Label Smarter, Not Harder: CleverLabel for Faster Annotation of Ambiguous Image Classification with Higher Quality

Lars Schmarje[1 [0000−0002−6945−5957]], Vasco Grossmann[1 [0000−0001−7560−3867]], Tim Michels[1 [0000−0001−8827−0838]], Jakob Nazarenus[1 [0000−0002−6800−2462]], Monty Santarossa[1 [0000−0002−4159−1367]], Claudius Zelenka[1 [0000−0002−9902−2212]], and Reinhard Koch[1 [0000−0003−4398−1569]]

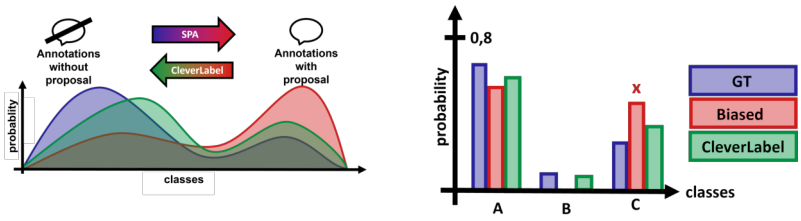
Kiel University {las,vgr,tmi,jna,msa,cze,rk}@informatik.uni-kiel.de

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 **C**ost-effective **L**ab**E**ling using **V**alidated proposal-guid**E**d annotations and **R**epaired **LABEL**s. 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 [54]. Therefore, current research efforts try to mitigate this issue by using unlabeled data [55,4,56] 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 [57,40,46,6] 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



(a) idea

(b) examples

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.

this issue which significantly limits their use in the development of methods that generalize well for ambiguous real-world data.

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 [46,50]. 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 **C**ost-effective **L**ab**E**ling using **V**alidated proposal-guid**E**d annotations and **R**epaired **LABEL**s (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 addition, based on a previously published user study [49], 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 [63,38]. 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 [5,60]. The reasons for this issue can arise for example from image artifacts like low resolution [43], inconsistent definitions [59], uncertainty of the data [1,45] or subjective interpretations [52,32].

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 [62,8], 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 [51]. [12] showed that proposal-based data labeling increases both accuracy and speed for their user study

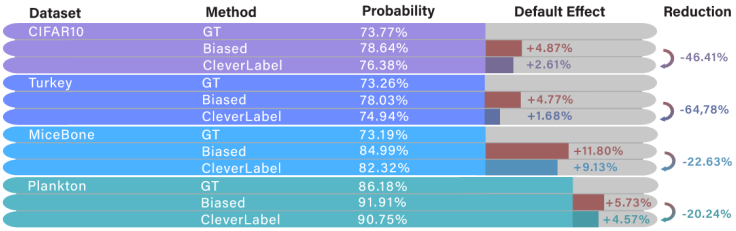


Fig. 2: Average annotation probability of a proposed class with the proposal unknown (GT, Unbiased) and known (Biased) to the annotators in four evaluated datasets. The proposal increases the probability in all observed cases, revealing a clear default effect in the investigated study. Its value is shown without any further processing (Biased) and with the contributed correction (CleverLabel) which consistently reduces the difference to the unbiased probabilities.

(n=54) which is in agreement with proof-of-concepts by [46,49]. The annotation suggestions for the classification in diagnostic reasoning texts had positive effects on label speed and performance without an introduction of a noteworthy bias [51]. 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 [53] 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 [61] 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 [58]. Consensus processes [1,42] where a joint statement is reached manually or with technical support are also out of scope.

2 Methods

Previous research on proposal-based systems [41,29,51] 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
  $r \leftarrow$ random(0,1)
  if $r \leq A$ **then** ▷ Accept proposal
    $a'^x_{i,\rho_x} \leftarrow 1$
  else ▷ Sample from remaining classes
    $k \leftarrow$ sampled from $P(L^x = k \mid \rho_x \neq k)$
    $a'^x_{i,k} \leftarrow 1$
  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 given 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^x_b 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 $P(L^x_b = k)$ 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 $P(L^x = k) \approx \sum_{i=0}^{N-1} \frac{a^x_{i,k}}{N}$ with $a^x_{i,k} \in \{0, 1\}$ 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, $P(L^x = k) \neq P(L^x_b = k)$ for at least one class k. Especially, for the proposed class ρ_x it can be expected that $P(L^x = \rho_x) < P(L^x_b = \rho_x)$.

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, $P(L^x = \rho_x)$. 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

$$A = \delta + (\mathbf{1}^* - \delta)P(L^x = \rho_x) \tag{1}$$

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 $a'^x_{i,k} \in \{0, 1\}$ as in Algorithm 1 and describe the biased distribution similar to the unbiased distribution via $P(L^x_b = k) \approx \sum_{i=0}^{N'-1} \frac{a'^x_{i,k}}{N'}$ 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 the supplementary.

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 $P(L^x_b = k)$ 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 $\mu P(L^x_b = k) + (1 - \mu)c(\hat{k}, k)$ with the most likely class $\hat{k} = \text{argmax}_{j \in \{1,..,K\}} P(L^x_b = j)$ 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 $P(L^x_b = k)$ 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	Synthethic	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

$$\begin{aligned} & P(L^x = k \mid L^x \neq \rho_x) = P(L_b^x = k \mid L^x \neq \rho_x) \\ \Leftrightarrow & \frac{P(L^x = k, L^x \neq \rho_x)}{P(L^x \neq \rho_x)} = P(L_b^x = k \mid L^x \neq \rho_x) \\ \Leftrightarrow & P(L^x \neq \rho_x) = (1 - B) P(L_b^x = k \mid L^x \neq \rho_x) \\ & \approx (1 - B) \cdot \sum_{i \notin M_{\rho_x}} \frac{a'_{i,k}}{N' - |M_{\rho_x}|}. \end{aligned}$$

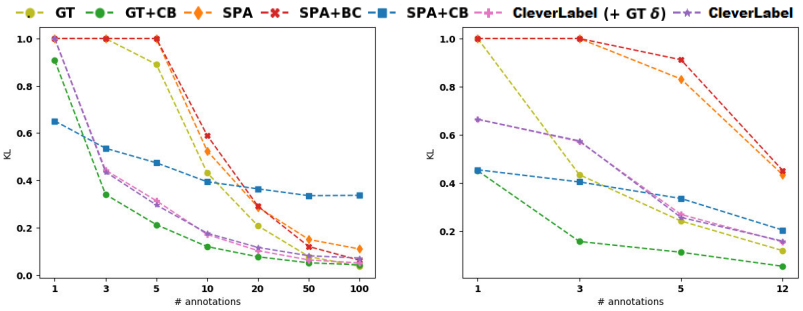
This results in an approximate formula for the original ground truth distribution which relies only on the annotations with proposals. The joint distribution is deducted in the supplementary. It is important to note that the quality of these approximations relies on a large enough number of annotations N'.

2.3 Implementation details

We use a small user study which was proposed in [49] 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 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 method DivideMix [27] and its implementation in [47] to generate the proposals.



(a) Synthetic (b) Real

Fig. 3: Label improvement (i.e. smaller Kullback-Leibler divergence (KL) [25]) in regard to the number of annotations per image. Annotations are created synthetically with SPA (a) or with proposals in a real user study (b). Results are clamped for visualization to the range 0 to 1.

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 two 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. Additionally, in the supplementary we provide a comparison between our proposed simulation proposal acceptance and other possible simulations. The analysis shows that our chosen simulation is optimal in most cases while remaining easy to reproduce.

3.1 Label Improvement

We show that CB and BC lead to similar improved results on both simulated and real biased distributions. This similarity between CB and BC illustrates the practical benefit of SPA.

Metrics & Comparison To measure label improvement, we use the Kullback-Leibler divergence (KL) [25] metric between the soft ground truth $P(L^x = k)$ and the estimated distribution. The input for the label improvement methods, i.e. the skewed distributions, are generated either by our method SPA or by use

of real proposal acceptance data from [49]. 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.

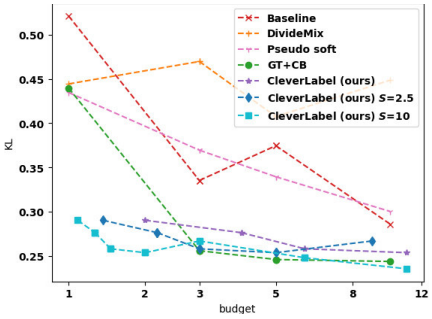
Results If we look at the results on the synthetic data created by SPA in Figure 3a, we see the expected trend that more annotations improve results. While using only CB (SPA+CB) is the best method for one annotation, it is surpassed by all other methods with enough annotations. The baseline (GT), especially in combination with blending (GT+CB), is the best method for higher number of annotations. Our label improvement (CleverLabel) is in most cases the second best method and blending is a major component (SPA+CB). The bias correction (SPA+BC) improves the results further if $\sim$20 or more annotations are available. Using the correct offset (CleverLabel + 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, wee 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.

Results on real annotations are shown in Figure 3b. For consistency we keep the previous notation for CleverLabel, even though SPA was not used here to generate biased distributions. The real results show similar trends to the synthetic data. However, the baseline method without blending (GT) 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 for real data, likely because the effect is diminished by the difference of simulation and reality.

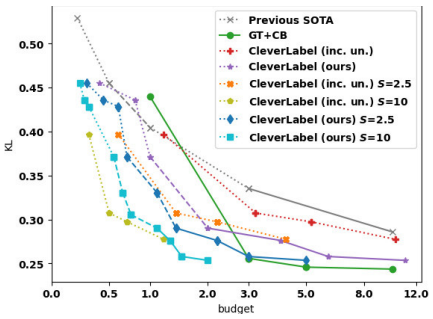
Overall, the results on synthetic and real data are similar. Thus, SPA can be used as a valid tool during method development. It is important to point out that the use of proposals will speed up annotations. Hence, different methods should not necessarily be compared for the same number of annotations. E.g. CleverLabel often performs slightly worse than GT in Figure 3b. Considering a speedup of 2, we would have to compare CleverLabel with 5 annotations to GT with around 3, as explained in the budget calculation in subsection 3.2. Due to the similarity of the generated proposals with SPA and the real proposal data from [49], we conclude that experiments can also be verified only on generated proposals with SPA.

3.2 Benchmark evaluation

Metrics & Comparison We show the results for CleverLabel on [47]. 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.



(a) Comparisons of benchmark results with 100% in. sup.



(b) Pareto front visualization of benchmark results with 20,50, 100% in. sup.

Fig. 4: Left: Compares previous state-of-the-art out of 20 evaluated methods reported in [47] (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, best viewed in color and digitally

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) as well as KL between ground truth and the input of the second stage ($\hat{KL}$) (for details about the stages please see the original benchmark [47]). We also provide an additional ablation where we replaced the fixed model in the second sage 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 [49] or 10 as in [50]. 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.

Results We present in Figure 4a 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 4b. 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,lightblue) 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 2, we conduct several ablations to investigate the impact of individual parts of our method. Comparing KL and $\hat{KL}$ scores, we see similar trends between each other and to subsection 3.1. 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. With CleverLabel (the combination of CB and BC) the scores are slightly decreased for example from 0.17 to 0.16.

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 which is reasonable based on previously reported values [49].

Table 2: Benchmark results ablation study across different number of annotations per image, the number of annotations is given in the top row, 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, the best results per block are marked bold per number of annotations

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 (μ=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 (μ=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

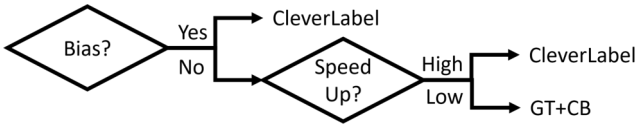


Fig. 5: Flowchart about how to annotate ambiguous data based on the questions if an introduced bias is acceptable and if the expected speedup S is high (> 3)

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 5 and was extended in [48]. Firstly, is a bias in the data acceptable? Be aware that in CleverLabel all labels are human validated and that many consensus processes already use an agreement system [1] with multiple reviewers. If a small bias is acceptable you can directly 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.

The main limitations of this work arise due to the fact that not more than four datasets could be evaluated. 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 3.1, 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.

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 J. Nazarenus by the *OP der Zukunft* project funded by the Business Development and Technlogy Transfer Corporation of Schleswig Holstein (WTSH, Germany) within the REACT-EU program. We also acknowledge the funding of M. Santarossa by the KI-SIGS project (grant number FKZ 01MK20012E) and 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

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: Proceedings of the 1st workshop on benchmarking: past, present and future. pp. 15–21 (2021)
4. 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)
5. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., van den Oord, A., van den Oord, A., van den Oord, A.: Are we done with ImageNet? arXiv preprint arXiv:2006.07159 (2020)
6. 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
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 pp. 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) (2022)
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. Transactions of the Association for Computational Linguistics **10**, 92–110 (2022). https://doi.org/10.1162/tacl_a_00449
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. pp. 392–401. Association for Computing Machinery (2021). 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 (2021)
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. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. pp. 407–422. Springer (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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 16000–16009 (2022)
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 (2014). https://doi.org/10.1609/hcomp.v2i1.13165
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). https://doi.org/10.1007/978-3-030-58558-7_29
23. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (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? In: International Conference on Machine Learning. pp. 6448—-6458. PMLR (2020)
31. Lukov, T., Zhao, N., Lee, G.H., Lim, S.N.: 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. pp. 648–665. Springer (2022). https://doi.org/10.1007/978-3-031-19775-8_38
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. Advances in Neural Information Processing Systems **35**, 18571–18585 (2022)

33. Misra, I., van der Maaten, L., van der Maaten, L.: Self-Supervised Learning of Pretext-Invariant Representations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 6707–6717 (2020)
34. Motamedi, M., Sakharnykh, 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** (2019)
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)
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
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., Mammarappallil, 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. Advances in Neural Information Processing Systems **35**, 33215—-33232 (2022)
48. Schmarje, L., Grossmann, V., Zelenka, C., Koch, R.: Annotating Ambiguous Images: General Annotation Strategy for High-Quality Data with Real-World Biomedical Validation. arXiv preprint arXiv:2306.12189 (2023)
49. 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)
50. Schröder, S.M., Kiko, R., Koch, R.: MorphoCluster: Efficient Annotation of Plankton images by Clustering. Sensors **20** (2020)
51. Schulz, C., Meyer, C.M., Kiesewetter, 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 (2019). https://doi.org/10.18653/v1/P19-1265
52. Schustek, P., Moreno-Bote, R.: Instance-based generalization for human judgments about uncertainty. PLOS Computational Biology **14**(6), e1006205 (2018). https://doi.org/10.1371/journal.pcbi.1006205
53. 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)
54. 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)
55. Singh, A., Nowak, R., Zhu, J.: Unlabeled data: Now it helps, now it doesn't. Advances in neural information processing systems **21** (2008)
56. 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)
57. Tarling, P., Cantor, M., Clapés, A., Escalera, S.: Deep learning with self-supervision and uncertainty regularization to count fish in underwater images. Plos one **17**(5), 1–22 (2021)
58. Tifrea, A., Clarysse, J., Yang, F.: Uniform versus uncertainty sampling: When being active is less efficient than staying passive. arXiv preprint arXiv:2212.00772 (2022)
59. Uijlings, J., Mensink, T., Ferrari, V.: The Missing Link: Finding label relations across datasets (2022)
60. Vasudevan, V., Caine, B., Gontijo-Lopes, R., Fridovich-Keil, S., Roelofs, R.: When does dough become a bagel? Analyzing the remaining mistakes on ImageNet. Advances in Neural Information Processing Systems **35**, 6720–6734 (2022)
61. Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., Liu, Y.: Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. ICLR pp. 1–23 (2021)
62. Wei, X., Cong, H., Zhang, Z., Peng, J., Chen, G., Li, J.: Faint Features Tell: Automatic Vertebrae Fracture Screening Assisted by Contrastive Learning (2022)
63. 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)