Journal of Data-centric Machine Learning Research (2024) Submitted 01/24; Revised 03/24; Published 05/24

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*

Editor:

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 im-

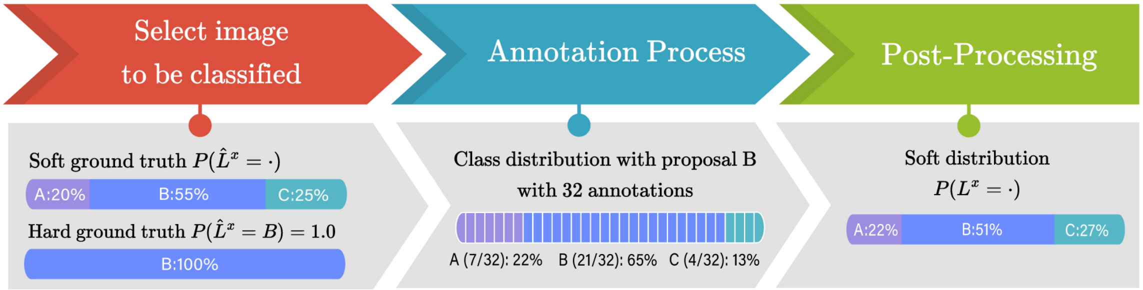


Figure 1: Illustration of the concept of hard and soft labels and how they can be created from annotations – The recommended process has three steps. In the first step, an image x is selected for annotation. The unknown ground-truth distribution ($P(\hat{L}^x = \cdot)$) could either be soft or hard as 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.

practical 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 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 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

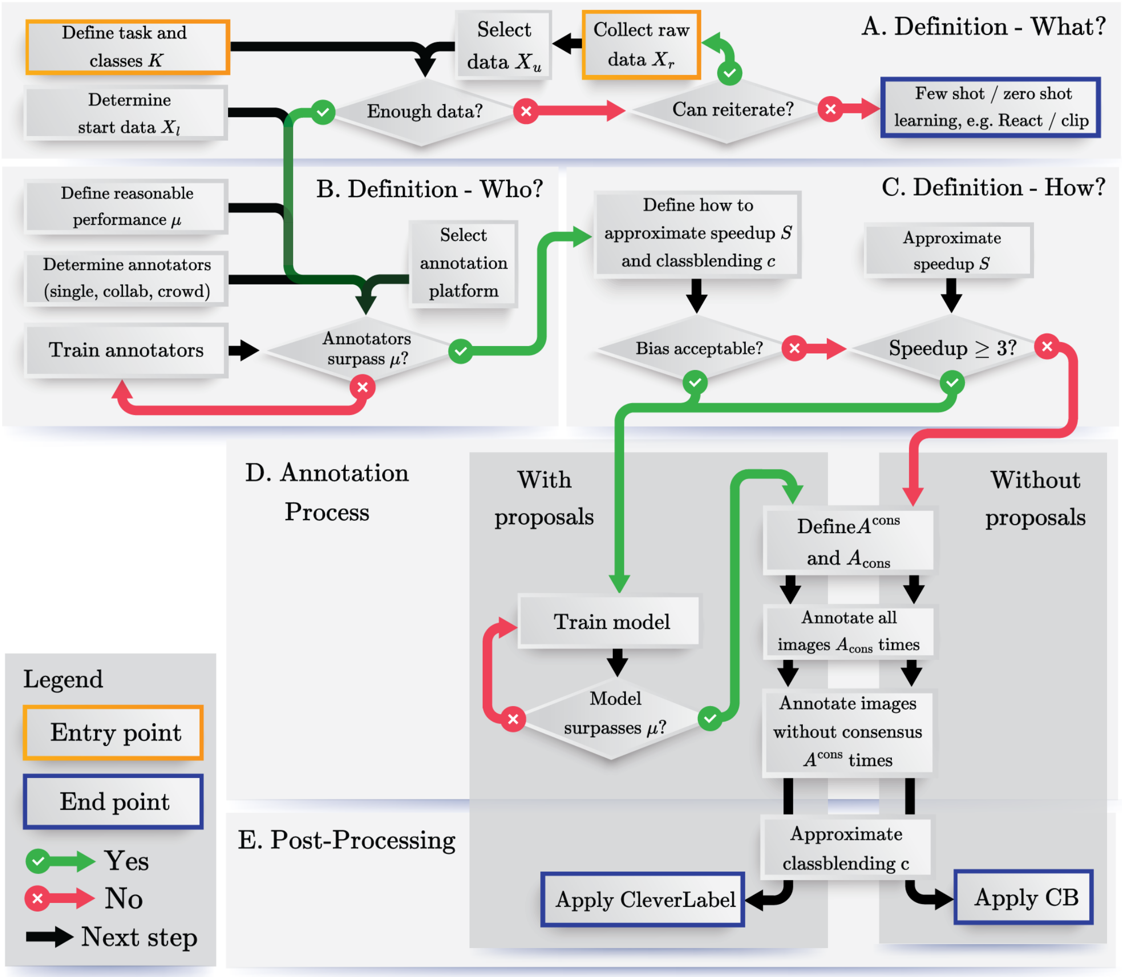

Define task and classes K
Select data X_u
Collect raw data X_r
A. Definition - What?
Determine start data X_l
Enough data?
Can reiterate?
Few shot / zero shot learning, e.g. React / clip
Define reasonable performance μ
B. Definition - Who?
Select annotation platform
Determine annotators (single, collab, crowd)
Train annotators
Annotators surpass μ?
C. Definition - How?
Define how to approximate speedup S and classblending c
Approximate speedup S
Bias acceptable?
Speedup ≥ 3?
D. Annotation Process
With proposals
Without proposals
Define A^{cons} and A_{cons}
Train model
Annotate all images A_{cons} times
Model surpasses μ?
Annotate images without consensus A^{cons} times
Legend
Entry point
End point
Yes
No
Next step
E. Post-Processing
Approximate classblending c
Apply CleverLabel
Apply CB

Figure 2: Flowchart with guidelines on how to annotate ambiguous data, best viewed in color.

selecting a representative subset (X_u) for annotation. The selection of X_u emphasizes 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. 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 Ap-

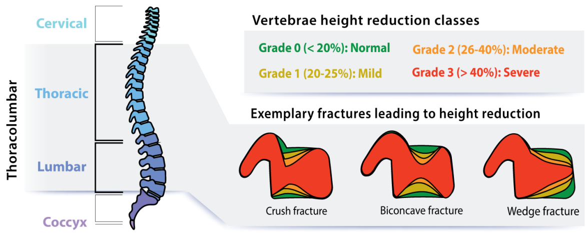


Figure 3: Illustration of a spine and definition of height reduction classes

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)

pendix C. Further details regarding data preprocessing, annotators, annotation platform, approximated variables, and used hyperparameters are provided in Appendix D.

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

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 ± 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%)

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 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 ($P(\hat{L}^x = \cdot)$) 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 research, 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-curated, 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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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. *Marine Ecology Progress Series*, 552:61–70, 2016. ISSN 01718630. doi: 10.3354/meps11775.

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.

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.

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.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV 2016*, pages 630–645, 2016.

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.

Victoria Hemming, Mark A. Burgman, Anca M. Hanea, Marissa F. McBride, and Bonnie 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.

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.

Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. The Principles of Data-Centric AI (DCAI). pages 1–14, 2022.

Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed 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.

Kristian Muri Knausgård, Arne Wiklund, Tonje Knutsen Sørdalen, Kim Tallaksen Halvorsen, Alf Ring Kleiven, Lei Jiao, and Morten Goodwin. Temperate fish detection and classification: a deep learning based approach. *Applied Intelligence*, 2021. ISSN 15737497. doi: 10.1007/s10489-020-02154-9.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

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.

Ujwal Krothapalli and A Lynn Abbott. Adaptive label smoothing. *arXiv preprint arXiv:2009.06432*, 2020.

Solomon Kullback and Richard A. Leibler. On Information and Sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

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.

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.

Yuan-Hong Liao, Amlan Kar, and Sanja Fidler. Towards Good Practices for Efficiently Annotating Large-Scale Image Classification Datasets. *CVPR*, pages 4350–4359, 2021.

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.

Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, Xuanyi Dong, and Chengqi Zhang. Isometric Propagation Network for Generalized Zero-shot Learning. *International Conference on Learning Representations*, 2021a.

Zac Yung-Chun Liu, Shoumik Roychowdhury, Scott Tarlow, Akash Nair, Shweta Badhe, and Tejas Shah. AutoDC: Automated data-centric processing. (NeurIPS):1–6, 2021b.

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.

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.

Ott R. Lyman. *An Introduction to Statistical Methods and Data Analysis.* 1993.

Dim P Papadopoulos, Ethan Weber, and Antonio Torralba. Scaling up instance annotation via label propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15364–15373, 2021.

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.

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.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li 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.

Christoph Sager, Christian Janiesch, and Patrick 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.

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.

Abhishek Singh Sambyal, Narayanan C. Krishnan, and Deepti R. Bathula. Towards Reducing Aleatoric Uncertainty for Medical Imaging Tasks. In *IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2022.

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 Regocnition*, volume 11824 LNCS, pages 374–386. Springer, 2019.

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, 2021a. ISSN 23318422. doi: 10.3390/s21196661.

Lars Schmarje, Yuan-Hong Liao, and Reinhard Koch. A Data-Centric Image Classification Benchmark. *NeurIPS 2021 Data-centric AI workshop*, 2021b.

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, 2022a.

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)*, 2022b.

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.

Simon-Martin Schröder, Rainer Kiko, and Reinhard Koch. MorphoCluster: Efficient Annotation of Plankton images by Clustering. *Sensors*, 20, 2020.

Anjany Sekuboyina, Malek E. Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Dalong 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 Olaguer, 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.

Viktoriia Sharmanska, Daniel Hernandez-Lobato, Jose Miguel Hernandez-Lobato, and Novi Quadrianto. Ambiguity Helps: Classification with Disagreements in Crowdsourced Annotations. *Proceedings of the IEEE Computer Society Conference on Computer Vi-*

sion and Pattern Recognition, 2016-Decem:2194–2202, 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.241.

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.

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.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *ICLR*, 2017.

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.

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.

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.

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.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

Xin Wei, Huaiwei Cong, Zheng Zhang, Junran Peng, Guoping Chen, and Jinpeng Li. Faint Features Tell: Automatic Vertebrae Fracture Screening Assisted by Contrastive Learning. 2022.

Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae Gil 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.

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.

Jichang Zhang, Yuanjie Zheng, and Yunfeng 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.

Appendix A. Full Explanation of Flowchart

A.1 Definition - What?

The first crucial step involves defining the classification task. Specifically, this entails identifying the K classes relevant to the given problem and dataset. Additionally, gathering or generating raw, unlabeled data, denoted as X_r, is necessary based on the use case.

The next step involves deciding the number of images to label. We consider use cases with several thousand to tens of thousands of images, a common scenario in real-world applications (Brünger et al., 2019; Schmarje et al., 2021a; Volkmann et al., 2021; Durden et al., 2016; Saleh et al., 2020). For datasets with hundreds of thousands or millions of images, self-supervised techniques like SimCLRv2 or OmniMAE (Chen et al., 2020; Girdhar et al., 2023) may be applicable. However, even these cases require a smaller labeled dataset for performance evaluation, which can be labeled using our strategy. We denote $X_u \subseteq X_r$ as the subset containing samples to be labeled.

Our focus is on common supervised or semi-supervised deep learning approaches (Sohn et al., 2020; Tarvainen and Valpola, 2017; Lee, 2013). Therefore, it's imperative to ensure enough samples per class are available post-annotation. The term "enough samples" varies depending on the use case. Based on previous results (Schmarje et al., 2022a), each class should represent at least 1% of the complete data, with several dozen images per class for robust evaluation. In cases of limited images, additional considerations for few or zero-shot learning approaches are necessary (Wang et al., 2020, 2019; Liu et al., 2021a), although this extends beyond our paper's scope.

For each image x in a small labeled dataset $x \in X_l \subseteq X_u$, at least one hard-encoded ground truth label is necessary. This data is crucial for annotator evaluation (see subsection A.2) and potentially for training a proposal generation network (see subsection A.3). An initial dataset size of about 20% of the total dataset size was beneficial in (Schmarje et al., 2023), though the exact size may vary depending on the specific use-case.

A.2 Definition - Who?

The choice of annotators is a pivotal aspect of the annotation process. Various approaches include using a single user, collaborating with several workers, or employing crowd-sourcing methods (Sager et al., 2021). The annotators' background dictates the extent of training or selection necessary for the task (Liao et al., 2021).

For evaluating annotation quality, two elements are essential alongside the small labeled dataset X_l: a performance threshold μ and an appropriate annotation platform. The threshold μ should align with the evaluation metrics used, such as macro accuracy or F1-score, which often show high correlation (Laine and Aila, 2017; He et al., 2022; Berthelot et al., 2019; Schmarje et al., 2022a; Knausgård et al., 2021; Li et al., 2020). (Schmarje et al., 2022b) suggests that a threshold between 60% and 80% is typically adequate if annotations are aggregated. However, the label quality in X_l influences the maximum achievable scores; for instance, scores higher than 80% might be unattainable due to intra- or inter-observer variability (Schmarje et al., 2019).

Based on the task definition (see subsection A.1), a browsing environment is recommended for tasks with a few classes having many occurrences, allowing for side-by-side image comparison (Sager et al., 2021). Alternative platforms like (Schröder et al., 2020) are also viable. Although these platforms may require setup overhead, they can significantly expedite the annotation process. Annotators should be trained to exceed the performance threshold μ on the labeled data X_l using the chosen platform. Another concern might be the annotator calibration, which is the alignment of annotations between the annotators. We actually do not require such alignment, because we measure these differences between the annotators. However, if an easy rule allows better alignment of annotators and yields less ambiguous results this direction can be pursued.

A.3 Definition - How?

A critical decision in this stage is whether to incorporate proposals in the annotation process. While proposals can significantly reduce annotation costs, they may introduce a bias (Jachimowicz et al., 2019; Schmarje et al., 2023). This bias results from annotators' tendency to favor the proposed class, skewing the label distribution $P(L^x = \cdot)$ towards it. For making this decision, several parameters, such as the class confusion matrix c and the expected annotation speedup S, need estimation or a defined approach for later approximation (see subsection B.2 and subsection B.3 for techniques).

The first question to address is whether introducing a bias is acceptable. In cases where bias is tolerable, proposals should be used. If not, the expected speedup S becomes the determining factor. Proposals are more advantageous with a higher speedup, as a higher speedup allows for easier reversal of the introduced bias with more annotations. (Schmarje et al., 2023) identifies a trade-off point for speedup around 3; below this, it's advisable to avoid proposals due to the difficulty in reversing the bias within a reasonable budget. The class confusion matrix c is utilized in post-processing (see subsection A.5) and not needed until then.

A.4 Annotation Process

This step focuses on creating the annotations, with various preparations depending on whether proposals are used. Literature indicates that proposals, particularly DC3 proposals, enhance annotation consistency, especially with ambiguous data (Papadopoulos et al., 2021; Desmond et al., 2021; Schmarje et al., 2021a). We recommend DC3 proposals (Schmarje et al., 2022b), which perform well with real-world ambiguous data and are compatible with many semi-supervised methods (Sohn et al., 2020; Li et al., 2020; Tarvainen and Valpola, 2017). Normal network predictions can also be used but might reduce overall performance. The labeled dataset X_l serves as training data for a neural network to generate these proposals, which should be trained until reaching the same quality threshold μ as the annotators.

Two important variables to define are the number of annotations required for consensus (A_{cons}) and the total number of annotations (A^{cons}). These variables balance between the best approximation of the ground-truth distribution $P(\hat{L}^x = \cdot)$ and the effort and cost of annotations. In subsection B.2 and subsection B.3, we discuss plausible numbers of annotations per hour and potential confidence guarantees for different combinations of A_{cons} and A^{cons}. The annotations can either be created by different individuals or by showing the same images multiple times to one or multiple annotators. For the latter case, it is important to think about a cooldown period, the time between showing the same image again. This period can be quite small, down to a few hours, as long as there is a high probability that the annotator will not remember a case between annotations.

A.5 Post-Processing

The use of proposals in annotation can introduce biases due to the default effect (Jachimowicz et al., 2019). Schmarje et al. (2023) propose CleverLabel, a combination of class blending and bias correction, to improve the approximated distribution if proposals were used. If no proposals were employed, enhancing the distribution $P(L^x = \cdot)$ by blending with the class distribution c is still recommended.

Appendix B. Implementation and approximation details

This subsection provides more in-depth descriptions for some of the previous parts. They are separated to allow a concentration on the most important information without diluting it with necessary but minor considerations. We start this section with a description of additional minor consideration which apply to the complete annotation strategy (see subsection B.1). Additionally, we will discuss how the expected workload for the human annotators can be calculated (see subsection B.2) and what confidence with regard to the soft label can be expected given a certain number of annotations (see subsection B.3). Lastly, we will reiterate the most important information about CleverLabel from (Schmarje et al., 2023).

B.1 Additional minor considerations for the Annotation Strategy

- All decisions should typically be supported by domain experts or the end users which will use the labeled data in the end.
- Constraints like collaboration modes and support for proposals should be considered. For example, if crowd-sourcing should be used for annotation, this restricts the appropriate platforms. Due to the fact the decision if proposals will be used is made later in subsection A.3, it is advisable to ensure that these future options are supported to avoid unnecessary backtracking.
- If the training of the annotators is not successful it might be necessary to step back and rethink previous decision to allow a successful training. For example, special training with experts could be held, more detailed description of the task could be given and general feedback of annotators could be considered.
- Proposals ρ_x with a low ground truth probability ($P(\hat{L}^x = \rho_x) = 0$) should be avoided and (Schmarje et al., 2022b) indicates only up to 10% of all proposals should be of such poor quality to allow successful training. The distribution $P(\hat{L}^x = \cdot)$ might not be known at this point but if available this should be checked in addition to the threshold μ.
- The variables c, δ and S may be approximated in various ways. One possible approach is defined in (Schmarje et al., 2023) by annotating about 100 images at least 10 times

for the estimation of c, δ and $P(\hat{L}^x = \cdot)$. The speedup S can for example be measured on a small subset with and without proposals during the training of the annotators. However, if the variables are not required during the decision process described above, they could also be approximated on the real annotation data or ignored entirely. For example, c can easily be approximated on the complete dataset if no proposals are used during the annotation process and an approximation of S is not required if a small introduction of a bias is acceptable. The variables could also be approximated with an educated guess either based on previous knowledge or experiments.

- The question is how A_{cons} and A^{cons} should be selected. We concentrate on two major aspects, what is a feasible workload for the annotators and what guarantees can be given. See for more details the approximations in subsection B.2 and subsection B.3.

- The class blending (CB or CleverLabel) in the post-processing step may not be necessary if the number of annotations is high (e.g. above 50), since the distribution is already approximated quite accurately.

B.2 Approximation of workload

The expected workload can be approximated as the expected number of annotations per hour done by one annotator as $\frac{a}{h}$. Depending on the expected portion of complete consensus of the dataset p_c and the total dataset size $N = |X_u|$, we can calculate the expected annotation time in hours with $A_N = N \cdot (p_c \cdot A_{\text{cons}} + (1 - p_c) \cdot A^{\text{cons}})/\frac{a}{h}$. In Figure 5 the required annotation time A_N in days for $N = 10,000, A_{\text{cons}} = 10, A^{\text{cons}} = 50$ with varying p_c is given as well as values reported in (Schmarje et al., 2022b). The reported values for $\frac{a}{h}$ are between 1,000 and 4,500 without proposals or 3,000 and 10,000 with proposals. Thus, even in the worst-case the expected workdays with proposals are less than 10 days, or two weeks, which should be feasible in most cases especially when multiple annotators are available. These values are examples and have of course to be adopted to the specific use-case.

B.3 Approximation of confidences

The confidences of the calculated label distributions should be considered when A^{cons} and A_{cons} are defined. Considering the binary case of the annotation task for class k for each image x, i.e. whether class k is annotated for x or not, and having a sufficient number of annotations A, the 95%-confidence interval can be calculated using the Wald confidence interval (Lyman, 1993). The interval is given by

$$P(\hat{L}^x = k) \pm Z_{95\%} \cdot \sqrt{\frac{p_k(1 - p_k)}{A}} \tag{1}$$

with 95%-confidence value, which is 1.96, and p_k is the unknown true GT probability for class k. The confidence interval is the largest for $p = 0.5$. The formula can be rearranged to calculate the required annotations A if a 95% confidence interval around the approximation

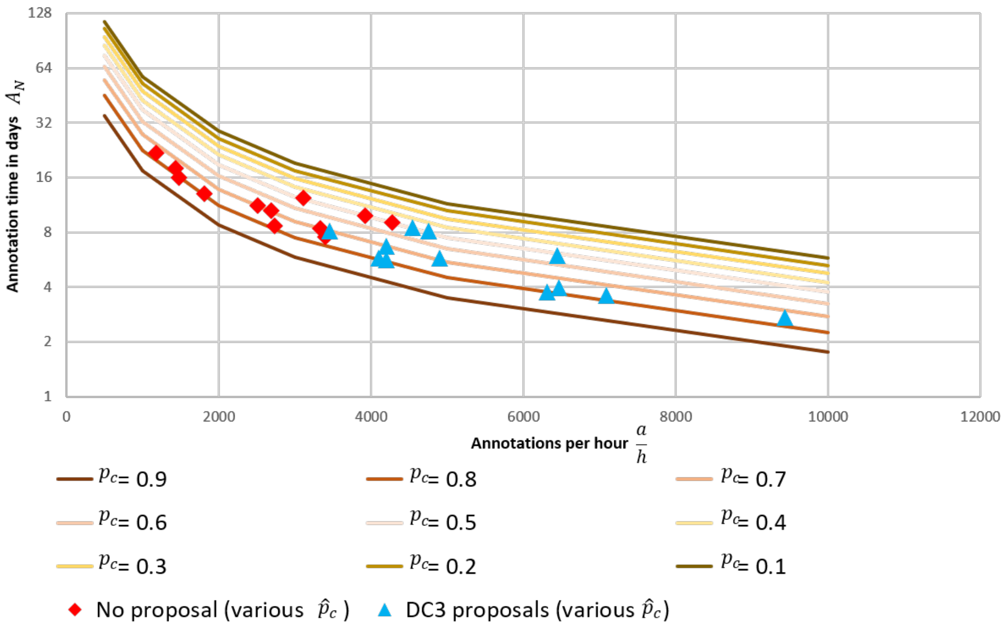


Figure 5: Annotation time in days based on expected consensus ratio p_c and annotations per hour $\frac{a}{h}$, reported values from (Schmarje et al., 2022b) and with 95% agreeing votes threshold $\hat{p}_c$

$P(\hat{L}^x = k)$ of width W and margin $W/2$ is to be guaranteed.

$$\begin{aligned}\frac{W}{2} &= Z_{95\%} \cdot \sqrt{\frac{p_k(1-p_k)}{A}} \\ \frac{W^2}{4} &= Z_{95\%}^2 \cdot \frac{p_k(1-p_k)}{A} \\ A &= \frac{4 \cdot Z_{95\%}^2 \cdot p_k(1-p_k)}{W^2}\end{aligned} \tag{2}$$

Figure 6 visualizes the required annotations A for various p and W. These calculations can guarantee with 95% confidence that the confidence interval has a width of $W \approx 1$ for $A = 3$, $W \approx 0.62$ for $A = 10$ and $W \approx 0.28$ for $A = 50$. However, a high number of A is indirectly assumed because the Wald confidence interval approximates the original binomial distribution with a normal distribution. This approximation is especially inaccurate for small numbers of A and p near 0 or 1, as pointed out by Lyman (1993). They suggest to approximate the 95% confidence intervals with $((0.25)^{\frac{1}{A}}, 1)$ when p is close to one. This would give a lower bound of the interval of 0.63 for $A = 3$, 0.87 for $A = 10$ and 0.97 for $A = 50$. All these calculations assume an idealized version of the annotators, which must be taken into account in their interpretation. For this reason, it will not be investigated further how the confidence intervals can be approximated more precisely, but conclude that a consensus decision is uncertain for $A = 3$ annotations, but reasonable for ten or more annotations. Similarly, approximations of p with a confidence interval width less than 0.3 are only reasonable for about 50 annotations.

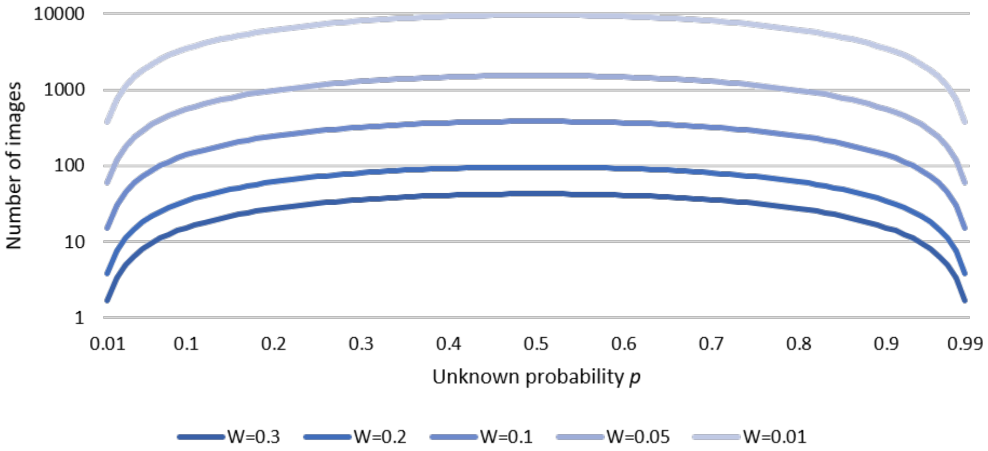


Figure 6: Number of images for confidence width W, logarithmic scale

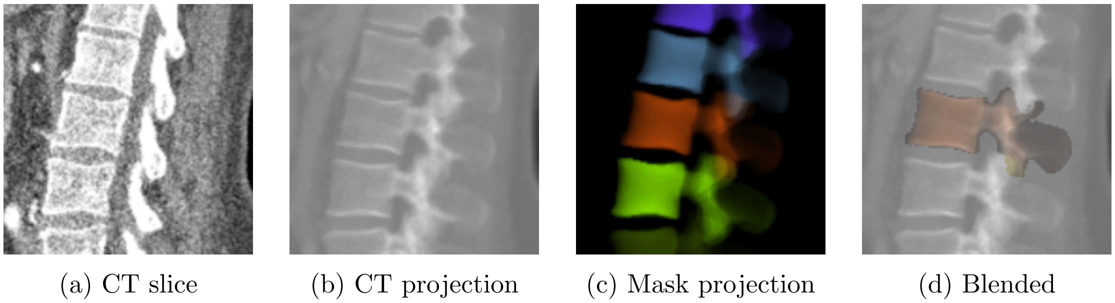
B.4 Details on using CleverLabel and CB

This section provides an overview about the methodology proposed in (Schmarje et al., 2023). These elements are reiterated here to keep this paper self-contained. Schmarje et al. presented in (Schmarje et al., 2023) two different approaches to improve the resulting distribution when using proposals. The approaches were called class blending (CB) and bias correction (BC) and formed together the method CleverLabel. Class blending uses a constant transition matrix c between classes and blends this fixed class distribution with the calculated distribution. The constant transition matrix is motivated by the fact that for example a car and truck are more easily confused than a car and a dog. The bias correction used the theoretical analysis of the introduced bias when using proposals to invert this process. A main variable for this calculation is the dataset specific offset δ. This offset describes how likely it is that a human will accept the proposal even if the proposed class would normally not been considered. A offset of zero means that the proposal has no influence while a value of one means that the proposal is always accepted regardless of the content. For details, please see the original publication (Schmarje et al., 2023).

Appendix C. Differences to previous work

Unlike to the original author of the Verse2019 (Löffler et al., 2020) dataset, we will not use medical experts for the annotation and will use 2D projections of the original 3D CT data.

Our decision to use non-medical experts has two main reasons. First, we want to show that non-experts can successfully annotate data with the help of proposals. Second, we plan to generate tens of thousands of annotations, which would not be feasible with medical experts, who are in high demand. While we will not use medical experts for the majority of the annotations, we have worked closely with the radiologist at University Medical Center Schleswig-Holstein. We discussed our design choices with them, particularly regarding the definition of the task and the representation of the vertebrae used. As mentioned above, we trained the hired workers on this specific task and ensured that they achieved a macro



(a) CT slice (b) CT projection (c) Mask projection (d) Blended

Figure 7: Illustration of different image modalities

accuracy and F1 score of at least $\mu = 0.6$ on the complete dataset with and without proposals for at least two scores before they could annotate the final scores.

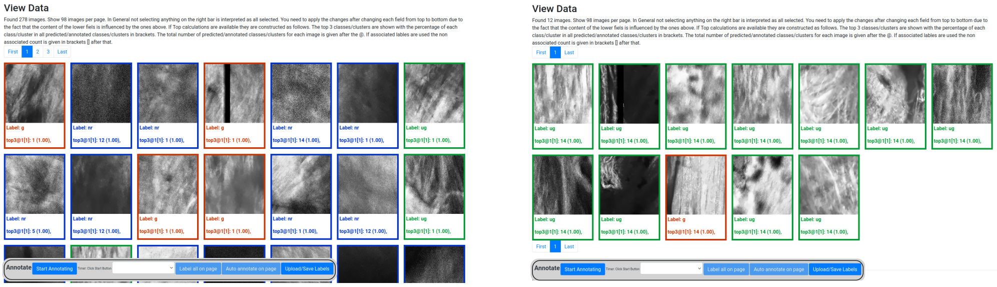
We do not use the original 3D CT data (see Figure 7a) for the annotation but rather a 2D projection of the central slices for each vertebra (see Figure 7b) and its segmentation (see Figure 7c). Additionally, we show the vertebrae above and below the target vertebrae to allow an evaluation of the height reduction. A blended example is shown in Figure 7d. There are three major benefits/reasons of using 2D projections over 3D CT data. First, the Genant scores were defined on 2D X-Ray and 2D projections of CT data are visually very similar to X-Ray and thus are better comparable. Second, we want to use the reported methods from the literature for the proposal generation (Schmarje et al., 2022b), which were only tested on 2D data. While it should be theoretically possible to extend the used networks to 3D, such major structural changes could lead to an inferior result which would dilute the results. But such experiments may be future work. Lastly, when it comes to diagnostics, radiologists mainly look at one central slice which we include and the major 3D structures can be detected on this projection.

Annotators and annotation platform

We hired 5 workers to annotate vertebrae in a web-based environment developed for this evaluation. Public annotation platforms do not support the combination of clusters and classes that we want to evaluate in this paper.

Figure 8 displays the used annotation website. It is important to note, that all images are displayed to the user in a grid. In the case without proposal (see Figure 8a), the shown classes are mixed and all have to be annotated manually by the user. In the second case with proposal (see Figure 8b), most of the images are of the same class and can be annotated simultaneously, while only the errors need to be corrected. As X_l we used the partially available labels from (Löffler et al., 2020).

The software is a custom implementation to handle various inputs like classification and clusters. We plan to release a version in the future and please contact us if you are interested in a preview.



View Data

Found 278 images. Show 98 images per page. In General not selecting anything on the right bar is interpreted as all selected. You need to apply the changes after changing each field from top to bottom due to the fact that the content of the lower fiels is influenced by the ones above. If Top calculations are available they are constructed as follows. The top 3 classes/clusters are shown with the percentage of each class/cluster in all predicted/annotated classes/clusters in brackets. The total number of predicted/annotated classes/clusters for each image is given after the @. If associated lables are used the non associated count is given in brackets [] after that.

(a) No Proposal

View Data

Found 12 images. Show 98 images per page. In General not selecting anything on the right bar is interpreted as all selected. You need to apply the changes after changing each field from top to bottom due to the fact that the content of the lower fiels is influenced by the ones above. If Top calculations are available they are constructed as follows. The top 3 classes/clusters are shown with the percentage of each class/cluster in all predicted/annotated classes/clusters in brackets. The total number of predicted/annotated classes/clusters for each image is given after the @. If associated lables are used the non associated count is given in brackets [] after that.

(b) With DC3 proposal cluster

Figure 8: Screenshots of the annotation website on the MiceBone dataset from (Schmarje et al., 2022a), The sidebar for data selection is hidden for data protection reasons.

Appendix D. Used parameters and their calculation

D.1 Approximation of desired variables

In a medical consensus process in an expert panel, the proposals of different people are unified, but influences between people due to social dynamics cannot be excluded. Therefore, we argue that a bias that may be introduced due to human approved annotations is comparable to a consensus process and therefore acceptable. This means that we will use proposals during the annotation and c and S only need to be approximated after the annotation process based on the defined strategy. To allow a validation of our decisions, we will also raise annotations without proposals and thus we can approximate the class distributions c and speedup S based on these annotations. Additional approaches to approximate S as well as the approximation of a data-specific offset δ (see (Schmarje et al., 2023)) are discussed in the supplement.

Based on the training annotations, we estimated the $\frac{a}{h}$ to be about 2,000 and 3,000 without and with proposals, respectively. These results are similar values previously reported in (Schmarje et al., 2022b). For $A_{\text{cons}} = 10$, $A^{\text{cons}} = 50$ and estimated $p_c = 0.5$, this would result in about 250,000 annotations with and without proposals. Furthermore, based on the estimated $\frac{a}{h}$ of about 2,500, this resulted in a workload of about 20 hours per annotator which was within our given budget. So we chose the number of annotations at the recommended values of $A_{\text{cons}} = 10$, $A^{\text{cons}} = 50$. Additionally, we planned to annotate the 3D data to evaluate the quality difference between 2D and 3D. We limited these annotations to the test set of the original Verse2019 dataset (Löffler et al., 2020).

We used annotations created during the training of the annotators to estimate δ. We randomly picked 20 examples without consensus from the annotations without proposals. Based on these samples, the proposal per image p_x, the probabilities $P(L_x = \rho_x)$ with and without proposal and the formula for simulated proposal acceptance of (Schmarje et al., 2023), we calculated $\delta \approx 0.11$.

Table 2: Used hyperparameters for the proposal and evaluation network, † approximated based on number of image iterations, epochs not counted

Parameter	Proposal network	Evaluation network
Architecture	DC3(Schmarje et al., 2022b) + Mean-Teacher(Tarvainen and Valpola, 2017)	ResNet50(He et al., 2016)
Learning rate	0.03	0.1
Epochs	600†	60
Optimizer	SGD	SGD
Batch size	64	128
Weight Decay	0.0005	0.0005
Pretrained Weights	N/A	ImageNet
Source code	(Schmarje et al., 2022b)	(Schmarje et al., 2022a)

D.2 Used hyperparameters

The used hyperparameters are defined in Table 2. For more detail see the respective source code [1].

Appendix E. Additional results of analysis

E.1 Learning effect

Regarding the learning effect of the annotators, the training iterations showed an increase in F1 scores over time. The improvement in F1 scores across all annotators between the training and the last three iterations was 6.79 without proposals and 9.19 with proposals. When comparing the benefit of DC3 proposals versus no proposals, the average improvement in F1 scores was 3.27 for the first three iterations and 0.41 for the last three iterations. The proposal still has a benefit in the end but the annotators can more successfully annotate event without them. Moreover, the introduction of proposal-based annotations resulted in a reduction in training time which also diminished over time. The first three iterations saw a reduction of 13.49 minutes, while the last three iterations experienced a decrease of 4.55 minutes. Overall, we can conclude that the proposals helped to improve the quality and reduce the annotation time. However, this effect was diminished over time while the performance with and without proposals increases. Thus, we see a learning effect of the annotators.

E.2 Evaluation of human performance

A comparison of the annotator and network performance on the Verse2019 (Löffler et al., 2020) test set is given in Table 3. When comparing the macro F1 scores in 2D between annotations performed with and without proposals and those with proposal-based guidance, the latter achieved a significantly higher F1 score of about 2%. This means that the employment of proposal-based guidance resulted in improved annotations by the annotators. The average macro F1 score for 3D annotations shows no significant difference compared to the 2D annotations, meaning that the modality difference was of no to little importance to the annotators as expected in subsection 3.1.

[1] Will be released after publication

Table 3: Comparison of human and network performance on Verse2019 (Löffler et al., 2020) – * based on ten randomly sampled annotations

(a) Human Performance

	$F1$
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

(b) Network

	$F1$
Sampled from Human*	0.69
Proposal Network	0.57
Wei et al. (Wei et al., 2022)	0.59
Ours*	0.63

The marginal class distribution on all vertebrae from the test set between the ground truth and the majority vote for 2D annotations was quite similar, with percentages of 80.33, 10.38, 6.01, and 3.28 versus 80.33, 10.93, 5.74, and 3.01, respectively. In addition, only 17% of the majority votes deviated from the ground truth. Comparing the average uncertainty based on $P(L^x = \cdot)$ for concurring and dissenting majority votes with the ground truth, we find a higher annotator uncertainty of 24.20% for dissenting scores than for concurring scores at 7.78%. These results indicate that the annotators produced annotations of similar quality to the original test data, with higher uncertainties for potentially difficult or incorrect majority votes.

Appendix F. Dataset Card

Attribute	**Details**
Dataset Name	VerSe 2D Projections with Soft Labels
Dataset URL	`https://zenodo.org/records/8115942` **Original datasets:** • **VerSe'19:** `https://osf.io/nqjyw/` • **VerSe'20:** `https://osf.io/t98fz/`
Description	The dataset contains 2D projection of the central slices for vertebra of the VerSe (Large Scale Vertebrae Segmentation Challenge) datasets. These are collections of multi-detector CT scans for vertebral labelling and segmentation, significant for its size, diversity, and detailed annotations. The data, collected in conjunction with MICCAI 2019 and 2020, is anonymized and includes a variety of normal and pathological cases.
Data Collection	**Domain:** Medical Imaging, 2D Projections of Spine Multi-detector Computed Tomography (MDCT) **Annotation Details:** Soft-labels by evaluating proposal-based human non-expert annotations **VerSe'19:** 160 CT scans from 141 patients, 1725 annotated vertebrae **VerSe'20:** 374 CT scans from 335 patients, 4505 annotations
Data Preprocessing	Scans and segementation masks have been anonymized and converted to NIFTI format (https://nifti.nimh.nih.gov/) in the VerSe dataset. The resolution of the CT images was adjusted to limit computational demands for deep learning algorithms. Coordinates of vertebral body centroids per vertebral level are stored in JSON format.
Annotations	The proposed multi-step algorithmic framework has been applied to generate annotations with or without proposals by trained non-experts, including post-processing with CleverLabel and class blending.
Usage	Development and evaluation of algorithms for vertebral labelling and segmentation in spine CT scans. VerSe'19 and VerSe'20 were used in challenges to evaluate accuracy, precision, and the effectiveness of segmentation and fracture detection as part of MICCAI 2019 and 2020.
Known Limitations	While including multiple sites, the datasets might still not represent all possible variations in spine pathology and patient demographics. **VerSe'19:** Exclusion of patients younger than 30 years and certain spinal conditions like bone metastasis
Ethical Considerations	Both VerSe studies were approved by the local institutional review board, which conducted a thorough review of the research methodology, ensuring adherence to ethical guidelines. All patient data included in this study were anonymized to protect the privacy and confidentiality of the individuals.
Acknowledgements	VerSe has been supported by the European Research Council under Horizon2020 (GA637164–iBack–ERC–2014–STG) and NVIDIA.

Table 4: Overview of the VerSe 2D Projections with Soft Labels