CONTRASTIVE LEARNING FROM CROWDS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the learning from multiple annotations (crowds) problem in classification settings and propose a contrastive learning approach, generalizing its (un)supervised counterparts. Based on *soft labels*, which are normalized histograms of annotations over classes, we exploit the disagreement signal in the annotations to drive the shaping of the representation space. Our key contribution is to write the common (un)supervised contrastive loss functions as a cross-entropy and generalize the target distribution to be the inverted distance over the soft labels. Blurred class information allows to perform easier hard mining without the need for large batch sizes, which we support theoretically and empirically. We provide a theoretical relationship between the risks of the contrastive stage and the downstream linear classification that also uses soft labels. Empirically, we experiment with real-world datasets of various flavors (image classification, sleep staging) to demonstrate the effectiveness of the proposed method.

1 INTRODUCTION

When one collects data for classification tasks, instances are relatively easy to obtain, e.g., one can download millions of images within hours; but it would take months to validate their labels with a single human (Deng et al., 2009). Since one needs more labeled data to have more generalizable classifiers, crowdsourcing is often employed where a lot of humans annotate a large dataset together (Russell et al., 2008). Furthermore, even on relatively small datasets, gathering multiple annotations may be warranted, in particular for harder labeling tasks where the amount of annotation errors can be substantial. In clinical settings, for example, multiple medical experts could be consulted for, e.g., a verdict on an X-ray image as it may be too risky to rely solely on one person's opinion in a critical setting (Reiß et al., 2021).

Be it experts or crowds online annotating datasets, the structure at hand is the same: multiple, possibly conflicting answers to the same question, e.g., "is this image of a cat or a dog?" or "what stage of sleeping is this?". Since traditional classifiers strictly expect a single answer (meaning label/annotation) per instance, they are simply unable to learn from multiple annotations. There are two main ways to tackle this learning problem (Ibrahim et al., 2025): one either has to aggregate the annotations to a single label (or to a probability distribution, i.e., a soft label (Peterson et al., 2019)) and then use a traditional classifier on this dataset; or design a method to directly learn a classifier from the multiply annotated data (e.g., transition matrix based models (Rodrigues & Pereira, 2018; Nguyen et al., 2024)).

Regardless of the approach, the starting point is that there is *signal* rather than noise in the conflicting annotations. Think of an image labeled as *cat* by 1 person versus 50 people unanimously: the latter case practically leaves no doubt that the image is indeed of a cat. What's more, if 24 of the latter group labeled the image as *dog* and the remaining 26 *cat*, there is now drastically more information than having a single label of *cat* (e.g., the image is possibly an ambiguous one such as of a Donskoy cat). This example also shows why hard label aggregation techniques such as majority voting is dangerous to use as it is a very lossy compression of label information at times.

This disagreement signal has been used in a myriad of ways in the literature. Earlier works focused on gauging inter-rater (annotator) reliability, i.e., expertise via mainly statistically motivated white-box models based on, e.g., conditional independence (of annotators given the ground truth) assumptions (Dawid & Skene, 1979; Whitehill et al., 2009). More recent works employed end-to-end deep learning based mechanisms to learn a fully-fledged classifier mainly based on annotator

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

confusion matrices (Tanno et al., 2019; Nguyen et al., 2024). We take a step back and make the following observation: if, under learning from multiple annotations setup, one cannot train a traditional classifier as there is no single ground truth, why not employ unsupervised learning? On the other hand, we do have *some* label information; it would be a waste to not utilize it. Therefore, we aim to reconcile these "no label" and "too many labels" perspectives.

In particular, we focus on contrastive learning: in its unsupervised flavor, instances are augmented to produce self-supervision for contrasts (Chen et al., 2020); in its supervised flavor, the (hard) label information is used as guidance to better shape the representation space (Khosla et al., 2020). Both of these approaches rely on the law of large numbers for *hard mining*, i.e., large batch sizes are needed for effective contrasts. We find the middle way and propose *softened disagreement-supervised contrastive learning* (SDSL) where we use *soft labels* (which are simply normalized histograms of annotations over classes) as our source of contrast, which generalizes its unsupervised and supervised counterparts. Critically, the "free" hard mining due to the disagreement signal in soft labels allows us to bypass large batch size dependence.

We summarize our main contributions as follows.

- We introduce a contrastive learning perspective to learning from multiple annotations in classification tasks via soft labels to exploit the disagreement information. The method has an easy to implement algorithm with a machine-agnostic nature.
- We observe a dramatic decrease in batch size dependence for hard instance mining, alluding to more effective contrasts between instances thanks to the disagreement signals. This implies gathering multiple annotations may help learning even if one has access to the ground "truth". We also avoid augmentation, which increases dataset applicability, i.e., not restricted to augmentable domains, e.g., image.
- Theoretically, we relate the risk of the proposed soft label-based contrastive learning to that of the downstream linear classification, which again uses soft labels, illustrating that optimizing for the former does help for the latter.
- Empirically, we validate our approach with real-world datasets of different domains (image classification, sleep staging) and different annotation schemes (dense, sparse), showing comparable/better results against the current approaches.

2 Related Work

Learning from multiple annotations. Earlier works have mainly sought to extract expertise information, i.e., inter-rater variability out of multiple annotations rather than learning a classifier per se. The seminal work of Dawid & Skene (1979) propose a probabilistic model to capture annotator error rates in a binary classification scenario using maximum likelihood estimation. Whitehill et al. (2009) propose a similar model but also took the instance difficulty into account. Building on top of them as an instance-dependent statistical modeling of annotators, Raykar et al. (2010) learn a classifier jointly with annotator expertise, targeting multiclass classification.

More recently, end-to-end models where the annotator confusion is integrated into a machine, e.g., a neural network while learning a classifier for the latent ground truth have been proposed. Rodrigues et al. (2014) adapt Gaussian processes to the multiple annotation scenario for binary classification. Guan et al. (2018) perform two-stage training: they first learn each annotator's brain, i.e., a mapping via a neural network for each annotator is learned, which are then combined with a learned global convex combination. Rodrigues & Pereira (2018) propose a joint learning of annotator confusion and latent truth via a neural network architecture. The network is augmented via a crowd layer, which bridges the latent truth to each annotator's response via, e.g., a matrix which acts as an instance-independent pseudo-confusion matrix of the annotator. While the identifiability issue (which is about recovering the latent truth from a matrix-vector multiplication) is addressed in Rodrigues & Pereira (2018) by abstaining from overparametrizing the bridge between the truth and the observed annotators, Tanno et al. (2019) formally attack this problem via regularizing the loss function that Rodrigues & Pereira (2018) use by penalizing the trace of the instance-independent confusion matrix. While the works above target instance-independent noise per annotator, Li et al. (2022) learn an instance impact matrix for each instance to additively combine with an instance-

independent confusion matrix. Nguyen et al. (2024) take a similar approach but imposes sparsity to the instance-dependent component, assuming it only takes effect on the outlier instances.

Cao et al. (2017) and Peterson et al. (2019) are prominent works that do not quite fit in to the categorization above. In the former, the authors learn two mappings: one from the instance space to the latent truth and one from the annotations to the latent truth, and maximize the mutual information between the two. In the latter, the authors propose to use soft labels as the target distribution in the classic learning of classification, which can be seen as a label aggregation technique.

Contrastive learning. Chopra et al. (2005) propose a supervised contrastive loss function that aims to drive "genuine" pairs together and "impostors" apart in the representation space. Goldberger et al. (2004) learn a linear embedding to maximize (soft) neighborhoods based on label information; the loss function is a softmax over the distances in the representation space, and this variant has been the most commonly used one since. Its nonlinear extension was proposed by Salakhutdinov & Hinton (2007), which was recently modified to have a temperature parameter by Frosst et al. (2019). With the advent of bigger data and computation, focus on unsupervised contrastive learning resulted in, e.g., Triplet loss (Schroff et al., 2015) and N-pair loss (Sohn, 2016) which offer techniques to judge dissimilarity in absence of labels, i.e., efficient hard mining. More recently, SimCLR proposed by Chen et al. (2020) augments images to provide self-supervision where the augmented siblings are the positive pairs sharing the same class, and all the others are assumed to be in a different class within a given batch. Its supervised counterpart, SupCon of Khosla et al. (2020) takes the logarithm in the loss function of Frosst et al. (2019) outside the inner summation of positive (i.e., of same class) pairs, which was shown to greatly outperform the variant with the logarithm inside.

To the best of our knowledge, there is only one work at the intersection of learning from multiple annotations and contrastive learning: that of Yang et al. (2023). The authors first assume a third-party method obeying a certain loss function form and while learning that method, they suffer a supervised contrastive loss *for each* annotator. Losses from each part are then combined with a hyperparameter, i.e., it essentially regularizes the third-party method. The main drawback of this procedure is that no inter-annotator structure is utilized; the contrastive loss from each annotator's labels are independently summed with uniform weights.

Unlike Yang et al. (2023) and the literature, our contrastive learning framework is a direct generalization of its unsupervised and supervised counterparts while the dependence on the batch size is decreased and the need for augmentation is lifted. As it is based on soft labels, annotations per instance are considered together, not independently, and the computational time does not depend on the number of annotators while the vast majority of the literature scales linearly, which is particularly an issue in sparsely annotated datasets (e.g., there are 100 annotators in total but an instance is annotated by 8 of them at maximum). Overall, we bring a fresh perspective to the problem of learning from multiple annotations by reconciling the "no label" and "too many labels" views through contrastive representation learning under the supervision of disagreement signals.

3 METHODOLOGY

In this section, we present the notation used and then introduce our contrastive learning framework and present its gradient and risk analysis.

3.1 PRELIMINARIES

We consider learning from multiple annotations (crowds) in multiclass classification tasks. We are handed a dataset with N instances and K classes. The n^{th} instance is annotated by $R_n \geq 1$ annotators; we denote this vector of raw annotations via $\widetilde{\boldsymbol{y}}_n \in [K]^{R_n}$ where $[K] \coloneqq \{1, 2, \ldots, K\}$. Note that the number of annotations per instance may differ. $\mathcal{X}$ is the instance space, e.g., $\mathbb{R}^m$; $\mathcal{A}$ is the annotator space, e.g., all humans enumerated; $\mathcal{P}$ is the $K-1$ dimensional probability simplex representing a categorical distribution; $\widetilde{\mathcal{Y}}$ is where an annotator's hard decision over a fixed instance lies, i.e., $[K]$; and $\mathcal{Y}$ is the space for the unobserved ground truth of a fixed instance, i.e., $[K]$. X, A, P, and Y denote random variables over their respective spaces and the shared domain is $\mathcal{X}$. Our goal is to learn a classifier $h : \mathcal{X} \to \mathcal{Y}$.

The *soft label* corresponding to each $\widetilde{\boldsymbol{y}}_n$ is a vector denoted by $\widetilde{\boldsymbol{s}}_n \in [0,1]^K$ whose k^{th} element is given by

$$\widetilde{s}_n[k] = \frac{1}{R_n} \sum_{r=1}^{R_n} \mathbb{1}\left[\widetilde{y}_n[r] = k\right], \tag{1}$$

where $\mathbb{1}$ is the indicator function. To exemplify, for $K = 4$ and $R_n = 5$, let the raw annotations $\widetilde{\boldsymbol{y}}_n = [1, 1, 3, 4, 3]$. Then the corresponding soft label is $\widetilde{\boldsymbol{s}}_n = [0.4, 0.0, 0.4, 0.2]$, i.e., the normalized histogram of annotations over each class. We note that while we derive the soft labels from the raw annotations here, all our framework needs is soft labels, however produced (e.g., see the second dataset in Section 4.1). As an extra, we present a consistent estimator of the Bayes risk based on soft labels to show their usefulness in Appendix A.1, generalizing the recent result of Ushio et al. (2026) to the multiclass case (with slightly improved bounds for the binary case).

3.2 Soft Contrastive Learning

Let us recall the commonly used loss functions for unsupervised (SimCLR; Chen et al. (2020)) and supervised (SupCon; Khosla et al. (2020)) contrastive learning over a batch of size B:

$$\mathcal{L}_{\text{SimCLR}}\left(f; \{x_b\}_{b=1}^B\right) = \sum_{b=1}^{B} -\log \frac{\exp\left(\langle f(x_b), f(x_{\text{aug(b)}})\rangle / \tau\right)}{\sum_{b' \neq b} \exp\left(\langle f(x_b), f(x_{b'})\rangle / \tau\right)} \tag{2}$$

$$\mathcal{L}_{\text{SupCon}}\left(f; \{x_b, y_b\}_{b=1}^B\right) = \sum_{b=1}^{B} \frac{1}{|P(b)|} \sum_{\substack{b' \neq b \\ y_{b'} = y_b}} -\log \frac{\exp\left(\langle f(x_b), f(x_{b'})\rangle / \tau\right)}{\sum_{b'' \neq b} \exp\left(\langle f(x_b), f(x_{b''})\rangle / \tau\right)}, \tag{3}$$

where $f\colon \mathcal{X} \to \mathcal{S}^d$ is the encoder (embedder) with $\mathcal{S}^d$ being the unit hypersphere of dimension d (L_2 norm), $\tau > 0$ is the temperature hyperparameter, aug(b) represents the index of the augmented sibling of the b^{th} instance in SimCLR and $|P(b)|$ is the cardinality of $\{b' \in [B]\colon y_{b'} = y_b, b' \neq b\}$, i.e., the number of non-anchor instances in the batch with same label as the anchor in SupCon. Our key observation is that these two loss functions can be written in a common form using cross-entropy:

$$\mathcal{L}_{\text{contrastive}}(f) = \sum_{b=1}^{B} H\left(\boldsymbol{\alpha}_b, \boldsymbol{p}_b(f)\right). \tag{4}$$

The predictive distribution $\boldsymbol{p}_b(f) \in [0,1]^{B-1}$ is the same for both: softmax of the similarities among the embeddings in the batch per anchor b, i.e., the argument of $\log$ in equations 2 and 3. The target distribution $\boldsymbol{\alpha}_b \in [0,1]^{B-1}$ is where they differ at:

$$\alpha_b[j] = \begin{cases} \begin{cases} 1 & \text{if } j = \text{aug(b)} \\ 0 & \text{otherwise} \end{cases} & \text{if SimCLR} \\ \begin{cases} 1/|P_b| & \text{if } y_j = y_b \\ 0 & \text{otherwise} \end{cases} & \text{if SupCon} \end{cases} \tag{5}$$

for $j \in [B] \setminus \{b\}$. In other words, the representations are driven to match the one-hot distribution marked by the augmented siblings in the unsupervised version (SimCLR) and the sparse-uniform distribution formed by the same-class instances in the supervised one (SupCon). The so-called *positive* pairs are therefore either the self-supervised instances under no labels or the instances with the same label in case of single labels; remaining ones are automatically the *negative* ones. Based on these categorizations, contrastive learning tries to put the positive (similar) ones closer and negative (dissimilar) ones apart. In our setting, the target distribution is a dense one: the positivity (similarity) of the instances in the batch with respect to the anchor is now *softened* — it is based on the distance of the instances' soft labels from that of the anchor. This gives a dense distribution and everyone is both positive and negative at the same time, i.e., a value in the continuum $[0,1]$ governs it instead of hard zeros and ones.

The target distribution we base on soft labels is simply the inverted L_1 distance between the soft labels:

$$\widetilde{\alpha_b}[j] = 1 - \frac{1}{2} \sum_{k=1}^{K} \left|\widetilde{s}_b[k] - \widetilde{s}_j[k]\right| \tag{6}$$

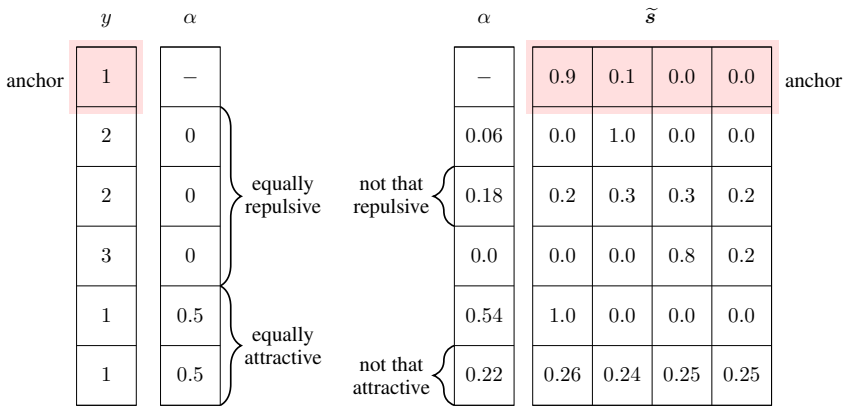


Figure 1: Illustration of how the disagreement signal can be helpful. Batch size is 6, number of classes is 4. The anchor being contrasted is the first one (shaded). On the left are the single label ground truth ys and the α corresponding to the anchor in SupCon. A given non-anchor is either equally repulsive or attractive depending on whether it shares the same label of 1; there is no middle ground. On the right are the soft labels $\widetilde{\boldsymbol{s}}$ and the α for the anchor derived using equation 6. When "zoomed in", some instances start to be not that repulsive / attractive. Note also that there is no instance with hard label 4 in the batch; this slows the convergence for SupCon, but SDSL is less affected since the class information is blurred, based on the probability distributions inferred from the multiple annotations.

for $j \in [B] \setminus \{b\}$. To make it sum to 1, we sum-normalize to arrive at

$$\alpha_b[j] = \frac{\widetilde{\alpha_b}[j]}{\sum_{j \neq b} \widetilde{\alpha_b}[j]}. \tag{7}$$

The choice of the L_1 distance allows us to fall back to SupCon *exactly* in case of a singly-labeled dataset. Indeed, when all the soft labels are one-hot (as is in case of single annotation), the L_1 distance between them will be 0 if they share the same class and 1 otherwise, which yields exactly the sparse-uniform target distribution of SupCon described above. We note that $\widetilde{\alpha_b}[j] = 0$ if and only if $\langle \widetilde{\boldsymbol{s}}_b, \widetilde{\boldsymbol{s}}_j \rangle = 0$, i.e., if the corresponding two soft labels are orthogonal such that there is no class k voted for in both of their annotations.

The softened contrastive loss function is then

$$\mathcal{L}_{\text{SDSL}}\left(f; \{x_b, \widetilde{\boldsymbol{y}}_b\}_{b=1}^{B}\right) = \sum_{b=1}^{B} \sum_{b' \neq b} -\alpha_b[b'] \log \frac{\exp\left(\langle f(x_b), f(x_{b'})\rangle / \tau\right)}{\sum_{b'' \neq b} \exp\left(\langle f(x_b), f(x_{b''})\rangle / \tau\right)}. \tag{8}$$

In Figure 1, we contrast our framework with the supervised counterpart SupCon to illustrate the usefulness of the disagreement signal in the annotations and how we utilize it [1]. As shown, we simply have more information with multiple annotations to shape the representation space in a better way such that the downstream linear classification has an easier job. For instance, in SDSL, thanks to the fractional push and pulls ($\alpha_b[j] \in [0, 1]$) in the representation space instead of hard ones ($\alpha_b[j] \in \{0, 1\}$), the ambiguous instances are more distinctly embedded between clusters, which are formed in the early parts of training thanks to the clear patterns (Han et al., 2025) of instances having (nearly) unanimously agreed votes. This surely helps for, e.g., calibration or classification with a reject option (Hendrickx et al., 2024) as the linear classifier will place them close to its boundary.

The implementation of the loss function in equation 8 is rather easy; the pseudocode of learning over a batch pass is as follows.

```
def sdsl_loss(embeddings, soft_labels):
```

[1] SupCon is not directly comparable to our method as we do not assume access to ground truth; ours is a generalization. The comparison aims to show how useful multiple annotations might be in contrastive learning.

```
    # Get the predictions: tempered and softmaxed similarities
    similarities = embeddings @ embeddings.T / temperature
    similarities *= (1 - eye_like(similarities))  # b' != b
    log_probs = similarities - log(exp(similarities).sum(1))

    # Get the targets: normalized inverted L1 distances
    alphas = 1 - pairwise_L1_distance(soft_labels)
    alphas /= alphas.sum(1)

    # Suffer cross-entropy between predictions and targets
    per_anchor_log_prob = (alphas * log_probs).sum(1)
    loss = -per_anchor_log_prob.mean()

    return loss
```

Note that if one passes one-hot hard labels to this subroutine in lieu of *soft_labels*, one exactly recovers SupCon without changing anything in the code.

3.2.1 Easier Hard-Mining

The recurring theme of the dense target distribution $\alpha_b[j] \in [0, 1]$ also helps in mining instances for contrast in a given batch. Given an anchor, the positive and negative pairs to contrast against is now a blurred concept; every instance is some amount of positive and some amount of negative depending on its soft label's distance to that of the anchor.

A facet of this is actually already seen in Figure 1. Since there is no "4" in the single label vector on the left side, the 4th class has no representative in this particular batch to contrast against. On the right side, however, the 4th class has representation in *every* instance under soft labels. This emphasizes the need for large batch sizes in (un)supervised contrastive learning (and lack thereof in our framework) and this need only gets amplified with increasing number of classes. Again, this should be seen as motivation to collect multiple annotations (and our method's effective way of utilizing them) rather than comparing contrastive learning paradigms.

We formally argue for this ease in mining from a gradient analysis perspective. Let i be the index of the anchor in a batch of size B. Let $\boldsymbol{z}_j \coloneqq f(x_j)$ be the embedding vectors for all $j \in [B]$. Let $\mathcal{L}_i$ denote the per-anchor loss of $\mathcal{L}_{\text{SDSL}}$, i.e., a summand in equation 8. The negative gradient vector of the i^{th} anchor's loss with respect to its embedding is given by

$$-\frac{\partial \mathcal{L}_i}{\partial \boldsymbol{z}_i} = \frac{1}{\tau} \sum_{j \neq i} \alpha_i[j] \left(\boldsymbol{z}_j - \sum_{k \neq i} w_{ik} \boldsymbol{z}_k \right), \tag{9}$$

where $w_{ik} = \frac{\exp(\langle \boldsymbol{z}_i, \boldsymbol{z}_k \rangle / \tau)}{\sum_{\substack{k'=1 \\ k' \neq i}}^{B} \exp(\langle \boldsymbol{z}_i, \boldsymbol{z}_{k'} \rangle / \tau)}$ are the current similarity-based weights. In case of SupCon, $\alpha_i[j]$ are 0 when i and j are not of the same class, i.e., no direct gradient information is attained from those instances. In our case, on the other hand, $\alpha_i[j]$ takes continuous values in $[0, 1]$, which allows the machine to also *directly* exploit the non-anchors. For example, in Figure 2, let the anchor be the image of a Donskoy cat in the middle for a cat/dog classification task and let the batch size be 4. The image is arguably an ambiguous one, i.e., it carries some features of a dog, e.g., the lower face part. Under hard labels, it is a cat; therefore, the image in the bottom right which is clearly of a cat is the *only* instance whose embedding vector will have a direct effect on the gradient, hence the dashed orange vector with high magnitude pulling it towards. Under soft labels, however, there is simply more information through which the apparent ambiguity is reflected, and thanks to the now-available contribution from the other two images at the top and bottom left, the solid blue vector with low magnitude is the resultant (negative) gradient vector, with which the ambiguous instance stays more distinctively apart from the herd of clear images, which helps the downstream classification. A corollary of this is saving from time or memory: fix the number of epochs, SDSL will require lower batch sizes to converge, helping with memory; fix the batch size, it will require less number of epochs to converge, helping with time. This could be considered significant in, e.g., hardware constrained machine learning scenarios, e.g., Herrmann et al. (2026).

3.3 Completing the Picture

The soft contrastive learning procedure yields an $f \colon \mathcal{X} \to \mathcal{S}^d$; to achieve our goal of learning an $h \colon \mathcal{X} \to \mathcal{Y}$, we simply train a linear classifier over the learned representation with cross-entropy

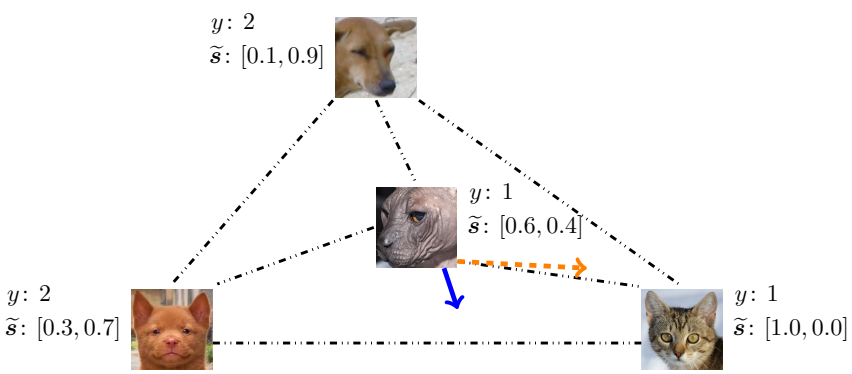


Figure 2: A hypothetical representation space with batch size 4, number of classes 2 (cat (1) versus dog (2)). Images represent their embedding vectors and the text next to the images indicate their hard (y) and soft labels ($\widetilde{\boldsymbol{s}}$). Let the one in the middle be the anchor, which is arguably an ambiguous one (can also think of a, e.g., blurred, cropped or noisy image). Equation 9 suggests the gradient updates for SupCon using hard labels is the dashed orange one, and for SDSL, the solid blue one. SupCon's relies too much on the mere fact of sharing the same hard label whereas with soft labels, SDSL's update is more cautious.

where the targets are the soft labels, i.e., $g\colon \mathcal{S}^d \to \mathbb{R}^K$ characterized by a matrix $\boldsymbol{W}_g \in \mathbb{R}^{K\times d}$. Then, $h(x) = \operatorname{argmax}_k\ g(f(x))$ is obtained.

But how does the contrastive learning with the soft labels help the linear classifier learned over the representations it produces? We show in the following proposition how one can relate the risks of these two stages that both use soft labels (its proof is provided in Appendix A.2).

Proposition 1. *Consider the random vector $\widetilde{\boldsymbol{S}} \in [0,1]^K$ whose components are the (normalized) Poisson-binomial random variables of equation 1 representing soft labels. Let $\boldsymbol{W}_g \in \mathbb{R}^{K\times d}$ defining the linear classifier g be such that $\boldsymbol{W}_g[k,:] = \mathbb{E}_{\widetilde{S}_x[k]}[f(x)]$, i.e., realized by $\sum_{n=1}^N \widetilde{s}_n[k] f(x_n) / \sum_{n=1}^N \widetilde{s}_n[k]$ as the soft label-weighted average embeddings per class $k \in [K]$ at its rows. Let the risk the embedder f suffers through $\boldsymbol{W}_g$ over soft labels in the linear classification be $R_{lin}(f) = \mathbb{E}_{\substack{x\sim X\\ \widetilde{\boldsymbol{s}}_x\sim\widetilde{\boldsymbol{S}}}}\left[-\sum_{k=1}^K \widetilde{s}_x[k]\log\frac{\exp(\boldsymbol{W}_g[k,:]f(x))}{\sum_{k'=1}^K\exp(\boldsymbol{W}_g[k',:]f(x))}\right]$, and the per-anchor per-batch SDSL risk with batch size B be $R_{SDSL}(f) = \mathbb{E}_{\substack{x,\{x_b\}\sim X^B\\ \widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}\sim\widetilde{\boldsymbol{S}}^B}}\left[\sum_{b=1}^{B-1} -\alpha_x[b]\log\frac{\exp(\langle f(x),f(x_b)\rangle/\tau)}{\sum_{b'=1}^{B-1}\exp(\langle f(x),f(x_{b'})\rangle/\tau)}\right]$. Then,*

$$R_{lin}(f) \le R_{SDSL}(f) + \log\left((B-1)^{-1}K^2\cosh^2(1/\tau)\max_{k\in[K]}P(Y=k)\right). \tag{10}$$

The bound suggests that R_{lin} decreases together with R_{SDSL}, which implies that optimizing for SDSL does help to get good representations to learn a linear classifier, if both stages leverage the soft label information derived from the multiple annotations. In Section 4.1, we will graphically show that the sharpness of the gap $R_{\text{lin}}(f) - R_{\text{SDSL}}(f)$ is stronger in our case than the supervised counterpart across the batch size, further supporting our conclusions from the gradient analysis above in Section 3.2.1 on the easier hard mining without the need for large batches.

4 Experiments

We first compare the batch size dependent convergence behavior of our method and SupCon and then experiment with five real-world datasets of image and time-series flavors.

4.1 Batch Size Dependence

Here we further support our claim based on the gradient analysis in Section 3.2.1 that our framework is less sensitive to the batch size compared to its supervised counterpart using the ground truth with

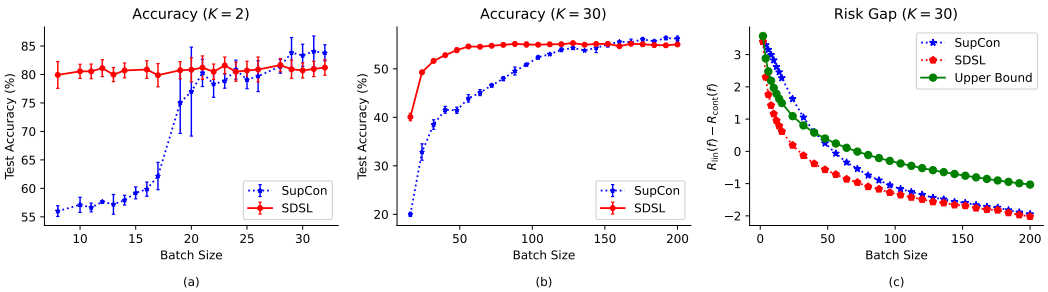


Figure 3: Different perspectives into the batch size dependence of contrastive learning paradigms. (a, b): Test accuracy trajectories of SupCon (blue, dashed) with gold ground truth and SDSL (red, solid) with soft labels on datasets with different number of classes. SupCon struggles to learn for lower batch sizes while SDSL is quite stable and able to mine informative gradient updates from the disagreement signal. (c): Risk gaps for both paradigms and the claimed upper bound (green, circled, solid) from Proposition 1 (R_{cont} in the y-label stands for either method). SDSL's always respects the bound even in the low batch size regimes and has consistently a lower gap.

synthetic experiments. Furthermore, we also allude to the risk relation in Section 3.3 by showing the stronger compliance to the upper bound for lower batch sizes.

We experiment with two datasets of different number of classes:

- 25,000 cat and dog images for binary classification from the dataset of Cukierski (2013), i.e., $K = 2$. First, 30 annotations per image are generated by instance-dependent label noise (Xia et al., 2020) over the gold labels (noise rates are uniformly and independently sampled from $[0.1, 0.45]$), then soft labels are generated based on them. The test set is of size 12,500.
- 50,000 synthetic instances with $K = 30$. Here we skip the raw annotation generation but instead directly produce soft labels: we train a shallow MLP over the gold labels and use its (softmaxed) predictions as soft labels[2]. The test set is of size 10,000.

We try two methods: SDSL (ours) and SupCon (with *gold* labels). Except for the batch size, other hyperparameters are kept constant: 20 epochs with a three-layer ReLU MLP as the encoder, $\tau = 0.1$. Both use a linear classifier on the representations, which is trained on soft labels for SDSL and on gold labels for SupCon.

We plot the test accuracy (with gold labels) of the models against the batch size B in Figure 3(a, b) over three trials. SDSL's trajectory is quite stable in both axes: it converges to the peak accuracy quite quickly with lower B and also has lower accuracy variance over trials. SupCon, on the other hand, struggles in both axes: with lower B, it considerably underperforms and has higher variance. This is expected as it relies on the law of large numbers to collect diverse instances in terms of labels, while soft labels simply offer more information already with small batches.

We also look at this batch size dependence in a different perspective by testing the risk relation in Proposition 1. We look at the risk gap $R_{\text{lin}}(f) - R_{\text{SDSL}}(f)$ against the batch size using the second dataset above with 30 classes. The setup is the same except we take $\tau = 1$. In the bound, $\max_{k \in [K]} P(Y = k) = 1/K$ as it is a balanced dataset. In Figure 3(c), we observe that for SDSL, the bound is respected for all B with achieving a lower risk gap than SupCon (especially in the low batch size regime) and eventually converging to the same value. These empirically confirm the easier hard mining without the need for large batches.

4.2 Real-world Datasets

Here, we compare our method against other multiple annotation methods in the literature over real-world datasets annotated by humans: CIFAR-10-N (Wei et al., 2022), a densely annotated CIFAR-10

[2]This implies that one does not strictly need annotations first to use our framework; all it needs is soft labels, however produced.

Table 1: Mean test accuracy (%) results with one standard deviation reported over three real-world datasets with human annotations. The best ones are highlighted in bold and second bests underlined, under statistical significance. The row of our method SDSL is shaded. "—" means the method was unable to converge to a non-constant solution even under extensive hyperparameter search.

	CIFAR-10-N	CIFAR-10-Human	IS-RC
Majority	88.45 ± 0.25	67.59 ± 0.39	80.24 ± 4.42
DoctorNet	$\underline{90.04 \pm 0.33}$	66.58 ± 0.08	82.78 ± 1.40
CrowdLayer	87.86 ± 1.36	68.06 ± 0.92	82.40 ± 3.18
TraceReg	87.06 ± 0.75	61.44 ± 0.33	81.87 ± 1.58
SoftDirect	85.59 ± 0.47	—	$\underline{84.50 \pm 1.96}$
Max-MIG	$\underline{90.52 \pm 0.17}$	71.89 ± 0.27	—
CrowdCons	84.57 ± 1.71	56.15 ± 2.48	83.67 ± 2.48
COINNet	90.37 ± 1.72	78.59 ± 0.25	77.48 ± 6.75
SDSL	$\mathbf{91.20 \pm 0.06}$	$\mathbf{79.15 \pm 0.10}$	$\mathbf{86.35 \pm 0.99}$

(Krizhevsky, 2009); CIFAR-10-Human (Peterson et al., 2019), a sparsely annotated CIFAR-10 with training and test sets flipped where we consider the 100 worst annotators in terms of deviation from the ground truth; and IS-RC (Kuna et al., 2013), a sleep-staging time series dataset annotated by medical experts. Details of the datasets are in Appendix A.3.

The methods from the literature we compare against are Max-MIG (Cao et al., 2017), DoctorNet (Guan et al., 2018), CrowdLayer (Rodrigues & Pereira, 2018), TraceReg (Tanno et al., 2019), SoftDirect (Peterson et al., 2019), CrowdCons (Yang et al., 2023) and COINNet (Nguyen et al., 2024). As machines, for all models, we use ResNet-34 for CIFAR-10-N, ResNet-18 for CIFAR-10-Human, and a simple U-Net for IS-RC. We emphasize that we do the hyperparameter validation using the *majority votes* instead of gold labels as it is more realistic unlike the previous works (e.g., LabelMe in Rodrigues & Pereira (2018), CIFAR-10-N in Nguyen et al. (2024)). Optimization details are in Appendix A.3.

The test accuracy results are presented in Table 1. Our method ranks the first in terms of generalization performance. The first two datasets share the same instance space, that of CIFAR-10 images. The first one is densely annotated, i.e., 3 annotations for *all* instances whereas the second one is a sparse one, i.e., an annotator pool of 100 people with $\max_n R_n = 8$. This shows that SDSL can handle sparse and dense datasets well. We note that thanks to soft labels, SDSL's computational resources are independent of the number of annotators. The third dataset IS-RC aims for sleep staging, a 5-class classification problem on medical time series. The literature mostly experiments with only image classification; this dataset brings a new perspective and our method shows it performs well under this scenario as well (and without augmentation). We also emphasize that SDSL has relatively lower variance than other methods, paralleling the stability observed in Figure 3(a, b).

Limitations. We found that a heavily augmented dataset that leaves little room for contrasts and one that has the vast majority of its instances labeled by a single annotator pose problems for the proposed method (exemplary datasets and results are presented in Appendix A.4).

5 Conclusion

We introduced a contrastive learning method for the classification with multiple annotations (crowds) problem by reconciling the "no label" and "too many labels" views through representation learning under the supervision of disagreement signals, a fresh perspective to the problem in the literature. We derived a soft labels-based target distribution to guide the forming of the representation space; this is a dense distribution based on the disagreement signal unlike the (un)supervised versions which are one-hot or sparse-uniform as special cases. Zooming in to the ground truth by soft labels also allowed us to lift the need for large batch sizes for hard instance mining. We theoretically supported the method by relating the risks of the contrastive and downstream linear classification phases, and the empirical findings over several real-world datasets confirmed the effectiveness of the method. The future work involves addressing the aforementioned limitations.

AI USE STATEMENT

No AI tool (generative or otherwise) is used in any part of this work.

REFERENCES

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 2015.

Han Bao, Yoshihiro Nagano, and Kento Nozawa. On the surrogate gap between contrastive and supervised losses. In *International Conference on Machine Learning*, 2022.

Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. Max-mig: an information theoretic approach for joint learning from crowds. In *International Conference on Learning Representations*, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

Will Cukierski. Dogs vs. Cats. Kaggle, 2013.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1979.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Nicholas Frosst, Nicolas Papernot, and Geoffrey Hinton. Analyzing and improving representations with the soft nearest neighbor loss. In *International Conference on Machine Learning*, 2019.

Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*, 2004.

Melody Y. Guan, Varun Gulshan, Andrew M. Dai, and Geoffrey E. Hinton. Who said what: modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*, 2018.

Andi Han, Wei Huang, Zhanpeng Zhou, Gang Niu, Wuyang Chen, Junchi Yan, Akiko Takeda, and Taiji Suzuki. On the role of label noise in the feature learning process. In *International Conference on Machine Learning*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: a survey. *Machine Learning*, 2024.

Nina Herrmann, Jan Stenkamp, Benjamin Karic, Stefan Oehmcke, and Fabian Gieseke. Boosted trees on a diet: Compact models for resource-constrained devices. In *International Conference on Learning Representations*, 2026.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

Shahana Ibrahim, Panagiotis A. Traganitis, Xiao Fu, and Georgios B. Giannakis. Learning from crowdsourced noisy labels: A signal processing perspective. *IEEE Signal Processing Magazine*, 2025.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Bogdan Kulynych, Yao-Yuan Yang, Yaodong Yu, Jarosł aw Bł asiok, and Preetum Nakkiran. What you see is what you get: Principled deep learning via distributional generalization. In *Advances in Neural Information Processing Systems*, 2022.

Samuel Kuna, Clete Kushida, James Walsh, Magdy Younes, Bethany Staley, Alex Hanlon, Allan Pack, Grace Pien, and Atul Malhotra. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep*, 2013.

Jingzheng Li, Hailong Sun, and Jiyi Li. Beyond confusion matrix: learning from multiple annotators with awareness of instance features. *Machine Learning*, 2022.

Tri Nguyen, Shahana Ibrahim, and Xiao Fu. Noisy label learning with instance-dependent outliers: Identifiability via crowd wisdom. In *Advances in Neural Information Processing Systems*, 2024.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 2010.

Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Every annotation counts: Multi-label deep supervision for medical image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Filipe Rodrigues and Francisco C. Pereira. Deep learning from crowds. In *AAAI Conference on Artificial Intelligence*, 2018.

Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Gaussian process classification and active learning with multiple annotators. In *International Conference on Machine Learning*, 2014.

Filipe Rodrigues, Mariana Lourenço, Bernardete Ribeiro, and Francisco C. Pereira. Learning supervised topic models for classification and regression from crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

Bryan Russell, Antonio Torralba, Kevin Murphy, and William Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 2008.

Ruslan Salakhutdinov and Geoff Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In *International Conference on Artificial Intelligence and Statistics*, 2007.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, 2016.

Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Ryota Ushio, Takashi Ishida, and Masashi Sugiyama. Practical estimation of the optimal classification error with soft labels and calibration. In *International Conference on Learning Representations*, 2026.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, 2009.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems*, 2020.

Hang Yang, Xunbo Li, and Witold Pedrycz. Learning from crowds with contrastive representation. *IEEE Access*, 2023.

A APPENDIX

A.1 SOFT LABELS ARE USEFUL

Here, we show the usefulness of soft labels which are central to our framework by proposing a (straightforward) estimator of the Bayes risk of a multiclass classification task. The Bayes risk (error) is given by

$$\mathcal{E}^{\text{Bayes}}(X, Y) = \mathbb{P}(Y \neq \operatorname*{argmax}_k \mathbb{P}(Y = k \mid X)), \tag{11}$$

i.e., the probability of the oracle getting the hard decision wrong due to stochasticity. Our estimator simply substitutes the average of soft labels per class in lieu of the unknown posterior of $Y \mid X$.

Proposition 2. *$\widehat{\mathcal{E}}^{\text{Bayes}}(\{\widetilde{\boldsymbol{y}}_n\}_{n=1}^N) = 1 - \frac{1}{N}\sum_{n=1}^N \max_k \widetilde{s}_n[k]$ is a biased but asymptotically (in the number of annotations per instance) unbiased and consistent estimator of $\mathcal{E}^{\text{Bayes}}(X, Y)$ with Bias $\left(\widehat{\mathcal{E}}^{\text{Bayes}}, \mathcal{E}^{\text{Bayes}}\right) := \mathbb{E}_{X,A,P}\left[\widehat{\mathcal{E}}^{\text{Bayes}} - \mathcal{E}^{\text{Bayes}}\right]$ and Var $\left(\widehat{\mathcal{E}}^{\text{Bayes}}\right) := \mathbb{E}_{X,A,P}\left[\left(\widehat{\mathcal{E}}^{\text{Bayes}} - \mathbb{E}_{X,A,P}\left[\widehat{\mathcal{E}}^{\text{Bayes}}\right]\right)^2\right]$ satisfying*

$$-\frac{\sqrt{2\log K}}{N}\sum_{n=1}^N \frac{1}{\sqrt{R_n}} \leq \text{Bias}\left(\widehat{\mathcal{E}}^{\text{Bayes}}, \mathcal{E}^{\text{Bayes}}\right) \leq 0 \tag{12}$$

$$\text{Var}\left(\widehat{\mathcal{E}}^{\text{Bayes}}\right) \leq \frac{K}{4N^2}\sum_{n=1}^N \frac{1}{R_n}. \tag{13}$$

Remark. *Ushio et al. (2026) recently proposed this estimator for binary classification; we generalize it to multiclass scenarios and provide its bias and variance bounds (slightly improved for $K = 2$).*

Remark. *If one has access to $\mathbb{P}(Y \mid X, A)$ for each annotator and instance, one can do better in terms of the bounds. For instance, the (exact) variance proxy of the (normalized) Poisson-binomial $\frac{1}{R_n}\sum_{r=1}^{R_n} \mathbb{1}\left[\widetilde{y}_n[r] = k\right]$ with expectations $\mathbb{P}(Y = k \mid X = x_n, A = r) =: p_{r,k,n}$ is given by*

$$\frac{1}{R_n^2}\sum_{r=1}^{R_n} \frac{\frac{1}{2} - p_{r,k,n}}{\log\left(\frac{1}{p_{r,k,n}} - 1\right)}, \tag{14}$$

which we simply upper bound by $1/4$ per each term, i.e., $p_{r,k,n} = 0.5\ \forall r, k, n$ (which is only sharp at $K = 2$). One heuristic could be assuming $p_{r,k,n} \in \{0, 1\}\ \forall r, k, n$ if $R_n \geq 2$ and all annotations unanimously agree on the label of the instance. This yields 0 for the corresponding terms in the summation above, which could dramatically reduce the bias, for example.

The proof of this proposition is provided in Appendix A.2.

The bias and variance bounds show how to scale dataset dimensions to get a good estimate of the Bayes risk, which gives an upper bound on the generalization accuracy any machine can achieve (one minus the Bayes risk). Note that in the case of single labels, i.e., the traditional classification scenarios, this upper bound on generalization attains its trivial value 100%, which is unrealistic. By using soft labels based on multiple annotations, we are able to obtain a more precise estimate of the best achievable performance. For example, assuming a good estimate of 75%, a machine achieving a 72% score is deemed near perfect which may otherwise be seen not good enough.

Relatedly, this supports the "truth is a lie" perspective (Aroyo & Welty, 2015). While there are certainly instances which clearly belong to a particular class, there are also instances in the gray area, e.g., those having features of more than one class such as a dog that looks like a cat or an artifact in a medical signal. It is then unsettling to force a single ground truth per instance and lose valuable information. Of course, classification with a reject option (Hendrickx et al., 2024), distributional generalization (Kulynych et al., 2022) etc. are concepts based on this discrepancy but none of them specifically focus on multiple annotations. In this paper, we base our contrastive learning framework on soft labels to drive the shaping of the representation space through this valuable information that is the disagreement signal.

A.2 PROOFS

Proof of Proposition 1. We first present a Lemma from Bao et al. (2022) without proof.

Lemma 1 (Bao et al. (2022), Lemma 3)**.** *For* $\boldsymbol{z} \in [-L^2, L^2]^N$,

$$2\log N \leq \mathrm{LSE}(\boldsymbol{z}) + \mathrm{LSE}(-\boldsymbol{z}) \leq 2\log(N\cosh(L^2)), \tag{15}$$

where $\mathrm{LSE}(\boldsymbol{z}) = \log\sum_{n=1}^{N}\exp(z_n)$ *is the log-sum-exp function.*

We have

$$R_{\mathrm{SDSL}}(f) = \mathbb{E}_{\substack{x,\{x_b\}\sim X^B \\ \widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}\sim\widetilde{\boldsymbol{S}}^B}}\left[\sum_{b=1}^{B-1} -\alpha_x[b]\log\frac{\exp(\langle f(x), f(x_b)\rangle/\tau)}{\sum_{b'=1}^{B-1}\exp(\langle f(x), f(x_{b'})\rangle/\tau)}\right] \tag{16}$$

$$\overset{\text{(i)}}{=} \mathbb{E}_{\substack{x,\{x_b\} \\ \widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\sum_{b=1}^{B-1} -\alpha_x[b]\Big(\langle f(x), f(x_b)\rangle/\tau - \log\sum_{b'=1}^{B-1}\exp(\langle f(x), f(x_{b'})\rangle/\tau)\Big)\right] \tag{17}$$

$$\begin{aligned}\overset{\text{(ii)}}{=} &-\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\,\mathbb{E}_{\substack{\{x_b\} \\ \mid\, x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\langle f(x), f(x_b)\rangle/\tau\right]\right] \\ &+\mathbb{E}_{\substack{x,\{x_{b'}\} \\ \widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_{b'}}\}}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\log\sum_{b'=1}^{B-1}\exp(\langle f(x), f(x_{b'})\rangle/\tau)\right]\end{aligned} \tag{18}$$

$$\overset{\text{(iii)}}{=} -\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\,\mathbb{E}_{\substack{\{x_b\} \\ \mid\, x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\langle f(x), f(x_b)\rangle/\tau\right]\right] + \mathbb{E}_{\substack{x,\{x_{b'}\} \\ \widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_{b'}}\}}}\left[\log\sum_{b'=1}^{B-1}\exp(\langle f(x), f(x_{b'})\rangle/\tau)\right] \tag{19}$$

$$\begin{aligned}\overset{\text{(iv)}}{\geq} &-\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\,\mathbb{E}_{\substack{\{x_b\} \\ \mid\, x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\langle f(x), f(x_b)\rangle/\tau\right]\right] \\ &+\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_{b'}}\}}\left[\log\sum_{b'=1}^{B-1}\exp\left(\mathbb{E}_{\{x_b\}\mid x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_{b'}}\}}\left[\langle f(x), f(x_{b'})\rangle/\tau\right]\right)\right]\end{aligned} \tag{20}$$

$$\overset{\text{(v)}}{=} -\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\,\mathbb{E}_{\substack{\{x_b\} \\ \mid\, x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\langle f(x), f(x_b)\rangle/\tau\right]\right] + \mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_{b'}}\}}\left[\log\sum_{b'=1}^{B-1}\exp\left(\langle f(x), \boldsymbol{W}_g[k_{b'},:]\rangle/\tau\right)\right] \tag{21}$$

$$\begin{aligned}\overset{\text{(vi)}}{\geq} &-\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\,\mathbb{E}_{\substack{\{x_b\} \\ \mid\, x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\langle f(x), f(x_b)\rangle/\tau\right]\right] \\ &-\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_{b'}}\}}\left[\log\sum_{b'=1}^{B-1}\exp\left(\langle -f(x), \boldsymbol{W}_g[k_{b'},:]\rangle/\tau\right)\right] + 2\log(B-1)\end{aligned} \tag{22}$$

$$\begin{aligned}\overset{\text{(vii)}}{\geq} &-\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\,\mathbb{E}_{\substack{\{x_b\} \\ \mid\, x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\langle f(x), f(x_b)\rangle/\tau\right]\right] \\ &-\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x}\left[\log\sum_{b'=1}^{B-1}\mathbb{E}_{\{\widetilde{\boldsymbol{s}}_{x_{b'}}\}\mid x,\widetilde{\boldsymbol{s}}_x}\left[\exp\left(\langle -f(x), \boldsymbol{W}_g[k_{b'},:]\rangle/\tau\right)\right]\right] + 2\log(B-1)\end{aligned} \tag{23}$$

$$\begin{aligned}\overset{\text{(viii)}}{\geq} &-\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\,\mathbb{E}_{\substack{\{x_b\} \\ \mid\, x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\langle f(x), f(x_b)\rangle/\tau\right]\right] \\ &-\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x}\left[\log\left(\sum_{b'=1}^{B-1}\Big(\max_{k\in[K]}P(Y=k)\Big)\sum_{k=1}^{K}\exp\left(\langle -f(x), \boldsymbol{W}_g[k,:]\rangle/\tau\right)\right)\right] + 2\log(B-1)\end{aligned} \tag{24}$$

$$\overset{\text{(ix)}}{=} -\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\,\mathbb{E}_{\substack{\{x_b\}\\ \mid\, x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\langle f(x),f(x_b)\rangle/\tau\right]\right] - \mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x}\left[\log\left((B-1)\Big(\max_{k\in[K]}P(Y=k)\Big)\sum_{k=1}^{K}\exp\left(\langle -f(x),\boldsymbol{W}_g[k,:]\rangle/\tau\right)\right)\right] + 2\log(B-1) \tag{25}$$

$$\overset{\text{(x)}}{=} -\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\,\mathbb{E}_{\substack{\{x_b\}\\ \mid\, x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\langle f(x),f(x_b)\rangle/\tau\right]\right] - \mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x}\left[\log\sum_{k=1}^{K}\exp\left(\langle -f(x),\boldsymbol{W}_g[k,:]\rangle/\tau\right)\right] + \log\frac{B-1}{\max_{k\in[K]}P(Y=k)} \tag{26}$$

$$\overset{\text{(xi)}}{\geq} -\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b]\,\mathbb{E}_{\substack{\{x_b\}\\ \mid\, x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}}\left[\langle f(x),f(x_b)\rangle/\tau\right]\right] + \mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x}\left[\log\sum_{k=1}^{K}\exp\left(\langle f(x),\boldsymbol{W}_g[k,:]\rangle/\tau\right)\right] + \log\frac{B-1}{\max_{k\in[K]}P(Y=k)} - 2\log(K\cosh^2(1/\tau)) \tag{27}$$

$$\overset{\text{(xii)}}{=} -\mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x,\{\widetilde{\boldsymbol{s}}_{x_b}\}}\left[\sum_{b=1}^{B-1}\alpha_x[b][\langle f(x),\boldsymbol{W}_g[k_b,:]\rangle/\tau\right] + \mathbb{E}_{x,\widetilde{\boldsymbol{s}}_x}\left[\log\sum_{k=1}^{K}\exp\left(\langle f(x),\boldsymbol{W}_g[k,:]\rangle/\tau\right)\right] + \log\frac{B-1}{\max_{k\in[K]}P(Y=k)} - 2\log(K\cosh^2(1/\tau)) \tag{28}$$

$$\overset{\text{(xiii)}}{=} -\sum_{k=1}^{K}\widetilde{s}[k]\mathbb{E}_x\left[\langle f(x),\boldsymbol{W}_g[k,:]\rangle/\tau\right] + \mathbb{E}_x\left[\sum_{k=1}^{K}\widetilde{s}_x[k]\log\sum_{k'=1}^{K}\exp\left(\langle f(x),\boldsymbol{W}_g[k',:]\rangle/\tau\right)\right] + \log\frac{B-1}{\max_{k\in[K]}P(Y=k)} - 2\log(K\cosh^2(1/\tau)) \tag{29}$$

$$= R_{\text{lin}}(f) - \log\Big((B-1)^{-1}K^2\cosh^2(1/\tau)\max_{k\in[K]}P(Y=k)\Big), \tag{30}$$

where we used

(i): log distributes over multiplication;

(ii): expectation is linear;

(iii): $\sum_{b=1}^{B-1}\alpha_x[b]=1$ and the inner summation is independent of b;

(iv): law of total expectation (over the non-anchors) and Jensen's inequality with the convex LSE;

(v): inner product is (bi)linear and the definition of $\boldsymbol{W}_g$

(vi): left side of Lemma 1 with $z[b']=\langle f(x),\boldsymbol{W}_g[k_{b'},:]\rangle/\tau$ for $b'\in[B-1]$;

(vii): law of total expectation (over the soft labels) and Jensen's inequality with the convex $-\log$;

(viii): the fact $\mathbb{E}_D[h(D)]\leq\sum_{d\in\mathcal{D}}\max_d\mathbb{P}(D=d)h(d)$ for a discrete D with its support on $\mathcal{D}$;

(ix): summation in the second expectation is independent of b';

(x): take the constants outside;

(xi): right side of Lemma 1 with $z[k] = \langle -f(x), \boldsymbol{W}_g[k,:] \rangle / \tau$ for $k \in [K]$ and since $||f(x)||_2 = 1$, then $||\boldsymbol{W}_g||_2 \leq 1$ due to Jensen's inequality with the convex norm, which yields $|\langle -f(x), \boldsymbol{W}_g[k,:] \rangle / \tau| \leq 1/\tau$ due to homogeneity of the inner product and the Cauchy-Schwarz inequality;

(xii): law of total expectation (over the non-anchors) and (bi)linearity of the inner product;

(xiii): law of total expectation (over the soft labels);

and the last equality follows from the definition of $R_{\text{lin}}(f)$. □

Proof of Proposition 2. Bounding the bias. We first start with the upper bound:

$$\mathbb{E}_{X,A,P}\left[\widehat{\mathcal{E}}^{\text{Bayes}}\right] = \mathbb{E}_{X,A,P}\left[1 - \frac{1}{N}\sum_{n=1}^{N}\max_k \widetilde{s}_n[k]\right] \tag{31}$$

$$= \mathbb{E}_{X,A,P}\left[1 - \frac{1}{N}\sum_{n=1}^{N}\max_k \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{1}\left[\widetilde{y}_n[r] = k\right]\right] \tag{32}$$

$$= 1 - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{X,A,P}\left[\max_k \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{1}\left[\widetilde{y}_n[r] = k\right]\right] \tag{33}$$

$$= 1 - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{X,A}\left[\mathbb{E}_{P|X,A}\left[\max_k \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{1}\left[\widetilde{y}_n[r] = k\right]\right]\right] \tag{34}$$

$$\leq 1 - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{X,A}\left[\max_k \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{E}_{P|X,A}\left[\mathbb{1}\left[\widetilde{y}_n[r] = k\right]\right]\right] \tag{35}$$

$$= 1 - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{X,A}\left[\max_k \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{P}(Y = k \mid X = x_n, A = r)\right] \tag{36}$$

$$= 1 - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{X}\left[\mathbb{E}_{A|X}\left[\max_k \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{P}(Y = k \mid X = x_n, A = r)\right]\right] \tag{37}$$

$$\leq 1 - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{X}\left[\max_k \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{E}_{A|X}\left[\mathbb{P}(Y = k \mid X = x_n, A = r)\right]\right] \tag{38}$$

$$= 1 - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{X}\left[\max_k \mathbb{P}(Y = k \mid X = x_n)\right] \tag{39}$$

$$= 1 - \frac{1}{N}\sum_{n=1}^{N}\int \max_k \mathbb{P}(Y = k \mid X = x)\, dp_X(x) \tag{40}$$

$$= 1 - \int \max_k \mathbb{P}(Y = k \mid X = x)\, dp_X(x) \tag{41}$$

$$= \mathcal{E}^{\text{Bayes}}, \tag{42}$$

where we repeatedly used the law of total expectation and Jensen's inequality with the convex $\max$. We obtained $\mathbb{E}_{X,A,P}\left[\widehat{\mathcal{E}}^{\text{Bayes}}\right] \leq \mathcal{E}^{\text{Bayes}}$, i.e., the bias is always nonpositive. Note that equality happens if and only if the categorical posterior distribution of $Y \mid X$ is almost surely constant, which never happens in practice but is still a technical possibility.

For the lower bound, we have:

$$-\text{Bias}(\widehat{\mathcal{E}}^{\text{Bayes}}, \mathcal{E}^{\text{Bayes}}) = \mathbb{E}_{X,A,P}\left[\frac{1}{N}\sum_{n=1}^{N}\left(\max_k \widetilde{s}_n[k] - \max_k \mathbb{P}(Y = k \mid X = x_n)\right)\right] \tag{43}$$

$$= \mathbb{E}_{X,A,P}\left[\frac{1}{N}\sum_{n=1}^{N}\left(\max_k \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{1}\left[\widetilde{y}_n[r]=k\right] - \max_k \mathbb{P}(Y=k \mid X=x_n)\right)\right] \tag{44}$$

$$\overset{(*)}{\leq} \mathbb{E}_{X,A,P}\left[\frac{1}{N}\sum_{n=1}^{N}\max_k\left(\frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{1}\left[\widetilde{y}_n[r]=k\right] - \mathbb{P}(Y=k \mid X=x_n)\right)\right] \tag{45}$$

$$= \mathbb{E}_{X,A,P}\left[\frac{1}{N}\sum_{n=1}^{N}\max_k\left(\left(\frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{1}\left[\widetilde{y}_n[r]=k\right] - \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{P}\left(Y=k \mid X=x_n, A=r\right)\right)\right.\right.$$
$$\left.\left. + \left(\frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{P}\left(Y=k \mid X=x_n, A=r\right) - \mathbb{P}(Y=k \mid X=x_n)\right)\right)\right] \tag{46}$$

$$\overset{(**)}{\leq} \mathbb{E}_{X,A,P}\left[\frac{1}{N}\sum_{n=1}^{N}\max_k\left(\frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{1}\left[\widetilde{y}_n[r]=k\right] - \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{P}\left(Y=k \mid X=x_n, A=r\right)\right)\right.$$
$$\left. + \max_k\left(\frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{P}\left(Y=k \mid X=x_n, A=r\right) - \mathbb{P}(Y=k \mid X=x_n)\right)\right] \tag{47}$$

$$= \mathbb{E}_{X,P}\left[\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{A|X,P}\left[\max_k\left(\frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{1}\left[\widetilde{y}_n[r]=k\right] - \frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{P}\left(Y=k \mid X=x_n, A=r\right)\right)\right]\right]$$
$$+ \mathbb{E}_{X,A}\left[\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{P|X,A}\left[\max_k\left(\frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{P}\left(Y=k \mid X=x_n, A=r\right) - \mathbb{P}(Y=k \mid X=x_n)\right)\right]\right], \tag{48}$$

where we used $\max_k f(k) - \max_k g(k) \leq \max_k(f(k)-g(k))$ at (*) and $\max_k(f(k)+g(k)) \leq \max_k f(k) + \max_k g(k)$ at (**) (maximums are attained since the search is over a finite set, $[K]$). We now have two expectations of the form $\mathbb{E}\left[\max_k Z_k - \mathbb{E}[Z_k]\right]$. Since both Z_k's have the range $[0,1]$, i.e., bounded, they are trivially also subgaussian random variables. Therefore, we may use the following maximal inequality over subgaussians: $\mathbb{E}\left[\max_k Z_k - \mathbb{E}[Z_k]\right] \leq \sigma\sqrt{2\log K}$, where σ^2 is an upper bound on the variance proxy of Z_k's. We note that each Z_k in our case is an average of R_n bounded and (conditionally) independent random variables. By Hoeffding's lemma, their variance proxy is upper bounded by $(1-0)^2/4 = 1/4$, yielding the upper bound $\sqrt{\frac{1}{4R_n}}\sqrt{2\log K}$ for each $n \in [N]$. Averaging over all n's yield the desired $\frac{\sqrt{2\log K}}{N}\sum_{n=1}^{N}\frac{1}{\sqrt{R_n}}$.

Bounding the variance.

$$\mathrm{Var}\left(\widehat{\mathcal{E}}^{\text{Bayes}}\right) = \mathrm{Var}\left(1 - \frac{1}{N}\sum_{n=1}^{N}\max_k \widetilde{s}_n[k]\right) \tag{49}$$

$$= \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{Var}\left(\max_k \widetilde{s}_n[k]\right) \tag{50}$$

$$\overset{(*)}{\leq} \frac{1}{N^2}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{Var}\left(\widetilde{s}_n[k]\right) \tag{51}$$

$$= \frac{1}{N^2}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{Var}\left(\frac{1}{R_n}\sum_{r=1}^{R_n}\mathbb{1}\left[\widetilde{y}_n[r]=k\right]\right) \tag{52}$$

$$= \frac{1}{N^2}\sum_{n=1}^{N}\frac{1}{R_n^2}\sum_{k=1}^{K}\sum_{r=1}^{R_n}\mathrm{Var}\left(\mathbb{1}\left[\widetilde{y}_n[r]=k\right]\right) \tag{53}$$

$$\overset{(**)}{\leq} \frac{1}{N^2}\sum_{n=1}^{N}\frac{1}{R_n^2}\sum_{k=1}^{K}\sum_{r=1}^{R_n}\frac{1}{4} \tag{54}$$

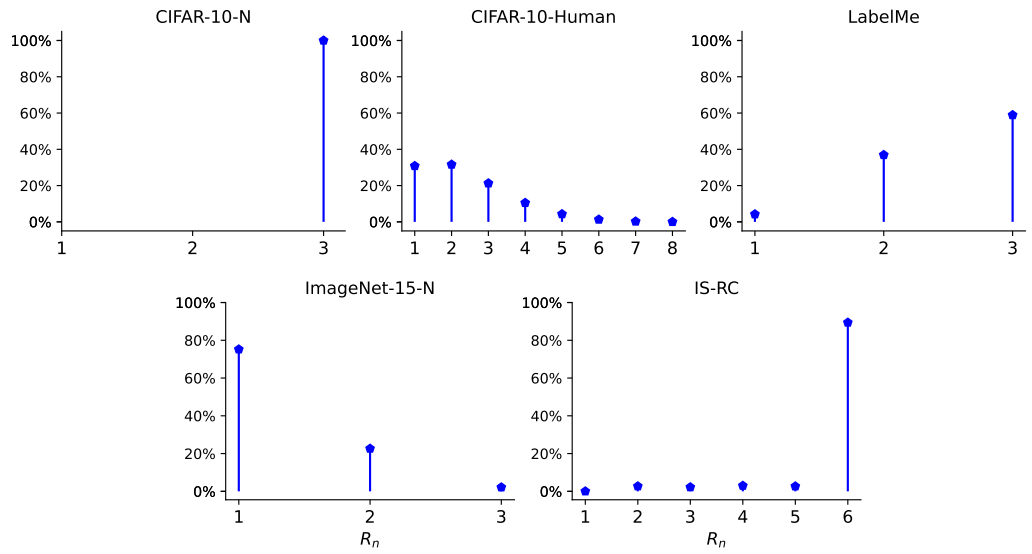


Figure 4: Histograms of the number of annotations per instance R_n over the datasets.

$$= \frac{K}{4N^2} \sum_{n=1}^{N} \frac{1}{R_n}, \tag{55}$$

where we used the Efron-Stein inequality at (*) and the fact that a Bernoulli random variable has variance at most $1/4$ at (**). □

A.3 Details of the Setup of the Experiments

Datasets.

- CIFAR-10-N (Wei et al., 2022): Based on the well-known CIFAR-10 dataset (Krizhevsky, 2009). 10-class classification of 32x32x3 RGB images with 50,000 training and 10,000 test samples. Each training instance has 3 annotations.
- CIFAR-10-Human (Peterson et al., 2019): Based on CIFAR-10 but training and test sets are flipped, i.e., 10,000 training and 50,000 test samples. We take the worst 100 annotators as a pool based on their discrepancy from the gold truth, as taking all of the 2,571 annotators makes it a less challenging problem in terms of learning from multiple annotations (e.g., when all are taken, the majority votes agree with the ground truth $99\%+$). This reduces the number of training samples to 8,673. Note that the dataset is not densely annotated, e.g., maximum annotation an instance gets is 8 (see also Figure 4).
- IS-RC (Kuna et al., 2013): 5-class classification task of sleep staging annotated by 6 medical experts (mostly densely annotated; see Figure 4). It consists of 70 patients' polysomnographic records with 7 channels. We use 68 patients' data (*AL_36_112108* and *AL_57_111009* were dismissed due to missing data) and all channels. Each recording is divided into 30 seconds epochs with sampling rate 128Hz. The categories to classify into are wakefulness, REM stage an N1-3 stages of sleep. 20 epochs using SGD with batch size of 1 patient is employed. 48 patients are spared for training, 10 for validation and 10 for testing. Note that there is no ground truth even in the testing split of this dataset; we used the majority voting as a proxy.
- LabelMe (Rodrigues et al., 2017): Based on the annotation tool LabelMe (Russell et al., 2008), it primarily consists of 1,000 images with a pool of 59 annotators (sparsely annotated, i.e., $\max_n R_n = 3$; see Figure 4) for 8-class classification. However, the dataset has been augmented by a factor of 10 to obtain 10,000 images, which are then preprocessed with a pretrained VGG-16 (Simonyan & Zisserman, 2015). This makes it a heavily augmented dataset, i.e., 90% of the images are copies.

Table 2: Mean test accuracy (%) results with one standard deviation reported over two real-world datasets with human annotations. The best ones are highlighted in bold and second bests underlined, under statistical significance. The row of our method SDSL is shaded.

	LabelMe	ImageNet-15-N
Majority	75.23 ± 0.81	66.64 ± 2.95
DoctorNet	$\underline{80.78 \pm 0.77}$	66.75 ± 0.25
CrowdLayer	$\mathbf{81.62 \pm 0.74}$	67.36 ± 2.66
TraceReg	53.59 ± 0.48	62.99 ± 2.29
SoftDirect	75.95 ± 0.28	70.02 ± 1.76
Max-MIG	79.85 ± 0.50	$\mathbf{77.47 \pm 0.91}$
CrowdCons	75.84 ± 4.04	64.16 ± 0.16
COINNet	73.39 ± 0.48	$\underline{69.03 \pm 9.18}$
SDSL	79.46 ± 0.40	$\underline{70.65 \pm 0.99}$

- ImageNet-15-N (Nguyen et al., 2024): Based on the well-known ImageNet dataset (Deng et al., 2009), the authors take a subset of it with 2,514 images with 15 classes and employ an annotator pool of size 100. However, the annotations are particularly sparse, e.g., $\sim 80\%$ of the dataset is singly-annotated (see Figure 4). The images are preprocessed with CLIP (Radford et al., 2021).

Machines.

- ResNet-34 (He et al., 2016) is used for all methods in CIFAR-10-N.
- ResNet-18 (He et al., 2016) is used for all methods in CIFAR-10-Human.
- A simple UNet (Ronneberger et al., 2015) of 12 down convolutions, an LSTM (Hochreiter & Schmidhuber, 1997) in the middle and 12 up convolutions back is used for all methods in IS-RC.
- A three-layer ReLU MLP (1024, 512, 256 neurons, respectively) is used for all methods in LabelMe and ImageNet-15-N (which come as preprocessed as described above).

Preprocessing. All datasets undergo standardization (subtract the mean, divide by the standard deviation; both statistics are obtained from the training split). For CIFAR-10-N/Human, we employ random crops and random horizontal flips (not for augmentation; in-place).

Optimization. For all datasets except IS-RC, 20% of the training data is spared for validation, for which we take the majority voting and monitor the accuracy as the early stopping metric (patience of 30 epochs) and best model to load back. Similarly, 200 epochs using SGD with a batch size of 128 is employed for all except IS-RC. (IS-RC's particulars are mentioned above in the *Datasets* part.) As for hyperparameters, the common parameter of learning rate is tuned for all models and datasets across the grid $\{\eta/100, \eta/10, \eta, 10\eta, 100\eta\}$ with $\eta = 0.01$. Noncritical hyperparameters, if any, are taken from the respective papers/source code as is (e.g., uniform distribution for $\boldsymbol{p}_0$ of Max-MIG). For our model, the only hyperparameter is τ, which we find to be fine at 0.01 across all tasks with insensitivity. Also, we repeat the trials for both the contrastive part and the linear classification part, e.g., for each 3 trials of the contrastive part, 3 linear classifiers are employed and the standard deviation reported is over the mean of each of these 3 linear classifier's trials. For the other methods, validation is done for more sensitive hyperparameters in addition to the learning rate, e.g., λ of TraceReg is searched in $\{1\mathrm{e}{-4}, 1\mathrm{e}{-3}, 1\mathrm{e}{-2}\}$ as well as μ_1, μ_2 of COINNet. We use CrowdLayer in CrowdCons while optimizing for their Equation 13. In certain cases (e.g., Max-MIG in IS-RC), we were unable to make it converge to a non-constant solution, i.e., one that does not diverge to always predict a certain class even though we extended the hyperparameter grid greatly, e.g., for Max-MIG in IS-RC, we tried for its "left" and "right" learning rates the set $\{1\mathrm{e}{-6}, 1\mathrm{e}{-5}, 1\mathrm{e}{-4}, 1\mathrm{e}{-3}, 5\mathrm{e}{-3}, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0\}$ producted by itself.

A.4 Limiting Cases

Here, we demonstrate limitations of our model using two real-world datasets LabelMe (Rodrigues et al., 2017) and ImageNet-15-N (Nguyen et al., 2024), whose descriptions are detailed in Appendix

A.3. The results are in Table 2. While our method does not perform considerably worse, it lacks behind in comparison. We conjecture about this as follows based on the inner workings of the method.

- LabelMe: this is a heavily augmented dataset. Going from 1,000 to 10,000 images by augmentation results in 90% of the dataset being copies. This means they share the exact same annotations as well, leading to the same soft label information. Leaving little room for contrasts then harms the proposed method as it is based on contrastive learning (as opposed to the competing methods).
- ImageNet-15-N: this is an extremely sparsely annotated dataset. From Figure 4, we see that $\sim 80\%$ of the dataset is singly-labeled. In other words, the problem is almost of a (single) label noise nature. However, when $R_n = 1 \ \forall n$, our method falls back to SupCon, which does nothing special for label noise, i.e., the small number of annotations dominating the dataset results in a negligible disagreement signal which largely disables the guidance via soft labels.