervised, and semi-supervised algorithms against our baseline. The baseline does not adjust the initialized dataset in the first phase in any way and just forwards these labels to the supervised training of the second phase. Thus it is equivalent to supervised learning in a model-centric benchmark. We selected the other methods based on their recency, access to authors code or reimplementations and if they are state-of-the-art or commonly used as comparisons in the literature. The *supervised* methods are Heteroscedastic [15], SNGP [43], and ELR+ [44]. The *semi-supervised* methods are Mean-Teacher [70], π -Model [36], FixMatch [67], DC3 [64], Pseudo-Label [38], and DivideMix [40]. The *self-supervised* methods are BYOL [21], MOCOv2 [13], SimCLR [11], and SWAV [10]. Detailed descriptions about most of them are given in [63] and their key characteristics are presented in Table 2. We use the reported hyperparameters for Imagenet [33] or Webvision [41] by the original authors to ensure a comparison out-of-the-box across different image domains. For DC3[64], we investigated the combinations with Mean-Teacher, π -Model, FixMatch, and Pseudo-Label. For Pseudo-Label, we used two different implementations (v1 and v2) and variants with or without pretraining and soft or hard labels as input. In total, this results in 20 investigated methods. For better referencing, we group them as described above but put methods that *use soft labels* into their own group.

³<https://arboretum.harvard.edu/research/data-resources/>

A. Own previous papers

Initialization Schemes We investigated a fixed set of initialization schemes and note them m - n for m annotations for a subset of data with a relative size n . For easier reference, we group them as

- *Supervised Learning (SL)* 01-1.00
- *Supervised Learning+ (SL+)* 03-1.00, 05-1.00, 10-1.00
- *Semi-Supervised Learning (SSL)* 01-0.10, 01-0.20, 01-0.50
- *Semi-Supervised Learning (SSL+)* 10-0.10, 05-0.20, 02-0.50

Implementation Details The final results depend on a good set of fixed hyperparameters like learning rate for the model Φ for each dataset during the evaluation. Therefore, we determined them by applying Hyperopt [7] with 100 search trials across the same grid of parameters for all datasets. The target was the minimization of KL between $\Phi(x)$ and l_x for the baseline experiment with exactly ten annotations per image across one slice. We executed these and later experiments on an Nvidia RTX 3090 with 24GB VRAM or comparable hardware. Some combinations of models and input sizes could not fit on this hardware and therefore were ignored to keep the needed hardware to a minimum. Details about the used parameter grid are given in the supplementary. We ensure that all folds are randomly generated, while restrictions about similar images are considered. Without these restrictions, similar images, e.g., frames from the same camera might lead to an information leakage between the folds which would negatively influence the interpretability of the results.

3 Analysis

The evaluation was conducted across combinations of all datasets, methods, initialization schemes, and slices. A complete cross-combination would result in 5400 experiments from which we selected 3456 experiments to save resources since some combinations would not add more insights e.g. due to inferior performance of similar methods. A detailed overview of the initialization schemes used per method can be found in Table 2. As shown in Table 1, we have a large variability between the datasets, especially in ACC and \hat{ACC} that range from 36% to 96% for the baseline. Due to the fact that the baseline does not adjust the initialization, \hat{ACC} can be seen as the performance of humans in improving the labels for the given budget. The datasets Benthic, MiceBone, Pig, Treeversity#6, and Turkey have an over 10% lower ACC than the expected \hat{ACC} , which marks them as particularly challenging for the model. Moreover, the datasets Benthic, MiceBone, Pig, QualitMRI, and Treeversity#6 have an \hat{ACC} of lower than 85% which marks them as difficult even for humans. The \hat{ACC} of Synthetic is even lower due to the artificially created labels. Due to this variability, an average across the scores can be misleading. For this reason, we report the median in this paper and report the full results including the standard error of the mean (SEM) in the supplementary.

What metrics should we use? We analyzed the correlations between our metrics to determine which contain the same or similar information and which are complementary. All calculated Pearson correlation coefficients have a p-value < 0.01 and the four strongest correlations are ACC vs. $F1$ (0.99), KL vs. ECE (0.68) $F1$ vs. κ (0.77) and ACC vs. κ (0.77). The other correlations are around -0.5. Selected correlations are illustrated in Figure 4 and additional graphics and analysis are given in the supplementary. $F1$ balances the precision and recall but in our experiments we see almost identical values to ACC which we credit to the averaging per class. This means we only need to concentrate on ACC as a classification score. Overconfident models are a problem of modern machine learning [23] and the higher correlation between KL and ECE compared to any of the two with ACC , $F1$ or κ indicates that our focus on classification metrics like ACC and $F1$ could be the issue. Only metrics like KL and ECE consider the complete distribution and which justifies using mainly KL for the evaluation of this benchmark.

One annotation is not enough It is to be expected that more and better data should lead to increased performance which we quantify in Figure 5a. It can be seen that all metrics improve with an increased budget in the form of more annotated data or more annotations per image. However, the impact is lower for more annotations per image. For example, ACC increases from around 55% to 69% for the full supervision. Up to 10 annotations per image increase the score only to around 72%. This difference can be explained by the fact that additional annotations are most valuable to improve uncertain labels. In alignment with previous research [71, 22, 6], these results across thousands of

A.1. Long papers

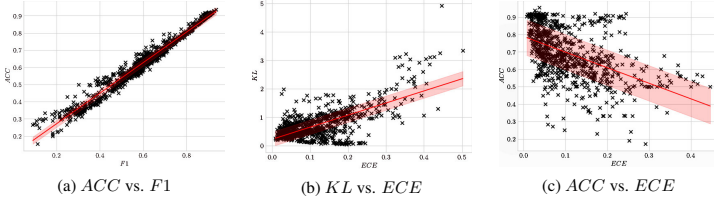


Figure 4: Correlations between selected metrics across all experiments. The red line represents the linear regression between the metrics and the light red area the mean absolute error of the regression.

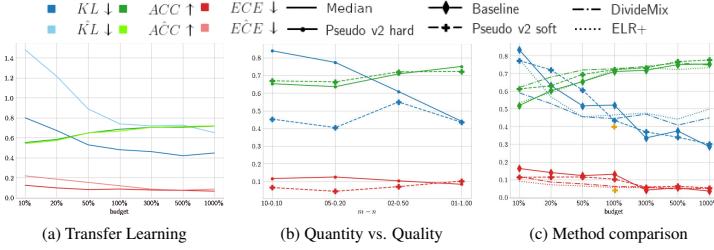


Figure 5: Analysis of all or selected methods across different budgets or initialization schemes. For details about the definitions see section 2. The orange crosses in (c) represent the best performance of Pseudo v2 soft with another initialization scheme see (b). The budget is increased by an raised portion of labeled data (n) until 1.00 and then increased further by using additionally multiple annotations per image (m).

experiments empirically justify that one annotation is not enough to capture the ground-truth of an item. Some improvements can be gained from correction annotation errors via a majority vote but the high disagreement and low ACC of the baseline on some datasets support the hypothesis that ambiguous annotations are a main source for the improvement. This ambiguity can not be described with a single hard majority vote and thus highlights the importance of using soft labels. Additionally, we see that KL is about half as small as \bar{KL} , ECE is around 3-10% lower than \bar{ECE} and ACC is 1-2% better than \bar{ACC} . As shown previously by Hinton et al. [26, 12] the knowledge distillation via a neural network into soft labels can be beneficial for ACC . We find the impact for metrics like KL and ECE even higher which supports our design of a two-phase benchmark.

Limits of current state-of-the-art To determine the best-performing state-of-the-art method, we gathered their relative improvements over the baseline in Table 3a. The best algorithm for each type of the soft, semi-supervised, supervised, self-supervised approaches based on the average performance across the budgets of 10%, 100%, and 1000% are Pseudo v2 soft, DivideMix, ELR+, and Mocov2, respectively. We visualize the best three of them in Figure 5c and give detailed results across the datasets for the budget 100% in Table 3b. The full results can be found in the supplementary. All top three methods are pretrained on ImageNet and outperform the rest in the field they were designed for. DivideMix is the best during partial supervision (budget < 100%), ELR+ is more noise robust (budget > 100%), and Pseudo v2 soft has the lowest KL score (budget > 100%). It is important to note that ImageNet pretraining leads to improvements on many datasets (see Pseudo v2 not in the supplementary) but also to worse results on others. It needs to be investigated if other pretraining such as CLIP [55] or unsupervised pretraining on larger datasets [12, 57] could improve on these results. Overall, the current state-of-the-art methods are insufficient for a label preprocessing across all domains. For a high budget any investigated method is worse than the supervised baseline without adjustments of the initialized dataset which shows the lack of appropriate algorithms for such budgets.

A. Own previous papers

Table 3: Results for the best performing methods – The best metric is marked bold while the 2nd and 3rd best are italic. Only methods with at least one top3 ranking across the budgets are presented. The full results are in the supplementary. (a) show the relative improvement over the baseline. (b) are detailed results for the budget of 100% across all datasets.

(a) Improvements							(b) Details 100% Budget										
Budget	10%		100%		1000%		Dataset	Benthic	CIFAR100	Microtime	Pig	Punkin	QualityMRI	Synthetic	Treeversity#1	Treeversity#6	Turkey
	Median	Mean \pm SEM	Median	Mean \pm SEM	Median	Mean \pm SEM											
ELR+	-0.16	-0.59 \pm 0.22	-0.43	-0.17 \pm 0.06	0.15	0.24 \pm 0.06	Baseline	1.17 \pm 0.04	0.41 \pm 0.02	0.25 \pm 0.06	0.75 \pm 0.09	0.54 \pm 0.02	1.73 \pm 0.48	0.00 \pm 0.01	0.40 \pm 0.04	1.02 \pm 0.03	0.40 \pm 0.03
SCNP	-0.26	-0.59 \pm 0.27	0.00	-0.15 \pm 0.09	0.20	0.26 \pm 0.04	ELR+	0.20 \pm 0.03	0.29 \pm 0.01	0.22 \pm 0.09	0.29 \pm 0.03	0.24 \pm 0.03	1.44 \pm 0.43	0.18 \pm 0.02	0.00 \pm 0.01	0.47 \pm 0.05	0.52 \pm 0.05
DisadvMts	-0.21	-0.78 \pm 0.25	-0.03	-0.17 \pm 0.08	0.16	0.21 \pm 0.07	SCNP	1.11 \pm 0.05	0.36 \pm 0.03	0.36 \pm 0.14	0.77 \pm 0.07	0.51 \pm 0.02	0.25 \pm 0.14	0.10 \pm 0.00	0.00 \pm 0.00	1.07 \pm 0.05	0.42 \pm 0.04
ϵ	-0.14	-0.59 \pm 0.22	-0.12	-0.22 \pm 0.09	N/A	0.21 \pm 0.07	DisadvMts	0.87 \pm 0.07	0.36 \pm 0.03	0.39 \pm 0.07	0.95 \pm 0.08	0.54 \pm 0.01	0.66 \pm 0.29	0.35 \pm 0.00	0.47 \pm 0.01	0.62 \pm 0.05	0.61 \pm 0.06
Pseudo v2 hard	-0.33	-0.63 \pm 0.19	-0.06	-0.17 \pm 0.05	N/A	0.28 \pm 0.04	ϵ	0.77 \pm 0.09	0.33 \pm 0.02	0.38 \pm 0.09	0.81 \pm 0.10	0.30 \pm 0.02	0.98 \pm 0.04	0.00 \pm 0.00	0.52 \pm 0.00	0.37 \pm 0.03	0.53 \pm 0.01
Pseudo v2 soft	-0.29	-0.60 \pm 0.19	-0.04	-0.17 \pm 0.05	0.01	0.00 \pm 0.02	Disadv	0.72 \pm 0.04	0.40 \pm 0.02	0.39 \pm 0.04	0.87 \pm 0.06	0.59 \pm 0.02	0.93 \pm 0.14	0.00 \pm 0.00	0.62 \pm 0.01	0.69 \pm 0.09	0.61 \pm 0.04
MOCNv2	-0.29	-0.63 \pm 0.22	-0.08	-0.13 \pm 0.10	N/A		Pseudo v2 hard	0.97 \pm 0.10	0.43 \pm 0.02	0.45 \pm 0.10	0.87 \pm 0.06	0.53 \pm 0.03	0.29 \pm 0.05	0.10 \pm 0.01	0.46 \pm 0.03	0.99 \pm 0.09	0.59 \pm 0.05
							Pseudo v2 soft	1.00 \pm 0.08	0.41 \pm 0.01	0.40 \pm 0.08	0.70 \pm 0.01	0.52 \pm 0.06	0.62 \pm 0.10	0.10 \pm 0.01	0.00 \pm 0.00	0.63 \pm 0.13	0.27 \pm 0.00
							MOCNv2	0.91 \pm 0.05	0.98 \pm 0.02	0.77 \pm 0.04	0.26 \pm 0.00	0.52 \pm 0.01	0.29 \pm 0.17	0.13 \pm 0.01	0.00 \pm 0.03	0.61 \pm 0.01	0.42 \pm 0.08

Moreover, an average better performance does not mean that the gains are equal across all datasets. For example, ELR+ has the lowest KL at a budget of 100% for five out of ten datasets but on the QualityMRI dataset, it is among the worst methods. This means while some methods might work on some datasets they might not generalize to other datasets. Overall, we see the highest KL for the datasets Benthic, Pig, QualityMRI and Treeversity#6 which also have the lowest agreement as seen in Table 1 except for the synthetic dataset. These datasets also show the largest variance in results across the methods. We conclude that the impact of the data is larger than the impact of the current preprocessing of state-of-the-art methods. This highlights the importance for investigating the data and label generation more if they are more impactful than the method itself.

While a higher budget leads to improved metrics, it also matters how it is used. In Figure 5b, we investigated the impact on KL and ACC for a budget of 100% for Pseudo-Label using hard or soft labels. We find that the accuracy is comparable between the methods and increases with a rising percentage of labeled data (m). For hard labels, the KL improves equally. If we use soft labels for training, we see lower results for 05-0.20 and 10-0.10. We conclude that we should investigate more how we distribute our budget if increasing it is not an option.

4 Discussion

Overall, we can confirm several previous research hypotheses while identifying missing information and thus new research opportunities with our novel datasets and benchmark.

We can demonstrate that data quality positively impacts the classifications scores like ACC and $F1$ and distribution-based scores like KL and ECE . Knowledge distillation can improve the approximation of the underlying distribution further. We agree with previous research [71, 17, 22, 6] that one annotation is not enough and we need to use soft labels to handle ambiguous data. KL and ECE are highly correlated (0.7) and are improved more when using soft labels. We believe that focusing on learning the real distribution and thus minimizing KL can lead to less overconfident models. Using soft labels as input seems to be crucial for achieving this since hard labels and classification metrics like ACC lead to models which slightly ignore the real ground truth distribution.

Nevertheless, most of the investigated state-of-the-art method do not use soft labels and often interpret noise only as errors in the annotation process. These issues need to be addressed in future research and a simple method like Pseudo v2 soft illustrates how the KL score can be lowered with this approach. For the largest budget, the baseline is the best model and even special noise estimation algorithms like ELR+ [44] and SNGP [43] can not achieve better results. We see a high variance across the datasets for different methods in our benchmark. However, we need methods which work across a variety of domains out-of-the-box to allow an easy application to current research question in other domains. Another practical issue is that we need to find solutions for acquiring soft labels even with a limited budget. In many research projects, it is difficult to annotate thousands of images with domain experts and annotating them multiple times would only increase the costs further. Thus, we need to bridge the research in semi-supervised learning and ambiguous and noise estimation. Such combinations could allow the usage of soft labels on a subset of images and simultaneously determine annotation errors. Our benchmark and datasets allowed the identification of these issues and thus could also be used to research new methods to solve these issues. We are confident that our benchmark and datasets can facilitate the bridged research on the topic of semi-supervised learning and ambiguous image estimation for real world image classification problems.

A.1. Long papers

Impact This work as a benchmark provides ten datasets and a detailed evaluation across 20 algorithms on this benchmark. The provided data can allow the investigation of research questions on the topics of e.g. noise estimation, data annotation scheme, or realistic semi-supervised learning. This work is intended to allow and spark future research and thus no direct social impacts are expected. However, this basic research is time and resource-consuming. For the final evaluation, we conducted experiments with about 5500 GPU hours which equals around 600kg CO_2 . For this reason, we limited the evaluation always to necessary elements when possible in order to not increase the needed GPU hours further.

Limitations The 20 investigated algorithms are only evaluated with one fixed set of hyperparameters across different datasets during the labeling phase. For optimal performance, a tuning per algorithm would have been required. We were interested in the general performance out-of-the-box and therefore neglected this issue due to resource minimization. The researched datasets are all below 15,000 images and the unsupervised learning potential on millions of images could not be investigated. We want to provide a detailed analysis in relation to the underlying distribution l_x per image which is only possible with multiple annotations per image. For larger datasets, this effort was just not feasible. Classification with hundreds or more classes are also not feasible because the annotation costs increase with the number of classes. As described above we conducted more than a thousand experiments but we had to select and combine several results in a comprehensive manner in this paper. These aggregations can not capture all details. Much more detailed analysis e.g., per dataset would be possible and thus we included all raw results in the supplementary. Due to the fixed initialization scheme, we can not investigate active learning approaches. However, this restriction is chosen to allow a better comparability and future researchers could decide against such a restriction.

Conclusion In alignment with previous research, we show that one annotation is not enough to handle ambiguous and noisy images and their underlying ground truth distribution. Multiple annotations and some kind of soft label are required to capture the difference in the images. Future research needs to investigate in more detail how annotations are being created, including annotation costs. We show that current state-of-the-art can help under certain budget or dataset constraints. However, methods with consistent results across a variety of datasets and budgets are missing. We release all datasets and the benchmark publicly to enrich future research on these topics.

Acknowledgments and Disclosure of Funding

We thank Mark Collier for his valuable feedback and discussion about the benchmark. We thank the annotators Kristina Ahlqvist, Daniel Grundig, Stine Heindorff, Jana Krambeck, Kathrin Körner, Richard Lange, Miina Tuominen-Brinkas and Emirhan Ustalar for their valuable work.

We acknowledge funding of L. Schmarje by the ARTEMIS project (Grant number 01EC1908E) funded by the Federal Ministry of Education and Research (BMBF, Germany). R. Kiko also acknowledges support via a “Make Our Planet Great Again” grant of the French National Research Agency within the “Programme d’Investissements d’Avenir”; reference “ANR-19-MPGA-0012”. V. Grossmann is employed with funds provided by Kiel Marine Science (KMS) and Future Ocean Network (FON) by Kiel University. Funds to conduct the PlanktonID project were granted to R. Kiko and R. Koch (CP1733) by the Cluster of Excellence 80 “Future Ocean” within the framework of the Excellence Initiative by the Deutsche Forschungsgemeinschaft (DFG) on behalf of the German federal and state governments. Turkey data set was collected as part of the project “RedAlert – detection of pecking injuries in turkeys using neural networks” which was supported by the “Animal Welfare Innovation Award” of the “Initiative Tierwohl”.

References

- [1] Martin Marten, Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudler, Josh Levensberg, Rajat Monga, Sherry Moore, Derek Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, and Others. TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

A. Own previous papers

- [2] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 15662535. doi: 10.1016/j.inffus.2021.05.008.
- [3] P. F.E. E E Addison, D. J. Collins, R. Trebilco, S. Howe, N. Bax, P. Hedge, G. Jones, P. Miloslavich, C. Roelfsema, M. Sams, R. D. Stuart-Smith, P. Scanes, P. Von Baumgarten, and A. McQuatters-Gollop. A new wave of marine evidence-based management: Emerging challenges and solutions to transform monitoring, evaluating, and reporting. *ICES Journal of Marine Science*, 75(3):941–952, 2018. ISSN 10959289. doi: 10.1093/icesjms/fsx216.
- [4] Görkem Algan and Ilkay Ulusoy. Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey. *Knowledge-Based Systems*, 2020. ISSN 23318422. doi: 10.1016/j.knosys.2021.106771.
- [5] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*, 2018.
- [6] Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. We Need to Consider Disagreement in Evaluation. In *BPPF*, 2021.
- [7] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.
- [8] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.
- [9] J Brünger, S Dippel, R Koch, and C Veit. ‘Tailception’: using neural networks for assessing tail lesions on pictures of pig carcasses. *Animal*, 13(5):1030–1036, 2019. ISSN 17517311. doi: 10.1017/S1751731118003038.
- [10] Mathilde Caron, Priya Goyal, Ishan Misra, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. ISSN 23318422.
- [11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2002.05709*, (PMLR): 1597–1607, 2020. ISSN 23318422.
- [12] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.
- [13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [14] Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. A Simple Probabilistic Method for Deep Classification under Input-Dependent Label Noise. *arXiv preprint arXiv:2003.06778*, 2020.
- [15] Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. Correlated Input-Dependent Label Noise in Large-Scale Image Classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (c):1551–1560, 2021.
- [16] Phil Culverhouse, Robert Williams, Beatriz Reguera, Vincent Herry, and Sonsoles González-Gil. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series*, 247:17–25, 2003. doi: 10.3354/meps247017.
- [17] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. 2021.

A.1. Long papers

- [18] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, Others, Brendan O'Donoghue, Daniel Visentin, and Others. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.
- [19] Laurens der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.
- [20] Zhengqi Gao, Fan-Keng Sun, Mingran Yang, Sucheng Ren, Zikai Xiong, Marc Engeler, Antonio Burazer, Linda Wildling, Luca Daniel, and Duane S. Boning. Learning from Multiple Annotator Noisy Labels via Sample-wise Label Fusion. 2022.
- [21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.
- [22] Vasco Grossmann, Lars Schmarje, and Reinhard Koch. Beyond Hard Labels: Investigating data label distributions. *ICML 2022 Workshop DataPerf: Benchmarking Data for Data-Centric AI*, 2022.
- [23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- [24] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5138–5147, 2019.
- [25] Tony Hey. The Fourth Paradigm – Data-Intensive Scientific Discovery. In *Communications in Computer and Information Science*, pages 1–1. 2012. doi: 10.1007/978-3-642-33299-9_1.
- [26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. 2015.
- [27] Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? *International Conference on Machine Learning*, pages 4629–4640, 2021.
- [28] Alain Jungo, Raphael Meier, Ekin Ermis, Marcela Blatti-Moreno, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *Medical Image Computing and Computer Assisted Interventions, MICCAI*, pages 682–690. Springer, 2018.
- [29] D Karimi, G Nir, L Fazli, P C Black, L Goldenberg, and S E Salcudean. Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images—Role of Multiscale Decision Aggregation and Data Augmentation. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1413–1426, 2020. doi: 10.1109/JBHI.2019.2944643.
- [30] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? 2017.
- [31] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019.
- [32] Alex Krizhevsky, Geoffrey Hinton, and Others. Learning multiple layers of features from tiny images. Technical report, 2009.
- [33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 60, pages 1097–1105. Association for Computing Machinery, 2012. doi: 10.1145/3065386.

A. Own previous papers

- [34] Ujwal Krothapalli and A Lynn Abbott. Adaptive label smoothing. *arXiv preprint arXiv:2009.06432*, 2020.
- [35] S Kullback and R A Leibler. On Information and Sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.
- [36] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.
- [37] Daniel Langenkämper, Robin van Kevelaer, Autun Purser, and Tim W Nattkemper. Gear-Induced Concept Drift in Marine Images and Its Effect on Deep Learning Classification. *Frontiers in Marine Science*, 7, 2020. ISSN 2296-7745. doi: 10.3389/fmars.2020.00506.
- [38] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 2, 2013.
- [39] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5456, 2018.
- [40] Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations*, pages 1–14, 2020.
- [41] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. WebVision Database: Visual Learning and Understanding from Web Data. 2017.
- [42] Zhenghua Li, Min Zhang, and Wenliang Chen. Ambiguity-aware Ensemble Training for Semi-supervised Dependency Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 457–467. Association for Computational Linguistics, 2014. doi: 10.3115/v1/P14-1043.
- [43] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.
- [44] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-Learning Regularization Prevents Memorization of Noisy Labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.
- [45] Adam Marcu, Antonia; Prugel-Bennett. On Data-centric Myths. *NeurIPS 2021 Data-centric AI workshop*, 2021.
- [46] Mary L McHugh. Interrater reliability: the kappa statistic. *PubMed, Biochemia*(3):276–82, 2012.
- [47] Mohammad Motamedi, Nikolay Sakharnykh, and Tim Kaldewey. A Data-Centric Approach for Training Deep Neural Networks with Less Data. *NeurIPS 2021 Data-centric AI workshop*, 2021.
- [48] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [49] Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W Dusenberry, Sebastian Farquhar, Angelos Filos, Marton Havasi, Rodolphe Jenatton, Ghassen Jerfel, and Others. Uncertainty Baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*, 2021.
- [50] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021.
- [51] Rafal Obuchowicz, Mariusz Oszust, and Adam Piorkowski. Interobserver variability in quality assessment of magnetic resonance images. *BMC Medical Imaging*, 20(1):109, 2020. ISSN 1471-2342. doi: 10.1186/s12880-020-00505-z.

A.1. Long papers

- [52] E.A. A Ooms, H.M. M Zonderland, M.J.C. J C Eijkemans, M. Kriege, B. Mahdavian Delavary, C.W. W Burger, and A.C. C Ansink. Mammography: Interobserver variability in breast density assessment. *The Breast*, 16(6):568–576, 2007. ISSN 09609776. doi: 10.1016/j.breast.2007.04.007.
- [53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [54] Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:9616–9625, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00971.
- [55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. 2021.
- [56] Sai Rajeswar, Pau Rodriguez, Soumye Singhal, David Vazquez, and Aaron Courville. Multi-label iterated learning for image classification with label ambiguity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4783–4793, 2022.
- [57] Pierre H. Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, and Michal Valko. BYOL works even without batch statistics. *arXiv preprint arXiv:2010.10241*, 2020.
- [58] Abhishek Singh Sambyal, Narayanan C, Krishnan, and Deepti R. Bathula. Towards Reducing Aleatoric Uncertainty for Medical Imaging Tasks. 2021.
- [59] Monty Santarossa, Ayse Kilic, Claus von der Burchard, Lars Schmarje, Claudius Zelenka, Stefan Reinhold, Reinhard Koch, and Johann Roeder. MedRegNet: unsupervised multimodal retinal-image registration with GANs and ranking loss. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 321–333. SPIE, 2022.
- [60] Lars Schmarje, Claudius Zelenka, Ulf Geisen, Claus-C. Glüer, and Reinhard Koch. 2D and 3D Segmentation of uncertain local collagen fiber orientations in SHG microscopy. *DAGM German Conference of Pattern Recognition*, 11824 LNCS(November):374–386, 2019. ISSN 23318422. doi: 10.1007/978-3-030-33676-9_26.
- [61] Lars Schmarje, Johannes Brünger, Monty Santarossa, Simon-Martin Schröder, Rainer Kiko, and Reinhard Koch. Fuzzy Overclustering: Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy. *Sensors*, 21(19):6661, 2021. ISSN 1424-8220. doi: 10.3390/s21196661.
- [62] Lars Schmarje, Yuan-Hong Liao, and Reinhard Koch. A Data-Centric Image Classification Benchmark. *NeurIPS 2021 Data-centric AI workshop*, 2021.
- [63] Lars Schmarje, Monty Santarossa, Simon-Martin Schroder, Reinhard Koch, Simon-Martin Schröder, Reinhard Koch, Simon-Martin Schroder, and Reinhard Koch. A Survey on Semi-, Self- and Unsupervised Learning for Image Classification. *IEEE Access*, pages 1–1, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3084358.
- [64] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, Claudius Zelenka, Rainer Kiko, Jenny Stracke, Nina Volkmann, and Reinhard Koch. A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.
- [65] T Schoening, A Purser, D Langenkämper, I Suck, J Taylor, D Cuvelier, L Lins, E Simon-Lledó, Y Marcon, D O B Jones, T Nattkemper, K Köser, M Zurowietz, J Greinert, and J Gomes-Pereira. Megafauna community assessment of polymetallic-nodule fields with cameras:

A. Own previous papers

- platform and methodology comparison. *Biogeosciences*, 17(12):3115–3133, 2020. doi: 10.5194/bg-17-3115-2020.
- [66] Philipp Schustek and Rubén Moreno-Bote. Instance-based generalization for human judgments about uncertainty. *PLOS Computational Biology*, 14(6):e1006205, jun 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006205. URL <https://dx.plos.org/10.1371/journal.pcbi.1006205>.
- [67] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.
- [68] Hwanjun Song, Minseok Kim, Dongmin Park, Jae-Gil Lee, Yooju Shin, and Jae-Gil Lee. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–19, 2022. ISSN 2162-237X. doi: 10.1109/TNNLS.2022.3152527.
- [69] Igor Stępień, Rafał Obuchowicz, Adam Piórkowski, and Mariusz Oszust. Fusion of Deep Convolutional Neural Networks for No-Reference Magnetic Resonance Image Quality Assessment. *Sensors*, 21(4), 2021. ISSN 1424-8220. doi: 10.3390/s21041043.
- [70] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *ICLR*, 2017.
- [71] Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72: 1385–1470, 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12752.
- [72] Matias Valdenegro-Toro and Daniel Saromo. A Deeper Look into Aleatoric and Epistemic Uncertainty Disentanglement. 2022.
- [73] Nina Volkmann, Johannes Brünger, Jenny Stracke, Claudius Zelenka, Reinhard Koch, Nicole Kemper, and Birgit Spindler. So much trouble in the herd: Detection of first signs of cannibalism in turkeys. In *Recent advances in animal welfare science VII Virtual UFAW Animal Welfare Conference*, page 82, 2020.
- [74] Nina Volkmann, Johannes Brünger, Jenny Stracke, Claudius Zelenka, Reinhard Koch, Nicole Kemper, and Birgit Spindler. Learn to train: Improving training data for a neural network to detect pecking injuries in turkeys. *Animals* 2021, 11:1–13, 2021. doi: 10.3390/ani11092655.
- [75] Nina Volkmann, Claudius Zelenka, Archana Malavalli Devaraju, Johannes Brünger, Jenny Stracke, Birgit Spindler, Nicole Kemper, and Reinhard Koch. Keypoint Detection for Injury Identification during Turkey Husbandry Using Neural Networks. *Sensors*, 22(14):5188, 2022. ISSN 1424-8220. doi: 10.3390/s22145188.
- [76] Zhuowei Wang, Jing Jiang, Bo Han, Lei Feng, Bo An, Gang Niu, and Guodong Long. SemiNLL: A Framework of Noisy-Label Learning by Semi-Supervised Learning. 2020.
- [77] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. 2021.
- [78] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Kwok. Searching to exploit memorization effect in learning from corrupted labels. *arXiv preprint arXiv:1911.02377*, 2019.
- [79] Xingrui Yu, Bo Han, Jiangechao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR, 2019.
- [80] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-Labeling ImageNet: From Single to Multi-Labels, From Global to Localized Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2340–2350, 2021.

A.1. Long papers

- [81] Xin Zhang, Zixuan Liu, Kaiwen Xiao, Tian Shen, Junzhou Huang, Wei Yang, Dimitris Samaras, and Xiao Han. CoDiM: Learning with Noisy Labels via Contrastive Semi-Supervised Learning. 2021.
- [82] Zizhao Zhang, Han Zhang, Sercan O. Arik, Honglak Lee, and Tomas Pfister. Distilling Effective Supervision from Severe Label Noise. *Conference on Computer Vision and Pattern Recognition*, 2020.

A. Own previous papers

Checklist

1. For all authors...
 - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [\[Yes\]](#) All claims are either introduced in Section 2 or are concluded in Section 4
 - (b) Did you describe the limitations of your work? [\[Yes\]](#) See Section 4
 - (c) Did you discuss any potential negative societal impacts of your work? [\[Yes\]](#) See Section 4
 - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [\[Yes\]](#) All raw data was acquired from open data or in agreement with ethical guidelines.
2. If you are including theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? [\[N/A\]](#) No theoretical claims made
 - (b) Did you include complete proofs of all theoretical results? [\[N/A\]](#) No theoretical claims made
3. If you ran experiments (e.g. for benchmarks)...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [\[Yes\]](#) See Supplementary or the main repository.
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [\[Yes\]](#) See Section 2
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [\[Yes\]](#) See Section 1, As explained in 3 most reported values are median values due to outliers, error bars would clutter the graphics and decrease their interpretability. We give all results (including SEM) in the supplementary for any interested reader.
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [\[Yes\]](#) See the implementation details in Section 2
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? [\[Yes\]](#) We credited all authors when we used or extended their datasets. See under datasets in Section 2
 - (b) Did you mention the license of the assets? [\[Yes\]](#) All license information are in the license file in the Git Repository
 - (c) Did you include any new assets either in the supplemental material or as a URL? [\[Yes\]](#) Multiple datasets are created or extended by us, also the source code for the benchmark is new. The source code and datasets are reachable as described in the access section in the supplementary or in the main repository.
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [\[Yes\]](#) We either used own datasets, public datasets or received personal permission from respective data owners.
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [\[N/A\]](#) The datasets are images of animals, objects or plants and thus can not identify people. Our data does not contain offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [\[N/A\]](#) We either used preliminary existing crowd sourcing materials or used only selected few workers as annotators and gave only oral instructions. We give credit and reference all previous instructions.
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [\[N/A\]](#) Not applicable, we see no potential risk for humans by annotating images.

A.1. Long papers

- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [\[Yes\]](#) We included hourly wages where applicable in the supplementary.

A. Own previous papers

A.1.6 Label Smarter, Not Harder: CleverLabel for Faster Annotation of Ambiguous Image Classification with Higher Quality

The original paper is not openly available. This is the camera ready version which I'm allowed to share in combination with the following statement: "This preprint has not undergone peer review (when applicable) or any post-submission improvements or corrections. The Version of Record of this contribution is published in Pattern Recognition, DAGM GCPR 2023, Lecture Notes in Computer Science, vol 14264, and is available online at https://doi.org/10.1007/978-3-031-54605-1_30"

Label Smarter, Not Harder: CleverLabel for Faster Annotation of Ambiguous Image Classification with Higher Quality

Lars Schmarje¹, Vasco Grossmann¹, Tim Michels¹, Jakob Nazarenius¹, Monty Santarossa¹, Claudius Zelenka¹, and Reinhard Koch¹

Kiel University {las,vgr,tmi,jna,msa,cze,rk}@informatik.uni-kiel.de

Abstract. High-quality data is crucial for the success of machine learning, but labeling large datasets is often a time-consuming and costly process. While semi-supervised learning can help mitigate the need for labeled data, label quality remains an open issue due to ambiguity and disagreement among annotators. Thus, we use proposal-guided annotations as one option which leads to more consistency between annotators. However, proposing a label increases the probability of the annotators deciding in favor of this specific label. This introduces a bias which we can simulate and remove. We propose a new method CleverLabel for Cost-effective LabEling using Validated proposal-guidEd annotations and Repaired LABELs. CleverLabel can reduce labeling costs by up to 30.0%, while achieving a relative improvement in Kullback-Leibler divergence of up to 29.8% compared to the previous state-of-the-art on a multi-domain real-world image classification benchmark. CleverLabel offers a novel solution to the challenge of efficiently labeling large datasets while also improving the label quality.

Keywords: Ambiguous · data-centric · data annotation

1 Introduction

Labeled data is the fuel of modern deep learning. However, the time-consuming manual labeling process is one of the main limitations of machine learning [53]. Therefore, current research efforts try to mitigate this issue by using unlabeled data [54,4,55] or forms of self-supervision [33,19,22,18]. Following the data-centric paradigm, another approach focuses on improving data quality rather than quantity [34,39,15]. This line of research concludes that one single annotation is not enough to capture ambiguous samples [8,10,3,47], where different annotators will provide different annotations for the same image. These cases are common in most real-world datasets [56,40,46,6] and would require multiple annotations per image to accurately estimate its label distribution. Yet, established benchmarks such as ImageNet or CIFAR [24,23] are currently not considering this issue which significantly limits their use in the development of methods that generalize well for ambiguous real-world data.

A. Own previous papers

2 L. Author et al.

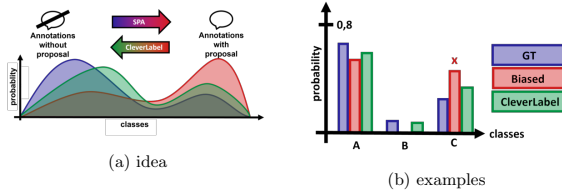


Fig. 1: Illustration of distribution shift – We are interested in the ground-truth label distribution (blue) which is costly to obtain due to multiple required annotations per image. Thus, we propose to use proposals as guidance during the annotation to approximate the distribution more cost efficiently (red). However, this distribution might be shifted toward the proposed class. We provide with CleverLabel (green) a method to improve the biased label distribution (red) to be closer to the original unbiased distribution (blue). Additionally, we provide with SPA an algorithm to simulate and analyze the distribution shift. The concrete effects are shown in the right example for the MiceBone dataset on a public benchmark [47] with the proposal marked by x.

Acquiring multiple annotations per sample introduces an additional labeling effort, necessitating a trade-off between label quality and quantity. While semi-supervised learning potentially reduces the amount of labeled data, the issue of label quality still arises for the remaining portion of labeled data [28]. One possible solution for handling ambiguous data is using proposal guided annotations [41,11] which have been shown to lead to faster and more consistent annotations [46,49]. However, this approach suffers from two potential issues: (1) Humans tend towards deciding in favor of the provided proposal [20]. This *default effect* introduces a bias, since the proposed class will be annotated more often than it would have been without the proposal. Thus, an average across multiple annotation results in a skewed distribution towards the proposed class as shown in Figure 1. (2) Real human annotations are required during development which prevents rapid prototyping of proposal systems.

We provide with CleverLabel and SPA two methods to overcome these two issues. Regarding issue (1), we propose Cost-effective **La**bEling using **V**alidated proposal-guidE**d** annotations and **R**epaired **La**bEls (CleverLabel) which uses a single class per image as proposal to speed-up the annotation process. As noted above, this might skew the label distribution towards the proposed class which can be corrected with CleverLabel. We evaluate the data quality improvement achieved by training a network on labels generated by CleverLabel by comparing the network’s predicted label probability distribution to the ground truth label distribution, which is calculated by averaging labels across multiple annotations as in [47]. Improved data quality is indicated by a reduction in the difference between the predicted distribution and the ground truth distribution. In addi-

tion, based on a previously published user study [48], we empirically investigate the influence of proposals on the annotator’s labeling decisions. Regarding issue (2), we propose Simulated Proposal Acceptance (SPA), a mathematical model that mimics the human behavior during proposal-based labeling. We evaluate CleverLabel and SPA with respect to their technical feasibility and their benefit when applied to simulated and real-world proposal acceptance data. Finally, we evaluate these methods on a real-world benchmark and we provide general guidelines on how to annotate ambiguous data based on the gained insights.

Overall, our contributions commit to three different areas of interest: (1) For improving label quality, we provide the novel method CleverLabel and show across multiple simulated and real world datasets a relative improvement of up to 29.8% with 30.0% reduced costs in comparison to the state of the art. (2) For annotating real-world ambiguous data, we provide annotation guidelines based on our analysis, in which cases to use proposals during the annotation. (3) For researching of countering the effect of proposals on human annotation behavior, we provide our simulation of proposal acceptance (SPA) as an analysis tool. SPA is motivated by theory and shows similar behavior to human annotators on real-world tasks. It is important to note that this research allowed us to achieve the previous contributions. We provide a theoretical justification for SPA and show that it behaves similarly to human annotators.

1.1 Related work

Data and especially high-quality labeled data is important for modern machine learning [62,38]. Hence, the labeling process is most important in uncertain cases or in ambiguous cases as defined by [47]. However, labeling is also not easy in these cases as demonstrated by the difficulties of melanoma skin cancer classification [36]. The issue of data ambiguity still remains even in large datasets like ImageNet [24] despite heavy cleaning efforts [5,59]. The reasons for this issue can arise for example from image artifacts like low resolution [43], inconsistent definitions [58], uncertainty of the data or annotators [1,45] or subjective interpretations [51,32].

It is important to look at data creation as part of the problem task because it can greatly impact the results. Recent works have shown that differences can depend on the aggregation of labels between annotators [61,8], the selection of image data sources on the web [37], if soft or hard labels are used as label representation [8,10,16,3] or the usage of label smoothing [30,31,35]. In this work we concentrate on the labeling step issues only. Simply applying SSL only partially solves the problem as it tends to overfit [2]. Hence labeling is necessary and the goal should be to label better and more.

A commonly used idea we want to focus on is proposal-based labeling. It is also known as verification-based labeling [41], label-spreading [11], semi-automatic labeling [29], or suggestion-based annotation [50]. [12] showed that proposal-based data labeling increases both accuracy and speed for their user study (n=54) which is in agreement with proof-of-concepts by [46,48]. The annotation suggestions for the segmentation and classification in diagnostic reasoning

A. Own previous papers

4 L. Author et al.

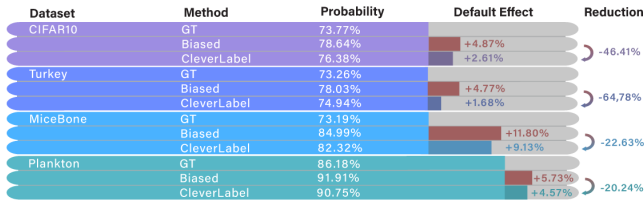


Fig. 2: Average annotation probability of a proposed class with the proposal unknown (GT, Unbiased) and known (Biased) to the annotators in four evaluated datasets. The proposal increases the probability in all observed cases, revealing a clear default effect in the investigated study. Its value is shown without any further processing (Biased) and with the contributed correction (CleverLabel) which consistently reduces the difference to the unbiased probabilities.

texts had positive effects on label speed and performance without an introduction of a noteworthy bias [50]. We continue this research for the problem of image classification and show that a bias is introduced and how it can be modeled and reversed.

Acceptance or rejection of a proposal was previously modeled e.g. for the review process of scientific publications [9]. They applied a Gaussian process model to simulate the impact of human bias on the acceptance of a paper, but rely on a per annotator knowledge. A simulation framework for instance-dependent noisy labels is presented in [17,14] by using a pseudo-labeling paradigm and [21] uses latent autoregressive time series model for label quality in crowd sourced labeling. Another aspect of labeling are annotation guidelines which can also have an impact on data quality as [52] demonstrate for app reviews. We do not consider guidelines as biases, instead they are a part of data semantics and use only real annotations per image. This has the benefit of avoiding unrealistic synthetic patterns as shown by [60] and simplifies the required knowledge which makes the process more easily applicable.

Note that active learning [44] is a very different approach, in which the model in the loop decides which data-point is annotated next and the model is incrementally retrained. It is outside the scope of this article and it might not be suited for a low number of samples with high ambiguity as indicated by [57]. Consensus processes [1,42] where a joined statement is reached manually or with technical support are also out of scope.

2 Methods

Previous research on proposal-based systems [41,29,50] suggests an influence of the default effect bias on the label distribution. While its impact is assessed as negligible in some cases, it circumvents the analysis of an unbiased annotation

Algorithm 1 Simulated Proposal Acceptance (SPA)

Require: Proposal ρ_x ; $a_i^x \in \{0\}^K$
 Calculate acceptance probability A
 $r \leftarrow \text{random}(0,1)$
if $r \leq A$ **then** \triangleright Accept proposal
 $a_{i,\rho_x}^x \leftarrow 1$
else \triangleright Sample from remaining classes
 $k \leftarrow \text{sampled from } P(L^x = k \mid \rho_x \neq k)$
 $a_{i,k}^x \leftarrow 1$
end if

distribution [20] which can be desirable, e.g. in medical diagnostics. As we can identify a significant bias in our own proposal-based annotation pipeline for several datasets (see Fig. 2), two questions arise: how to mitigate the observed default effect and how it was introduced?

In this section, we provide methods to answer both questions. Before we can mitigate the observed default effect, we have to understand how it was introduced. Thus, we introduce simulated proposal acceptance (SPA) with the goal of reproducing the human behavior for annotating images with the guidance of proposals. SPA can be used to simulate the labeling process and allow experimental analysis and algorithm development before conducting large scale human annotations with proposals. Building on this understanding, we propose CleverLabel which uses two approaches for improving the biased label distribution to mitigate the default effect: 1. a heuristic approach of class distribution blending (CB) 2. a theoretically motivated bias correction (BC). CleverLabel can be applied to biased distributions generated by humans or to simulated results of SPA.

For a problem with $K \in \mathbb{N}$ classes let L^x and L_b^x be random variables mapping an unbiased or biased annotation of an image x to the selected class k . Their probability distributions $P(L^x = k)$ and $P(L_b^x = k)$ describe the probability that image x is of class k according to a set of unbiased or biased annotations. As discussed in the literature [30,31,35,10], we do not restrict the distribution of L_x further e.g. to only hard labels and instead assume, that we can approximate it via the average of N annotations by $P(L^x = k) \approx \sum_{i=0}^{N-1} \frac{a_{i,k}^x}{N}$ with $a_{i,k}^x \in \{0,1\}$ the i -th annotation for the class k which is one if the class k was selected by the i -th annotator or zero, otherwise. The default effect can cause a bias, $P(L^x = k) \neq P(L_b^x = k)$ for at least one class k . Especially, for the proposed class ρ_x it can be expected that $P(L^x = \rho_x) < P(L_b^x = \rho_x)$.

2.1 Simulated Proposal Acceptance

Given both unbiased as well as biased annotations for the same datasets, we analyze the influence of proposals on an annotator's choice. We notice that a main characteristic is that the acceptance probability increases almost linearly with the ground truth probability of the proposal, $P(L^x = \rho_x)$, as shown in

A. Own previous papers

6 L. Author et al.

the main diagonal in Figure 3. If a proposal was rejected, the annotation was mainly influenced by the ground truth probability of the remaining classes. This observation leads to the following model: For a given proposal ρ_x , we calculate the probability A that it gets accepted by an annotator as

$$A = \delta + (\mathbf{1}^* - \delta)P(L^x = \rho_x) \quad (1)$$

with $\delta \in [0, 1]$. $\mathbf{1}^*$ is an upper-bound for the linear interpolation which should be close to one. The offset parameter δ can be explained due to the most likely higher probability for the proposed class. We also find that this parameter is dataset dependent because for example with a lower image quality the annotator is inclined to accept a more unlikely proposal. In subsection 2.3, we provide more details on how to calculate these values.

With this acceptance probability we can now generate simulated annotations $a'_{i,k} \in \{0, 1\}$ as in Algorithm 1 and describe the biased distribution similar to the unbiased distribution via $P(L_b^x = k) \approx \sum_{i=0}^{N'-1} \frac{a'_{i,k}}{N'}$ with N' describing the number of simulated annotations. The full source-code is in the supplementary and describes all corner cases e.g. $P(L^x \rho_x) = 1$. An experimental validation of this method can be found in subsection 3.1.

2.2 CleverLabel

Class distribution Blending (CB) A label of an image is in general sample dependent but [7] showed that certain classes are more likely to be confused than others. Thus, we propose to blend the estimated distribution $P(L_b^x = k)$ with a class dependent probability distribution $c(\hat{k}, k)$ to include this information. This class probability distribution describes how likely \hat{k} can be confused with any other given class k . These probabilities can either be given by domain experts or approximated on a small subset of the data as shown in subsection 2.3. The blending can be calculated as $\mu P(L_b^x = k) + (1 - \mu)c(\hat{k}, k)$ with the most likely class $\hat{k} = \operatorname{argmax}_{j \in \{1, \dots, K\}} P(L_b^x = j)$ and blending parameter $\mu \in [0, 1]$. This approach can be interpreted as a smoothing of the estimated distribution which is especially useful in cases with a small number of annotations.

Bias Correction (BC) In subsection 2.1, we proposed a model to use the knowledge of the unbiased distribution $P(L^x = k)$ to simulate the biased distribution $P(L_b^x = k)$ under the influence of the proposals ρ_x . In this section, we formulate the reverse direction for correcting the bias to a certain degree.

According to Equation 1, for $k = \rho_x$ we can approximate

$$B := P(L^x = \rho_x) = \frac{A - \delta}{\mathbf{1}^* - \delta} \approx \frac{\frac{|M_{\rho_x}|}{N'} - \delta}{\mathbf{1}^* - \delta},$$

with $M_{\rho_x} = \{i \mid i \in \mathbb{N}, i \leq N', a'_{i,\rho_x} = 1\}$ the indices of the annotations with an accepted proposal. Note that we have to clamp the results to the interval $[0, 1]$ to receive valid probabilities for numerical reasons.

Table 1: Used offsets (δ) for proposal acceptance

dataset	Benthic	CIFAR10H	MiceBone	Pig	Plankton
User Study	N/A	9.73%	36.36%	N/A	57.84%
Calculated	40.17%	0.00%	41.03%	25.72%	64.81%
dataset	QualityMRI	Synthetic	Treeversity#1	Treeversity#6	Turkey
User Study	N/A	N/A	N/A	N/A	21.64%
Calculated	0.00%	26.08%	26.08%	20.67%	14.17%

For $k \neq \rho_x$ we deduce the probability from the reject case of Algorithm 1

$$\begin{aligned}
P(L^x = k \mid L^x \neq \rho_x) &= P(L_b^x = k \mid L^x \neq \rho_x) \\
\Leftrightarrow \frac{P(L^x = k, L^x \neq \rho_x)}{P(L^x \neq \rho_x)} &= P(L_b^x = k \mid L^x \neq \rho_x) \\
&\Leftrightarrow P(L^x \neq \rho_x) = (1 - B)P(L_b^x = k \mid L^x \neq \rho_x) \\
&\approx (1 - B) \cdot \sum_{i \notin M_{\rho_x}} \frac{a'_{i,k}}{N' - |M_{\rho_x}|}.
\end{aligned}$$

This results in a approximate formula for the original ground truth distribution which relies only on the annotations with proposals. The joined distribution is deducted in the supplementary. It is important to note that the quality of these approximations relies on a large enough number of annotations N' .

2.3 Implementation details

We use a small user study which was proposed in [48] to develop / verify our proposal acceptance on different subsets. The original data consists of four dataset with multiple annotations per image. We focus on the no proposal and class label proposal annotation approaches but the results for e.g. specific DC3 cluster proposals are similar and can be found in the supplementary.

We calculated the ground-truth dataset dependent offset δ with a light weight approximation described in the supplementary. An overview about the calculated offsets is given in Table 1 in combination with the values of the user study where applicable. Due to the fact, that it can not be expected, that this parameter can be approximated in reality with a high precision we use for all experiment except otherwise stated, a balancing threshold $\mu = 0.75$, $\mathbf{1}^* = 0.99$ and $\delta = 0.1$. More details about the selection of these parameters are given in the supplementary.

The class distributions used for blending are approximated on 100 random images with 10 annotations sampled from the ground truth distribution. For a better comparability, we do not investigate different amounts of images and annotations for different datasets but we believe a lower cost solution is possible especially on smaller datasets such as QualityMRI. For this reason, we ignore this static cost in the following evaluations. If not otherwise stated, we use the

A. Own previous papers

8 L. Author et al.

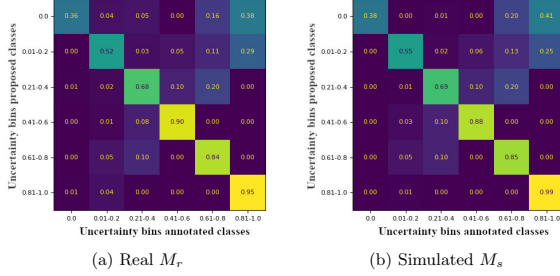


Fig. 3: Visual comparison the uncertainty bins for real vs. simulated proposal acceptance on the MiceBone dataset, Normalized per row / per proposal uncertainty bin, meaning that e.g. in Real M_r that if a class with soft ground truth probability 0.21-0.4 is proposed, in 0.10 of cases a class with ground truth probability .041-0.6 is annotated. Hence, some cells are 0 by default.

method DivideMix [27] and its implementation in [47] to generate the proposals. With other methods the results are very similar and thus are excluded because they do not add further insights. We include the original labels which are used to train the method in the outputted label distribution by blending it with the output in proportion to the used number of annotations. Please see the supplementary for more details about the reproducibility.

3 Evaluation

We show that SPA and our label improvements can be used to create / reverse a biased distribution, respectively. In three subsections, we show that both directions are technically feasible and are beneficial in practical applications. Each section initially gives a short motivation, describes the evaluation metrics and provides the actual results.

3.1 Simulated Proposal Acceptance

We need to verify that the our proposed method SPA is a good approximation of the reality in comparison to other methods and that an implementation is technically feasible.

Metrics & Comparisons We know for the evaluation of every image x the proposed class p_x , the annotated class a_x and the soft ground truth distribution $P(L^x = k)$ for class k in the study [48]. For the evaluation , we calculate a matrix between the proposed class probability and the actually annotated class

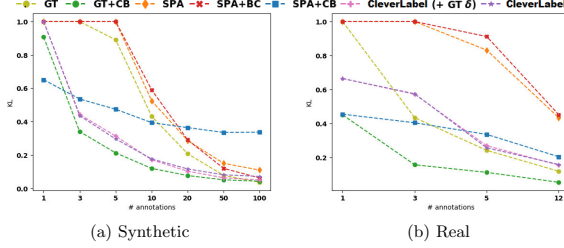


Fig. 4: Evaluation of label improvement on synthetic data created with SPA and real user study across different amounts of annotations. Results are clamped for visualization to the range 0 to 1.

probability by aggregating the probabilities into uncertainty bins with a size of 0.2 and a special bin for 0.0, which results in total in 6 bins. Normalized examples are in Figure 3.

Our proposed method is composed of two parts: *ACCEPT* with offset δ the proposal, otherwise use the *GT* distribution for selection as defined in subsection 2.1. Other possible components would be *RANDOMly* annotate a label, use the most *LIKELY* class label or any combination of the before. We compare our proposed method (ACCEPT+GT) against 6 other methods: ACCEPT+LIKELY, 2*ACCEPT+GT, 2*ACCEPT+RANDOM, RANDOM, GT, LIKELY. Two ACCEPTs mean that we use an offset acceptance of the proposal and then an offset acceptance of the most likely class if it was not the original proposal. As metric, we use half of the sum of differences (SOD) between the real matrix M_r and the simulated matrix M_s or as formula $SOD(M_r, M_s) = 0.5 * (\sum_{i,j} \text{abs}(M_{r_{i,j}} - M_{s_{i,j}}))$ which is the number of differently assigned images to all bins asides from duplicates. We report the normalized SOD by the total number of entries in M_r as the average with standard deviation across three repetitions and include more results e.g. proposals based on DC3 clusters in the supplementary. The δ of the User Study was used for the simulation. We developed our method and the comparisons on the datasets Turkey and Plankton and only verified on MiceBone and CIFAR10H.

Results A visual comparison of the real and simulated results for all uncertainty bins can be seen in Figure 3. The main diagonal line contains the accepted proposals while the rest, especially the upper right corner are the rejected images. We see that the presented matrices are very similar, even in overlapping regions between accepted and rejected proposals as for uncertainty bin 0.41-0.6 of the proposed and annotated classes.

In Table 2, we compare the proposed method with six other possible algorithms. We see that our proposed method is for all datasets one of the best

A. Own previous papers

10 L. Author et al.

Table 2: Comparison of investigated methods for the proposal acceptance, results within a one percent boundary of the best results are marked bold [†] datasets used only for validation,

<i>SOD</i> in [%]	CIFAR10H [†]	MiceBone [†]	Plankton	Turkey
ACCEPT+GT (Ours)	1.97 ± 0.06	2.57 ± 0.16	3.62 ± 0.05	4.77 ± 0.13
ACCEPT+LIKELY	1.20 ± 0.25	7.88 ± 0.06	5.50 ± 0.15	7.21 ± 0.32
2*ACCEPT+GT	2.79 ± 0.22	3.38 ± 0.26	3.25 ± 0.13	3.78 ± 0.28
2*ACCEPT+RANDOM	3.03 ± 0.11	4.76 ± 0.11	4.00 ± 0.31	4.64 ± 0.28
RANDOM	86.37 ± 0.13	50.92 ± 0.64	81.51 ± 0.18	54.83 ± 0.93
GT	5.00 ± 0.03	10.34 ± 0.19	10.59 ± 0.11	5.69 ± 0.30
LIKELY	4.08 ± 0.00	16.41 ± 0.00	11.85 ± 0.00	11.47 ± 0.00

methods. However, some methods e.g. 2*ACCEPT+GT can sometimes even be better. This data allows two main conclusions. SPA is clearly better than some naive approaches like RANDOM or GT. SPA is not optimal. It can neither reproduce the real results completely nor is the best method across all datasets. However, it shows very strong performance and is less complex than e.g. 2*ACCEPT + RANDOM. We conclude that SPA is at a sweet spot between simplicity and correctness.

3.2 Label Improvement

We show that CB and BC lead to similar increased results on simulated and real biased distributions while the similarity illustrates the practical benefit of SPA.

Metrics & Comparison As a metric, we use the Kullback-Leibler divergence [25] between the soft ground truth $P(L^x = k)$ and the estimated distribution. We generate the skewed distributions either by our method SPA or use real proposal acceptance data from [48]. The reported results are the median performance across different annotation offsets or datasets for the synthetic and real data, respectively. For the real data, we used the calculated δ defined in Table 1 for the simulation but as stated above $\delta = 0.1$ for the correction in CleverLabel. The method *GT* is the baseline and samples annotations directly from $P(L^x = k)$. The full results are in the supplementary.

Results If we look at the results on the synthetic data created by SPA in Figure 4a, we see the expected trend of improved results with more annotations. While using only CB is the best method with one annotation, the performance is surpassed by all other methods with enough annotations. The baseline (GT) is especially with the combination of blending (+CB) the best method for most number of annotations. Our label improvement (CleverLabel) is in most cases the second best method and blending is a major component (+CB). The bias correction (+BC) improves the results further for higher number of annotations at around 20+. Using the correct offset (+ δ GT) during the correction which was used in the simulation of SPA, is of lower importance. When we look at the

full results in the supplementary, we see benefits of a better δ at an offset larger than 0.4 and more annotations than 5. We conclude that label improvement is possible for synthetic and real data and that the combination of CB and BC with an offset of 0.1 is in most cases the strongest improvement.

The real results in Figure 4b show the similar trends as in the synthetic data. However, the baseline method without blending performs stronger and some trends are not observable because we only have up to 12 annotations. The correct value for the offsets is even less important in the real data, most likely because the effect is diminished by the difference of the simulation and reality. It is important to note that we keep the same notation with CleverLabel but in this case SPA was not used to generate the biased distribution but real annotations with proposals. Overall, the results analysis on synthetic and real data is similar and thus SPA can be used as a valid tool during method development.

It should be pointed out that the cost of labeling is not equivalent to the number annotations as we can expect a speedup of annotations when using proposals. For example, CleverLabel often performs slightly worse than GT in Figure 4b. Considering a speedup of 2, we actually have to compare CleverLabel with 5 annotations to GT at around 3, as explained in the budget calculation in subsection 3.3.

3.3 Benchmark evaluation

We show the results for CleverLabel on [47].

Metrics & Comparison We compare against the top three benchmarked methods: *Baseline*, *DivideMix* and *Pseudo v2 soft*. *Baseline* just samples from the ground-truth but still performed the best with a high number of annotations. *DivideMix* was proposed by [27] and *Pseudo v2 soft* (*Pseudo soft*) uses Pseudo-Labels [26] of soft labels to improve the labels. We evaluate the Kullback-Leibler divergence (KL) [25] between the ground truth and the output of the second stage (the evaluation with a fixed model) and KL between ground truth and the input of the second stage (\hat{KL}). We also provide an additional ablation where we replaced the fixed model in the second stage with a visual transformer [13]. The hyperparameters of the transformer were not tuned for each dataset but kept to common recommendations. The speedup S which can be expected due to using proposals depends on the dataset and used approach. For this reason, we include this parameter in our comparison with the values of 1 (no speedup), 2.5 as in [48] or 10 as in [49]. S is used to calculate the *budget* as *initial supervision per image* (*in. sup.*) + (*percentage annotated of* $X \cdot$ *number of annotations per image*)/ S . *In. sup.* describes the percentage of labeled data that is annotated once in phase one of the benchmark. For the skewed distribution generation which is correct by CleverLabel, we used SPA with the calculated δ in Table 1. For CleverLabel a heuristically chosen $\delta = 0.1$ was used if not otherwise stated ($+GT\delta$). The results are the median scores of all datasets of the averages across three folds. Full results including mean scores are in the supplementary.

A. Own previous papers

12 L. Author et al.

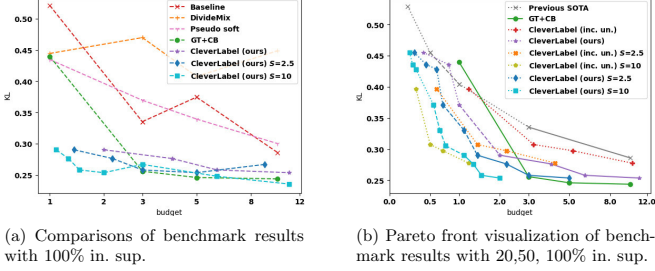


Fig. 5: Left: Compares previous state-of-the-art (first three), new baseline (GT+CB) and our method (CleverLabel) including different speedups S . Right: The marker and color define the method, while the line style to the next node on the x-axis visualizes the initial supervision, logarithmic scaled budgets

Results We present in Figure 5a a comparison of our method CleverLabel with previous state-of-the-art methods on the benchmark with an initial supervision of 100%. Even if we assume no speedup, we can achieve lower KL scores than all previous methods, regardless of the used number of annotations. Our proposed label improvement with class blending can also be applied to samples from the ground truth distribution (GT + CB) and achieves the best results in many cases. Due to the fact that it does not leverage proposals it can not benefit from any speedups S . If we take these speedups into consideration, CleverLabel can achieve the best results across all budgets except for outliers.

We investigate lower budgets where the initial supervision could be below 100% in Figure 5b. The full results can be found in the supplementary. If we compare our method to the combined Pareto front of all previous reported results, we see a clear improvement regardless of the expected speedup. Two additional major interesting findings can be taken from our results. Firstly, the *percentage of labeled data* which is equal to the *initial supervision* for CleverLabel (violet, blue, light blue) is important as we see improved results from *initial supervision* of 20 to 50 to 100%. This effect is mitigated with higher speedups because then CleverLabel can achieve lower budget results not possible by other initial supervisions. Secondly, we can improve the results further by using proposals also on the unlabeled data (inc. un., red, orange, yellow) after this initialization. This increases the budget because the *percentage of labeled data* is 100% regardless of the *initial supervision* but results in improved scores. With $S = 10$ we can even improve the previous state of the art (Pseudo soft, in. sup 20%, 5 annotations) at the budget of 1.0 from 0.40/0.47 to 0.30/0.33 at a budget of 0.7 which is a relative improvement of 25%/29.8% with median/ mean aggregation.

In Table 3, we conduct several ablations to investigate the impact of individual parts of our method. Comparing KL and \hat{KL} scores, we see similar trends

Table 3: Benchmark results ablation study, first block of rows KL result on normal benchmark, second block of rows \hat{KL} on benchmark, last block of rows KL results on benchmark but with ViT as backbone, all results are median aggregations across the datasets

method	1	3	5	10	20	50	100
CleverLabel (ours)	0.29 ± 0.02	0.28 ± 0.01	0.26 ± 0.02	0.25 ± 0.01	0.27 ± 0.02	0.25 ± 0.02	0.24 ± 0.01
CleverLabel (+ GT δ)	0.30 ± 0.02	0.28 ± 0.01	0.27 ± 0.01	0.25 ± 0.02	0.24 ± 0.01	0.24 ± 0.01	0.24 ± 0.01
Only CB	0.34 ± 0.03	0.28 ± 0.01	0.27 ± 0.01	0.25 ± 0.02	0.25 ± 0.01	0.25 ± 0.02	0.25 ± 0.02
Only CB ($\mu=0.25$)	0.33 ± 0.02	0.28 ± 0.01	0.33 ± 0.02	0.29 ± 0.01	-	-	-
Only BC	-	0.30 ± 0.02	0.29 ± 0.02	0.26 ± 0.02	-	-	-
CleverLabel (ours)	0.68 ± 0.03	0.32 ± 0.01	0.25 ± 0.01	0.16 ± 0.00	0.11 ± 0.00	0.08 ± 0.00	0.07 ± 0.00
CleverLabel (+ GT δ)	0.68 ± 0.03	0.41 ± 0.01	0.29 ± 0.01	0.16 ± 0.00	0.10 ± 0.00	0.05 ± 0.00	0.04 ± 0.00
Only CB	0.68 ± 0.03	0.32 ± 0.01	0.25 ± 0.01	0.16 ± 0.01	0.12 ± 0.00	0.09 ± 0.00	0.08 ± 0.00
Only CB ($\mu=0.25$)	0.55 ± 0.02	0.33 ± 0.01	0.29 ± 0.01	0.24 ± 0.01	-	-	-
Only BC	-	1.22 ± 0.04	0.78 ± 0.02	0.36 ± 0.02	-	-	-
CleverLabel (+ GT δ)	-	0.22 ± 0.01	-	0.18 ± 0.01	-	-	0.16 ± 0.01
Only CB	-	0.20 ± 0.01	-	0.18 ± 0.01	-	-	0.17 ± 0.01

between each other and to subsection 3.2. Class blending (CB) is an important part of improved scores but the impact is stronger for \hat{KL} . A different blending threshold ($\mu = 0.25$) which prefers the sample independent class distribution leads in most cases to similar or worse results than our selection of 0.75. Bias Correction (BC) and the correct GT offset have a measurable impact on the \hat{KL} while on KL we almost no difference but a saturation at around 0.24 for all approaches most likely due the used network backbone. With a different backbone e.g. a transformer [13] we can verify that BC positively impacts the results.

4 Discussion

In summary, we analyzed the introduced bias during the labeling process when using proposals by developing a simulation of this bias and provided two methods for reducing the proposal-introduced bias. We could show that our methods outperform current state of the art methods on the same or even lower labeling budgets. For low annotation budgets, we have even surpassed our newly proposed baseline of class blending in combination with annotation without proposals. Cost is already a limiting factor when annotating data and thus only results with a better performance for a budget of less than one (which equals the current annotation of every image once) can be expected to be applied in real world applications. We achieved this goal with CleverLabel with speedups larger than 4 with is reasonable based on previously reported values [48].

Based on our research, how should one annotate ambiguous image classification data? While there currently is no strategy for every case, the problem can be broken down into the two major questions as depicted in Figure 6. Firstly, is a bias in the data acceptable? Be aware that in CleverLabel all labels are human validated and that many consensus process already use an agreement system [1] with multiple reviewers. If a small bias is acceptable you can directly

A. Own previous papers

14 L. Author et al.

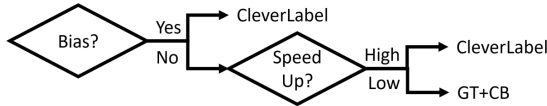


Fig. 6: Flowchart about how to annotate ambiguous data based on the questions if an introduced bias is acceptable and if the expected speedup S is high (> 3)

use proposals and an optional correction like *CleverLabel*. However, if a bias is not acceptable, the second major question is the expected speedup by using proposals for annotating your specific data. In case of a high expected speedup, the trade-off between the introduced bias and the ability to mitigate it with BC and CB favors *CleverLabel*. For a low speedup, we recommend avoiding proposals and to rely on class blending which is applicable to any dataset if you can estimate the class transitions as described in subsection 2.3. It is difficult to determine the exact trade-off point, because CB improves the results with fewer (10-) annotations, BC improves the results at above (20+) and both each other. Based on this research, we recommend a rough speedup threshold of around three for the trade-off.

4.1 Limitations

We aim at a general approach for different datasets but this results in non-optimal solutions for individual datasets. Multiple extensions for SPA like different kinds of simulated annotators would be possible but would require a larger user study for evaluation. In subsection 3.2, we compared our simulation with real data on four datasets, but a larger comparison was not feasible. It is important to note that SPA must not replace human evaluation but should be used for method development and hypothesis testing before an expensive human study which is needed to verify results. We gave a proof of concept about the benefit of bias correction with higher annotation counts with a stronger backbone like transformers. A full reevaluation of the benchmark was not feasible and it is questionable if it would lead to new insights because the scores might be lower but are expected to show similar relations.

5 Conclusion

Data quality is important but comes at a high cost. Proposals can reduce this cost but introduce a bias. We propose to mitigate this issue by simple heuristics and a theoretically motivated bias correction which makes them broader applicable and achieve up to 29.8% relative better scores with reduced cost of 30%. This analysis is only possible due to our new proposed method SPA and results in general guidelines for how to annotate ambiguous data.

Ethical Statement

While proposal-based labeling offers several benefits, it generally introduces a bias which might have ethical implications depending on the use case. We believe that it is important to take steps to improve proposal-guided annotations and introduce CleverLabel in order to enhance their quality by mitigating, however not eliminating, the effects of bias. Every operator must consciously decide whether the resulting reduced bias has a negative effect for their own application. CleverLabel aims at the utilization of all available data. As ambiguous labels can introduce additional data uncertainty into a model, while it is common to exclude these data from training. However, we expect that their consideration can also provide more nuanced information, potentially allowing for more accurate and fair decision-making. By this means, ambiguous labels may result in less overconfident models.

This work aims to facilitate and encourage further research in this field. While the contributions can be used to investigate a variety of research questions in order to improve the accuracy of predictions, we cannot identify any direct negative ethical impacts.

References

1. Addison, P.F.E.E., Collins, D.J., Trebilco, R., Howe, S., Bax, N., Hedge, P., Jones, G., Miloslavich, P., Roelfsema, C., Sams, M., Stuart-Smith, R.D., Scanes, P., Von Baumgarten, P., McQuatters-Gollop, A.: A new wave of marine evidence-based management: Emerging challenges and solutions to transform monitoring, evaluating, and reporting. *ICES Journal of Marine Science* **75**(3), 941–952 (2018). <https://doi.org/10.1093/icesjms/fsx216>
2. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2020). <https://doi.org/10.1109/IJCNN48605.2020.9207304>
3. Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., Uma, A.: We Need to Consider Disagreement in Evaluation. In: BPPF (2021)
4. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems* **32** (2019)
5. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., van den Oord, A.: Are we done with ImageNet? *arXiv preprint arXiv:2006.07159* (2020)
6. Brünner, J., Dippel, S., Koch, R., Veit, C.: ‘Tailception’: using neural networks for assessing tail lesions on pictures of pig carcasses. *Animal* **13**(5), 1030–1036 (2019). <https://doi.org/10.1017/S1751731118003038>
7. Collier, M., Mustafa, B., Kokiopoulou, E., Jenatton, R., Berent, J.: Correlated Input-Dependent Label Noise in Large-Scale Image Classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (c)*, 1551–1560 (2021)
8. Collins, K.M., Bhatt, U., Weller, A.: Eliciting and Learning with Soft Labels from Every Annotator. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. **10**(1) (jul 2022), <http://arxiv.org/abs/2207.00810>

A. Own previous papers

16 L. Author et al.

9. Cortes, C., Lawrence, N.D.: Inconsistency in conference peer review: revisiting the 2014 neurips experiment. arXiv preprint arXiv:2109.09774 (2021)
10. Davani, A.M., Díaz, M., Prabhakaran, V.: Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations (2021)
11. Desmond, M., Duesterwald, E., Brimijoin, K., Brachman, M., Pan, Q.: Semi-automated data labeling. In: NeurIPS 2020 Competition and Demonstration Track. pp. 156–169. PMLR (2021)
12. Desmond, M., Muller, M., Ashktorab, Z., Dugan, C., Duesterwald, E., Brimijoin, K., Finegan-Dollak, C., Brachman, M., Sharma, A., Joshi, N.N., Pan, Q.: Increasing the speed and accuracy of data labeling through an ai assisted interface. In: 26th International Conference on Intelligent User Interfaces. p. 392–401. IUI '21, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3397481.3450698>, <https://doi.org/10.1145/3397481.3450698>
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR (oct 2021), <http://arxiv.org/abs/2010.11929>
14. Gao, Z., Sun, F.K., Yang, M., Ren, S., Xiong, Z., Engeler, M., Burazer, A., Wildling, L., Daniel, L., Boning, D.S.: Learning from Multiple Annotator Noisy Labels via Sample-wise Label Fusion (2022)
15. Gordon, M.L., Zhou, K., Patel, K., Hashimoto, T., Bernstein, M.S.: The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–14. ACM (2021). <https://doi.org/10.1145/3411764.3445423>
16. Grossmann, V., Schmarje, L., Koch, R.: Beyond Hard Labels: Investigating data label distributions. ICML 2022 Workshop DataPerf: Benchmarking Data for Data-Centric AI (2022)
17. Gu, K., Masotto, X., Bachani, V., Lakshminarayanan, B., Nikodem, J., Yin, D.: An instance-dependent simulation framework for learning with label noise. Machine Learning pp. 1–26 (2022)
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners (nov 2021), <http://arxiv.org/abs/2111.06377>
19. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. Advances in neural information processing systems **32** (2019)
20. Jachimowicz, J.M., Duncan, S., Weber, E.U., Johnson, E.J.: When and why defaults influence decisions: A meta-analysis of default effects. Behavioural Public Policy **3**(2), 159–186 (2019)
21. Jung, H., Park, Y., Lease, M.: Predicting next label quality: A time-series model of crowdwork. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing **2**(1), 87–95 (Sep 2014). <https://doi.org/10.1609/hcomp.v2i1.13165>, <https://ojs.aaai.org/index.php/HCOMP/article/view/13165>
22. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big Transfer (BiT): General Visual Representation Learning. In: Lecture Notes in Computer Science, pp. 491–507 (2020)
23. Krizhevsky, A., Hinton, G., Others: Learning multiple layers of features from tiny images. Tech. rep. (2009)

24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. vol. 60, pp. 1097–1105. Association for Computing Machinery (2012). <https://doi.org/10.1145/3065386>
25. Kullback, S., Leibler, R.A.: On Information and Sufficiency. *Ann. Math. Statist.* **22**(1), 79–86 (1951). <https://doi.org/10.1214/aoms/1177729694>
26. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: *Workshop on challenges in representation learning, ICML*. vol. 3, p. 2 (2013)
27. Li, J., Socher, R., Hoi, S.C.H.: DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In: *International Conference on Learning Representations*. pp. 1–14 (2020)
28. Li, Y.F., Liang, D.M.: Safe semi-supervised learning: a brief introduction. *Frontiers of Computer Science* **13**(4), 669–676 (2019)
29. Lopresti, D., Nagy, G.: Optimal data partition for semi-automated labeling. In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. pp. 286–289. IEEE (2012)
30. Lukasik, M., Bhojanapalli, S., Menon, A.K., Kumar, S.: Does label smoothing mitigate label noise? *International Conference on Machine Learning (PMLR)*, 6448–6458 (mar 2020), <http://arxiv.org/abs/2003.02819>
31. Lukov, T., Zhao, N., Lee, G.H., Lim, S.N.: Teaching with Soft Label Smoothing for Mitigating Noisy Labels in Facial Expressions. pp. 648–665 (2022)
32. Mazeika, M., Tang, E., Zou, A., Basart, S., Chan, J.S., Song, D., Forsyth, D., Steinhardt, J., Hendrycks, D.: How Would The Viewer Feel? Estimating Wellbeing From Video Scenarios (NeurIPS) (2022), <http://arxiv.org/abs/2210.10039>
33. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6707–6717 (2020)
34. Motamed, M., Sakharykh, N., Kaldewey, T.: A Data-Centric Approach for Training Deep Neural Networks with Less Data. *NeurIPS 2021 Data-centric AI workshop* (2021)
35. Müller, R., Kornblith, S., Hinton, G.: When Does Label Smoothing Help? *Advances in neural information processing systems* **32** (jun 2019), <http://arxiv.org/abs/1906.02629>
36. Naeem, A., Farooq, M.S., Khelifi, A., Abid, A.: Malignant melanoma classification using deep learning: datasets, performance measurements, challenges and opportunities. *IEEE Access* **8**, 110575–110597 (2020)
37. Nguyen, T., Ilharco, G., Wortsman, M., Oh, S., Schmidt, L.: Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP pp. 1–46 (2022), <http://arxiv.org/abs/2208.05516>
38. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks* (2021)
39. Northcutt, C.G., Jiang, L., Chuang, I.L.: Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research* **70**, 1373–1411 (2021). <https://doi.org/10.1613/JAIR.1.12125>
40. Ooms, E.A., Zonderland, H.M., Eijkemans, M.J.C., Kriege, M., Mahdavian Delavary, B., Burger, C.W., Ansink, A.C.: Mammography: Interobserver variability in breast density assessment. *The Breast* **16**(6), 568–576 (2007). <https://doi.org/10.1016/j.breast.2007.04.007>

A. Own previous papers

18 L. Author et al.

41. Papadopoulos, D.P., Weber, E., Torralba, A.: Scaling up instance annotation via label propagation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15364–15373 (2021)
42. Patel, B.N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., Halabi, S., Langlotz, C., Lo, E., Mammarpallil, J., Mariano, A.J., Riley, G., Seekins, J., Shen, L., Zucker, E., Lungren, M.: Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine* **2**(1), 1–10 (2019). <https://doi.org/10.1038/s41746-019-0189-7>
43. Peterson, J., Battleday, R., Griffiths, T., Russakovsky, O.: Human uncertainty makes classification more robust. Proceedings of the IEEE International Conference on Computer Vision **2019-Octob**, 9616–9625 (2019). <https://doi.org/10.1109/ICCV.2019.00971>
44. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. *ACM computing surveys (CSUR)* **54**(9), 1–40 (2021)
45. Saleh, A., Laradji, I.H., Konovalov, D.A., Bradley, M., Vazquez, D., Sheaves, M.: A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports* **10**(1), 1–10 (2020). <https://doi.org/10.1038/s41598-020-71639-x>
46. Schmarje, L., Brünger, J., Santarossa, M., Schröder, S.M., Kiko, R., Koch, R.: Fuzzy Overclustering: Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy. *Sensors* **21**(19), 6661 (2021). <https://doi.org/10.3390/s21196661>
47. Schmarje, L., Grossmann, V., Zelenka, C., Dippel, S., Kiko, R., Oszust, M., Pastell, M., Stracke, J., Valros, A., Volkmann, N., Koch, R.: Is one annotation enough? A data-centric image classification benchmark for noisy and ambiguous label estimation. 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks (2022)
48. Schmarje, L., Santarossa, M., Schröder, S.M., Zelenka, C., Kiko, R., Stracke, J., Volkmann, N., Koch, R.: A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering. Proceedings of the European Conference on Computer Vision (ECCV) (2022)
49. Schröder, S.M., Kiko, R., Koch, R.: MorphoCluster: Efficient Annotation of Plankton images by Clustering. *Sensors* **20** (2020)
50. Schulz, C., Meyer, C.M., Kieseewetter, J., Sailer, M., Bauer, E., Fischer, M.R., Fischer, F., Gurevych, I.: Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2761–2772. Association for Computational Linguistics, Florence, Italy (Jul 2019). <https://doi.org/10.18653/v1/P19-1265>, <https://aclanthology.org/P19-1265>
51. Schustek, P., Moreno-Bote, R.: Instance-based generalization for human judgments about uncertainty. *PLOS Computational Biology* **14**(6), e1006205 (jun 2018). <https://doi.org/10.1371/journal.pcbi.1006205>, <https://dx.plos.org/10.1371/journal.pcbi.1006205>
52. Shah, F.A., Sirts, K., Pfahl, D.: The impact of annotation guidelines and annotated data on extracting app features from app reviews. arXiv preprint arXiv:1810.05187 (2018)
53. Sheng, V.S., Provost, F.: Get Another Label ? Improving Data Quality and Data Mining Using Multiple , Noisy Labelers Categories and Subject Descriptors. New York pp. 614–622 (2008)

54. Singh, A., Nowak, R., Zhu, J.: Unlabeled data: Now it helps, now it doesn't. *Advances in neural information processing systems* **21** (2008)
55. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)* (2020)
56. Tarling, P., Cantor, M., Clapés, A., Escalera, S.: Deep learning with self-supervision and uncertainty regularization to count fish in underwater images pp. 1–22 (2021)
57. Tifrea, A., Clarysse, J., Yang, F.: Uniform versus uncertainty sampling: When being active is less efficient than staying passive (i) (2022), <http://arxiv.org/abs/2212.00772>
58. Uijlings, J., Mensink, T., Ferrari, V.: The Missing Link: Finding label relations across datasets (jun 2022), <http://arxiv.org/abs/2206.04453>
59. Vasudevan, V., Caine, B., Gontijo-Lopes, R., Fridovich-Keil, S., Roelofs, R.: When does dough become a bagel? Analyzing the remaining mistakes on ImageNet (2022), <http://arxiv.org/abs/2205.04596>
60. Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., Liu, Y.: Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations (2021)
61. Wei, J., Zhu, Z., Luo, T., Amid, E., Kumar, A., Liu, Y.: To Aggregate or Not? Learning with Separate Noisy Labels (jun 2022), <http://arxiv.org/abs/2206.07181>
62. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-Labeling ImageNet: From Single to Multi-Labels, From Global to Localized Labels. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2340–2350 (2021)

A. Own previous papers

A.1.7 Annotating Ambiguous Images: General Annotation Strategy for Image Classification with Real-World Biomedical Validation on Vertebral Fracture Diagnosis

Annotating Ambiguous Images: General Annotation Strategy for High-Quality Data with Real-World Biomedical Validation

Lars Schmarje

SCIENCE@SCHMARJE-SH.DE *Kiel University*

Vasco Grossmann

VGR@INFORMATIK.UNI-KIEL.DE *Kiel University*

Claudius Zelenka

CZE@INFORMATIK.UNI-KIEL.DE *Kiel University*

Johannes Brünger

JOHANNES.BRUENGER@IBAK.DE *IBAK GmbH*

Reinard Koch

RK@INFORMATIK.UNI-KIEL.DE *Kiel University*

Reviewed on OpenReview: <https://openreview.net/forum?id=gii9qgtvsL>

Abstract

In the field of image classification, existing methods often struggle with biased or ambiguous data, a prevalent issue in real-world scenarios. Current strategies, including semi-supervised learning and class blending, offer partial solutions but lack a definitive resolution. Addressing this gap, our paper introduces a novel strategy for generating high-quality labels in challenging datasets. Central to our approach is a clearly designed flowchart, based on a broad literature review, which enables the creation of reliable labels. We validate our methodology through a rigorous real-world test case in the biomedical field, specifically in deducing height reduction from vertebral imaging. Our empirical study, leveraging over 250,000 annotations, demonstrates the effectiveness of our strategies decisions compared to their alternatives.

Keywords: Data quality management, ambiguous data, annotation data application, data-centric AI, vertebral fracture

1 Introduction

Deep learning methods are at the leading methods for image classification, dependent on the availability of substantial high-quality data (Beyer et al., 2020; Yun et al., 2021). While techniques like self- or semi-supervision can reduce the need for labeled data, high-quality labels remain a necessity, especially in domain-specific tasks (Liu et al., 2023).

A critical challenge lies in the reliability and consistency of human annotations, especially for ambiguous or complex classification tasks. The consensus in current research suggests that single annotations are inadequate for ensuring label quality in such cases (Davani et al., 2022; Grossmann et al., 2022; Basile et al., 2021; Sharmanska et al., 2016). To mitigate this, the adoption of soft labels, derived from averaging multiple annotations, has been proposed (Schmarje et al., 2022a). Soft labels can capture the inherent data uncertainty, which is different from the model uncertainty often assumed by uncertainty estimation methods. However, obtaining multiple annotations per image is resource-intensive and impractical for large datasets or when expertise is scarce (Krizhevsky et al., 2012). Schmarje et al. (2023) introduces proposal-guided annotations, where a pre-trained network provides

A. Own previous papers

GENERAL ANNOTATION STRATEGY FOR HIGH-QUALITY DATA

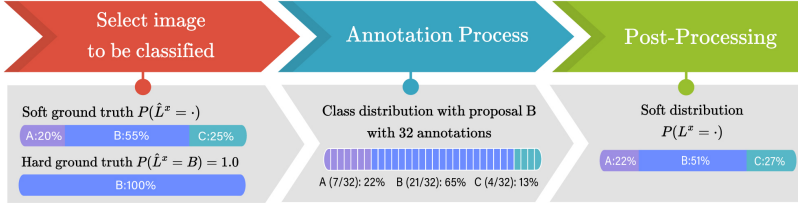


Figure 1: Illustration of the concept of hard and soft labels and how they can be created from annotations – The recommended process has three steps. In the first step, an image x is selected for annotation. The unknown ground-truth distribution ($P(\hat{L}^x = \cdot)$) could either be soft or hard as shown by the examples in the lower half. During the annotation, multiple annotations are created either with or without proposal. A proposal means that one class is recommended during the annotation process. In the example, class B is proposed and 32 annotations are generated. The average across these annotations could already be used as an approximation of the soft-label $P(L^x = \cdot)$, however it might be biased towards the proposal since it is more likely to accept a proposal Schmarje et al. (2023). Our post-processing step enhances the approximated distribution ($P(L^x = \cdot)$) from the second step by reducing this bias. In the provided example, the probability of class B is reduced since it was most likely overestimated due to the used proposal of class B.

preliminary class estimates to guide annotators, thereby enhancing annotation efficiency and quality (Desmond et al., 2021; Schmarje et al., 2021a, 2022b). An illustrative example of this process is shown in Figure 1.

In this paper, we present a comprehensive strategy for generating high-quality data in ambiguous classification scenarios. High quality data means that we improve the label quality in comparison to one annotation only. We validate our approach in vertebral fracture diagnosis, a field with significant challenges due to data ambiguity, which is ideal for testing our strategy’s efficacy.

The diagnosis of vertebral fractures holds critical importance in medicine. As noted by Haczynski and Jakimiuk (2001), vertebral fractures significantly increase mortality risks and recurrence rates. Classifying these fractures, primarily based on vertebral height reduction, is challenging due to degenerative changes, leading to inconsistent annotations and impacting neural network training. This scenario underscores the necessity of our approach, especially when domain experts are scarce.

Our primary contribution lies in synthesizing literature insights into a practical, step-by-step annotation guide for ambiguous real-world data. While individual annotation techniques have been explored previously, our unified strategy addresses the entire annotation process comprehensively. We aim to make this approach universally applicable to various image classification tasks, ensuring high-quality data. Furthermore, we will release software guidelines to assist in creating high-quality annotations following our approach.

Our second key contribution is empirically validating our strategy on a vertebral fracture dataset (Löffler et al., 2020; Sekuboyina et al., 2021). By defining the guidelines prior to

A.1. Long papers

ANNOTATING AMBIGUOUS IMAGES

dataset testing, we provide an objective evaluation, underscoring the real-world applicability of our method.

By focusing on practical applications, this work seeks to promote a data-centric perspective in research and application, contributing to the discourse on data quality best practices.

2 Practical Guidelines: How to Annotate Ambiguous Data?

Before detailing our approach, it's essential to understand why these guidelines are crucial, especially in the context of ambiguous data. Typical labeling guidelines (Sager et al., 2021; Chang et al., 2017) rely on definitive, hard-encoded labels and often overlook the disagreement among annotators about a given image's class. Such discrepancies lead to ambiguous labels or, in broader terms, ambiguous data – a prevalent issue in numerous real-world applications (Tarling et al., 2021; Sambyal et al., 2022; Karimi et al., 2020; Brünger et al., 2019; Jiang et al., 2021; Zhang et al., 2023). The probability distribution of a hard label for an image x can be represented as $P(L^x = \cdot), P(\hat{L}^x = \cdot) \in \{0, 1\}^K$, where K is the number of classes and L^x, \hat{L}^x are random variables mapping an image x to its class probability. We distinguish between L^x and \hat{L}^x , which represent the estimated and ground-truth probability distributions for image x , respectively. The literature underlines that a single annotation per image is inadequate for capturing data ambiguity (Uma et al., 2021; Schmarje et al., 2022a, 2023; Davani et al., 2022; Grossmann et al., 2022; Basile et al., 2021; Sharmanska et al., 2016), necessitating multiple annotations for high-quality data.

Davani et al. (2022) demonstrated that averaging multiple annotations to form a soft label, as proposed in (Schmarje et al., 2022a; Hemming et al., 2018), is superior to using a majority vote in ambiguous cases. Moreover, label smoothing, a technique that tempers hard-coded labels with a constant factor, has been established as a method for enhancing network performance (Vaswani et al., 2017; Krothapalli and Abbott, 2020; Lukasik et al., 2020). A soft-label probability distribution for image x can be expressed as $P(L^x = \cdot), P(\hat{L}^x = \cdot) \in [0, 1]^K$.

It is crucial to recognize that ambiguous data is not merely an error, as suggested in (Park and Chung, 2022), but a characteristic of the data itself. Such data inherently possess uncertainty and should be treated accordingly. As such, the soft ground truth distribution $P(\hat{L}^x = \cdot)$ is generally unknown and can only be approximated for validating predicted results. This approach aligns with the growing emphasis on data-centric deep learning (Liu et al., 2021b; Jarrahi et al., 2022; Whang et al., 2023; Grossmann et al., 2022), which prioritizes data quality over model architecture.

We present a comprehensive overview of our practical guidelines through a flowchart in Figure 2. The guidelines encompass five main steps: defining the specifics of what, who, and how the annotation process will be executed, the annotation process itself, and its post-processing. The subsequent sections provide insights into the pipeline, with detailed discussions in Appendix A and Appendix B.

1. **Definition - What?** This stage outlines the classification task, specifying the classes (K) and gathering raw, unlabeled image data (X_r). A crucial element is selecting a representative subset (X_u) for annotation. The selection of X_u emphasizes

A. Own previous papers

GENERAL ANNOTATION STRATEGY FOR HIGH-QUALITY DATA

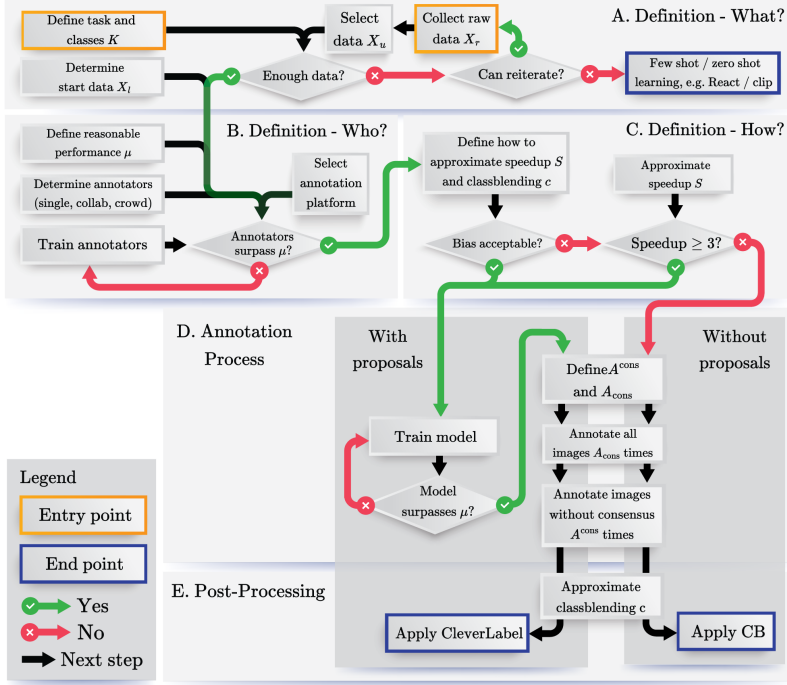


Figure 2: Flowchart with guidelines on how to annotate ambiguous data, best viewed in color.

the importance of having an adequate sample size per class for effective training and evaluation or considers the use of self-supervised techniques for datasets without enough representative information. The definition of X_l describes a smaller, precisely labeled dataset for evaluating annotators and training proposal generation networks.

2. Definition - Who?

This stage focuses on identifying suitable annotators for image labeling, ranging from individuals to crowdsourcing. It is important to train the annotators according to the task's specific needs and setting a quality threshold for annotations (μ), typically recommended between 60% and 80%. For classes with higher occurrence rates, a browsing environment for annotation is advised to improve process efficiency.

3. Definition - How?

In this stage, the decision to include proposals in the annotation process is discussed.

A.1. Long papers

ANNOTATING AMBIGUOUS IMAGES

Proposals can accelerate annotation but may introduce bias as discussed in (Schmarje et al., 2023). Key considerations include the acceptability of bias, the extent of speedup (S), and a class confusion matrix (c) for later processing. Proposals are recommended if the bias is manageable and the speedup significant (typically above a threshold of 3), otherwise, standard annotations are preferred.

4. Annotation Process

This phase involves conducting the annotations, with or without proposals. For ambiguous data, overclustering (Schmarje et al., 2021b,a) is advised, and DC3 (Schmarje et al., 2022b) proposals are recommended for better performance. We propose to separate the annotations process into the number of annotations needed for early consensus (A_{cons}) and the total annotations for difficult cases (A^{cons}). This separation balances ground-truth accuracy against the effort and cost of obtaining annotations.

5. Post-Processing

The concluding phase addresses the bias potentially introduced by proposals in the annotation process. CleverLabel (Schmarje et al., 2023), combining class blending and bias correction, is proposed for refining the approximated distribution if proposals were used. Even without proposals, blending the distribution with the class distribution (c) is recommended to enhance it.

3 Evaluation

Our objective is to address the challenges of annotating real-world application tasks, such as classifying vertebral fractures in medical images. To demonstrate the practicality of our proposed strategy, we apply it to the task using publicly available datasets Verse2019 (Löffler et al., 2020) and Verse2020 (Sekuboyina et al., 2021).

3.1 Applying the Strategy

The Verse datasets were chosen for their reproducibility and suitability in validating our workflow for classifying osteoporotic vertebral fractures. Following Sekuboyina et al. (2021); Löffler et al. (2020), we focus on thoracolumbar vertebrae, the primary site for osteoporotic fractures, as illustrated in Figure 3. Our classification approach is based on an adapted version of the Genant semi-quantitative score (Genant et al., 1996), which categorizes fractures into four classes based on the degree of height reduction in vertebrae compared to their neighbors. However, we exclude degenerative deformities, such as minor height reductions (up to 20%), which are often indistinguishable in the given CT images due to lower resolution compared to standard radiographs. As detailed in Figure 4, this results in a dataset of 3,761 individual vertebrae, with a notable class imbalance skewed towards class zero.

In contrast to the original authors of the Verse2019 dataset (Löffler et al., 2020), we employ non-medical experts for annotation and utilize 2D projections from the 3D CT data. The rationale and specifics of this approach are elaborated in the appendix Appendix C. Further details regarding data preprocessing, annotators, annotation platform, approximated variables, and used hyperparameters are provided in Appendix D.

A. Own previous papers

GENERAL ANNOTATION STRATEGY FOR HIGH-QUALITY DATA

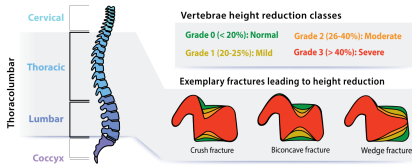


Figure 3: Illustration of a spine and definition of height reduction classes

Name	δ	\hat{p}_c	# Classes	Largest Class [%]	S
Verse (Ours)	0.1143	0.6833	4	90.11	1.1636
CIFAR10-H	0.0973	0.7766	10	10.16	2.4665
MiceBone	0.3636	0.4775	3	70.48	1.4471
Plankton	0.5784	0.7368	10	30.37	4.4319
Turkey	0.2164	0.6863	3	75.95	1.4877

Figure 4: Comparison of dataset-specific variables with previously reported values (Schmarje et al., 2023)

3.2 Analysis

Our annotators completed over 100 iterations, half with proposals and half without. However, due to the failure of one annotator to meet the acceptance threshold μ , their annotations, including training iterations, were excluded, leaving us with 80 valid iterations or about 250,000 annotations. An additional 11 iterations employed 3D data from the test dataset for comparative analysis. The average annotation rate was calculated at 3,259.36 annotations per hour, indicating that approximately two hours are needed to complete the 3,761 annotations, accounting for technical overhead and breaks. Comparison of dataset-specific variables with those in previous studies (Schmarje et al., 2022b) is provided in Figure 4. δ represents the data-specific offset when annotating with proposals as described in (Schmarje et al., 2023) and represents how strongly annotators are influenced in their decision by a proposal. 0 means no influence while 1 means following the proposal completely. S denotes the speedup between annotating with and without proposals. \hat{p}_c indicates the percentage of data with at least 95% consistent annotations. Notably, the dataset offset δ and near-consensus rate \hat{p}_c are similar to prior reports. We observed a larger dominant class and a smaller speedup S in our dataset, as shown in Figure 4. This finding is theoretically consistent with the notion that a proposal is less impactful when the majority of images already belong to a specific class, as is the case with the high class imbalance towards class 0 in the Verse dataset. Our human annotator analysis indicated comparable quality to original test data, with increased uncertainties in cases of challenging or incorrect majority votes. Further details are provided in subsection E.2.

EVALUATION OF NETWORK PERFORMANCE WITH HARD LABELS

Firstly, we evaluate the performance of our approach on the originally defined test data by (Löffler et al., 2020). Our best performing network for proposal generation achieved a macro F1 score of 0.57. This network was a semi-supervised Mean-Teacher model (Tarvainen and Valpola, 2017) with overclustering and default hyperparameters from (Schmarje et al., 2022b) and was trained on the original annotations on the subset X_l as defined in Appendix C. A standard ResNet50 model (He et al., 2016) with default hyperparameters (details in the supplementary) and cross-entropy loss function attained a score of 0.63 if it was trained not on our newly annotated data. This outperformed the more complex proposal network and prior results by Wei et al. (2022) for non-specialized networks. Our data-centric approach enabled these improvements without model modifications, purely based on

A.1. Long papers

ANNOTATING AMBIGUOUS IMAGES

Table 1: Results of KL divergence across different methods and number of annotations – The first method follows our recommended guidelines. The next two methods utilize proposals, while the last three do not, leading to a slower annotation time compared to the recommendation. Therefore, when comparing, one should consider a higher number of annotations for DC3 proposals, as they could potentially be achieved in the same or less time. The results are presented as $KL \pm STD$ (Relative Change in % compared to the recommended method).

Method			Number of Annotations			
Proposal	Blending	Correction	1	3	5	10
DC3	Balanced	Yes	0.4481 ± 0.1957 (0.00%)	0.2425 ± 0.0741 (0.00%)	0.1832 ± 0.0095 (0.00%)	0.1615 ± 0.0129 (0.00%)
DC3	Balanced	No	1.0309 ± 0.1750 (130.07%)	0.2704 ± 0.0760 (11.53%)	0.2105 ± 0.0256 (14.85%)	0.4546 ± 0.4343 (181.49%)
DC3	No	Yes	1.7382 ± 0.1553 (287.91%)	0.2499 ± 0.0374 (3.08%)	0.2081 ± 0.0171 (13.57%)	0.2798 ± 0.1325 (73.25%)
No	Only Blends	No	0.2311 ± 0.0224 (-48.42%)	0.2394 ± 0.0930 (-1.27%)	0.1904 ± 0.0209 (3.88%)	0.1644 ± 0.0062 (1.79%)
No	Balanced	No	0.8565 ± 0.4466 (91.14%)	0.1966 ± 0.0245 (-18.90%)	0.1678 ± 0.0224 (-8.44%)	0.1568 ± 0.0165 (-2.92%)
No	No	No	0.5578 ± 0.1259 (24.49%)	0.2451 ± 0.0127 (1.10%)	0.5435 ± 0.4285 (196.61%)	0.1898 ± 0.0380 (17.52%)

input data quality. Future research could explore further enhancements by incorporating specialized loss functions and advanced backend models into our workflow.

EVALUATION OF NETWORK PERFORMANCE WITH SOFT LABELS

In Table 1, we compare the network predictions on the improved distributions with blending and correction ($(P(L^x = \cdot))$) against the approximated soft ground truth distribution ($(P(\tilde{L}^x = \cdot))$) using Kullback-Leibler divergence (Kullback and Leibler, 1951). Comparing DC3 proposals with and without the proposed blending and correction steps in (Schmarje et al., 2023), we find improved results using both methods. Therefore, we confirm the efficacy of CleverLabel (a combination of these improvements) in the post-processing phase. Additionally, we can validate that applying balanced blending to annotations without proposals enhances results. Particularly, class blending based on the majority class proved effective for this task (Only Blends). However, the benefit diminishes with more than five annotations, at which point it becomes less effective than our strategy’s recommended method. In applying our proposed annotation strategy (subsection 3.1), we noted that bias introduction was not a major concern. Yet, our results affirm that the strategy leads to optimal outcomes even in scenarios where bias is unacceptable. Given a speedup of about 1.2, our flow chart would recommend to annotate without proposals, using only balanced blending, which emerged as the most effective method for multiple annotations. Theoretically, with higher speedups, we could compare a larger number of annotations with proposals to fewer annotations without proposals. For instance, at a decision boundary speedup of 3, five annotations with a proposal have a lower annotation cost than three without proposal ($5 \cdot 0.33$ vs. 3) and yield a reduced KL score. Hence, our strategy recommends the correct approach even in scenarios with higher speedups.

3.3 Limitations and Future Work

The use of a single dataset for validating our strategy may appear as a limitation, and the reliance on existing literature could be seen as a weakness. However, we contend that these aspects are, in fact, complementary. Our strategy, derived from a synthesis of recent re-

A. Own previous papers

GENERAL ANNOTATION STRATEGY FOR HIGH-QUALITY DATA

search, leverages verified methodologies to ensure the validity of the overall approach, rather than just its individual components. Moreover, all potential decision paths in the graph were empirically tested. Thus, we believe the validation using the Verse datasets is adequate, chosen for their real-world relevance and abundance of ambiguous data, necessitating complex annotations and yielding insightful results. While the Verse datasets are smaller and less widely used than ImageNet (Russakovsky et al., 2015) or CIFAR10 (Krizhevsky and Hinton, 2009), their advantage lies in providing non-curved, class-distinct datasets. Although curated datasets like ImageNet include ambiguous classes (Beyer et al., 2020; Vasudevan et al., 2022; Peterson et al., 2019), they do not exhibit the level of noise and ambiguity found in real-world data as demonstrated in (Schmarje et al., 2021a, 2022a). The requirement for multiple annotations under various setups for this study necessitated over 250,000 annotations for roughly 4,000 images, making such an extensive evaluation unfeasible for larger datasets at present. While it is easier to annotate any classification task with our strategy, we still need many more annotations for evaluation, and thus evaluation is currently limited to smaller datasets. Future research aims to apply and validate this approach on a larger scale, and we invite fellow researchers to join us in this endeavor.

4 Conclusion

This study introduces a comprehensive strategy for annotating and processing ambiguous data, with a specific focus on vertebral fracture diagnosis. Our analysis of human annotator performance underscored the efficacy of proposal-guided annotation, which not only enhanced F1 scores but also streamlined the training process. A neural network trained on the newly annotated data exhibited superior classification performance on the original test set, potentially halving the Kullback-Leibler (KL) divergence score compared to our new baseline. Through thorough evaluation across all possible scenarios, we have substantiated the effectiveness of our strategy, rooted in literature, for this specific application. By providing practical guidelines and demonstrating their successful application in a real-world context, this research significantly contributes to the creation of high-quality datasets and the advancement of data-centric deep learning.

5 Reproducibility Statement

To ensure reproducibility, we have documented and made accessible all essential resources of this study. Our source code, data processing methods, model training, and evaluation procedures, along with a model card, are provided for transparency. We utilized the publicly available Verse'19 (Löffler et al., 2020) and Verse'20 (Sekuboyina et al., 2021) datasets, concentrating on thoracolumbar vertebrae for classifying osteoporotic fractures. Detailed preprocessing instructions are included to help with data reconstruction.

The human annotation process was executed on a web-based platform, employing non-medical expert annotators who received specific training for this task. This project entailed over 250,000 annotations across approximately 4,000 images. To minimize variability inherent in human annotation, we established comprehensive guidelines and criteria for a consistent approach.

A.1. Long papers

ANNOTATING AMBIGUOUS IMAGES

For model training, we adopted a semi-supervised Mean-Teacher (Tarvainen and Valpola, 2017) model with overclustering, utilizing a ResNet50 architecture. Key hyperparameters included a learning rate of 0.03 for the proposal network and 0.1 for the evaluation network, batch sizes of 64 and 128, and a weight decay of 0.0005. The proposal network underwent training for approximately 600 epochs, while the evaluation network was trained for about 60 epochs.

Acknowledgments

We acknowledge funding of L. Schmarje by the ARTEMIS project (Grant number 01EC1908E) funded by the Federal Ministry of Education and Research (BMBF, Germany). We further acknowledge the funding of V. Grossmann by the Marispace-X project (grant number 68GX21002E), both funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK, Germany).

References

- V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, and A. Uma. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21, 2021.
- D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060, 2019.
- L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, A. van den Oord, A. van den Oord, and A. van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.
- J. Brünger, S. Dippel, R. Koch, and C. Veit. ‘Tailception’: using neural networks for assessing tail lesions on pictures of pig carcasses. *Animal*, 13(5):1030–1036, 2019. ISSN 17517311. doi: 10.1017/S1751731118003038.
- J. C. Chang, S. Amershi, and E. Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. *Conference on Human Factors in Computing Systems - Proceedings*, 2017-May:2334–2346, 2017. doi: 10.1145/3025453.3026044.
- T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.
- A. M. Davani, M. Díaz, and V. Prabhakaran. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. ISSN 2307387X.
- M. Desmond, E. Duesterwald, K. Brimijoin, M. Brachman, and Q. Pan. Semi-automated data labeling. In *NeurIPS 2020 Competition and Demonstration Track*, pages 156–169. PMLR, 2021.

A. Own previous papers

GENERAL ANNOTATION STRATEGY FOR HIGH-QUALITY DATA

- J. M. Durden, B. J. Bett, T. Schoening, K. J. Morris, T. W. Nattkemper, and H. A. Ruhl. Comparison of image annotation data generated by multiple investigators for benthic ecology. *Marine Ecology Progress Series*, 552:61–70, 2016. ISSN 01718630. doi: 10.3354/meps11775.
- H. K. Genant, M. Jergas, L. Palermo, M. Nevitt, R. S. Valentin, D. Black, and S. R. Cummings. Comparison of semiquantitative visual and quantitative morphometric assessment of prevalent and incident vertebral fractures in osteoporosis. *Journal of Bone and Mineral Research*, 11(7):984–996, 1996. ISSN 08840431. doi: 10.1002/jbmr.5650110716.
- R. Girdhar, A. El-Nouby, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. OmniMAE: Single Model Masked Pretraining on Images and Videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10406–10417, 2023.
- V. Grossmann, L. Schmarje, and R. Koch. Beyond Hard Labels: Investigating data label distributions. *ICML 2022 Workshop DataPerf: Benchmarking Data for Data-Centric AI*, 2022.
- J. Haczynski and A. Jakimiuk. Vertebral fractures: a hidden problem of osteoporosis. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 7(5):1108–1117, 2001.
- K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV 2016*, pages 630–645, 2016.
- K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.
- V. Hemming, M. A. Burgman, A. M. Hanea, M. F. McBride, and B. C. Wintle. A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution*, 9(1):169–180, 2018. ISSN 2041210X. doi: 10.1111/2041-210X.12857.
- J. M. Jachimowicz, S. Duncan, E. U. Weber, and E. J. Johnson. When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, 3(2): 159–186, 2019.
- M. H. Jarrahi, A. Memariani, and S. Guha. The Principles of Data-Centric AI (DCAI). pages 1–14, 2022.
- J. A. Jiang, M. K. Scheuerman, C. Fiesler, and J. R. Brubaker. Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, 16(8):e0256762, 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0256762.
- D. Karimi, G. Nir, L. Fazli, P. C. Black, L. Goldenberg, and S. E. Salcudean. Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images—Role of Multiscale Decision Aggregation and Data Augmentation. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1413–1426, 2020. doi: 10.1109/JBHI.2019.2944643.

A.1. Long papers

ANNOTATING AMBIGUOUS IMAGES

- K. M. Knausgård, A. Wiklund, T. K. Sordalen, K. T. Halvorsen, A. R. Kleiven, L. Jiao, and M. Goodwin. Temperate fish detection and classification: a deep learning based approach. *Applied Intelligence*, 2021. ISSN 15737497. doi: 10.1007/s10489-020-02154-9.
- A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 60, pages 1097–1105. Association for Computing Machinery, 2012. doi: 10.1145/3065386.
- U. Krothapalli and A. L. Abbott. Adaptive label smoothing. *arXiv preprint arXiv:2009.06432*, 2020.
- S. Kullback and R. A. Leibler. On Information and Sufficiency. *Ann. Math. Statist.*, 22(1): 79–86, 1951. doi: 10.1214/aoms/1177729694.
- S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.
- D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 2, 2013.
- J. Li, R. Socher, and S. C. H. Hoi. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations*, pages 1–14, 2020.
- Y.-H. Liao, A. Kar, and S. Fidler. Towards Good Practices for Efficiently Annotating Large-Scale Image Classification Datasets. *CVPR*, pages 4350–4359, 2021.
- H. Liu, K. Son, J. Yang, C. Liu, J. Gao, Y. J. Lee, and C. Li. Learning Customized Visual Models with Retrieval-Augmented Knowledge. *arXiv preprint arXiv:2301.07094*, 2023.
- L. Liu, T. Zhou, G. Long, J. Jiang, X. Dong, and C. Zhang. Isometric Propagation Network for Generalized Zero-shot Learning. *International Conference on Learning Representations*, 2021a.
- Z. Y.-C. Liu, S. Roychowdhury, S. Tarlow, A. Nair, S. Badhe, and T. Shah. AutoDC: Automated data-centric processing. (NeurIPS):1–6, 2021b.
- M. T. Löffler, A. Sekuboyina, A. Jacob, A. L. Grau, A. Scharr, M. E. Husseini, M. Kallweit, C. Zimmer, T. Baum, and J. S. Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):1–6, 2020. ISSN 26386100. doi: 10.1148/ryai.2020190138.
- M. Lukasik, S. Bhojanapalli, A. K. Menon, and S. Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020.

A. Own previous papers

GENERAL ANNOTATION STRATEGY FOR HIGH-QUALITY DATA

- O. R. Lyman. *An Introduction to Statistical Methods and Data Analysis*. 1993.
- D. P. Papadopoulos, E. Weber, and A. Torralba. Scaling up instance annotation via label propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15364–15373, 2021.
- K. Park and H. Chung. Uncertainty Guided Pseudo-Labeling: Estimating Uncertainty on Ambiguous Data for Escalating Image Recognition Performance. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, volume 2, pages 541–551. SCITEPRESS - Science and Technology Publications, 2022. doi: 10.5220/0010901600003116.
- J. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:9616–9625, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00971.
- O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- C. Sager, C. Janiesch, and P. Zschech. A survey of image labelling for computer vision applications. *Journal of Business Analytics*, 4(2):91–110, 2021. ISSN 25732358. doi: 10.1080/2573234X.2021.1908861.
- A. Saleh, I. H. Laradji, D. A. Konovalov, M. Bradley, D. Vazquez, and M. Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1):1–10, 2020. ISSN 20452322. doi: 10.1038/s41598-020-71639-x.
- A. S. Sambyal, N. C. Krishnan, and D. R. Bathula. Towards Reducing Aleatoric Uncertainty for Medical Imaging Tasks. In *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2022.
- L. Schmarje, C. Zelenka, U. Geisen, C.-C. Glier, and R. Koch. 2D and 3D Segmentation of uncertain local collagen fiber orientations in SHG microscopy. In *DAGM German Conference of Pattern Recognition*, volume 11824 LNCS, pages 374–386. Springer, 2019.
- L. Schmarje, J. Brünger, M. Santarossa, S.-M. Schröder, R. Kiko, and R. Koch. Fuzzy Overclustering: Semi-supervised classification of fuzzy labels with overclustering and inverse cross-entropy. *Sensors*, 21(19):6661, 2021a. ISSN 23318422. doi: 10.3390/s21196661.
- L. Schmarje, Y.-H. Liao, and R. Koch. A Data-Centric Image Classification Benchmark. *NeurIPS 2021 Data-centric AI workshop*, 2021b.
- L. Schmarje, V. Grossmann, C. Zelenka, S. Dippel, R. Kiko, M. Oszust, M. Pastell, J. Stracke, A. Valros, N. Volkmann, and R. Koch. Is one annotation enough? A data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35:33215—33232, 2022a.

A.1. Long papers

ANNOTATING AMBIGUOUS IMAGES

- L. Schmarje, M. Santarossa, S.-M. Schröder, C. Zelenka, R. Kiko, J. Stracke, N. Volkmann, and R. Koch. A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022b.
- L. Schmarje, V. Grossmann, T. Michels, J. Nazarenus, M. Santarossa, C. Zelenka, and R. Koch. Label Smarter, Not Harder: CleverLabel for Faster Annotation of Ambiguous Image Classification with Higher Quality. *arXiv preprint arXiv:2305.12811*, 2023.
- S.-M. Schröder, R. Kiko, and R. Koch. MorphoCluster: Efficient Annotation of Plankton images by Clustering. *Sensors*, 20, 2020.
- A. Sekuboyina, M. E. Husseini, A. Bayat, M. Löffler, H. Liebl, H. Li, G. Tetteh, J. Kukačka, C. Payer, D. Štern, M. Urschler, M. Chen, D. Cheng, N. Lessmann, Y. Hu, T. Wang, D. Yang, D. Xu, F. Ambellan, T. Amiranashvili, M. Ehlke, H. Lamecker, S. Lehnert, M. Lirio, N. P. de Olague, H. Ramm, M. Sahu, A. Tack, S. Zachow, T. Jiang, X. Ma, C. Angerman, X. Wang, K. Brown, A. Kirszenberg, É. Puybareau, D. Chen, Y. Bai, B. H. Rapazzo, T. Yeah, A. Zhang, S. Xu, F. Hou, Z. He, C. Zeng, Z. Xiangshang, X. Liming, T. J. Netherton, R. P. Mumme, L. E. Court, Z. Huang, C. He, L. W. Wang, S. H. Ling, L. D. Huynh, N. Boutry, R. Jakubicek, J. Chmelik, S. Mulay, M. Sivaprakasam, J. C. Paetzold, S. Shit, I. Ezhov, B. Wiestler, B. Glocker, A. Valentinitisch, M. Rempfler, B. H. Menze, and J. S. Kirschke. VERSE: A Vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical Image Analysis*, 73, 2021. ISSN 13618423. doi: 10.1016/j.media.2021.102166.
- V. Sharmanska, D. Hernandez-Lobato, J. M. Hernandez-Lobato, and N. Quadrianto. Ambiguity Helps: Classification with Disagreements in Crowdsourced Annotations. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:2194–2202, 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.241.
- K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.
- P. Tarling, M. Cantor, A. Clapés, and S. Escalera. Deep learning with self-supervision and uncertainty regularization to count fish in underwater images. *Plos one*, 17(5):1–22, 2021.
- A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *ICLR*, 2017.
- A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12752.
- V. Vasudevan, B. Caine, R. Gontijo-Lopes, S. Fridovich-Keil, and R. Roelofs. When does dough become a bagel? Analyzing the remaining mistakes on ImageNet. *Advances in Neural Information Processing Systems*, 35:6720–6734, 2022.

A. Own previous papers

GENERAL ANNOTATION STRATEGY FOR HIGH-QUALITY DATA

- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. *Advances in neural information processing systems*, 30, 2017.
- N. Volkmann, J. Brünger, J. Stracke, C. Zelenka, R. Koch, N. Kemper, and B. Spindler. Learn to train: Improving training data for a neural network to detect pecking injuries in turkeys. *Animals* 2021, 11:1–13, 2021. doi: 10.3390/ani11092655.
- W. Wang, V. W. Zheng, H. Yu, and C. Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.
- Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.
- X. Wei, H. Cong, Z. Zhang, J. Peng, G. Chen, and J. Li. Faint Features Tell: Automatic Vertebrae Fracture Screening Assisted by Contrastive Learning. 2022.
- S. E. Whang, Y. Roh, H. Song, and J. G. Lee. Data collection and quality challenges in deep learning: a data-centric AI perspective. *VLDB Journal*, 2023. ISSN 0949877X. doi: 10.1007/s00778-022-00775-9.
- S. Yun, S. J. Oh, B. Heo, D. Han, J. Choe, and S. Chun. Re-Labeling ImageNet: From Single to Multi-Labels, From Global to Localized Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2340–2350, 2021.
- J. Zhang, Y. Zheng, and Y. Shi. A Soft Label Method for Medical Image Segmentation with Multirater Annotations. *Computational Intelligence and Neuroscience*, 2023:1–11, 2023. ISSN 1687-5265. doi: 10.1155/2023/1883597.

A.2 Short papers

A.2.1 Life is not black and white – Combining Semi-Supervised Learning with fuzzy labels

A. Own previous papers

Life Is Not Black and White - Combining Semi-supervised Learning With Fuzzy Labels

A concept paper

Lars Schmarje¹, Reinhard Koch¹

¹Multimedia Information Processing Group, University of Kiel, Christian-Albrechts-Platz 4, 24118 Kiel, Germany

Abstract

The required amount of labeled data is one of the biggest issues in deep learning. Semi-Supervised Learning can potentially solve this issue by using additional unlabeled data. However, many datasets suffer from variability in the annotations. The aggregated labels from these annotations are not consistent between different annotators and thus are considered fuzzy. These fuzzy labels are often not considered by Semi-Supervised Learning. This leads either to an inferior performance or to higher initial annotation costs in the complete machine learning development cycle. We envision the incorporation of fuzzy labels into Semi-Supervised Learning and give a proof-of-concept of the potential lower costs and higher consistency in the complete development cycle. As part of our concept, we discuss current limitations, future research opportunities and potential broad impacts.

Keywords

Semi-Supervised Learning, uncertainty, variability, fuzzy, classification, clustering, annotation, consensus

1. Introduction

Deep Learning was successfully applied to many computer vision problems over the last years. One of the biggest issues with deep learning is the required amount of labeled data for training. Thus, many Semi- and Self-Supervised algorithms have been proposed which can decrease the required labeled data by using additional unlabeled data [1, 2, 3, 4, 5, 6]. These methods aim for the clear vision that we feed the complete data into the network and additionally provide only few labeled samples per class. This step can then be used in a general machine learning development cycle (MLDC, Figure 1a) to reduce the annotation cost. We describe the MLDC as a three step cycle (data collection, model training and model evaluation) based on [7].

However, this vision is missing two important facts because life is not just black and white. Firstly, we as humans have to provide such labels. We will make mistakes and suffer from inconsistency in the form of intra- and interobserver variability [8, 9]. This means that the labels we provide may be wrong or even differ over time or between annotators. Secondly, in the real-world, we often encounter an ambiguous situation where a true label is either difficult to obtain or not existing. For example, the cross bread of two different dogs can not be classified as one of its parents. These two issues have been summarized before as *fuzzy labels* [10, 4]

LWDA'21: Lernen, Wissen, Daten, Analysen September 01–03, 2021, Munich, Germany

✉ las@informatik.uni-kiel.de (L. Schmarje); rk@informatik.uni-kiel.de (R. Koch)

📞 0000-0002-6945-5957 (L. Schmarje); 0000-0003-4398-1569 (R. Koch)

© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

A.2. Short papers

and have been discussed as problematic for many Semi-Supervised Learning algorithms. The issue of fuzzy labels is common in many life science [11, 12, 13] or medical [14] datasets. Even in curated datasets like Imagenet [15] and CIFAR10 [16, 17] these issues can exist for example due to correlated label noise with visual similar classes [18]. In current research, these issues are often countered by expensive label cleaning [19] or with noise estimation [20]. We believe these perspectives are often too narrow and lead to unnecessary prerequisites of other steps in the MLDC and/or are expensive with regard to the required annotation time. In this paper, we will give an alternative perspective of incorporating Semi-Supervised Learning into a MLDC which is aware of the issue of fuzzy labels (Figure 1c).

1.1. The issue of fuzzy labels

A *fuzzy* label l is a probability distribution over the classes C for an image x . A label is the aggregation of the annotations a_1, \dots, a_n for x . These annotations are one-hot encoded estimates of the label l by human annotators. Fuzzy labels are aggregated annotations which are not consistent with each other (e.g. $\forall i : a_i \neq \frac{1}{n} \sum_j a_j$). For example, if you have 10 annotations and 5 annotations are for class A while the other 5 are for class B, the aggregated fuzzy label could be 0.5 for each class. If all annotations are consistent with each other, we call the label *consistent*. This definition is similar to soft and hard labels. However, every hard label could also be represented by a one-hot encoded soft-label but a label is either fuzzy or consistent. Throughout this paper, we will call also the corresponding image x fuzzy or consistent depending on its label. A dataset is called consistent if all images are consistent otherwise it is fuzzy.

These fuzzy labels pose two issues we need to overcome to achieve the vision of Semi-Supervised Learning. The first issue is that we need to incorporate the knowledge about the existence of fuzzy labels into the training. Many Semi-Supervised Learning (SSL) algorithms assume that only consistent labels exist in their dataset. Schmarje et al. [10] showed that SSL algorithms have a inferior performance if they do not consider the existence of fuzzy labels. An assumption for the reason for this performance was that fuzzy images are not easily classifiable into the existing classes and thus confuse the algorithm. The fuzzy labels can be incorporate in the training procedure with for example S2C2 which is an extension to most existing SSL algorithms [4]. The second issue is that most algorithms aim at good hard classifications as output for the neural network. This hard classification makes sense for consistent labels but does not work properly for fuzzy labels as they may have no best hard label. If we want to integrate the output of an algorithm into a MLDC, we need to consider this. For example, it could be necessary to resort fuzzy data either into an existing class, a new class or exclude the data.

1.2. Concept

Our main concept is illustrated in Figure 1c. We will first repeat the issue of most Semi-Supervised Learning (SSL) in the MLDC on uncurated data and than show the difference to our idea.

As stated above, SSL aims at reducing the annotation cost by leveraging unlabeled data which is often expected to be consistent or resulting in inferior performance [10]. However, if we have

A. Own previous papers

a uncurated dataset, we argue that the requirement of consistent labels is not given. It is costly to detect these consistent labels in the given dataset and determine their label. The costs are so high because we need to use strategies e.g. a consensus process or strict protocols like [21] to counter intra- and interobserver variability. Moreover, we need a post-processing to decide what we do with difficult, fuzzy images. If we have for example an image with a label of 50% for two classes we have to select a hard label or ignore this image. After this costly dataset cleaning, we can apply any SSL algorithm with a relative low cost (Figure 1b). The generated predictions can then be used to annotate more data more easily. The main issue is that the SSL algorithms expect also cleaned unlabeled data e.g. no fuzzy intermediate images between classes as [10] hypothesises. Otherwise, some main assumptions like the separation of classes in a higher feature space might not be valid.

We envision a machine learning cycle that includes the knowledge of fuzzy labels into the SSL algorithm (Figure 1c). Thus, the initial annotation could then be limited to a small portion of the data which we leverage as labeled data. No assumptions and annotations are then need for the unlabeled data. This could for example be achieved by additional clustering of the data or an automatic distinction between consistent and fuzzy images. A consensus step could still be required in the MLDC to use these output predictions as additional input. But this step could be implemented based on predictions of the network and thus make the annotation and consensus cheaper. Overall, we expect a lower number of required annotations and thus cost while achieving a higher consistency in the data for our envisioned development cycle in comparison to the previous one. We give a proof-of-concept in the next section.

2. Proof-of-Concept

2.1. Dataset

We use the Mice Bone dataset proposed in [4] and show examples in Figure 1d. This dataset consists of gray-scale images of collagen fibers of mice bone. The classification problem has three classes: similar, dissimilar and not relevant fiber structures. We follow [4, 10] and use only consistent images for training and validation and enforce a class balance in this data. We call this consistent and balanced training data *seed* and use the rest of the data as unlabeled data.

We use two different seeds for the Mice Bone data. The first seed is based on the segmentation masks from [22]. Due to the dimension reduction from a segmentation to a classification label and the uncertain segmentation, the labels show high variability and only one annotation is available. We treat all labels as consistent and this leads to an inconsistent seed. The second seed is based on three independent annotations of each image. These annotations allow us to estimate the fuzziness of each image and thus lead to a valid seed.

2.2. Metrics

We aim at faster and more consistent annotations throughout the complete MLDC. Every presented experiment is executed on the same raw data with one of above-described seeds. An experiment consists of three independent annotations of the same person over time for the complete dataset. For ease of referencing, we view these independent annotations over time as

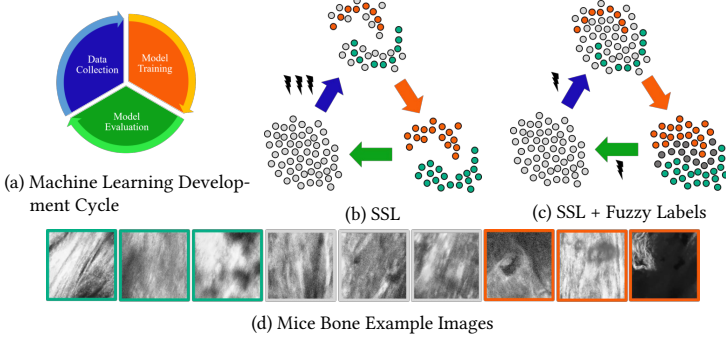


Figure 1: The left image (a) illustrates a simplified machine learning development cycle. The middle (b) and right (c) image how this process can be visualized with Semi-Supervised Learning (SSL) without and with the consideration of fuzzy labels. The light gray circles represent unlabeled images. The colored circles represent labeled images or the prediction of these labeled images. Dark gray circles represent datapoints which are identified as fuzzy. The number of thunderbolts indicates the number of required annotations by humans in each step. In the lower figure (d), we show 9 real-world Mice Bone examples from [4]. The images of the classes similar (green) and dissimilar (orange) fiber orientations are easy to consistently annotate. The middle images (grey) are more difficult to consistently annotate and thus are fuzzy.

three (intra-)annotators. The annotations were either done without any additional information or with different outputs of an SSL algorithm given as *support (predictions)*. These support predictions could be accepted or manually corrected. This support can also be an overclustering of the data. An overclustering is a clustering with more cluster than used classes [10, 4]. We give in total 6 different evaluation metrics which are either average over the three annotators or the cross-combinations between them.

Cohen's *kappa* coefficient (κ) is a statistical metric that is often used to measure the intra- and inter-observer variability [23]. The coefficient measures the agreement between two annotators for a classification task. *Accuracy* (Acc.) measures the true positives in relation to all annotations between two annotators. *F1-Score* aggregate the precision and recall between the annotation of two annotators. *Total Accuracy* (T.Acc.) counts the consistent annotations between all annotators and divides them by the total number. This metric is like Acc. but not for a pair of annotators but for all three annotators in parallel. *Consistency* (Cons.) divides how many predictions were considered consistent by the total number of images. An image is consistent with the given support predictions, if no manual correction was applied. *Time* is the time it took the annotator to label the complete dataset.

2.3. Results

We discuss six different experimental setup with support predictions as proof-of-concept results in Table 1. The first experiments used no support predictions. These results of this experiment

A. Own previous papers

Table 1

The first three rows describe the experiment in each row. An x marks the usage of a clean seed and — means not applicable because no support predictions were used. The abbreviations for the other columns are defined in subsection 2.2. The best result per column is marked bold.

Method for support predictions	Valid Seed	Used Output	Scores					
			$\kappa \uparrow$ [%]	Acc. \uparrow [%]	F1 \uparrow [%]	T.Acc. \uparrow [%]	Cons. \uparrow [%]	Time \downarrow [min]
None		None	71.35 ± 2.57	84.53 ± 1.23	80.23 ± 2.05	77.21	—	13.95 ± 2.25
FOC [10]		p_n	48.14 ± 11.9	71.78 ± 7.37	64.48 ± 8.25	59.24	65.15	9.09 ± 0.57
FOC [10]		p_o	56.4 ± 8.67	73.76 ± 5.55	68.69 ± 6.54	62.85	69.66	7.08 ± 0.44
S2C2 [4]	x	p_n	79.95 ± 1.03	87.92 ± 0.56	85.66 ± 1.14	82.01	85.26	5.15 \pm 0.60
S2C2 [4]	x	p_o	74.82 ± 0.17	84.15 ± 0.15	83.57 ± 0.24	76.74	84.96	5.51 ± 0.32
S2C2 [4]	x	$p_n \& p_o$	83.17 \pm 2.16	89.59 \pm 1.29	88.08 \pm 1.62	84.58	84.70	5.18 ± 0.63

were used to calculate the above-mentioned clean seed. The second and third experiments used support predictions from the inconsistent seed with the method FOC [10]. Either a classification head p_n or an overclustering head p_o was used as output resulting in a classification or overclustering of the data respectively. The last three experiments used a support calculated on the clean seed with the method S2C2 [4]. As output either classification head p_n , a overclustering head p_o or both heads ($p_n \& p_o$) based on the fuzziness estimation were used.

We want to highlight three important aspects. Firstly, the usage of support reduces the required annotation time while not necessarily improving its consistency. Secondly, the consistency with support is worse with an inconsistent seed and the method FOC in comparison to the others. Due to the similar architecture of FOC and S2C2, we attribute this worse performance rather to the inconsistent seed than to the method. Thirdly, S2C2 (with both outputs) more than halves the required annotation time while increases almost all other metrics.

3. Discussion

3.1. Future Research

Our proof-of-concept experiments show that using the output SSL algorithms that are aware of fuzzy labels can improve the consistency of acquired labels and reduce their annotation time. However, this benefit is not always achieved when using network predictions which enforces the importance of careful seed selection and/or SSL algorithm. As a first proof-of-concept, these results lack additional experiments such as different cross-combinations and comparisons to noise estimation and other algorithms. While these issues have to be addressed in future research, the results illustrate the potential benefits of our concept for a MLDC that considers fuzzy labels. Future research could also incorporate uncertainty estimates from humans and neural network predictions into the annotation process. A promising research direction would also be large user studies and the simulation of such studies for detailed ablations.

A.2. Short papers

3.2. Broader Impact

Semi-Supervised Learning (SSL) aims at decreasing the required amount of labeled data to a few samples per class. We envision a MLDC which allows us to apply SSL to most classification problems with relatively low cost. Especially, in domains like medical imaging, the required labeled data is a severe limiting factor. If we could resolve this issue with SSL which considers fuzzy labels, we would open a vast variety of new research opportunities at a large scale which is currently not feasible.

If we incorporate model predictions into our annotation process, we need to be aware that we might suffer from a confirmation bias. In the worst cases scenario, we could create a self-fulfilling prophecy which would lead to a degeneration of the complete MLDC. However, by carefully monitoring our processes as part of the development cycle we can detect these issues early. Additionally, we assume that cautiously leveraging this bias can lead to more high-quality data which can improve future algorithms and research.

Moreover, we advocate a change of evaluation perspective. While it is important to evaluate on highly curated datasets for algorithm development, we also have to look at the complete MLDC. Otherwise we can not detect the issues described in this paper that fuzzy labels are a limiting factor for applying SSL to different real-world data. We already see great benefits when applying deep learning to other research fields but decreasing such limiting factors further can potentially increase the impact of deep learning even more.

4. Conclusion

Semi-Supervised Learning (SSL) has great potential by solving one big issue of deep learning: The required amount of labeled data and its cost. However, many SSL algorithms do not take fuzzy labels into account and thus require expensive preprocessing steps of the data. By including the knowledge of fuzzy labels into our machine learning development cycle (MLDC), we envision that we can decrease the overall annotation time and thus its costs while achieving a higher consistency. We give a proof-of-concept of this concept on a Mice Bone dataset with fuzzy labels. We highlight that a lot of future research opportunities exist to validate and improve the presented ideas. We advocate to broaden our view from the algorithm development to the the complete MLDC to detect and overcome issues like fuzzy labels. This could spark a variety of before-infeasible research opportunities and thus lead to new breakthroughs in science in general.

References

- [1] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, arXiv preprint arXiv:2002.05709 (2020) 1597–1607. arXiv:2002.05709.
- [2] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big Self-Supervised Models are Strong Semi-Supervised Learners, Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020) (2020). arXiv:2006.10029.

A. Own previous papers

- [3] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, C. Raffel, FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence, *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)* (2020). [arXiv:2001.07685](#).
- [4] L. Schmarje, M. Santarossa, S.-M. Schröder, C. Zelenka, R. Kiko, J. Stracke, N. Volkmann, R. Koch, S2C2 - An orthogonal method for Semi-Supervised Learning on fuzzy labels (2021).
- [5] J.-B. Grill, F. Strub, F. Althché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent: A new approach to self-supervised Learning, *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)* (2020). [arXiv:2006.07733](#).
- [6] L. Schmarje, M. Santarossa, S.-M. Schroder, R. Koch, A Survey on Semi-, Self-and Unsupervised Learning for Image Classification, *IEEE Access* (2021) 1–1. [arXiv:2002.08721](#).
- [7] S. Biderman, W. J. Scheirer, Pitfalls in Machine Learning Research: Reexamining the Development Cycle (2020). [arXiv:2011.02832](#).
- [8] P. F. E. Addison, D. J. Collins, R. Trebilco, S. Howe, N. Bax, P. Hedge, G. Jones, P. Miloslavich, C. Roelfsema, M. Sams, R. D. Stuart-Smith, P. Scanes, P. Von Baumgarten, A. McQuatters-Gollop, A new wave of marine evidence-based management: Emerging challenges and solutions to transform monitoring, evaluating, and reporting, *ICES Journal of Marine Science* 75 (2018) 941–952.
- [9] D. Karimi, H. Dou, S. K. Warfield, A. Gholipour, Deep learning with noisy labels: exploring techniques and remedies in medical image analysis, *Medical Image Analysis* 65 (2020). [arXiv:1912.02911](#).
- [10] L. Schmarje, J. Brünger, M. Santarossa, S.-M. Schröder, R. Kiko, R. Koch, Beyond Cats and Dogs: Semi-supervised Classification of fuzzy labels with overclustering, *arXiv preprint arXiv:2012.01768* (2020). [arXiv:2012.01768](#).
- [11] P. Culverhouse, R. Williams, B. Reguera, V. Herry, S. González-Gil, Do experts make mistakes? A comparison of human and machine identification of dinoflagellates, *Marine Ecology Progress Series* 247 (2003) 17–25.
- [12] J. Brünger, S. Dippel, R. Koch, C. Veit, ‘Tailception’: using neural networks for assessing tail lesions on pictures of pig carcasses, *Animal* 13 (2019) 1030–1036.
- [13] J. M. Durden, B. J. Bett, T. Schoening, K. J. Morris, T. W. Nattkemper, H. A. Ruhl, Comparison of image annotation data generated by multiple investigators for benthic ecology, *Marine Ecology Progress Series* 552 (2016) 61–70.
- [14] E. Ooms, H. Zonderland, M. Eijkemans, M. Kriege, B. Mahdavian Delavary, C. Burger, A. Ansink, Mammography: Interobserver variability in breast density assessment, *The Breast* 16 (2007) 568–576.
- [15] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: *Advances in neural information processing systems*, volume 60, Association for Computing Machinery, 2012, pp. 1097–1105.
- [16] A. Krizhevsky, G. Hinton, Others, Learning multiple layers of features from tiny images, Technical Report, Citeseer, 2009.
- [17] J. Peterson, R. Battleday, T. Griffiths, O. Russakovsky, Human uncertainty makes classification more robust, *Proceedings of the IEEE International Conference on Computer Vision*

A.2. Short papers

- 2019-Octob (2019) 9616–9625. [arXiv:1908.07086](#).
- [18] M. Collier, B. Mustafa, E. Kokiopoulou, R. Jenatton, J. Berent, Correlated Input-Dependent Label Noise in Large-Scale Image Classification, *CVPR* (2021) 1551–1560. [arXiv:2105.10305](#).
 - [19] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, A. van den Oord, Are we done with ImageNet?, *arXiv preprint arXiv:2006.07159* (2020). [arXiv:2006.07159](#).
 - [20] J. Li, R. Socher, S. C. H. Hoi, DivideMix: Learning with Noisy Labels as Semi-supervised Learning, in: *International Conference on Learning Representations*, 2020, pp. 1–14. [arXiv:2002.07394](#).
 - [21] V. Hemming, M. A. Burgman, A. M. Hanea, M. F. McBride, B. C. Wintle, A practical guide to structured expert elicitation using the IDEA protocol, *Methods in Ecology and Evolution* 9 (2018) 169–180.
 - [22] L. Schmarje, C. Zelenka, U. Geisen, C.-C. Glüer, R. Koch, 2D and 3D Segmentation of Uncertain Local Collagen Fiber Orientations in SHG Microscopy, in: *DAGM German Conference of Pattern Recognition*, volume 11824 LNCS, Springer, 2019, pp. 374–386. [arXiv:1907.12868](#).
 - [23] M. L. McHugh, Interrater reliability: the kappa statistic, *PubMed Biochemia* (2012) 276–82.

A. Own previous papers

A.2.2 A Data-Centric Image Classification Benchmark

A Data-Centric Image Classification Benchmark

Lars Schmarje *

Multimedia Information Processing Group
University of Kiel
las@informatik.uni-kiel.de

Yuan-Hong Liao

University of Toronto, Vector Institute
andrew@cs.toronto.edu

Reinhard Koch

Multimedia Information Processing Group
University of Kiel
rk@informatik.uni-kiel.de

Abstract

High-quality labeled datasets are critical to the advances in machine learning and tend to benefit all kinds of model-centric algorithms, such as novel architectures and loss functions. The labeling process is usually label-intensive and time-consuming since it includes many turns of data selection, data cleaning, and data analysis. There are tons of work that aim to solve each specific step, but it lacks an understanding of how to combine them and, most importantly, a standard testbed for different dataset improving techniques. We, therefore, present the concept of a multi-domain benchmark for *acquiring consistent labels* with limited budgets. In contrast to most benchmarks that encourage novel model-centric algorithms, our multi-domain data-centric benchmark encourages algorithms to improve the provided dataset. The proposed benchmark consists of different resolutions, class distributions and domains ranging from biological to medical domains.

1 Introduction

Various benchmarks are used to gauge the advances in machine learning. Most benchmarks usually introduce a set of new tasks and a corresponding fixed dataset, such as ImageNet [11]. Machine learning researchers are encouraged to compete by devising novel algorithms on model architectures, loss functions, optimization techniques, etc., while discouraged to improve the data, such as relabeling the existing dataset or utilizing other in-domain datasets. Given that label errors are prevalent in machine learning datasets [25], this constraint can lead to overfitting the label errors in the train set. Recent work [3, 43] echo the importance of the data quality by showing that almost every model architecture improves after cleaning the train set labels.

There are many ways to improve datasets. Many data cleaning algorithms [1, 7, 14, 6] have been translated into tools and possibly repair errors. Removing biased [40, 23, 23, 39] or spurious correlations [38] from train set is particularly important in safety-critical applications, such as health care system. Active learning aims to sample new labeled data from the unlabeled pool [28, 12, 36] The advances in semi-supervised learning show that adding in-domain unlabeled data increases model performances [37] and model robustness [5]. While there are many different ways to improve datasets, it still lacks a systematic way to combine them, and most importantly, a standard testbed for different dataset improving techniques.

Most dataset improving techniques acquire additional human annotations, hampering direct comparisons between different approaches. To provide a reproducible and reliable benchmark for different

*Corresponding Author

A. Own previous papers

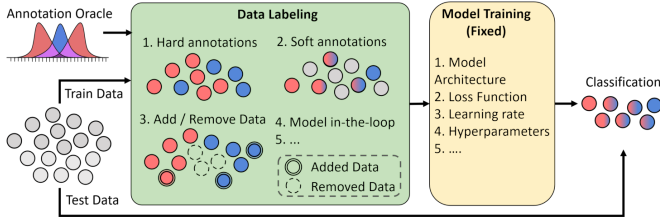


Figure 1: Illustration of data-centric Benchmark – The input for the benchmark a set of unlabeled images (grey dots) which are separated into training and test data. The unlabeled training data is labeled (green block) and a classifier is trained on the labeled data (yellow block). The target challenge is a classification of the test data with the trained model. Common benchmarks [18, 8, 17] focus on the model training and expect the labeled data to be given. In our benchmark, we fix the used model and focus on better data labeling e.g. by roughly labeling all samples, by detailed labeling of few samples or removing (dashed circles) or adding training data (double stroked circles).

dataset improving approaches, we need to include the worker simulations. Worker simulation has been widely discussed in many different domains, such as emergency and evacuation simulations [2], competitive crowd marketplaces [29], team performance [30], etc. In this work, we take the first step by focusing on image classification.

A major reason for label errors in image classification is the intra- and interobserver variability [27, 9, 26, 16, 31, 4, 10, 32, 15]. Following previous work [21], we simulate the worker responses from the underlying label distribution. Specifically, we model the underlying labels by assuming that every image has an unknown soft probability distribution $l \in [0, 1]^k$ for classification task over k classes. The assumption is motivated by two reasons. Firstly, the mislabeled images exist due to annotators' subjective opinions, e.g., the grade of an illness. A hard label $l \in \{0, 1\}^k$ could not model such a difference over the complete annotator population. Secondly, if we look for examples of biological processes, there are images of intermediate transition stages between two classes, e.g., the degeneration of living plankton to dead biomass.

The literature describes a variety of method to solve the classification task of images. Many of these algorithms are non-data-centric and thus either expect high quality hard labels [37] or consider some image as falsely classified [20]. These algorithms do not consider the unknown soft label distribution and thus ignore valuable information in the data. A consensus processes can be used to get more consistent labels but this requires many annotations and thus is often not viable on a large scale. Data-centric algorithms like [31, 32] consider the soft label distribution to create proposals for more consistent annotations e.g. by clustering visual similar images and therefore can solve the task most likely with better and fewer labels. However, no general testbed exists to compare these algorithms which hinders the research in this area.

As shown in Figure 1, we present the concept of a multi-domain benchmark for *acquiring consistent labels* with limited budgets. In contrast to most benchmarks, our benchmark includes the worker annotations simulations, providing participants noisy annotations at a fixed cost. The benchmark expects a fixed number of images and labels from the data-centric algorithms. The performance is determined by the test set performance of the neural network trained on the provided labels.

2 Benchmark

The main goal of our benchmark is acquiring consistent labels with as few annotations as possible. The benchmark can be used to compare a wide variety of data-centric methods. With the amount of annotation being fixed, the participants need to decide which image to annotate and how much they trust each annotation.

In general, the described task is a two dimensional problem with regard to the quality labels and the amount of annotations. A two-dimensional optimization is difficult to evaluate across different

A.2. Short papers

Table 1: Overview of possible usable dataset – # is an abbreviation for number. The class imbalance is given as the percentage of the smallest and largest class with regard to the complete dataset. n is the average of annotations per image

Name	# classes	Input size $[px]$	# Images	Class Imbalance [%]		n
				Smallest	Largest	
Plankton [31]	10	96x96	12280	4.16	30.37	24
Turkey [41]	2	96x96	8040	9.66	90.33	3
Mice Bone [34]	3	224x224	724	10.81	63.98	3
CIFAR10-H [27]	10	32x32	10000	9.88	10.16	51



Figure 2: Example images for possible datasets – The red and blue rows are examples of two different with a high agreement between the annotators. The grey center row show images with a high annotation variability between the red and blue class. Images taken from [32] with authors permission.

methods and thus we decided to simplify the task by introducing an annotation budget b . An annotation budget $b \in \mathbb{R}_{>0}$ describes how many annotations can be used relative to the size of the dataset. For example a budget of 0.5 for a dataset with 1000 images means that we can use 500 annotations. However, we can decide how we distribute the annotations between the images e.g. 5 images with 100 annotations each or 100 images with 5 annotations each. An annotation $a \in \{0, 1\}^k$ is a hard approximation of l from a human or an oracle in an automatic environment. The annotations can be automatically be acquired from an annotation oracle which is described in subsection 2.1.

The benchmark challenge is to relabel the data with a given budget b so that an given model Φ could be trained more successfully. The usage of one or few annotations will be insufficient to approximate the unknown soft probability distribution l in all cases. Thus the participants of the benchmark have to decide how to use their budget. Possible decisions could be the following: labeling many images roughly, labeling few images in detail, add or remove data points. A main goal is to identify and relabel poor approximation of l which are misleading during the training of Φ and thus improve the final classification [32, 43]. We can evaluate the consistency of the generated labels with the κ metric [22] and the prediction accuracy of Φ on the test set. We will describe suitable datasets the benchmark in subsection 2.2.

2.1 Annotation Oracle

For the systematic evaluation of our benchmark, we need ideally the soft label distribution l for every image and humans which provide new annotations during the relabeling. However, the soft label distribution is unknown and including humans in an automatic benchmark is not feasible. Thus, we need to approximate l and provide an annotation oracle based on the approximation of l .

If we have n annotations $a_1, \dots, a_n \in \{0, 1\}^k$ per image, we can approximate l with the average of the annotations $\hat{l} = \frac{1}{n} \sum_{i \in \{1, \dots, n\}} a_i$. We assume that for $n \rightarrow \infty$ the approximation \hat{l} is identical to l because it represent the consensus of different annotations. The annotation oracle models the human annotations by sampling from the approximated underlying label distributions \hat{l} . We might

A. Own previous papers

need add some simple thresholds to model elements like confirmation bias if the annotation should also consider a given proposal.

We are aware that this annotation oracle is a simplification of the reality and discuss possible benefits and issues in section 3.

2.2 Datasets

The main requirement for using a dataset in the benchmark is that we have n annotations per image with n as high as possible to approximate l better. We focus on 2D image classification with hundreds to thousands of images because otherwise the required annotations would not be obtainable. Preferable we would use already existing published [27] or unpublished [32, 4, 41, 24, 35] datasets with multiple annotations. An non-extensive overview about possible datasets is given in Table 1 and some example images are given in Figure 2. Overall, we aim at 5 to 10 different datasets with a wide variety. The variety should include the domains, resolutions, image modalities, number of classes, number of approximations and class distributions.

3 Discussion

In this part, we will discuss the beneficial effects, limitations and open questions of the our benchmark.

Evaluation – The general image classification task is very flexible and can be a proxy task for many use cases depending on the evaluation metric. For example we can use the Kullback-Leiber divergence [19] to measure the similarity between the predicted and the approximated ground-truth distribution \hat{l} instead of classification score like accuracy or F1-Score. This score would focus more on understanding the underlying distribution than to focus on possibly overconfident predictions [13] and could improve even the classification performance [42].

Simulation – A limitation of this benchmark is the simulation of human annotations. We can only approximate the unknown distribution l with an average over multiple annotations. The annotation oracle uses random samples from our approximated distribution \hat{l} as annotations and thus do not consider the difference between annotators. These simplifications could lead to a simulation of humans which is incorrect and thus would lead to ill-defined evaluations. However, this assumption is often used in the literature [32, 31] which makes us confident that the oracle can simulate the reality sufficiently.

Dataset Size – Moreover, we can not consider datasets with million of images because multiple annotations for approximated distribution \hat{l} would not be feasible.

Some open questions remain which need to be experimentally verified or be discussed in the community. It is currently unclear how much improvement can be gained from only tuning the data with a fixed model for training. It might be of interest to allow more recent models for training over time but this would limit the comparability. If we evaluate on fixed test set we need to ensure perfect labels. However, as we argued above we can only approximate the unknown ground-truth distribution and this will make mistakes which leads to misleading results.

Conclusion – Overall, we believe the benchmark to be beneficial for data-centric research as it considers the complete label distribution and does not ignore several aspects. Some restrictions have to be considered but the potential gain for machine learning research in general is invaluable.

Acknowledgements

We acknowledge funding of L. Schmarje by the ARTEMIS project (Grant number 01EC1908E) funded by the Federal Ministry of Education and Research (BMBF, Germany).

References

- [1] Abedjan, Z., Chu, X., Deng, D., Fernandez, R.C., Ilyas, I.F., Ouzzani, M., Papotti, P., Stonebraker, M., Tang, N.: Detecting data errors: Where are we and what needs to be done? *Proc. VLDB Endow.* **9**(12), 993–1004 (Aug 2016). <https://doi.org/10.14778/2994509.2994518>, <https://doi.org/10.14778/2994509.2994518>
- [2] Almeida, J.E., Rosseti, R.J.F., Coelho, A.L.: Crowd simulation modeling applied to emergency and evacuation simulations using multi-agent systems (2013)
- [3] Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., van den Oord, A.: Are we done with imagenet? (2020)
- [4] Brünner, J., Dippel, S., Koch, R., Veit, C.: ‘Tailception’: using neural networks for assessing tail lesions on pictures of pig carcasses. *Animal* **13**(5), 1030–1036 (2019)
- [5] Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P., Duchi, J.C.: Unlabeled data improves adversarial robustness (2019)
- [6] Chu, X., Ilyas, I., Papotti, P.: Holistic data cleaning: Putting violations into context. pp. 458–469 (04 2013). <https://doi.org/10.1109/ICDE.2013.6544847>
- [7] Chu, X., Morcos, J., Ilyas, I.F., Ouzzani, M., Papotti, P., Tang, N., Ye, Y.: Katara: A data cleaning system powered by knowledge bases and crowdsourcing. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. p. 1247–1261. SIGMOD ’15, Association for Computing Machinery, New York, NY, USA (2015). <https://doi.org/10.1145/2723372.2749431>, <https://doi.org/10.1145/2723372.2749431>
- [8] Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. pp. 215–223 (2011)
- [9] Culverhouse, P., Williams, R., Reguera, B., Herry, V., González-Gil, S.: Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series* **247**, 17–25 (2003)
- [10] De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O’Donoghue, B., Visentin, D., Others: Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* **24**(9), 1342–1350 (2018)
- [11] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 248–255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
- [12] Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data (2017)
- [13] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: *International Conference on Machine Learning*. pp. 1321–1330. PMLR (2017)
- [14] Kandel, S., Paepcke, A., Hellerstein, J., Heer, J.: Wrangler: Interactive visual specification of data transformation scripts. In: *ACM Human Factors in Computing Systems (CHI)* (2011), <http://vis.stanford.edu/papers/wrangler>
- [15] Karimi, D., Nir, G., Fazli, L., Black, P.C., Goldenberg, L., Salcudean, S.E.: Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images—Role of Multiscale Decision Aggregation and Data Augmentation. *IEEE Journal of Biomedical and Health Informatics* **24**(5), 1413–1426 (2020)
- [16] Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Medical Image Analysis* **65** (2020)
- [17] Krizhevsky, A., Hinton, G., Others: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)
- [18] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. vol. 60, pp. 1097–1105. Association for Computing Machinery (2012)
- [19] Kullback, S., Leibler, R.A.: On Information and Sufficiency. *Ann. Math. Statist.* **22**(1), 79–86 (1951). <https://doi.org/10.1214/aoms/117729694>, <https://doi.org/10.1214/aoms/117729694>

A. Own previous papers

- [20] Li, J., Socher, R., Hoi, S.C.H.: DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In: International Conference on Learning Representations. pp. 1–14 (2020)
- [21] Liao, Y.H., Kar, A., Fidler, S.: Towards good practices for efficiently annotating large-scale image classification datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4350–4359 (June 2021)
- [22] McHugh, M.L.: Interrater reliability: the kappa statistic. *PubMed Biochemia* (3), 276–82 (2012), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/>
- [23] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**(6) (Jul 2021). <https://doi.org/10.1145/3457607>, <https://doi.org/10.1145/3457607>
- [24] Niemeyer, F., Galbusera, F., Tao, Y., Kienle, A., Beer, M., Wilke, H.J.: A deep learning model for the accurate and reliable classification of disc degeneration based on mri data. *Investigative Radiology* **56**(2), 78–85 (2021)
- [25] Northcutt, C.G., Jiang, L., Chuang, I.L.: Confident learning: Estimating uncertainty in dataset labels (2021)
- [26] Ooms, E., Zonderland, H., Eijkemans, M., Kriege, M., Mahdavian Delavary, B., Burger, C., Ansink, A.: Mammography: Interobserver variability in breast density assessment. *The Breast* **16**(6), 568–576 (dec 2007)
- [27] Peterson, J., Battleday, R., Griffiths, T., Russakovsky, O.: Human uncertainty makes classification more robust. Proceedings of the IEEE International Conference on Computer Vision **2019-Octob**, 9616–9625 (2019)
- [28] Riccardi, G., Hakkani-Tur, D.: Active learning: Theory and applications to automatic speech recognition. *Speech and Audio Processing, IEEE Transactions on* **13**, 504 – 511 (08 2005). <https://doi.org/10.1109/TSA.2005.848882>
- [29] Saremi, R., Yang, Y., Vesonder, G., Ruhe, G., Zhang, H.: Crowdsim: A hybrid simulation model for failure prediction in crowdsourced software development (2021)
- [30] Saremi, R.L., Yang, Y., Ruhe, G., Messinger, D.: Leveraging crowdsourcing for team elasticity: an empirical evaluation at topcoder. In: 2017 IEEE/ACM 39th International Conference on Software Engineering: Software Engineering in Practice Track (ICSE-SEIP). pp. 103–112 (2017). <https://doi.org/10.1109/ICSE-SEIP.2017.2>
- [31] Schmarje, L., Brünger, J., Santarossa, M., Schröder, S.M., Kiko, R., Koch, R.: Beyond Cats and Dogs: Semi-supervised Classification of fuzzy labels with overclustering (2020)
- [32] Schmarje, L., Santarossa, M., Schröder, S.M., Zelenka, C., Kiko, R., Stracke, J., Volkman, N., Koch, R.: S2C2 - An orthogonal method for Semi-Supervised Learning on fuzzy labels (2021)
- [33] Schmarje, L., Zelenka, C., Geisen, U., Glüer, C.C., Koch, R.: 2D and 3D Segmentation of Uncertain Local Collagen Fiber Orientations in SHG Microscopy. In: DAGM German Conference of Pattern Recognition, vol. 11824 LNCS, pp. 374–386. Springer (2019)
- [34] Schmarje, L., Zelenka, C., Geisen, U., Glüer, C.C., Koch, R.: Dataset of 2D and 3D Segmentation of uncertain local collagen fiber orientations in SHG microscopy (2019)
- [35] Schoening, T., Bergmann, M., Ontrup, J., Taylor, J., Dannheim, J., Gutt, J., Purser, A., Nattkemper, T.W.: Semi-automated image analysis for the assessment of megafaunal densities at the Arctic deep-sea observatory HAUSGARTEN. *PLoS ONE* **7**(6), 1–14 (2012). <https://doi.org/10.1371/journal.pone.0038179>
- [36] Shen, Y., Yun, H., Lipton, Z., Kronrod, Y., Anandkumar, A.: Deep active learning for named entity recognition. In: Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 252–256. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). <https://doi.org/10.18653/v1/W17-2630>, <https://aclanthology.org/W17-2630>
- [37] Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)* (2020)
- [38] Srivastava, M., Hashimoto, T., Liang, P.: Robustness to spurious correlations via human annotations (2020)
- [39] Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T.: A deeper look at dataset bias (2015)

A.2. Short papers

- [40] Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528 (2011). <https://doi.org/10.1109/CVPR.2011.5995347>
- [41] Volkmann, N., Brünger, J., Stracke, J., Zelenka, C., Koch, R., Kemper, N., Spindler, B.: SO MUCH TROUBLE IN THE HERD: DETECTION OF FIRST SIGNS OF CANNIBALISM IN TURKEYS. In: Recent advances in animal welfare science VII Virtual UFAW Animal Welfare Conference. p. 82 (2020)
- [42] Xue, Y., Hauskrecht, M.: Efficient learning of classification models from soft-label information by binning and ranking. In: The Thirtieth International Flairs Conference (2017)
- [43] Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-labeling imagenet: From single to multi-labels, from global to localized labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2340–2350 (June 2021)

A. Own previous papers

A.2.3 Beyond hard labels: investigating data label distributions

Beyond Hard Labels: Investigating data label distributions

Vasco Grossmann^{*1} Lars Schmarje^{*1} Reinhard Koch¹

Abstract

High-quality data is a key aspect of modern machine learning. However, labels generated by humans suffer from issues like label noise and class ambiguities. We raise the question of whether hard labels are sufficient to represent the underlying ground truth distribution in the presence of these inherent imprecision. Therefore, we compare the disparity of learning with hard and soft labels quantitatively and qualitatively for a synthetic and a real-world dataset. We show that the application of soft labels leads to improved performance and yields a more regular structure of the internal feature space.

1. Motivation

Modern machine learning relies on high-quality data, but even large and manually cleaned datasets like ImageNet contain errors and uncertainties (Yun et al., 2021; Northcutt et al., 2021). In a model-centric view (Santarossa et al., 2022; Damm et al., 2021), we could try to increase the robustness of models to overcome such issues. However, in this work, we take a data-centric perspective and investigate the possibilities of improving data-quality (Marcu, Antonia; Prugel-Bennett, 2021; Schmarje et al., 2021b). In machine learning, we compare model predictions with ground-truth labels to measure the model performance and inherently also the data-quality. As ground-truth, we often use class labels created by humans and indirectly expect them to be perfect for the evaluation.

This approach has two major shortcomings: errors and uncertainties. Firstly, when humans create labels, errors and mistakes are unavoidable. Even extensive cleaning does not remove all of these issues as recent works showed (Beyer et al., 2020; Northcutt et al., 2021). Thus, we still have to

expect incorrect labels in our ground truth data. We call these errors *noise* (Li et al., 2020) and they lead to partially false evaluation results. Secondly, the human perception of image content can vary. Recent work (Peterson et al., 2019; Wei et al., 2021) showed that human provide different labels e.g., even for classifications of cats and dogs. This uncertainty can arise from different factors like subjective interpretations (Karimi et al., 2020), imperfect image qualities (Peterson et al., 2019) or arbitrary class distinctions (Beyer et al., 2020). We call this issue of data uncertainty *ambiguity*.

In the literature, we find more robust methods (Liu et al., 2020) or improved datasets (Beyer et al., 2020; Northcutt et al., 2021) to resolve the issues of noise and ambiguity. Most approaches share the common assumption that one hard label per image as ground truth is sufficient to capture the image information. In theory, this approach can remove all noise in our labels with sufficient effort, but the literature also shows that ambiguity is present in many real world datasets (Peterson et al., 2019; Wei et al., 2021; Schmarje et al., 2022a; Durden et al., 2016; Schmarje et al., 2022b; 2021a). Thus, we must use a consensus process or a majority vote if we want to collapse the annotations to one hard label. We raise the question whether such a collapse can represent the inherent ambiguity in the images.

We investigate the label distributions of one synthetic and one real-world dataset to answer this question. The focus lies on the comparison between hard and soft labels as input for the model. We evaluate the impact of these two representations quantitatively and qualitatively for estimating the label distribution. We discuss the implications of our findings for future data-quality estimations.

Several comparisons on different label types have been published. Tyrväinen showed that models trained on their own soft-labeled CIFAR10 dataset (Peterson et al., 2019) are more robust against adversarial attacks than models trained on hard-labeled data (Tyrväinen, 2021). Geng et al. introduced a association classification on soft labels to characterize their imprecisions and lead to more robust classification results (Geng et al., 2021). A survey on classifications with different label types is given in (Song et al., 2022).

Our main contributions are that we illustrate and discuss (1) the negative impact on data-quality of ambiguous images

^{*}Equal contribution. ¹Multimedia Information Processing Group, University of Kiel, Germany, cor. Correspondence to: Vasco Grossmann <vg@informatik.uni-kiel.de>, Lars Schmarje <las@informatik.uni-kiel.de>.

A. Own previous papers

Beyond Hard Labels: Investigating data label distributions

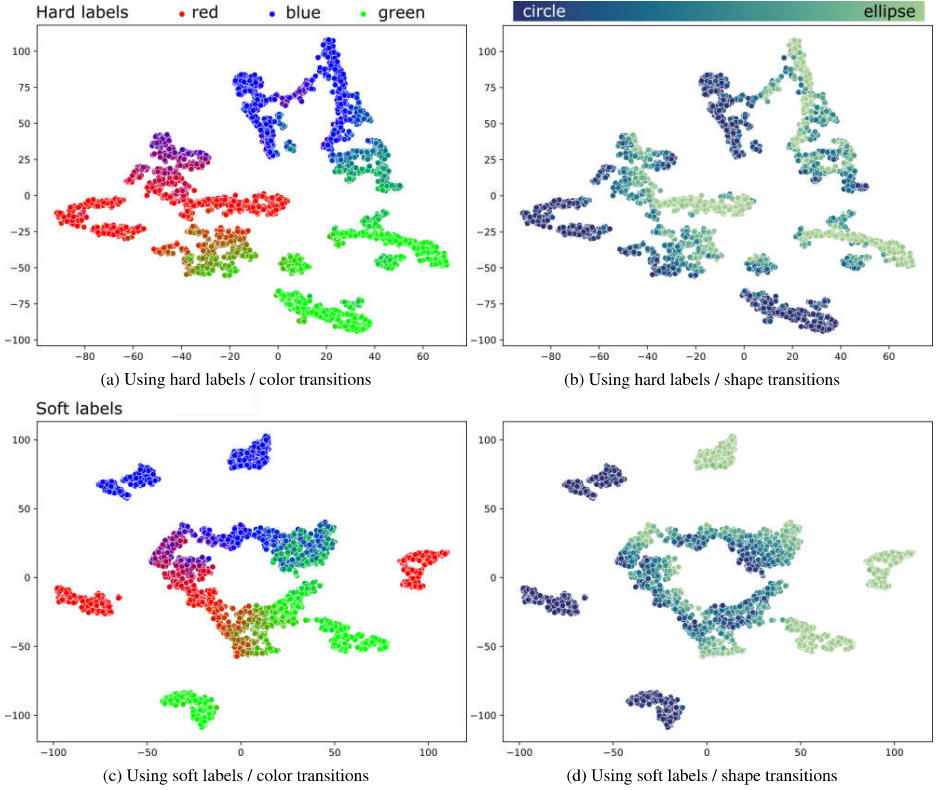


Figure 1. T-SNE plots for the synthetic dataset using hard / soft labels – The left images show the ground truth color interpolation, while the right images show the ground truth shape interpolation.

even without noise and (2) the insufficient representation of hard labels for ambiguous data.

2. Method

We compare the effect of hard and soft labels on one synthetic and one real-world dataset for image classification using deep learning. For a classification problem with k classes and for every image x , we use multiple human annotations $a_i(x) \in \{0, 1\}^k$ to create the hard and soft label. The hard label is the (relative) majority vote across all N annotations ($l_h(x) = \operatorname{argmax}_k \sum_{i=0}^N a_i(x) \in \{0, \dots, k\}$) and soft label is the average across all annotations ($l_s(x) = \frac{1}{N} \sum_{i=0}^N a_i(x) \in [0, 1]^k$).

As a synthetic dataset, we use red, blue and green circles or

ellipses and their color and shape interpolations on a black background. We generated 15,000 images and used 60% for training, 20% for validation and 20% for testing. The generated images are either directly one of the six classes (40%) or an interpolation of the classes (60%). Because of our knowledge of the used interpolation, we can estimate the labels perfectly. This means that we do not have any noise in our data but still have ambiguity. As a real-world dataset, we used the MiceBone dataset from (Schmarje et al., 2022b) but added more annotations per image for a better label distribution estimation. We have about 15 annotations on average per image for 7,240 images total and use the same training, validation and test split proportions as for the synthetic dataset. We have three classes: directed collagen fibers, undirected collagen fibers and not relevant structures. We have a class imbalance of about 70% for the non-relevant

A.2. Short papers

Beyond Hard Labels: Investigating data label distributions

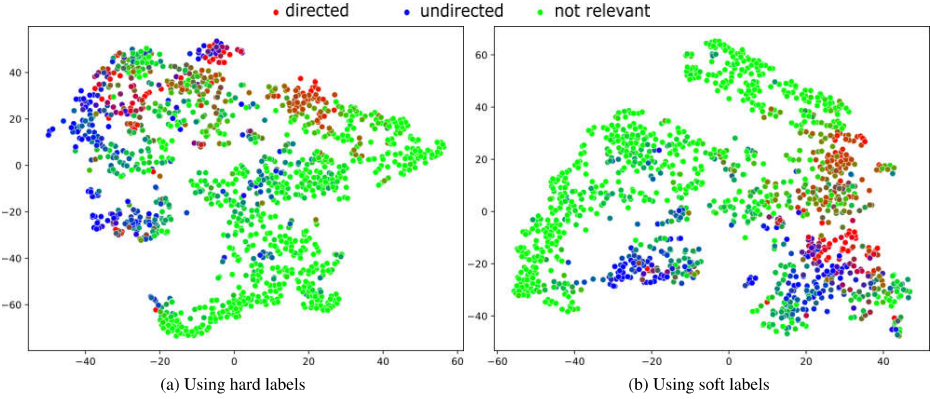


Figure 2. T-SNE plots for the MiceBone dataset using hard / soft labels – The different colors represent the three classes and the interpolated color directly corresponds to the ground truth class distribution.

class and 15% for the directed and undirected fibers. We can estimate an image label only based on the collected annotations and therefore have noise and ambiguity in our dataset.

We determined the hyperparameters like model, batch size and learning rate heuristically across the same predefined parameter grid. For the synthetic dataset, we used a pre-trained DenseNet121 (Huang et al., 2017) with a batch size of 128, a learning rate of 0.1 for 60 epochs and SGD with cosine learning rates (Loshchilov & Hutter, 2016) and weight decay of 0.0005. For the MiceBone dataset, we used a pre-trained ResNet50v2 (He et al., 2015) with a batch size of 128, a learning rate of 0.1 for 60 epochs, SGD with cosine learning rates, warm restarts (Loshchilov & Hutter, 2016) and weight decay of 0.001. All experiments were executed on an RTX 3090 Ti with 24GB VRAM.

We report the macro accuracy (ACC) and Kullback Leibler Divergence (KL) between the model predictions and the soft ground truth label distributions on the test set. The macro accuracy is the per class average of true positives divided by the samples of the class. For the synthetic dataset, this value is equivalent to the accuracy over the complete dataset but is more robust to the class imbalance on the MiceBone dataset. The Kullback Leibler Divergence is an established metric for measuring the difference between two distributions (Murphy, 2012). This metric allows a better insight into the difference between the ground truth and predicted distribution, since the accuracy only looks at the most likely class from the distribution. We report the average and standard deviation across three randomly generated training, validation, and test splits.

Table 1. Quantitative comparison between hard and soft labels – A small arrow after the metrics indicates if lower or higher values are better. The best results per metric and dataset are marked bold.

Dataset	Using Hard Labels		Using Soft Labels	
	$ACC \uparrow$	$KL \downarrow$	$ACC \uparrow$	$KL \downarrow$
Synthetic	0.8247 \pm 0.05002	0.4137 \pm 0.0836	0.9096 \pm 0.0137	0.0394 \pm 0.0009
Mice Bone	0.6217 \pm 0.04658	0.7884 \pm 0.1089	0.6903 \pm 0.0669	0.2280 \pm 0.0227

For the qualitative comparison, we used the internal global average pooling (GAP) layer features of the trained models. We calculated the T-SNE (der Maaten & Hinton, 2008) plots by aligning their feature distances with a perplexity of 30 over 5,000 iterations.

3. Discussion

We give in Table 1 a quantitative comparison for both datasets between using hard and soft labels. Across both datasets and metrics, we see that soft labels result in superior results. We see an average increase of 7-8% for ACC , while KL is reduced by a manifold. Due to the fact KL measures the difference in distribution and hard labels collapse the whole training distribution to one label, this difference can be expected. However, the increase in ACC indicates that the model can also obtain better majority votes if the input distribution contains the image ambiguity. We credit this to the possibility of differentiating between images with a high or low ambiguity with the same majority class. During back propagation with hard labels, both images are enforced to be treated equally, while with soft labels we allow a soft distribution of the error.

Qualitative evaluations using T-SNE plots are given in Fig-

A. Own previous papers

Beyond Hard Labels: Investigating data label distributions

ure 1 and Figure 2. If we compare the color interpolation in the embedding space for the synthetic dataset, we see smooth transitions for the hard and soft labels. However, with the soft labels, we see six distinct clusters surrounding a cyclic interpolation region. This structure exactly represents the data generation process where we generated non-interpolated images and interpolated images. If we look at the shape interpolations, we see that hard labels generate more individual clusters, while soft labels have a more connected interpolation space. Only with soft labels we give the model an exact representation of the expected feature space. With hard labels, the model must create an appropriate feature space based on the majority votes. These results indicate that the model automatically detects ambiguity in hard labels and tries to shape the feature space accordingly. Soft labels help structure the feature space. When we look at the MiceBone results, we can confirm this hypothesis. For hard labels as input, we see a feature space with clusters of low ambiguity and interpolations for images with high ambiguity. However, the separation and transitions are not as clear as in the synthetic dataset. When using soft labels as input, we see a better structure and more well-defined transitions between the ambiguous classes.

The found results are of high importance for the data-quality estimation. On the synthetic dataset, we do not have any noisy and only ambiguity. Thus, we can credit the performance difference completely to the better representation of soft labels in contrast to hard labels. This difference can also be confirmed on a real-world dataset. In the qualitative analysis, we showed that soft labels help structure the internal feature space for a better representation of the expected label distribution. If we look at our motivation, we know that label images suffer from ambiguity and thus it also theoretically impossible to capture this information in just one hard label. We conclude that soft labels are potentially more suitable representations of label distributions than hard labels. We should further investigate soft labels and their potential to create higher quality data. Of special interest is the question, how to obtain such soft labels. In many cases it is not feasible to annotate all images multiple times. Possible solutions for this issue include semi-supervised learning approaches (Tarvainen & Valpola, 2017; Sohn et al., 2020; Chen et al., 2020a) and proposal system (Schmarje et al., 2021a; 2022b).

A limitation of our work is that we only used two datasets and standard supervised learning. We must check whether our conclusions generalize to other datasets and different training protocols. We also neglected the issue of acquiring the annotations for estimating soft labels. For many datasets, it is not feasible to create multiple annotations for thousand or even millions of images. We are confident that combinations of recent developments in the field of semi-supervised / self-supervised learning (Chen et al., 2020b; Sohn et al.,

2020) and soft labels could be used to close the gap of required labeled images to the investigated fully supervised setting.

References

- Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and van den Oord, A. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2002.05709*, (PMLR): 1597–1607, 2020a. ISSN 23318422.
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big Self-Supervised Models are Strong Semi-Supervised Learners. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020b.
- Damm, T., Schmarje, L., Koser, N., Reinhold, S., Yilmaz, E., Krekieh, N., Lui, L.-Y., Cummings, S. R., Koch, R., and Glueer, C.-C. Artificial intelligence-driven hip fracture prediction based on pelvic radiographs exceeds performance of DXA: the “Study of Osteoporotic Fractures” (SOF). *Journal of Bone and Mineral Research*, 37: 193–193, 2021.
- der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.
- Durden, J. M., Bett, B. J., Schoening, T., Morris, K. J., Nattkemper, T. W., and Ruhl, H. A. Comparison of image annotation data generated by multiple investigators for benthic ecology. *Marine Ecology Progress Series*, 552: 61–70, 2016. ISSN 01718630. doi: 10.3354/meps11775.
- Geng, X., Liang, Y., and Jiao, L. ARC-SL: Association rule-based classification with soft labels. *Knowledge-Based Systems*, 225:107116, 2021.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.
- Karimi, D., Nir, G., Fazli, L., Black, P. C., Goldenberg, L., and Salcudean, S. E. Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images—Role of Multiscale Decision Aggregation and Data Augmentation. *IEEE Journal of Biomedical*

A.2. Short papers

Beyond Hard Labels: Investigating data label distributions

- and *Health Informatics*, 24(5):1413–1426, 2020. doi: 10.1109/JBHI.2019.2944643.
- Li, J., Socher, R., and Hoi, S. C. H. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations*, pp. 1–14, 2020.
- Liu, S., Niles-Weed, J., Razavian, N., and Fernandez-Granda, C. Early-Learning Regularization Prevents Memorization of Noisy Labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.
- Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- Marcu, Antonia; Prugel-Bennett, A. On Data-centric Myths. *NeurIPS 2021 Data-centric AI workshop*, 2021.
- Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021.
- Peterson, J., Battleday, R., Griffiths, T., and Russakovsky, O. Human uncertainty makes classification more robust. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:9616–9625, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00971.
- Santarossa, M., Kilic, A., von der Burchard, C., Schmarje, L., Zelenka, C., Reinhold, S., Koch, R., and Roider, J. MedRegNet: unsupervised multimodal retinal-image registration with GANs and ranking loss. In *Medical Imaging 2022: Image Processing*, volume 12032, pp. 321–333. SPIE, 2022.
- Schmarje, L., Brünger, J., Santarossa, M., Schröder, S.-M., Kiko, R., and Koch, R. Fuzzy Overclustering: Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy. *Sensors*, 21(19):6661, 2021a. ISSN 1424-8220. doi: 10.3390/s21196661.
- Schmarje, L., Liao, Y.-H., and Koch, R. A Data-Centric Image Classification Benchmark. *NeurIPS 2021 Data-centric AI workshop*, 2021b.
- Schmarje, L., Grossmann, V., Zelenka, C., Dippel, S., Kiko, R., Oszust, M., Pastell, M., Stracke, J., Valros, A., Volkmann, N., and Koch, R. Is one annotation enough? A data-centric image classification benchmark for noisy and ambiguous label estimation. *36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks*, 2022a.
- Schmarje, L., Santarossa, M., Schröder, S.-M., Zelenka, C., Kiko, R., Stracke, J., Volkmann, N., and Koch, R. A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022b.
- Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.
- Song, H., Kim, M., Park, D., Lee, J.-G., Shin, Y., and Lee, J.-G. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–19, 2022. ISSN 2162-237X. doi: 10.1109/TNNLS.2022.3152527.
- Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *ICLR*, 2017.
- Tyrväinen, S. *Soft labels and supervised image classification*. PhD thesis, 2021.
- Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., and Liu, Y. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. 2021.
- Yun, S., Oh, S. J., Heo, B., Han, D., Choe, J., and Chun, S. Re-Labeling ImageNet: From Single to Multi-Labels, From Global to Localized Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2340–2350, 2021.

A. Own previous papers

A.3 Miscellaneous papers

A.3.1 Artificial intelligence-driven hip fracture prediction based on pelvic radiographs exceeds performance of DXA: the Study of Osteoporotic Fractures (SOF)

A.3. Miscellaneous papers

Artificial intelligence-driven hip fracture prediction based on pelvic radiographs exceeds performance of DXA: the “Study of Osteoporotic Fractures” (SOF)

Timo Damm¹, Lars Schmarje², Niklas Koser¹, Stefan Reinhold², Eren Yilmaz¹, Nicolai Krekieh¹, Li-Yung Lui³, Steven R. Cummings³, Reinhard Koch², Claus-C. Glüer¹

1) Section Biomedical Imaging, Intelligent Imaging Lab (IIL), Department of Radiology and Neuroradiology, University Hospital Schleswig-Holstein (UKSH), Kiel University, Germany

2) Department of Computer Science, Multimedia Information Processing Group (MIP), Kiel University, Germany

3) California Pacific Medical Center, San Francisco, CA, USA

Background/Aim: Prediction of hip fracture (HF) risk using automated analysis of plain radiographs using artificial intelligence (AI) methods. Such methods could expand availability of diagnostic tests, automate, and potentially improve the overall identification of patients at risk. **Methods:** In the Study of Osteoporotic Fracture (SOF) AI methods based on deep convolutional neural networks (DCNNs) were used to analyse digitized pelvic radiographs from visit 1. Data were split into a training and validations dataset and an independent test dataset. In Step 1 preprocessing steps were performed that included placement of DXA-like region of interest (ROIs) to guide the DCNNs. These ROIs were placed automatically based on anatomical landmarks using a key-point-detector CNNs (CenterNet). Additional quality control measures were performed. For risk classification in step 2 we developed 3 different AI models. Two AI models were based on Resnet architecture. Resnet1 included additional image preprocessing steps while ResNet2 required only reduced-supervision during training. The third AI model was based on DenseNet architecture. Dual X-ray Absorptionmetry data was used as the reference standard, based on the femoral neck region (aBMD_FN). Cox proportional hazard models incorporating aBMD_FN or AI based risk estimates without and with age & BMI adjustment were evaluated and compared for difference in Harrell’s C. **Results:** Of a total of 7964 women (age 71.6±5.1 at baseline) preprocessing resulted in a dataset of 6338 women for training and validation of the DCNNs (with 924 incident HF during 14.0±6.3 years of follow-up) and of 1252 women for the test dataset (with 184 incident HF during 15.0±5.7 years of follow-up). aBMD_FN and all of the AI predictors were significantly associated with HF incidence in univariate and in age & BMI adjusted models (all p<0.001, table). Age & BMI adjusted DenseNet based predictors showed significantly better predictive power than age-adjusted aBMD_FN on same subjects (p<0.05). **Future refinement options:** DCNN models including the confounders, enhanced spatial resolution and gray level depth, and complete automation of the preprocessing steps. **Conclusion:** Automated AI driven analysis of pelvic radiographs based on a DenseNet model predicts HF better than DXA based aBMD of the femoral neck. AI of plain radiography demonstrates potential for predictive power better than DXA and high quality HF prediction at sites without access to DXA.

Table: Prediction of hip fracture

Predictor	univariate			age- & BMI-adjusted		
	sHR	C-statistic	AICc	adj sHR	C-statistic	AICc
aBMD FN	2.1(1.8-	.682±.020	2373.9	1.9(1.5-2.3)	.711±.020	2342.9
AI ResNet1	1.9(1.6-	.667±.021	2386.1	1.7(1.4-2.1)	.707±.020	2348.3
AI ResNet2	2.0(1.7-	.688±.019	2370.5	1.8(1.5-2.1)	.720±.018	2339.1
AI DenseNet	2.2(1.8-	.705±.019	2347.2	2.1(1.7-2.4)	.741±.018*	2314.6

*: significantly better than age& BMI adjusted aBMD model on same subjects

A. Own previous papers

A.3.2 Opportunistic hip fracture risk prediction in Men from X-ray: Findings from the Osteoporosis in Men (MrOS) Study

The original paper is not openly available. This is the camera ready version which I'm allowed to share in combination with the following statement: "This version of the contribution has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at https://doi.org/10.1007/978-3-031-16919-9_10. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use <https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms>."

Opportunistic hip fracture risk prediction in Men from X-ray: Findings from the Osteoporosis in Men (MrOS) Study

Lars Schmarje^{1,*}[0000–0002–6945–5957], Stefan Reinhold¹[0000–0003–3117–1569],
Timo Damm²[0000–0002–5595–5205], Eric Orwoll³[0000–0002–8520–7355], Claus-C.
Glüer²[0000–0003–3539–8955], and Reinhard Koch¹[0000–0003–4398–1569]

¹ MIP, Computer Science, Kiel University {las,sre,rk}@informatik.uni-kiel.de

² MOINCC, Kiel University, Germany {timo.damm,glueer}@rad.uni-kiel.de

³ Oregon Health & Science University, United States orwoll@ohsu.edu

* Corresponding author: las@informatik.uni-kiel.de

Abstract. Osteoporosis is a common disease that increases fracture risk. Hip fractures, especially in elderly people, lead to increased morbidity, decreased quality of life and increased mortality. Being a silent disease before fracture, osteoporosis often remains undiagnosed and untreated. Areal bone mineral density (aBMD) assessed by dual-energy X-ray absorptiometry (DXA) is the gold-standard method for osteoporosis diagnosis and hence also for future fracture prediction (prognostic). However, the required special equipment is not broadly available everywhere, in particular not to patients in developing countries. We propose a deep learning classification model (FORM) that can directly predict hip fracture risk from either plain radiographs (X-ray) or 2D projection images of computed tomography (CT) data. Our method is fully automated and therefore well suited for opportunistic screening settings, identifying high risk patients in a broader population without additional screening. FORM was trained and evaluated on X-rays and CT projections from the Osteoporosis in Men (MrOS) study. 3108 X-rays (89 incident hip fractures) or 2150 CTs (80 incident hip fractures) with a 80/20 split (training / validation) were used. We show that FORM can correctly predict the 10-year hip fracture risk with a validation AUC of $81.44\% \pm 3.11\%$ / $81.04\% \pm 5.54\%$ (mean \pm STD) including additional information like age, BMI, fall history and health background across a 5-fold cross validation on the X-ray and CT cohort, respectively. Our approach significantly ($p < 0.01$) outperforms previous methods like Cox Proportional-Hazards Model and FRAX[®] with 70.19 ± 6.58 and 74.72 ± 7.21 respectively on the X-ray cohort. Our model outperform on both cohorts hip aBMD based predictions (validation AUC $82.67\% \pm 0.21\%$ vs. $71.82\% \pm 0.50\%$ and $78.41\% \pm 0.33$ vs. $76.55\% \pm 0.89\%$). We are confident that FORM can contribute on improving osteoporosis diagnosis at an early stage.

Keywords: fracture risk prediction · osteoporosis · opportunistic screening

A. Own previous papers

2 L. Schmarje et al.

1 Introduction

Osteoporosis is a wide-spread systemic disease that leads to deterioration of bone mass and micro structure and subsequently to decreased bone strength inducing an increased fracture risk [23]. According to the United States Preventive Services Task Force, the lifetime risk of an osteoporotic fracture is about 50% in women and about 20% - 25% in men [34,22]. While osteoporosis affects all bones, fractures of the spine and hip are the most frequent. Especially hip fractures lead to increased morbidity, decreased quality of life and increased mortality — 20% of osteoporotic hip fractures lead to death within six month[7]. Being a silent disease before fracture, osteoporosis often remains undiagnosed and consequently untreated. Especially in men, only about 2% are diagnosed before fracture [22].

The gold-standard method for osteoporosis diagnosis is based on areal bone mineral density (aBMD) assessed by dual-energy X-ray absorptiometry (DXA). This modality is in general broadly available to patients in many countries worldwide - with some degree of uneven distribution among industrial nations. In developing countries in African and South America and the Middle East, the availability is poor [17,11]. More elaborate methods like volumetric bone mineral density (vBMD) assessed by quantitative computed tomography (QCT) or finite element modeling (FEM) of bone strength, either based on QCT or DXA, have shown to be superior to standard aBMD [39,20,29,35,2]. However, all these method either require special equipment, protocols or domain experts and the prognosis of osteoporotic fractures is an even more challenging and labor-intensive task.

In this paper, we focus on fracture prognosis in an opportunistic screening scenario: whenever radiographic imaging is available an automated method inspects the image for indicators of possible future fractures. Patients with high fracture risk could be advised to see a specialist to confirm the risk and possibly initiate preventive actions. Due to their outstanding capacity to learn task-relevant image features such methods - in particular convolutional neural networks (CNN) - have outperformed “classical” machine learning algorithms in many image analysis tasks[31,25,24,27,10,26]. We predict the risk of future fractures (prognostics), in contrast to detecting acute osteoporosis or incident fractures (diagnostics).

The goal is to develop a pipeline that can be used for opportunistic screening and hence beneficially leverage additional risk factors. For this purpose we propose a two-stage deep learning based classification method that is able to predict the 10-year fracture risk using only X-ray or CT scans and optionally case history as inputs. We train and evaluate our method on a dataset from the Osteoporotic Fractures in Men (MrOS⁴) study. We restrict our main evaluation to information (e.g. age, weight, height, etc.) that would be collectible in this setting; other information (e.g. aBMD) is only included for comparison.

Our key contributions are: (1) a fully automated system that can be used in an prognostic opportunistic screening scenario, (2) val AUC results of 81.44% and 81.04% on the X-ray and CT cohort respectively. (3) we signifi-

⁴ The Osteoporotic Fractures in Men (MrOS) Study: <https://mrosonline.ucsf.edu>

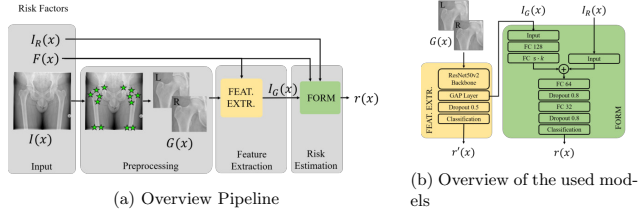


Fig. 1: Illustration⁵of proposed pipeline. (a) Inputs and pipeline stages: preprocessing, feature extraction and risk estimation. (b) Detailed view of models in (a) (yellow, green). Parameter of fully connected (FC) layers: number of hidden neurons. Further information can be found in section 2.

cantly outperform previous methods like Cox Proportional-Hazards Model and FRAX[®] in the opportunistic use case and achieve improved or competitive results for non-opportunistic settings, (4) beneficial integration of clinical risk factors with image based features into a deep learning pipeline.

1.1 Related Work

In the past numerous risk factors for osteoporotic fractures were identified. Among them are increased age, low body mass index (BMI), previous fragility fractures, smoking or alcohol intake. While aBMD alone has shown to be a not sufficiently sensitive predictor for screening applications[18], the combination with other risk factors (RF) is more promising. In [28] Schousboe et al. found that additional RF are better than a model using only aBMD and age for vertebrae fracture prediction. Elaborate statistical shape and density modeling based on volumetric QCT data has proven to be superior to DXA based aBMD models[3] for hip fracture prognosis. Hippisley-Cox et al. proposed the QFractureScores algorithm [13] to predict the 10-year fracture risk. FRAX[®] [18] is a fracture risk assessment tool that uses various RF with or without additional aBMD measurements to predict a 10-year fracture risk. The National Osteoporosis Foundation included FRAX[®] in its guidelines to recommend aBMD measurements or even treatment based on predicted fracture risk[36]. Su et al. [32] used classification and regression trees (CART) on common RF to predict fracture risk on the MoROS data and found a slight improvement over FRAX[®] based predictions. Treece et al. [33] used cortical bone mapping (CBM) to predict osteoporotic fractures in the MoROS study. They found that adding CBM to aBMD can improve fracture prognosis performance. Most of these methods do, however, require special protocols, modalities or in-depth interviews of the patient that might not be applicable to an opportunistic screening setting.

In [21], Pickhardt et al. were able to discriminate manually between patients with osteoporosis and with normal BMD using opportunistic abdomen CTs.

A. Own previous papers

4 L. Schmarje et al.

Recently, several related deep learning based approaches for semi-automated osteoporosis diagnosis have been presented. Ho et al. [14] and Hsieh et al. [15] used a deep learning architecture to predict DXA based aBMD from hip X-ray. Other works like [16] or [38] detect osteoporosis directly from image features of X-ray using end-to-end classification networks. They achieve high classification performance ($AUC > 0.9$) which could even be slightly improved [38] by incorporating clinical risk factors ($AUC > 0.92$). However, this diagnosis task is not comparable to the prognosis task that we target in our work. For prognosis, Hsieh et al. [15] used their predicted aBMD as input to FRAX[®] to predict a 10-year fracture risk. However, since the performance of this combination is limited by the performance of DXA-based aBMD, they were unable to achieve any improvement over baseline FRAX[®] + (DXA-based) aBMD.

Recently, Damm et al. proposed a fully automatic deep learning method to predict hip fractures in women using X-rays from the Study of Osteoporotic Fractures (SOF⁶)[6]. They showed that deep learning based methods are able to improve the prognostic performance of classical aBMD based models while maintaining a high degree of automation. However, they did not investigate additional risk factors and other image modalities as input such as CT.

2 Method

We propose an automatic image processing pipeline for the prediction of **F**uture **O**steoporotic **F**ractures **R**isk in **M**en for **H**ips (FORM). The pipeline consists of a preprocessing, a feature extraction and a risk estimation stage for each patient x . An overview is given in Figure 1a.

2.1 Preprocessing

The proposed method should be able to process 2D X-rays as well as 3D CT scans. To share both architecture and hyperparameters for both input modalities, we compute 2D projections from the 3D CT scans. This way, however, most of the 3D structural information from the CT scans is lost. To fully exploit the 3D information, a native 3D CNN could have been used, but this would have resulted in a much larger memory footprint and thus higher hardware requirements. Therefore, in this work, we have focused at first on confirming the usefulness of CT image data for predicting future fractures. In the following, the CT projections will be referred to simply as CT.

A Hough transform is used to detect the QCT calibration phantom that is present in all scans in order to remove it and the underlying table from the image⁷. Since the phantom is always located beneath the patient, the scans

⁶ The Study of Osteoporotic Fractures (SOF): <https://sofonline.ucsf.edu>

⁷ Example image and key points only for illustrative purpose; image source <https://radiopaedia.org/cases/normal-hip-x-rays>

⁸ In the MrOS study, the phantoms are used to calibrate HU to BMD. In this work no BMD calibration is performed for a more realistic opportunistic screening setting

can easily be cropped to exclude the phantom and the table. The cropped 3D scans are then projected onto the coronal plane and re-scaled by a constant value to achieve pixel value range of $[0, 1]$. We also investigated a re-scaling per patient to investigate the influence of the scanner HU calibration; results on these normalized CTs (CTN) are reported in the supplement.

CT projections and X-ray images $I(x)$ were split into two halves depicting the right and the left hip, respectively; images of the left hip were vertically flipped. A key point detection CNN inspired by [6] were used to detect 12 key points located around the femur. The key point detector was trained jointly on 1797 X-ray images and 208 CT projections (104 CT, 104 CTN) with manually annotated key point positions. The key point CNN classifies the image halves into three classes: *complete* (full proximal femur is visible), *incomplete* (proximal femur not completely visible) and *implant*. A selection of key points is used to crop the image to the proximal femur region (including the trochanter minor and the femoral head). This automatic selection of the region of interest leads to a more suitable input size for the neural networks without any loss of quality. These cropped images $G(x)$ are included in the dataset if the predicted class is *complete* with a confidence above 0.01 (X-ray) or 0.2 (CT).

The risk factors (RF) are additional information about the patient which might improve the hip fracture risk prediction. As we are not interested on the impact of a single risk factor, but rather whether the information is helpful in combination with image data, we grouped the RF for better referencing: *Base*, *Multiple*, *aBMD*, *FRAX*[®] and *TBS*[30]. The details are summarized in the supplementary. The *Base* group contains basic patient information like age and BMI. The *Multiple* group extends *base* and adds additional information from the case history and health background. This information might not be present in clinical routine but could be acquired from every patient via a questionnaire (non-densitometric). The other groups consists of other well-known risk factors, also including densitometry. For densitometry, additional imaging and evaluation is required and thus is not suitable for opportunistic screening. We included these risk factors as a comparison. 52 patients were excluded from the dataset due to missing data for at least one risk factor.

2.2 Feature Extraction

We train a CNN as a Feature Extractor with output $r'(x)$ on the cropped femur images $G(x)$ and extract the predicted Global Average Pooling (GAP) Features of the network. These GAP features $I_G(x) \in \mathbb{R}^{2048}$ are used as input for the next pipeline stage. For the training of the CNN, a ground truth label $F_t(x)$ is needed indicating whether the patient x will fracture by time horizon t (e.g. 10 years) or not. All patients with unknown fracture status, e.g., due to death before time horizon t , were excluded from the dataset. This renders all predicted risks conditional on the patient survival to time horizon t . This is acceptable for an opportunistic screening because we need to screen all patients regardless of whether or not they would survive to t [4] but might introduce a bias. FORM has a ResNet50v2 [12] backbone pretrained on ImageNet[19] with three additional

A. Own previous papers

6 L. Schmarje et al.

layers depicted in Figure 1b. Data was augmented with random flips, zooms and color changes; classes were weighted. Input image dimensions were 96x96 (CT) or 224x224 (X-ray) pixels. Training was performed for 50 epochs, a batch size of 36, learning rate of 10^{-4} , dropout $(0.5)^8$. A cross-entropy loss was used and input samples were weighted based on their class distribution. GAP features $G(x)$ were extracted after early stopping (see subsection 3.1 for details about the metrics).

2.3 Risk Estimation

For the risk estimation $r(x)$, we train a multi layer perceptron (MLP) to predict hip fractures up to the horizon t with the target $F_t(x)$. Its input can be varied: GAP-Features $I_G(x)$, RF $I_R(x)$ or both. Categorical data is one hot encoded and concatenated with normalized continuous data into the input vector $I_R(x)$. Using many RF together with the high-dimensional GAP features might make the model more prone to overfitting because individual datapoints become more distinct [9,1]. This is counteracted by a high dropout rate. To prevent imbalancing between $I_G(x)$ and $I_R(x)$ due to high dimensional differences (e.g. 2048 vs. 4) an MLP with 128 and $s \cdot k$ hidden nodes is used to reduce the dimension of the GAP features before concatenation. Here k is the dimension of $I_R(x)$ and s is a scaling hyperparameter set to 5. Hyperparameters are mostly shared in Feature Extraction and Risk Estimation and the differences are illustrated in Figure 1b. An ablation study to inspect the impact of the hyperparameters is included in the supplementary material and discussed in subsection 3.3.

3 Evaluation

3.1 Datasets and Baseline Methods

For training and evaluation, we used the dataset from the Osteoporosis in Men (MrOS) study. Patients were followed for more than 10 years. We used the first hip fracture that occurred after the baseline visit as our primary outcome. A detailed overview of the datasets statistics for the X-ray and the CT cohort in comparison to the complete study population can be found in Table 1. The number of included patients n is decreased by about one third of the overall number of patients with available image data due to censoring and excluded image halves (e.g. due to implants). During a 5 / 10 year follow-up 1.45% and 3% of the men suffered a hip fracture, respectively. This low number of cases limits the generalizability but it is possible to identify trends which repeat across different modalities, horizons and settings. Therefore, we use the same training validation split based on the patient IDs across all experiments for the respective cohorts. We used area under the receiver-operator curve (AUC) as the main metric. A

⁸ Implemented in Tensorflow 2.4, source code will be release on publication, experiments executed on Nvidia RTX 3090, inference < 1 second per image

⁸ <https://mrosonline.ucsf.edu>, Update august 2021

Table 1: Statistics for three different cohorts (all, CT or X-Ray data available). For each cohort we give the statistics for the complete cohort and with respect to the hip fracture horizon with (w) and without (w/o) fracture (fx). n represents the number of samples. Age is given as relative percentage for a specific range from n . Training / Validation refer to the number of used images in the respective sets in the first fold. - means that no data was used for training or validation. All other risk factors are given as mean (STD in brackets).

		Cohort All			Cohort CT						Cohort X-Ray		
		$t = 10$ year			$t = 10$ year			$t = 5$ year			$t = 10$ year		
		all	w/o fx	w fx	all	w/o fx	w fx	w/o fx	w fx		all	w/o fx	w fx
n		5994	4004	185	3165	2150	80	2198	32	3895	3108	89	
Age [%]	64-69	29.51	36.69	10.27	30.74	37.86	16.25	37.44	12.50	32.99	37.32	15.73	
	70-74	28.50	31.87	16.22	28.44	32.00	15.00	31.71	9.38	30.40	31.66	19.10	
	75-79	24.11	21.98	33.51	23.70	21.44	37.50	21.84	34.38	23.90	22.30	39.33	
	80+	17.88	9.47	40.00	17.12	8.70	31.25	9.01	43.75	12.71	8.72	25.84	
Height [m]		1.74	1.75	1.73	1.74	1.75	1.74	1.75	1.72	1.74	1.74	1.74	
		(0.07)	(0.07)	(0.06)	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)	(0.07)	(0.07)	(0.06)	
BMI [$\frac{kg}{m^2}$]		25.90	26.00	25.11	25.80	25.89	24.77	25.86	24.68	25.86	25.90	24.86	
		(3.65)	(3.54)	(3.63)	(3.60)	(3.48)	(3.38)	(3.49)	(3.02)	(3.57)	(3.48)	(3.65)	
Femoral aBMD [$\frac{g}{cm^3}$]		0.78	0.79	0.66	0.78	0.79	0.65	0.79	0.62	0.79	0.79	0.68	
		(0.13)	(0.13)	(0.11)	(0.13)	(0.13)	(0.08)	(0.13)	(0.08)	(0.13)	(0.12)	(0.10)	
Spine aBMD [$\frac{g}{cm^3}$]		1.07	1.07	1.01	1.07	1.06	1.00	1.06	0.98	1.07	1.07	1.00	
		(0.19)	(0.18)	(0.19)	(0.19)	(0.18)	(0.16)	(0.18)	(0.19)	(0.18)	(0.18)	(0.20)	
Avg. TBS		1.23	1.23	1.19	1.23	1.24	1.20	1.24	1.17	1.24	1.24	1.19	
		(0.13)	(0.13)	(0.13)	(0.13)	(0.12)	(0.12)	(0.12)	(0.13)	(0.12)	(0.12)	(0.12)	
FRAX* [%]		4.14	3.14	7.04	4.01	3.04	6.29	3.07	8.53	3.54	3.07	5.76	
		(4.39)	(3.07)	(5.67)	(4.28)	(2.93)	(5.85)	(2.94)	(7.97)	(3.56)	(2.89)	(4.87)	
FRAX* (w. aBMD) [%]		4.45	3.38	10.86	4.33	3.35	10.47	3.45	14.38	3.85	3.34	9.41	
		(5.54)	(4.04)	(9.09)	(5.41)	(3.95)	(8.92)	(4.12)	(9.41)	(4.74)	(3.95)	(8.56)	
Training		-	-	-	-	3403	128	3478	53	-	4353	107	
Validation		-	-	-	-	790	27	810	7	-	1086	35	

5-fold cross-validation was used to ascertain the validity of the comparison with the established baselines; across folds validation means and standard deviations (STD) are reported. To ensure reproducibility training was repeated 10 times for deep learning models. In the ablation studies, we analyzed only one fold across 10 repetitions and report means with their standard errors (SE). A two-sided Welch-Test[37] was used to compare the calculated means.

As baselines a Cox Proportional-Hazards Model (Cox) [5] and FRAX* was used. For the Cox model the same input as to our model FORM was used. However, the low variance of the high dimensional GAP features lead to a numerical degeneration of the Cox model. This was circumvented by performing a dimensionality reduction using Principal Component Analysis (PCA) [8] of the GAP feature space. The Cox model was fitted on the training data and used for prediction on the validation data. Best performing number of PCA components are reported based on the validation set.

3.2 Results

The proposed method (FORM), a Cox Proportional-Hazards Model (Cox) and FRAX* are compared using a five-fold cross-validation analysis in Table 2. It can be seen, that the proposed method outperforms Cox and FRAX* on both

A. Own previous papers

8 L. Schmarje et al.

Table 2: Cross-validation results (mean val. AUC \pm STD) – columns: different methods / inputs; rows: cohort. Bold: results within a one percent margin of the best for each cohort.

Cohort	FORM			Cox			FRAX [*]
	GAP	GAP + Base	GAP + Multiple	GAP	GAP + Base	GAP + Multiple	
X-Ray	81.57 \pm 3.13	81.09 \pm 3.18	81.44 \pm 3.11	61.14 \pm 16.80	70.26 \pm 5.71	70.19 \pm 6.58	74.72 \pm 7.21
CT	77.53 \pm 5.81	80.66 \pm 3.75	81.04 \pm 5.54	67.56 \pm 23.97	73.69 \pm 9.22	75.35 \pm 9.11	74.74 \pm 5.70

Table 3: Comparison fracture risk prediction – columns: different inputs which were used to train FORM; rows: different cohorts. All scores are given as mean val AUC \pm SE. Significant differences between the first and the other columns are marked italic ($p < 0.05$) or bold ($p < 0.01$). † input not used for FORM

Cohort	GAP + Multiple	Base	Multiple	GAP	GAP + Base	FRAX [*] †
X-Ray	78.41 \pm 0.33	66.38 \pm 1.76	69.67 \pm 0.99	<i>77.24 \pm 0.30</i>	77.81 \pm 0.38	77.43
CT	82.67 \pm 0.21	60.89 \pm 0.73	67.03 \pm 0.93	82.58 \pm 0.21	82.48 \pm 0.24	75.94

(a) Non-densitometric Settings

Cohort	GAP + Multiple	aBMD + Base	FRAX [*] + aBMD + Base	TBS + Base	FRAX [*] + aBMD†
X-Ray	78.41 \pm 0.33	76.55 \pm 0.89	81.50 \pm 0.83	72.66 \pm 1.34	80.92
CT	82.67 \pm 0.21	71.82 \pm 0.50	81.08 \pm 0.34	71.56 \pm 0.39	79.19

(b) Densitometric Settings

cohorts by around 6%. In general, using more risk factors in the GAP feature input improves the prediction; this benefit is slightly higher on the CT cohort. Especially for Cox, a high variance without risk factors can be observed which can be credited to one or two folds with significant lower performance (e.g. around 35% on one fold for the X-ray Cohort). On one fold the power to predict hip fractures were analyzed further by adding comparisons without GAP features and evaluations including densitometric inputs variables in Table 3. Across both cohorts, a significant improvement of around 20% to only using the risk factors group Base or Multiple can be seen. Moreover, a significant improvement of up to one percent is achieved when adding information about risk factors on X-ray and the vanilla FRAX^{*} is worse for both cohorts. Overall the image-based result are all similar and within a range of around three percent of 80%. For the densitometric settings, only the usage of Base risk factors is reported, because further information did not improve the results. In the comparison of the best non-densitometric model, with densitometric settings FORM still outperform most risk factors or the vanilla FRAX^{*} + aBMD predictor. Only X-ray based imaging is up to three percent worse than FRAX^{*} based predictions. We see an improvement of using the FRAX^{*} + Base as input to FORM in comparison to the vanilla FRAX^{*} predictor. Further results and ablations are in the supplementary.

3.3 Discussion

We conclude three major results: (1) FORM outperforms Cox and FRAX[®] on two image cohorts and performs similarly or better even if we include densitometric inputs as comparison, (2) only image information can be used for fracture risk prediction but additional risk factors can help the risk estimation and (3) FORM can leverage the combined information of image information and risk factors better than Cox. Across all experiments FORM outperforms the other non-densitometric models. Only the densitometric FRAX[®] predictor (including aBMD) performs better on the X-ray cohort than FORM. However, our models do not require additional imaging with DXA. Future research could highlight important image regions for the risk estimation or the importance of additional risk factors which could improve the interpretability and therefore the acceptance of the system in clinical routine. For Cox and FORM in Table 2, it can be seen that a fracture risk estimation only based on image information is possible and even outperforms predictions only based on risk factors in Table 3. Using additional risk factors as input can improve the results significantly by up to four percent. This shows that risk factors are a valid source for additional information but also that a majority of the information is already encoded in a patient's X-ray or CT. The Cox model performs in the best case similar or worse than FRAX[®] but is outperformed by FORM. While FRAX[®] might use other input variables, the Cox model is trained with the same inputs as FORM. We conclude that our model can learn from the high dimensional data better than Cox due to two reasons: The Cox model required preprocessed inputs via PCA to prevent degeneration. The overfitting prevents adding more than one or two PCA components as input. In subsection 2.2, we explained that patients were excluded due to early death. The censored patients cannot be directly evaluated, but their subgroup which survived for at least the first 5 years without a fracture. The number of false positive predicted patients across 20 repetitions are $3.63\% \pm 0.51\%$ SE and $5.01\% \pm 0.80\%$ SE for the validation and censored subset, respectively. We conclude that our model is performing at least plausible on this subset.

3.4 Limitations and Future Work

This study is based solely on the MrOS dataset, which consists only of men and contains an expected low number of incident (future) fractures. The identified trends are supported across different cohorts and settings but have to be confirmed on other studies. This study can only analyze the benefits of image data for opportunistic screening in a proof-of-concept fashion, since the strict imaging protocols were imposed for the study. A long term study in clinical routine is required to evaluate the practicability and questions about sensitivity / specificity calibration.

3.5 Conclusion

We have shown that X-ray and CT data can be automatically analyzed and processed by our method FORM for opportunistic hip fracture prognosis. We

A. Own previous papers

10 L. Schmarje et al.

achieved a mean validation AUC of greater than 80% for 10-year hip fracture risk in a five-fold cross-validation in both cohorts based on radiographic and CT data. This is significantly better than previous methods like Cox or FRAX[®] on the same or comparable input. Even in most cases, with additional densitometric RF, our method is significantly better. Overall, we are confident that our method FORM and image input in general are promising candidates for improving the identification of men at high risk of future osteoporotic hip fractures.

Acknowledgements We acknowledge funding of Lars Schmarje and Stefan Reinhold by the ARTEMIS project (grant no. 01EC1908E) funded by the Federal Ministry of Education and Research (BMBF), Germany.

References

1. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Inc. (1996)
2. Black, D.M., Bouxsein, M.L., Marshall, L.M., Cummings, S.R., Lang, T.F., Cauley, J.A., Ensrud, K.E., Nielson, C.M., Orwoll, E.S.: Proximal femoral structure and the prediction of hip fracture in men: a large prospective study using QCT. *Journal of Bone and Mineral Research* **23**(8), 1326–1333 (2008)
3. Bredbenner, T.L., Mason, R.L., Havill, L.M., Orwoll, E.S., Nicoletta, D.P., in Men (MrOS) Study, O.F.: Fracture risk predictions based on statistical shape and density modeling of the proximal femur. *Journal of bone and mineral research* **29**(9), 2090–2100 (2014)
4. Camacho, P.M., Petak, S.M., Binkley, N., Diab, D.L., Eldeiry, L.S., Farooki, A., Harris, S.T., Hurley, D.L., Kelly, J., Lewiecki, E.M., Pessah-Pollack, R., McClung, M., Wimalawansa, S.J., Watts, N.B.: American Association of Clinical Endocrinologists/American College of Endocrinology Clinical Practice Guidelines for the Diagnosis and Treatment of Postmenopausal Osteoporosis—2020 Update. *Endocrine Practice* **26**, 1–46 (2020)
5. Cox, D.R.: Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* (2), 187–220
6. Damm, T., Schmarje, L., Koser, N., Reinhold, S., Yilmaz, E., Krekieh, N., Lui, L.Y., Cummings, S.R., Koch, R., Glueck, C.C.: Artificial intelligence-driven hip fracture prediction based on pelvic radiographs exceeds performance of DXA: the “Study of Osteoporotic Fractures” (SOF). *Journal of Bone and Mineral Research* **37**, 193–193 (2021)
7. Ebeling, P.R.: Osteoporosis in men: Why change needs to happen. *World Osteoporosis Day Thematic Report International Osteoporosis Foundation*, Nyon (2014)
8. F.R.S., K.P.: LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572 (1901)
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
10. Grossmann, V., Schmarje, L., Koch, R.: Beyond Hard Labels: Investigating data label distributions. *arXiv preprint arXiv:2207.06224* (2022)
11. Hamidi, Z.: What’s BMD and What We Do in a BMD Centre? pp. 225–246 (2012)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 770–778 (2015)

A.3. Miscellaneous papers

Opportunistic hip fracture risks prediction 11

13. Hippisley-Cox, J., Coupland, C.: Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. *Bmj* **339** (2009)
14. Ho, C.S., Chen, Y.P., Fan, T.Y., Kuo, C.F., Yen, T.Y., Liu, Y.C., Pei, Y.C.: Application of deep learning neural network in predicting bone mineral density from plain X-ray radiography. *Archives of Osteoporosis* **16**(1), 1–12 (2021)
15. Hsieh, C.I., Zheng, K., Lin, C., Mei, L., Lu, L., Li, W., Chen, F.P., Wang, Y., Zhou, X., Wang, F., Others: Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. *Nature communications* **12**(1), 1–9 (2021)
16. Jang, R., Choi, J.H., Kim, N., Chang, J.S., Yoon, P.W., Kim, C.H.: Prediction of osteoporosis from simple hip radiography using deep learning algorithm. *Scientific reports* **11**(1), 1–9 (2021)
17. Johnell, O., Kanis, J.A.: An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporosis international* **17**(12), 1726–1733 (2006)
18. Kanis, J.A., Johnell, O., Odén, A., Johansson, H., McCloskey, E.: FRAX™ and the assessment of fracture probability in men and women from the UK. *Osteoporosis international* **19**(4), 385–397 (2008)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. vol. 60, pp. 1097–1105. Association for Computing Machinery (2012)
20. Langsetmo, L., Peters, K.W., Burghardt, A.J., Ensrud, K.E., Fink, H.A., Cawthon, P.M., Cauley, J.A., Schousboe, J.T., Barrett-Connor, E., Orwoll, E.S., Others: Volumetric bone mineral density and failure load of distal limbs predict incident clinical fracture independent of FRAX and clinical risk factors among older men. *Journal of Bone and Mineral Research* **33**(7), 1302–1311 (2018)
21. Pickhardt, P.J., Pooler, B.D., Lauder, T., del Rio, A.M., Bruce, R.J., Binkley, N.: Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Annals of internal medicine* **158**(8), 588–595 (2013)
22. Prasad, D., Nguyen, M.H.: Chronic hepatitis, osteoporosis, and men: under-recognised and underdiagnosed. *The Lancet Diabetes & Endocrinology* **9**(3), 141 (2021)
23. Salari, N., Ghasemi, H., Mohammadi, L., Rabieenia, E., Shohaimi, S., Mohammadi, M., Others: The global prevalence of osteoporosis in the world: a comprehensive systematic review and meta-analysis. *Journal of orthopaedic surgery and research* **16**(1), 1–20 (2021)
24. Santarossa, M., Kilic, A., von der Burchard, C., Schmarje, L., Zelenka, C., Reinhold, S., Koch, R., Roider, J.: MedRegNet: unsupervised multimodal retinal-image registration with GANs and ranking loss. In: *Medical Imaging 2022: Image Processing*. vol. 12032, pp. 321–333. SPIE (2022)
25. Schmarje, L., Brünger, J., Santarossa, M., Schröder, S.M., Kiko, R., Koch, R.: Fuzzy Overclustering: Semi-Supervised Classification of Fuzzy Labels with Overclustering and Inverse Cross-Entropy. *Sensors* **21**(19), 6661 (2021)
26. Schmarje, L., Grossmann, V., Zelenka, C., Dippel, S., Kiko, R., Oszust, M., Pastell, M., Stracke, J., Valros, A., Volkmann, N., Koch, R.: Is one annotation enough? A data-centric image classification benchmark for noisy and ambiguous label estimation. *arXiv preprint arXiv:2207.06214* (2022)

A. Own previous papers

12 L. Schmarje et al.

27. Schmarje, L., Santarossa, M., Schröder, S.M., Zelenka, C., Kiko, R., Stracke, J., Volkmann, N., Koch, R.: A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering. arXiv preprint arXiv:2106.16209 (2022)
28. Schousboe, J.T., Rosen, H.R., Vokes, T.J., Cauley, J.A., Cummings, S.R., Nevitt, M.C., Black, D.M., Orwoll, E.S., Kado, D.M., Ensrud, K.E., Others: Prediction models of prevalent radiographic vertebral fractures among older men. *Journal of Clinical Densitometry* **17**(4), 449–457 (2014)
29. Schousboe, J.T., Vo, T., Taylor, B.C., Cawthon, P.M., Schwartz, A.V., Bauer, D.C., Orwoll, E.S., Lane, N.E., Barrett-Connor, E., Ensrud, K.E., Others: Prediction of incident major osteoporotic and hip fractures by trabecular bone score (TBS) and prevalent radiographic vertebral fracture in older men. *Journal of Bone and Mineral Research* **31**(3), 690–697 (2016)
30. Schousboe, J.T., Vo, T.N., Langsetmo, L., Taylor, B.C., Kats, A.M., Schwartz, A.V., Bauer, D.C., Cauley, J.A., Ensrud, K.E.: Predictors of change of trabecular bone score (TBS) in older men: results from the Osteoporotic Fractures in Men (MrOS) Study. *Osteoporosis International* **29**(1), 49–59 (2018)
31. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems* 33 pre-proceedings (NeurIPS 2020) (2020)
32. Su, Y., Kwok, T.C.Y., Cummings, S.R., Yip, B.H.K., Cawthon, P.M.: Can classification and regression tree analysis help identify clinically meaningful risk groups for hip fracture prediction in older American men (the MrOS cohort study)? *JBMR plus* **3**(10), e10207 (2019)
33. Treece, G.M., Gee, A.H., Tonkin, C., Ewing, S.K., Cawthon, P.M., Black, D.M., Poole, K.E.S., in Men (MrOS) Study, O.F.: Predicting hip fracture type with cortical bone mapping (CBM) in the osteoporotic fractures in men (MrOS) study. *Journal of Bone and Mineral Research* **30**(11), 2067–2077 (2015)
34. US Preventive Services Task Force, .: Screening for osteoporosis: US preventive services task force recommendation statement. *Annals of internal medicine* **154**(5), 356–364 (2011)
35. Wang, X., Sanyal, A., Cawthon, P.M., Palermo, L., Jekir, M., Christensen, J., Ensrud, K.E., Cummings, S.R., Orwoll, E., Black, D.M., Others: Prediction of new clinical vertebral fractures in elderly men using finite element analysis of CT scans. *Journal of Bone and Mineral Research* **27**(4), 808–816 (2012)
36. Watts, N.B.: The fracture risk assessment tool (FRAX®): Applications in clinical practice. *Journal of Women's Health* **20**(4), 525–531 (2011)
37. Welch, B.L.: The generalization of 'student's' problem with several different population variances are involved. *Biometrika* **34**(1-2), 28–35 (1947)
38. Yamamoto, N., Sukegawa, S., Kitamura, A., Goto, R., Noda, T., Nakano, K., Takabatake, K., Kawai, H., Nagatsuka, H., Kawasaki, K., Others: Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates. *Biomolecules* **10**(11), 1534 (2020)
39. Yang, L., Parimi, N., Orwoll, E.S., Black, D.M., Schousboe, J.T., Eastell, R.: Association of incident hip fracture with the estimated femoral strength by finite element analysis of DXA scans in the Osteoporotic Fractures in Men (MrOS) study. *Osteoporosis International* **29**(3), 643–651 (2018)

A.3.3 Dulling while judging? Veränderte Beurteilung von Fotos zu Pickverletzungen bei Puten durch Wiederholungen

A. Own previous papers Dulling while judging? Veränderte Beurteilung von Fotos zu Pickverletzungen bei Puten durch Wiederholungen

Nina Volkmann^{1,2}, Lars Schmarje³, Reinhard Koch³ und Nicole Kemper²

¹Wissenschaft und Innovation für Nachhaltige Geflügelwirtschaft (WING), Stiftung Tierärztliche Hochschule Hannover
²Institut für Tierhygiene, Tierschutz und Nutztierethologie (ITTN), Stiftung Tierärztliche Hochschule Hannover
³Institut für Informatik, Christian-Albrechts-Universität zu Kiel

Fragestellungen

- Neigen Beobachter dazu, bei einer wiederholten Betrachtung von potentiellen Pickverletzungen, ihre Beurteilung zu ändern?
- Tritt ein solcher ("Gewöhnungs-") Effekt bei verschiedenen Beobachtern gleichermaßen auf?
- Bearbeitet im Rahmen einer Forschungsarbeit zu verschiedenen Annotationstypen (Schmarje et al., 2022)



Tiere, Material und Methoden

- Ausschnitte von Bildaufnahmen, erstellt unter Praxisbedingungen in einem Putenstall
- Bildausschnitte zur Bewertung auf Web-Server (Abb. 1)

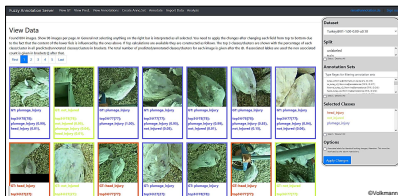
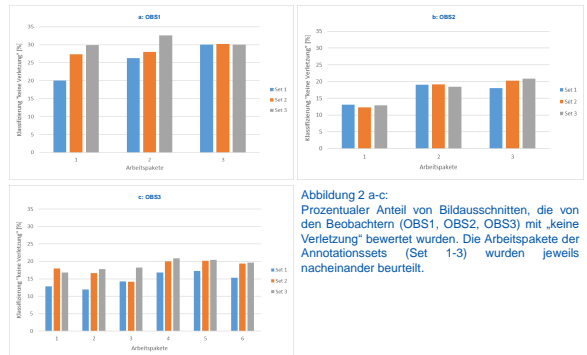


Abbildung 1: Screenshot des Web-Servers, auf dem die Bildausschnitte zur Bewertung zur Verfügung gestellt wurden.

- Bewertung durch drei Beobachter:
 - OBS1 = Person thematisch erfahren bzgl. Pickverletzungen bei Puten
 - OBS2 und OBS3 = Studenten der Informatik
- Beurteilung von Annotationssets mit insgesamt 2.076 Bildausschnitten, aus drei einzelnen Arbeitspaketen (804/636/636 Bilder)
- OBS1 und OBS2 bewerteten die Annotationssets je dreimal, OBS3 sechsmal
- Klassifizierung der Bilder:
 - Kopfverletzung, Verletzung des Gefieders oder keine Verletzung
- Auswertung des prozentualen Anteils der Klassifizierung in zeitlicher Abfolge der Beobachtungen

Ergebnisse

- OBS1 [Ø]:
 - 60% Gefiederverletzung
 - 12% Kopfverletzung
 - 28% keine Verletzung
- OBS2 und OBS3 [Ø]:
 - 70% Gefiederverletzung
 - 13% Kopfverletzung
 - 17% keine Verletzung
- bei wiederholter Bewertung stieg Anteil der Bilder, klassifiziert als „keine Verletzung“ (Abb. 2 a-c)
- Effekt am deutlichsten bei OBS1, bei 3. Wiederholung durchschnittlich > 5% der Bilder zusätzlich als „keine Verletzung“ gewertet



Rückschlüsse

- wiederholte Durchführung von Bewertungen kann zu veränderten Urteilen führen („Abstumpfung“?)
- auch andere Beurteilungen/Klassifizierungen eventuell nicht frei von Effekten wie Gewöhnung, Zeit, Müdigkeit oder Abnutzung

Weiterer Klärungsbedarf:

Können unbedarfte, fachfremde Beobachter objektiver beurteilen?
Ab welcher Anzahl von Bildern findet eine solche „Abstumpfung“ statt?

Bibliography

- [1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Reza-zadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Information Fusion*, 76:243–297, 2020. ISSN 15662535. doi: 10.1016/j.inffus.2021.05.008.
- [2] P. F.E. E E Addison, D. J. Collins, R. Trebilco, S. Howe, N. Bax, P. Hedge, G. Jones, P. Miloslavich, C. Roelfsema, M. Sams, R. D. Stuart-Smith, P. Scanes, P. Von Baumgarten, and A. McQuatters-Gollop. A new wave of marine evidence-based management: Emerging challenges and solutions to transform monitoring, evaluating, and reporting. *ICES Journal of Marine Science*, 75(3):941–952, 2018. ISSN 10959289. doi: 10.1093/icesjms/fsx216.
- [3] Ben Adlam, Jaehoon Lee, Lechao Xiao, Jeffrey Pennington, and Jasper Snoek. Exploring the Uncertainty Properties of Neural Networks’ Implicit Priors in the Infinite-Width Limit. *arXiv preprint arXiv:2010.07355*, 2020.
- [4] Konstantinos Panagiotis Alexandridis, Jiankang Deng, Anh Nguyen, and Shan Luo. Long-tailed Instance Segmentation using Gumbel Optimized Loss. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 353–369. Springer, 2022.
- [5] Gökrem Algan and Ilkay Ulusoy. Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey. *Knowledge-Based Systems*, 2020. ISSN 23318422. doi: 10.1016/j.knosys.2021.106771.
- [6] Dana Angluin and Philip Laird. Learning From Noisy Examples. *Machine Learning*, 2:343–370, 1988. doi: 10.1023/A:1022873112823.

Bibliography

- [7] Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, Andrew Gordon Wilson, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average. In *International Conference on Learning Representations*, pages 1–22, 2019.
- [8] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning Representations by Maximizing Mutual Information Across Views. In *Advances in Neural Information Processing Systems*, pages 15509–15519, 2019.
- [9] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*, 2018.
- [10] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Bala-guer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176—38189, 2022.
- [11] Martin Balunović, Mislav and Vechev. Adversarial training and provable defenses: Bridging the gap. *8th International Conference on Learning Representations (ICLR 2020)(virtual)*, 2020.
- [12] Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to See by Looking at Noise. *Advances in Neural Information Processing Systems*, 4(NeurIPS):2556–2569, 2021. ISSN 10495258.
- [13] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12):1061–1070, 2021. ISSN 25225839. doi: 10.1038/s42256-021-00423-x.

- [14] Ana M. Barragán-Montero, Melissa Thomas, Gilles Defraene, Steven Michiels, Karin Haustermans, John A. Lee, and Edmond Sterpin. Deep learning dose prediction for IMRT of esophageal cancer: The effect of data quality and quantity on model performance. *Physica Medica*, 83:52–63, 2021. ISSN 11201797. doi: 10.1016/j.ejmp.2021.02.026.
- [15] Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21, 2021.
- [16] A. Bäuerle, H. Neumann, and T. Ropinski. Classifier-Guided Visual Correction of Noisy Labels for Image Classification Tasks. *Computer Graphics Forum*, 39(3):195–205, 2020. ISSN 14678659. doi: 10.1111/cgf.13973.
- [17] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. *International Conference on Learning Representations*, 2019.
- [18] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060, 2019.
- [19] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, Aäron van den Oord, Aäron van den Oord, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.
- [20] J Brünger, S Dippel, R Koch, and C Veit. ‘Tailception’: using neural networks for assessing tail lesions on pictures of pig carcasses. *Animal*, 13(5):1030–1036, 2019. ISSN 17517311. doi: 10.1017/S1751731118003038.

Bibliography

- [21] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised Vision Transformers at Scale. *arXiv preprint arXiv:2208.05688*, 2022.
- [22] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.
- [23] Mathilde Caron, Priya Goyal, Ishan Misra, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. ISSN 23318422.
- [24] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. *Proceedings of the IEEE International Conference on Computer Vision*, pages 9630–9640, 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.00951.
- [25] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep Adaptive Image Clustering. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5880–5888, 2017.
- [26] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. *Conference on Human Factors in Computing Systems - Proceedings*, 2017-May:2334–2346, 2017. doi: 10.1145/3025453.3026044.
- [27] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [28] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-

- Supervised Learners. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.
- [29] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*, 2020.
 - [30] Yixin Chen and James Z Wang. Image categorization by learning and reasoning with regions. *The Journal of Machine Learning Research*, 5:913–939, 2004.
 - [31] Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 1–19. Springer, 2022. doi: 10.1007/978-3-031-20074-8_1.
 - [32] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.
 - [33] Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. A Simple Probabilistic Method for Deep Classification under Input-Dependent Label Noise. *arXiv preprint arXiv:2003.06778*, 2020.
 - [34] Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. Correlated Input-Dependent Label Noise in Large-Scale Image Classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1551–1560, 2021.
 - [35] Katherine M. Collins, Umang Bhatt, and Adrian Weller. Eliciting and Learning with Soft Labels from Every Annotator. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing.*, 10(1), 2022.

Bibliography

- [36] Phil Culverhouse, Robert Williams, Beatriz Reguera, Vincent Herry, and Sonsoles González-Gil. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series*, 247:17–25, 2003. doi: 10.3354/meps247017.
- [37] Timo Damm, Lars Schmarje, Niklas Koser, Stefan Reinhold, Eren Yilmaz, Nicolai Krekielehn, Li-Yung Lui, Steven R Cummings, Reinhard Koch, and Claus-C. Glueer. Artificial intelligence-driven hip fracture prediction based on pelvic radiographs exceeds performance of DXA: the Study of Osteoporotic Fractures (SOF). *Journal of Bone and Mineral Research*, 37:193–193, 2021.
- [38] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. ISSN 2307387X. doi: 10.1162/tacl_a_00449.
- [39] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O’Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Machkinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cian O Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T. Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A. Keane, and Olaf Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.
- [40] Philip de Rijk, Lukas Schneider, Marius Cordts, and Darius M Gavrilă. Structural Knowledge Distillation for Object Detection. *Advances in Neural Information Processing Systems*, 35:3858–3870, 2022.
- [41] Michael Desmond, Evelyn Duesterwald, Kristina Brimijoin, Michelle Brachman, and Qian Pan. Semi-automated data labeling. In *NeurIPS 2020 Competition and Demonstration Track*, pages 156–169. PMLR, 2021.

- [42] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, and Qian Pan. Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. In *26th International Conference on Intelligent User Interfaces*, pages 392–401. Association for Computing Machinery, 2021. doi: 10.1145/3397481.3450698.
- [43] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised Visual Representation Learning by Context Prediction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430. IEEE, 2015. doi: 10.1109/ICCV.2015.167.
- [44] Francisco J Candido dos Reis, Stuart Lynn, H Raza Ali, Diana Eccles, Andrew Hanby, Elena Provenzano, Carlos Caldas, William J Howat, Leigh-Anne McDuffus, Bin Liu, and Others. Crowdsourcing the general public for large scale molecular pathology studies in cancer. *EBioMedicine*, 2(7):681–689, 2015.
- [45] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1734–1747, 2015. ISSN 1939-3539. doi: 10.1109/tpami.2015.2496141.
- [46] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*, 2021.
- [47] John J. Dudley and Per Ola Kristensson. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems*, 8(2):1–37, 2018. ISSN 2160-6455. doi: 10.1145/3185517.
- [48] Jennifer M. Durden, Brian J. Bett, Timm Schoening, Kirsty J. Morris, Tim W. Nattkemper, and Henry A. Ruhl. Comparison of image annotation data generated by multiple investigators for benthic ecology.

Bibliography

- Marine Ecology Progress Series*, 552:61–70, 2016. ISSN 01718630. doi: 10.3354/meps11775.
- [49] Michael W. Dusenberry, Ghassen Jerfel, Yeming Wen, Yi An Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and Scalable Bayesian Neural Nets with Rank-1 Factors. *37th International Conference on Machine Learning, ICML 2020, Part F*16814:2762–2772, 2020.
- [50] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *33rd International Conference on Machine Learning, ICML 2016*, 3:1651–1660, 2016.
- [51] Francis Galton. Vox populi. *Nature*, 75(1949):450–451, 1907. ISSN 00280836. doi: 10.1038/075450a0.
- [52] Zhengqi Gao, Fan-Keng Sun, Mingran Yang, Sucheng Ren, Zikai Xiong, Marc Engeler, Antonio Burazer, Linda Wildling, Luca Daniel, and Duane S. Boning. Learning from Multiple Annotator Noisy Labels via Sample-wise Label Fusion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 407–422. Springer, 2022.
- [53] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A Survey of Uncertainty in Deep Neural Networks. *arXiv preprint arXiv:2107.03342*, 2021.
- [54] Harry K. Genant, Michael Jergas, Lisa Palermo, Michael Nevitt, Ria San Valentin, Dennis Black, and Steven R. Cummings. Comparison of semiquantitative visual and quantitative morphometric assessment of prevalent and incident vertebral fractures in osteoporosis. *Journal of Bone and Mineral Research*, 11(7):984–996, 1996. ISSN 08840431. doi: 10.1002/jbmr.5650110716.

- [55] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:4053–4065, 2019.
- [56] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*, pages 1–16, 2018.
- [57] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. OmniMAE: Single Model Masked Pretraining on Images and Videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10406–10417, 2023.
- [58] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *Proceedings of the International Conference on Neural Information Processing Systems*, pages 2672–2680, 2014.
- [59] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14. ACM, 2021. doi: 10.1145/3411764.3445423.
- [60] Julia Grabinski, Paul Gavrikov, Janis Keuper, and Margret Keuper. Robust Models are less Over-Confident. *Advances in Neural Information Processing Systems*, 35, 2022.
- [61] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020. ISSN 15564967. doi: 10.1002/rob.21918.
- [62] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tal-lec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Ghesh-laghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal

Bibliography

- Valko. Bootstrap your own latent: A new approach to self-supervised Learning. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.
- [63] Vasco Grossmann, Lars Schmarje, and Reinhard Koch. Beyond Hard Labels: Investigating data label distributions. *ICML 2022 Workshop DataPerf: Benchmarking Data for Data-Centric AI*, 2022.
- [64] Sebastian Gruber and Florian Buettner. Better Uncertainty Calibration via Proper Scores for Classification and Beyond. *Advances in Neural Information Processing Systems*, 35, 2022.
- [65] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.
- [66] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5138–5147, 2019.
- [67] Marton Havasi, Jasper Snoek, Dustin Tran, Jonathan Gordon, and José Miguel Hernández-Lobato. Refining the Variational Posterior Through Iterative Optimization. *4th workshop on Bayesian Deep Learning (NeurIPS 2019)*, pages 1–11, 2019.
- [68] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M. Dai, and Dustin Tran. Training independent subnetworks for robust prediction. *International Conference on Learning Representations*, pages 1–13, 2021.
- [69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV 2016*, pages 630–645, 2016.
- [70] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

- [71] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.
- [72] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-Efficient Image Recognition with Contrastive Predictive Coding. *Proceedings of the 37th International Conference on Machine Learning*, PMLR:4182–4192, 2020.
- [73] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, pages 1–24, 2019.
- [74] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.
- [75] Julia Hornauer and Vasileios Belagiannis. Gradient-based Uncertainty for Monocular Depth Estimation. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 613–630. Springer, 2022.
- [76] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning Discrete Representations via Information Maximizing Self-Augmented Training. *Proceedings of the 34th International Conference on Machine Learning—Volume 70*, pages 1558–1567, 2017.
- [77] Shih Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3(1), 2020. ISSN 23986352. doi: 10.1038/s41746-020-00341-z.

Bibliography

- [78] Jon M Jachimowicz, Shannon Duncan, Elke U Weber, and Eric J Johnson. When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, 3(2):159–186, 2019.
- [79] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative Models as a Data Source for Multiview Representation Learning. *arXiv preprint arXiv:2106.05258*, pages 1–22, 2021.
- [80] Xu Ji, João F. Henriques, Andrea Vedaldi, Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019.
- [81] Longlong Jing and Yingli Tian. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [82] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65, 2020. ISSN 23318422.
- [83] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in neural information processing systems* 30, 2017.
- [84] Ian D. Kivlichan, Zi Lin, Jeremiah Liu, and Lucy Vasserman. Measuring and Improving Model-Moderator Collaboration using Uncertainty Estimation. *WOAH 2021 - 5th Workshop on Online Abuse and Harms, Proceedings of the Workshop*, pages 36–53, 2021. doi: 10.18653/v1/2021.woah-1.5.
- [85] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019.

- [86] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active Surrogate Estimators: An Active Learning Approach to Label-Efficient Model Evaluation. *Advances in Neural Information Processing Systems*, 35(NeurIPS):24557–24570, 2022.
- [87] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- [88] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 60, pages 1097–1105. Association for Computing Machinery, 2012. doi: 10.1145/3065386.
- [89] Ujwal Krothapalli and A Lynn Abbott. Adaptive label smoothing. *arXiv preprint arXiv:2009.06432*, 2020.
- [90] Solomon Kullback and Richard A. Leibler. On Information and Sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.
- [91] Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Butoi, Paul Bertin, Jarrod Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: Direct Epistemic Uncertainty Prediction. *arXiv preprint arXiv:2102.08501*, 2021.
- [92] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.
- [93] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 2017 (Nips):6403–6414, 2017. ISSN 10495258.
- [94] Daniel Langenkämper, Robin van Kevelaer, Autun Purser, and Tim W Nattkemper. Gear-Induced Concept Drift in Marine Images and Its Effect on Deep Learning Classification. *Frontiers in Marine Science*, 7, 2020. ISSN 2296-7745. doi: 10.3389/fmars.2020.00506.

Bibliography

- [95] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 2, 2013.
- [96] Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations*, pages 1–14, 2020.
- [97] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective Kernel Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 510–519, 2019.
- [98] Yuan-Hong Liao, Amlan Kar, and Sanja Fidler. Towards Good Practices for Efficiently Annotating Large-Scale Image Classification Datasets. *CVPR*, pages 4350–4359, 2021.
- [99] Bin Liu, Zhirong Wu, Han Hu, and Stephen Lin. Deep Metric Transfer for Label Propagation with Limited Annotated Data. *2019 IEEE/CVF International Conference on Computer Vision Workshop (IC-CVW)*, 2019. doi: 10.1109/iccvw.2019.00167.
- [100] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning Customized Visual Models with Retrieval-Augmented Knowledge. *arXiv preprint arXiv:2301.07094*, 2023.
- [101] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.
- [102] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-Learning Regularization Prevents Memorization of Noisy Labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.
- [103] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International conference on machine learning*, pages 6226–6236. PMLR, 2020.

- [104] Maximilian T. Löffler, Anjany Sekuboyina, Alina Jacob, Anna Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S. Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):1–6, 2020. ISSN 26386100. doi: 10.1148/ryai.2020190138.
- [105] Daniel Lopresti and George Nagy. Optimal data partition for semi-automated labeling. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 286–289. IEEE, 2012.
- [106] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448—6458. PMLR, 2020.
- [107] Tohar Lukov, Na Zhao, Gim Hee Lee, and Ser-Nam Lim. Teaching with Soft Label Smoothing for Mitigating Noisy Labels in Facial Expressions. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 648–665. Springer, 2022. doi: 10.1007/978-3-031-19775-8_38.
- [108] Fan Ma, Deyu Meng, Xuanyi Dong, and Yi Yang. Self-paced Multi-view Co-training. *Journal of Machine Learning Research*, 21(57):1–38, 2020.
- [109] Andrey Malinin and Mark Gales. Predictive Uncertainty Estimation via Prior Networks. In *Advances in neural information processing systems*, volume 31, 2018.
- [110] Zelda Mariet, Rodolphe Jenatton, Florian Wenzel, and Dustin Tran. Distilling Ensembles Improves Uncertainty Estimates. *3rd Symposium on Advances in Approximate Bayesian Inference*, pages 1–10, 2020.
- [111] Mantas Mazeika, Eric Tang, Andy Zou, Steven Basart, Jun Shern Chan, Dawn Song, David Forsyth, Jacob Steinhardt, and Dan Hendrycks. How Would The Viewer Feel? Estimating Wellbeing From Video Scenarios. *Advances in Neural Information Processing Systems*, 35:18571–18585, 2022.

Bibliography

- [112] Mary L McHugh. Interrater reliability: the kappa statistic. *PubMed, Biochemia*(3):276–82, 2012.
- [113] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation Learning via Invariant Causal Mechanisms. *arXiv preprint arXiv:2010.07922*, pages 1–21, 2020.
- [114] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, pages 1–16, 2018.
- [115] Mohammad Motamedi, Nikolay Sakharnykh, and Tim Kaldewey. A Data-Centric Approach for Training Deep Neural Networks with Less Data. *NeurIPS 2021 Data-centric AI workshop*, 2021.
- [116] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When Does Label Smoothing Help? *Advances in neural information processing systems*, 32, 2019.
- [117] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [118] Stephen Mussmann and Percy Liang. On the relationship between data efficiency and error for uncertainty sampling. *35th International Conference on Machine Learning, ICML 2018*, 8:5910–5931, 2018.
- [119] Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D’Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating Prediction-Time Batch Normalization for Robustness under Covariate Shift. *arXiv preprint arXiv:2006.10963*, 2020.
- [120] Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W Dusenberry, Sebastian Farquhar, Angelos Filos, Marton Havasi, Rodolphe Jenatton, Ghassen Jerfel, Jeremiah Liu, Zeld Mariet, Jeremy Nixon, Shreyas Padhy, Jie Ren, Tim G. J. Rudner, Faris Sbah, Yeming Wen, Kevin Murphy, D. Sculley, Balaji Lakshminarayanan, Jasper Snoek, Yarin Gal, and Dustin Tran. Uncertainty Baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*, 2021.

- [121] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.
- [122] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. SELF: Learning to Filter Noisy Labels with Self-Ensembling. In *International Conference on Learning Representations*, pages 1–16, 2020.
- [123] Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring Calibration in Deep Learning. *CVPR workshops*, 2(7), 2019.
- [124] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.
- [125] Rafal Obuchowicz, Mariusz Oszust, and Adam Piorkowski. Inter-observer variability in quality assessment of magnetic resonance images. *BMC Medical Imaging*, 20(1):109, 2020. ISSN 1471-2342. doi: 10.1186/s12880-020-00505-z.
- [126] E.A. A Ooms, H.M. M Zonderland, M.J.C. J C Eijkemans, M. Kriege, B. Mahdavian Delavary, C.W. W Burger, and A.C. C Ansink. Mammography: Interobserver variability in breast density assessment. *The Breast*, 16(6):568–576, 2007. ISSN 09609776. doi: 10.1016/j.breast.2007.04.007.
- [127] OpenAI. Introducing ChatGPT. *Blog Entry*, 2022.
- [128] OpenAI. GPT-4 Technical Report. *arXiv:2303.08774*, 2023.
- [129] Shreyas Padhy, Zachary Nado, Jie Ren, Jeremiah Liu, Jasper Snoek, and Balaji Lakshminarayanan. Revisiting One-vs-All Classifiers for Predictive Uncertainty and Out-of-Distribution Detection in Neural Networks. *arXiv preprint arXiv:2007.05134*, 2020.
- [130] Dim P Papadopoulos, Ethan Weber, and Antonio Torralba. Scaling up instance annotation via label propagation. In *Proceedings of the*

Bibliography

- IEEE/CVF International Conference on Computer Vision*, pages 15364–15373, 2021.
- [131] Kyung Park and HyunHee Chung. Uncertainty Guided Pseudo-Labeling: Estimating Uncertainty on Ambiguous Data for Escalating Image Recognition Performance. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*, volume 2, pages 541–551. SCITEPRESS - Science and Technology Publications, 2022. doi: 10.5220/0010901600003116.
- [132] Bhavik N. Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, Curtis Langlotz, Edward Lo, Joseph Mammarappallil, A. J. Mariano, Geoffrey Riley, Jayne Seekins, Luyao Shen, Evan Zucker, and Matthew Lungren. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine*, 2(1):1–10, 2019. ISSN 23986352. doi: 10.1038/s41746-019-0189-7.
- [133] Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:9616–9625, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00971.
- [134] Francesco Pinto, Harry Yang, Ser-Nam Lim, Philip H. S. Torr, and Puneet K. Dokania. RegMixup: Mixup as a Regularizer Can Surprisingly Improve Accuracy and Out Distribution Robustness. *Neurips*, 2022.
- [135] Janis Postels, Mattia Segu, Tao Sun, Luca Sieber, Luc Van Gool, Fisher Yu, and Federico Tombari. On the Practicality of Deterministic Epistemic Uncertainty. *arXiv preprint arXiv:2107.00649*, 2021.
- [136] Guo-Jun Qi and Jiebo Luo. Small Data Challenges in Big Data Era: A Survey of Recent Progress on Unsupervised and Semi-Supervised Methods. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 44, pages 2168–2187. IEEE, 2020.

- [137] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [138] Sai Rajeswar, Pau Rodriguez, Soumye Singhal, David Vazquez, and Aaron Courville. Multi-label iterated learning for image classification with label ambiguity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4783–4793, 2022.
- [139] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection. *arXiv preprint arXiv:2106.09022*, 2021.
- [140] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.
- [141] Stephanie L. Rosenthal and Anind K. Dey. Towards maximizing the accuracy of human-labeled sensor data. *International Conference on Intelligent User Interfaces, Proceedings IUI*, pages 259–268, 2010. doi: 10.1145/1719970.1720006.
- [142] Noveen Sachdeva, Mehak Preet Dhaliwal, Carole-Jean Wu, and Julian McAuley. Infinite Recommendation Networks: A Data-Centric Approach. *Advances in Neural Information Processing Systems*, 35: 31292–31305, 2022.
- [143] Alzayat Saleh, Issam H. Laradji, Dmitry A. Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1):1–10, 2020. ISSN 20452322. doi: 10.1038/s41598-020-71639-x.
- [144] Abhishek Singh Sambyal, Narayanan C. Krishnan, and Deepti R. Bathula. Towards Reducing Aleatoric Uncertainty for Medical Imag-

Bibliography

- ing Tasks. In *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2022.
- [145] Monty Santarossa, Lukas Schneider, Claudius Zelenka, Lars Schmarje, Reinhard Koch, and Uwe Franke. Learning Stixel-based Instance Segmentation. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 427–434. IEEE, 2021. doi: 10.1109/IV48863.2021.9575565.
- [146] Monty Santarossa, Ayse Kilic, Claus von der Burchard, Lars Schmarje, Claudius Zelenka, Stefan Reinhold, Reinhard Koch, and Johann Roider. MedRegNet: unsupervised multimodal retinal-image registration with GANs and ranking loss. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 321–333. SPIE, 2022.
- [147] Lars Schmarje and Reinhard Koch. Life is not black and white – Combining Semi-Supervised Learning with fuzzy labels. *LWDA’21: Lernen, Wissen, Daten, Analysen September 01–03, 2021, Munich, Germany*, 2993(CEUR Workshop Proceedings):183–190, 2021.
- [148] Lars Schmarje, Claudius Zelenka, Ulf Geisen, Claus-C. Glüer, and Reinhard Koch. 2D and 3D Segmentation of uncertain local collagen fiber orientations in SHG microscopy. In *DAGM German Conference of Pattern Recognition*, volume 11824 LNCS, pages 374–386. Springer, 2019. doi: 10.1007/978-3-030-33676-9_26.
- [149] Lars Schmarje, Johannes Brünger, Monty Santarossa, Simon-Martin Schröder, Rainer Kiko, and Reinhard Koch. Fuzzy Overclustering: Semi-supervised classification of fuzzy labels with overclustering and inverse cross-entropy. *Sensors*, 21(19):6661, 2021. ISSN 23318422. doi: 10.3390/s21196661.
- [150] Lars Schmarje, Yuan-Hong Liao, and Reinhard Koch. A Data-Centric Image Classification Benchmark. *NeurIPS 2021 Data-centric AI workshop*, 2021.
- [151] Lars Schmarje, Monty Santarossa, Simon-Martin Schroder, and Reinhard Koch. A Survey on Semi-, Self- and Unsupervised Learning for Image Classification. *IEEE Access*, 9:82146–82168, 2021. ISSN 21693536. doi: 10.1109/ACCESS.2021.3084358.

- [152] Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, and Reinhard Koch. Is one annotation enough? A data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35:33215—33232, 2022.
- [153] Lars Schmarje, Stefan Reinhold, Eric Orwoll, Claus-C. Glüer, and Reinhard Koch. Opportunistic hip fracture risk prediction in Men from X-ray: Findings from the Osteoporosis in Men (MrOS) Study. *Predictive Intelligence in Medicine. PRIME 2022. Lecture Notes in Computer Science*, vol 13564, MICCAI 202:103–114, 2022. doi: 10.1007/978-3-031-16919-9_10.
- [154] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, Claudius Zelenka, Rainer Kiko, Jenny Stracke, Nina Volkmann, and Reinhard Koch. A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.
- [155] Lars Schmarje, Vasco Grossmann, Tim Michels, Jakob Nazarenus, Monty Santarossa, Claudius Zelenka, and Reinhard Koch. Label Smarter, Not Harder: CleverLabel for Faster Annotation of Ambiguous Image Classification with Higher Quality. *arXiv preprint arXiv:2305.12811*, 2023.
- [156] Lars Schmarje, Vasco Grossmann, Claudius Zelenka, and Reinhard Koch. Annotating Ambiguous Images: General Annotation Strategy for Image Classification with Real-World Biomedical Validation on Vertebral Fracture Diagnosis. *arXiv preprint arXiv:2306.12189*, 2023.
- [157] Timm Schoening, Autun Purser, Daniel Langenkämper, Inken Suck, James Taylor, Daphne Cuvelier, Lidia Lins, Erik Simon-Lledó, Yann Marcon, Daniel O. B. Jones, Tim Nattkemper, Kevin Köser, Martin Zurowietz, Jens Greinert, and Jose Gomes-Pereira. Megafauna community assessment of polymetallic-nodule fields with cameras: platform and methodology comparison. *Biogeosciences*, 17(12):3115–3133, 2020. doi: 10.5194/bg-17-3115-2020.

Bibliography

- [158] Simon-Martin Schröder, Rainer Kiko, and Reinhard Koch. MorphoCluster: Efficient Annotation of Plankton images by Clustering. *Sensors*, 20, 2020.
- [159] Claudia Schulz, Christian M Meyer, Jan Kiesewetter, Michael Sailer, Elisabeth Bauer, Martin R Fischer, Frank Fischer, and Iryna Gurevych. Analysis of Automatic Annotation Suggestions for Hard Discourse-Level Tasks in Expert Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2761–2772. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1265.
- [160] Philipp Schustek and Rubén Moreno-Bote. Instance-based generalization for human judgments about uncertainty. *PLOS Computational Biology*, 14(6):e1006205, 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006205.
- [161] Anjany Sekuboyina, Malek E. Hussein, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Da-long Cheng, Nikolas Lessmann, Yujin Hu, Tianfu Wang, Dong Yang, Daguang Xu, Felix Ambellan, Tamaz Amiranashvili, Moritz Ehlke, Hans Lamecker, Sebastian Lehnert, Marilia Lirio, Nicolás Pérez de Olague, Heiko Ramm, Manish Sahu, Alexander Tack, Stefan Zachow, Tao Jiang, Xinjun Ma, Christoph Angerman, Xin Wang, Kevin Brown, Alexandre Kirszenberg, Élodie Puybareau, Di Chen, Yiwei Bai, Brandon H. Rapazzo, Timyoas Yeah, Amber Zhang, Shangliang Xu, Feng Hou, Zhiqiang He, Chan Zeng, Zheng Xiangshang, Xu Liming, Tucker J. Netherton, Raymond P. Mumme, Laurence E. Court, Zixun Huang, Chenhang He, Li Wen Wang, Sai Ho Ling, Lê Duy Huynh, Nicolas Boutry, Roman Jakubicek, Jiri Chmelik, Supriti Mulay, Mohanasankar Sivaprakasam, Johannes C. Paetzold, Suprosanna Shit, Ivan Ezhov, Benedikt Wiestler, Ben Glocker, Alexander Valentinitsch, Markus Rempfler, Björn H. Menze, and Jan S. Kirschke. VERSE: A Vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical Image Analysis*, 73, 2021. ISSN 13618423. doi: 10.1016/j.media.2021.102166.

- [162] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the Existence of Simpler Machine Learning Models. *ACM International Conference Proceeding Series*, pages 1827–1858, 2022. doi: 10.1145/3531146.3533232.
- [163] Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014. ISSN 00200255. doi: 10.1016/j.ins.2013.07.030.
- [164] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating Machine Accuracy on ImageNet. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8634–8644. PMLR, 2020.
- [165] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.
- [166] Igor Stępień, Rafał Obuchowicz, Adam Piórkowski, and Mariusz Oszust. Fusion of Deep Convolutional Neural Networks for No-Reference Magnetic Resonance Image Quality Assessment. *Sensors*, 21(4), 2021. ISSN 1424-8220. doi: 10.3390/s21041043.
- [167] Fariborz Taherkhani, Hadi Kazemi, Ali Dabouei, Jeremy Dawson, and Nasser M Nasrabadi. A weakly supervised fine label classifier enhanced by coarse supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6459–6468, 2019.
- [168] Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. Active Learning Helps Pretrained Models Learn the Intended Task. *Advances in Neural Information Processing Systems*, 35, 2022.

Bibliography

- [169] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:10691–10700, 2019.
- [170] Penny Tarling, Mauricio Cantor, Albert Clapés, and Sergio Escalera. Deep learning with self-supervision and uncertainty regularization to count fish in underwater images. *Plos one*, 17(5):1–22, 2021.
- [171] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *ICLR*, 2017.
- [172] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multi-view Coding. *European conference on computer vision*, 2019.
- [173] Alexandru Tifrea, Jacob Clarysse, and Fanny Yang. Uniform versus uncertainty sampling: When being active is less efficient than staying passive. *arXiv preprint arXiv:2212.00772*, 2022.
- [174] Nenad Tomasev, Ioana Bica, Brian McWilliams, Lars Buesing, Razvan Pascanu, Charles Blundell, and Jovana Mitrovic. Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet? *arXiv preprint arXiv:2201.05119*, 2022.
- [175] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy: FixEfficientNet. *Advances in neural information processing systems*, 32, 2019.
- [176] Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards Reliability using Pretrained Large Model Extensions. *arXiv preprint arXiv:2207.07411*, 2, 2022.
- [177] Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from Disagreement:

- A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12752.
- [178] Matias Valdenegro-Toro and Daniel Saromo. A Deeper Look into Aleatoric and Epistemic Uncertainty Disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1508–1516. IEEE, 2022.
- [179] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [180] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision*, pages 268–285, 2020.
- [181] Vijay Vasudevan, Benjamin Caine, Raphael Gontijo-Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? Analyzing the remaining mistakes on ImageNet. *Advances in Neural Information Processing Systems*, 35:6720–6734, 2022.
- [182] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in neural information processing systems*, 30, 2017.
- [183] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, David Lopez-Paz, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation Consistency Training for Semi-Supervised Learning. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019. doi: 10.24963/ijcai.2019/504.
- [184] Nina Volkmann, Johannes Brünger, Jenny Stracke, Claudius Zelenka, Reinhard Koch, Nicole Kemper, and Birgit Spindler. Learn to train: Improving training data for a neural network to detect pecking injuries in turkeys. *Animals* 2021, 11:1–13, 2021. doi: 10.3390/ani11092655.

Bibliography

- [185] Nina Volkmann, Lars Schmarje, Reinhard Koch, and Nicole Kemper. Dulling while judging? Veränderte Beurteilung von Fotos zu Pickverletzungen bei Puten durch Wiederholungen. *Aktuelle Arbeiten zur artgemäßen Tierhaltung 2022: Vorträge anlässlich der 54. Internationalen Arbeitstagung Angewandte Ethologie bei Nutztieren der Deutschen Veterinärmedizinischen Gesellschaft e.V. (DVG) Fachgruppe "Ethologie und Tierhaltung" am 24. und*, pages 282–284, 2022.
- [186] Nina Volkmann, Claudius Zelenka, Archana Malavalli Devaraju, Johannes Brünger, Jenny Stracke, Birgit Spindler, Nicole Kemper, and Reinhard Koch. Keypoint Detection for Injury Identification during Turkey Husbandry Using Neural Networks. *Sensors*, 22(14): 5188, 2022. ISSN 1424-8220. doi: 10.3390/s22145188.
- [187] Johannes von Oswald, Seijin Kobayashi, Alexander Meulemans, Christian Henning, Benjamin F. Grewe, and João Sacramento. Neural networks with late-phase weights. *arXiv preprint arXiv:2007.12927*, 2020.
- [188] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to Unseen Domains: A Survey on Domain Generalization. In *IEEE Transactions on Knowledge and Data Engineering*. IEEE, 2022.
- [189] Mei Wang and Weihong Deng. Deep Visual Domain Adaptation: A Survey. *Neurocomputing*, 312:135–153, 2018. doi: 10.1016/j.neucom.2018.05.083.
- [190] Xiao Wang, Daisuke Kihara, Jiebo Luo, and Guo-Jun Qi. EnAET: Self-Trained Ensemble AutoEncoding Transformations for Semi-Supervised Learning. *arXiv preprint arXiv:1911.09265*, 2019.
- [191] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning. *arXiv preprint arXiv:2205.07246*, 2022.

- [192] Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. To Smooth or Not? When Label Smoothing Meets Noisy Labels. *Learning*, 1.1, 2021.
- [193] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. *ICLR*, pages 1–23, 2021.
- [194] Jiaheng Wei, Zhaowei Zhu, Tianyi Luo, Ehsan Amid, Abhishek Kumar, and Yang Liu. To Aggregate or Not? Learning with Separate Noisy Labels. *arXiv preprint arXiv:2206.07181*, 2022.
- [195] Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An Alternative Approach to Efficient Ensemble and Lifelong Learning. *arXiv preprint arXiv:2002.06715*, 2020.
- [196] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. *Advances in Neural Information Processing Systems*, 2020-Decem (NeurIPS), 2020. ISSN 10495258.
- [197] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised Data Augmentation for Consistency Training. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.
- [198] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, Quoc V. Le, Minh-Thang Luong, Quoc V. Le, Eduard Hovy, and Quoc V. Le. Self-Training With Noisy Student Improves ImageNet Classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695. IEEE, 2020. doi: 10.1109/CVPR42600.2020.01070.
- [199] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the Whole Rashomon Set of Sparse Decision Trees. *Advances in Neural Information Processing Systems*, 35: 14071–14084, 2022.

Bibliography

- [200] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Kwok. Searching to exploit memorization effect in learning from corrupted labels. *arXiv preprint arXiv:1911.02377*, 2019.
- [201] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR, 2019.
- [202] Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. *Proceedings of the British Machine Vision Conference 2016*, pages 87.1–87.12, 2016. doi: 10.5244/c.30.87.
- [203] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4L: Self-Supervised Semi-Supervised Learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019.
- [204] Yunfeng Zhang, Q. Vera Liao, and Rachel K.E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020. doi: 10.1145/3351095.3372852.
- [205] Chen Zhou, Mohit Prabhushankar, and Ghassan AlRegib. On the Ramifications of Human Label Uncertainty. *NeurIPS 2022 Workshop on Human in the Loop Learning*, 2022.
- [206] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5:44–53, 2018.