Machine Learning (2023) 112:1823–1845
https://doi.org/10.1007/s10994-022-06185-w

Learning from crowds with sparse and imbalanced annotations

Ye Shi[1] · Shao-Yuan Li[1] · Sheng-Jun Huang[1]

Received: 29 October 2021 / Revised: 1 March 2022 / Accepted: 9 May 2022 /
Published online: 14 June 2022

Abstract

Traditional supervised learning requires ground truth labels for training, whose collection however is difficult in many cases. Recently, crowdsourcing has established itself as an efficient labeling solution by resorting to non-expert crowds. To reduce the labeling error effects, one common practice is to distribute each instance to multiple workers, whereas each worker only annotates a subset of data, resulting in the sparse annotation phenomenon. In this paper, we show that when meeting with class-imbalance, i.e., even when the groundtruth labels are slightly imbalanced, the sparse annotations are prone to be skewly distributed and would bias the learning algorithm severely. To combat this issue, we propose one Distribution Aware Self-training based Crowdsourcing learning (DASC) approach, which supplements the sparse annotations by adding confident pseudo-annotations and at the same time re-balancing the annotation distribution. Specifically, we propose one distribution aware confidence measure to select the most confident pseudo-annotations, with minority/majority classes selected more/less frequently. As a universal framework, DASC is applicable to various crowdsourcing methods for consistent performance gains. We conduct extensive experiments over real-world crowdsourcing benchmarks, from slight to heavy imbalance ratio, with various annotation sparsity levels, and show that DASC substantially improves previous crowdsourcing models by 2%-20% absolute test accuracy, and yields much more balanced annotations.

Keywords Crowdsourcing · Sparse annotations · Class-imbalance · Self-training

Editors: Bo Han, Tongliang Liu, Quanming Yao, MingmingGong, Gang Niu, Ivor W. Tsang, Masashi Sugiyama.

✉ Shao-Yuan Li
lisy@nuaa.edu.cn

Ye Shi
shiye1998@nuaa.edu.cn

Sheng-Jun Huang
huangsj@nuaa.edu.cn

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

1 Introduction

Recently deep neural networks (DNNs) have achieved significant success in various domains, whose achievements rely heavily on large curated datasets with manually verified labels (Girshick et al., 2014; He et al., 2016; Krizhevsky et al., 2012). In many applications, instead of perfect supervision, weak supervision is often available due to the labeling cost or difficulty, such as incomplete supervision (Xu and Guo, 2021), inexact supervision (Li et al., 2021; Feng et al., 2021), cross-domain supervision (Pang et al., 2021), only positive-and-unlabeled supervision (Su et al., 2021) and inaccurate supervision (Li et al., 2021). Fortunately, datasets with low quality annotations such as web crawling images with automatic label extractions (Krause et al., 2016; Xiao et al., 2015) or crowdsourcing annotations(Horvitz, 2007; Thierry et al., 2010) are extensive. In this paper, we concern ourselves with developing methodologies for crowdsourcing annotated tasks. To alleviate the labeling difficulty issue, crowdsourcing distributes the labeling task to multiple easy-to-access crowd workers and aggregates over them to alleviate the labeling error. With the advent of crowdsourcing services such as Amazon Mechanical Turk (AMT),[1] Crowdflower[2] and reCAPTCHA,[3] crowdsourcing has been used in various fields including sentiment classification (Snow et al. 2008), medical diagnosis (Raykar et al., 2010; Shadi et al., 2016) and vision tagging (Filipe and Francisco, 2018; Welinder et al., 2010).

As the crowds can make mistakes, one core task is to deal with the annotation noise, for which purpose many approaches have been proposed (Filipe and Francisco, 2018; Dawid and Skene, 1979; Raykar et al., 2010; Zhou et al., 2012). However, there are two implicit assumptions made behind the existing methods: (1) the annotations are sufficiently collected for effective learning; (2) the class distribution of the instances and annotations are balanced. In many real-world applications, such assumptions are not true, which have posed further challenges for crowdsourcing learning and caused performance degeneration.

In crowdsourcing, annotation sparsity is common. To reduce the error effects, repetitive labeling is often exploited which employs multiple workers for labeling. At the same time, due to time and cost efficiency concerns, it's common that each worker is only assigned with a relatively small subset of data. This results in sparse annotations for each worker. The sparsity on one hand makes estimating the crowds' expertise challenging; on the other hand, when encountering imbalanced data, the annotations are prone to be imbalanced over classes. Traditional supervised class imbalance learning has widely shown that models trained on imbalanced data are biased towards majority classes and away from minority classes (Buda et al., 2018). In this paper, we show that in crowdsourcing scenarios with sparse annotations, even slight class imbalance can result in severe learning bias and performance degeneration. However, existing work rarely paid attention to this issue.

Figure 1 shows one illustrative example. Here one real-world image crowdsourcing dataset LabelMe (Filipe and Francisco, 2018) is used, which consists of 8 classes of 1, 000 training data and 1, 688 testing data. 2, 547 annotations for the training data are collected from 59 workers through the AMT platform. Figure 1a shows the annotation sparsity phenomenon. On average, each worker is assigned with only 43.169 images. Figure 1b shows the class distribution of groundtruth labels and annotations. Compared with

[1] https://www.mturk.com.

[2] http://crowdflower.com.

[3] http://recaptcha.net/.

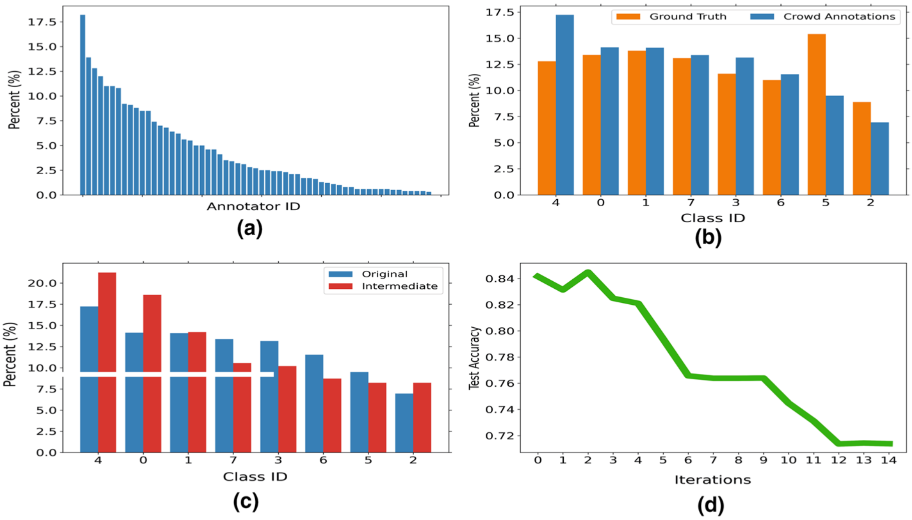


Fig. 1 Inspections of one real-world crowdsourcing dataset LabelMe. **a** shows the labeling sparsity issue of crowdsourcing workers. The vertical coordinate represents the percent of assigned instances for each workers among the whole instances. **b** shows the imbalanced class distribution of ground truth labels and crowdsourcing annotations. For ground truth labels, the vertical coordinate represents $\frac{N_c}{N}$, with N_c the number of instances with latent true class label c, N the total number of instances. For crowd annotations, the vertical coordinate represents $\frac{\overline{N}_c}{\overline{N}}$, with $\overline{N}_c$ the number of annotations labeled as class c, $\overline{N}$ the total number of crowd annotations. **c** shows the imbalanced class distribution of original crowd annotations (Original) and the combination of original crowds annotations and predicted pseudo-annotations at 5-th intermediate self-training iteration (Intermediate). **d** shows the test accuracy on balanced data for normal confidence based self-training method during learning process

the groundtruth labels, the annotations are slightly more skewed with a higher standard deviation (2.95% vs 1.85%).

Figure 1c and d show the effects of sparsity encountering class-imbalance. To alleviate the annotation sparsity, we apply self-training on one sota deep crowdsourcing learning approach Crowdlayer (Filipe and Francisco, 2018), and iteratively predict the unknown annotations and retrain the model.[4] Figure 1c shows the distribution of annotations generated by the model during one intermediate step, which becomes more skewed and more biased towards the majority classes. Figure 1d shows the accuracy on test data. As learning proceeds, the annotation imbalance is intensified and leads to severely deteriorated model quality. However, this issue has been rarely paid attention to and touched by previous crowdsourcing learning.

In this paper, to deal with the annotation sparsity and class imbalance challenges, we propose one approach called Distribution Aware Self-training based Crowdsourcing learning (DASC). At a high level, we iteratively predict the pseudo-annotations and select some of them to supplement the original annotations, which are expected to be confident and rebalance the annotation distribution. Within each iteration, a base model is trained using available annotations, and then acts as a teacher model to generate pseudo-annotations. To alleviate the imbalance issue, we propose to select the most confident pseudo-annotations

[4] The experimental details can be found in Sect. 4.

using resampling strategies, i.e., we undersample the majority classes and oversample minority classes. Then the learning model is retrained on the combination of observed and pseudo-annotations. DASC is a universal framework that can be applied to different crowdsourcing methods. We conduct extensive experiments on real-world datasets LabelMe (Filipe and Francisco, 2018), Music (Rodrigues et al., 2014) and CIFAR10-H (Peterson et al., 2019), showing that DASC delivers 2%-20% higher accuracies for the sota methods.

2 Related work

Crowdsourcing learning To estimate true labels from crowdsourcing annotations, one straightforward strategy is Majority Voting (MV), which however ignores the workers' expertises variance in reality. To counter this issue, probabilistic models which treat instances' true labels as unknown latent variables have been developed. One pioneer in this line is the DS model (Dawid and Skene, 1979) which exploits labeling error rates to parameterize the workers' expertise, and maximizes the annotation likelihood using expectation-maximization (EM). DS has served as the basis for a large number of crowdsourcing methods, which specialize on difference aspects including advanced levels of annotation generation processes (Welinder et al. 2010; Whitehill et al. 2009; Zhou et al. 2012), efficient inference algorithms (Liu et al., 2012), classifier learning (Raykar et al. 2010), Bayesian extensions (Kim and Ghahramani 2012; Simpson et al. 2013; Venanzi et al., 2014), and worker correlation modeling (Moreno et al., 2015; Li et al., 2019).

Recently, to make use of the strength of DNNs, deep crowdsourcing learning has been studied to build a DNN classifier from the noisy annotations and set up the new SOTAs. Guan et al. (2018) devoted itself to learning better voting weights for DNN workers using an EM-like procedure. Shadi et al. (2016) used a convolutional neural network (CNN) classifier as the latent true label prior and applied the EM optimization procedure. Filipe and Francisco (2018) treated the unknown true labels as one hidden layer of the DNN model and mapped them to the annotations through a linear mapping, then perform end-to-end SGD optimization. Tanno et al. (2019) and Chu et al. (2021) proposed mapping functions different from Filipe and Francisco (2018) for explanation and robustness. Atarashi et al. (2018) and Li et al. (2021) proposed fully Bayesian deep models to encode interpretable probabilistic structures.

However, existing works mainly assume the annotations are sufficiently collected, ignoring the intrinsic sparsity of crowd annotations. Moreover, when meeting class-imbalance, learning becomes more challenging. There are few works separately handling sparse or imbalanced annotations, e.g., Li et al. (2014), Li and Jiang (2018) and Chu and Wang (2021) considered annotation sparsity by supplementing them to improve performance, Zhang et al. (2015) considered class-imbalance and proposed one PLAT algorithm to estimate the threshold of the positive label frequency, but only for binary classification scenario. They only focus on a single aspect of read-world crowd annotations, which are not suitable for the sparse and class-imbalanced problem.

Imbalanced learning Learning with imbalanced data has rich literature. Classical methods include resampling and reweighting. Resampling focuses on oversampling the data of minority classes (Buda et al. 2018; Shen et al. 2016; Byrd and Lipton 2019) or undersampling the data of majority classes (He and Garcia 2009; Japkowicz and Stephen 2002; Shen et al. 2016). In contrast, reweighting adjusts weights during training for different classes

(Cui et al., 2019; Khan et al., 2018) or different samples (Cui et al., 2019; Lin et al., 2020). Other ideas including developing transfer learning (Jamal et al., 2020), meta-learning (Shu et al., 2019) and decoupling representation and classification (Kang et al., 2020; Tang et al., 2020; Zhou et al., 2020) have also been explored. The common evaluation protocol assumes that the ground truth labels are available during training, which may not reflect realistic settings. In our work, we consider the class imbalance for crowdsourcing tasks whose annotations are both noisy and sparse, which provides a new setup for imbalanced learning.

Self-training Self-training is one classical method for semi-supervised learning, which generates pseudo-labels for unlabeled samples with the model trained on labeled data. Lee (2013) proposed that utilizing the generated pseudo-labels is equivalent to entropy minimization. As the pseudo labels can be noisy, various works have devoted themselves to filtering the noisy labels by using measures based on confidence (Shi et al., 2018), neighborhood graphs (Iscen et al., 2019), uncertainty and calibration (Rizve et al., 2021). Recently, Berthelot et al. (2019, 2020) and Sohn et al. (2020) have achieved state-of-art semi-supervised learning performances by integrating pseudo-labeling with consistency regularization, which outputs consistent predictions over perturbations of instances Rasmus et al. (2015) and Tarvainen and Valpola (2017). Most semi-supervised learning mainly assumes that the labeled data is clean and rarely considers the imbalance issue. Recently, Wei et al. (2021) and Nassar et al. (2021) have considered the class-imbalance issue in labeled data and the pseudo labels explicitly. Compared with semi-supervised learning, our problem faces a more challenging scenario with corrupted labeled data.

3 Sparse and imbalanced crowdsourcing learning

3.1 Problem setup

We use $X = \{x_1, \cdots, x_N\}$ to denote the set of N training instances, $\overline{Y} \in \{0, 1, \cdots, C\}^{N\times K}$ their annotations collected from K crowd workers. $x_i \in \mathbb{R}^d$ means the d-dimensional feature values of the i-th instances, with a set of K annotations $\overline{y}_i = \{\overline{y}_i^k\}_{k=1}^K$. Here $\overline{y}_i^k = \overline{Y}_{ik}$ represents the label assignment of x_i given by worker k. $\overline{y}_i^k = 0$ indicates that worker k didn't tag x_i; $\overline{y}_i^k = c$ $(c \neq 0)$ means that x_i is categorized as c-th class by the k-th worker.

Using $\overline{N}_c$ to denote the number of annotations in class c for $\overline{Y}$, N_c the number of instances with latent true class label c. The annotation sparsity and class-imbalance issues can be formulated as:

$$\sum_c \overline{N}_c \ll N * K, \quad R = \frac{\max_c \overline{N}_c}{\min_c \overline{N}_c} \gg 1, \quad R_g = \frac{\max_c N_c}{\min_c N_c} \gg 1. \tag{1}$$

Here R and R_g are respectively the ratio between size of crowd annotations and groundtruth labels for the most frequent and least frequent class. Our goal is that, given the sparse and imbalanced training data X and $\overline{Y}$, learning a classifier $f : \mathbb{R}^d \rightarrow \{1, \cdots, C\}$ that generalizes well for unseen test data under class-balanced performance criteria.

Existing crowdsourcing work mainly focus on workers' expertise modeling and usage (Filipe and Francisco 2018; Dawid and Skene 1979; Shadi et al. 2016; Tanno et al. 2019), ignoring that annotation sparsity and class-imbalance are common in many

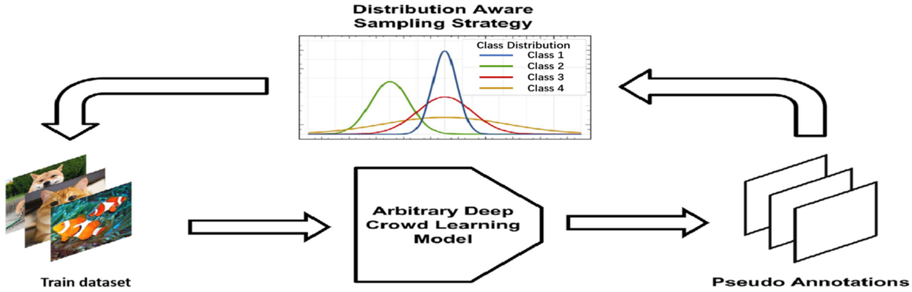


Fig. 2 Overview of the proposed DASC framework

real applications. As we have shown in the introduction, the two aspects pose new challenges and lead to poor performance for existing methods. In this paper, we propose one distribution aware self-training based crowdsourcing (DASC) learning framework to combat the two issues. We leverage the self-training strategy commonly used by modern state-of-the-art semi-supervised learning methods to predict the unknown annotations, and make use of them to retrain the model. Next, we will first introduce the proposed DASC framework, then introduce three state-of-the-art deep crowdsourcing learning base models that are implemented for DASC.

3.2 Distribution aware self-training based crowdsourcing learning (DASC)

Semi-supervised learning (SSL) provides one promising way to improve the performance of learning models by using unlabeled data in case of limited labeled data. One common approach used by modern SSL algorithms is the self-training method, which predicts pseudo-labels for unlabeled data and incorporates the most helpful ones to retrain the model (McLachlan, 1975; Xie et al. 2020). In this paper, to combat the annotation sparsity and class-imbalance, we use existing crowdsourcing methods as base models, and propose one distribution aware confidence measure to conduct self-training.

Figure 2 depicts the framework of the proposed approach. It consists of two main components: (1) the crowdsourcing learning base model, and (2) the pseudo-annotations selection strategy. During the training, we progressively predict pseudo-annotations for the unannotated instances for each worker, and add some of them into the training data, then update the learning model. The most confident pseudo-annotations which contribute to rebalancing the annotation distribution are selected. Next, we will explain the measure in detail.

3.2.1 Distribution aware confidence measure

Confidence Confidence is a commonly used measure in self-training, which measures how confident the prediction of the current model is for some instances. Using $\hat{p}(\overline{y}^k)$, e.g., defined in Eqs. (5, 8), to denote the pseudo-annotations probability of worker k on some unannotated instance x, entropy is often used to measure confidence:

$$entropy(\overline{y}^k) = -\sum_{c=1}^{C} \hat{p}(\overline{y}^k) \cdot \log \hat{p}(\overline{y}^k). \tag{2}$$

The pseudo-annotations with lower entropy values are considered to be more confident and more likely to be correct.

Traditional self-training would prefer to select pseudo-annotations with the least entropy values as authentic ones. However, as we discussed in the introduction, without taking the class-imbalance issue into account, the learning algorithm would be biased towards selecting majority class annotations and ignore the minority annotations. More seriously, this bias can accumulate throughout the training process, which will inevitably damage the performance. In the following, we propose our distribution aware confidence measure.

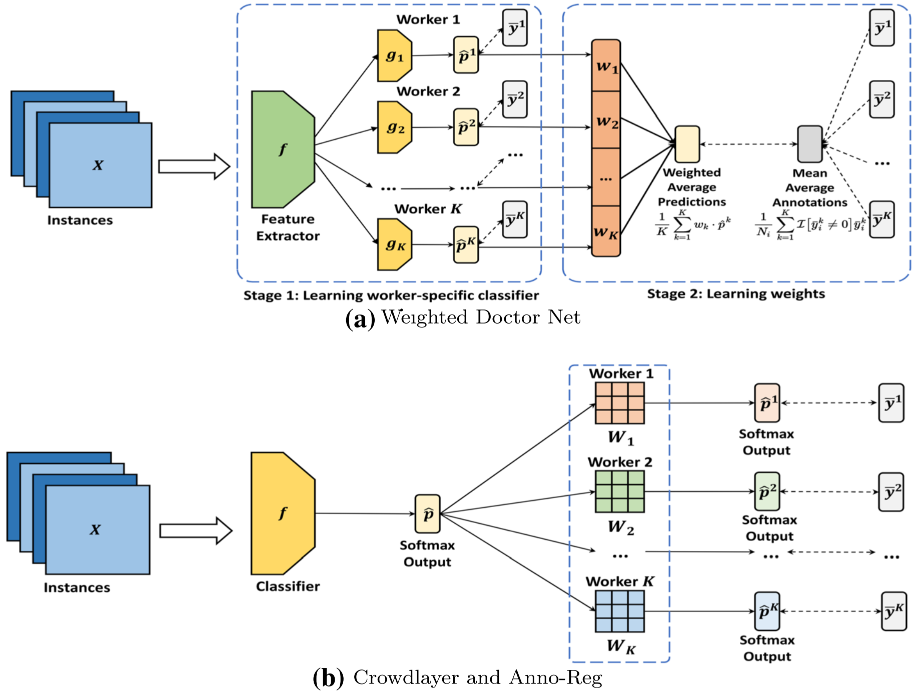
Distribution aware confidence Resampling is a common strategy for addressing the class-imbalance problem. It intuitively oversamples the majority classes or undersamples the minority classes to avoid the dominant effect of majority data. In this paper, we adopt the resampling strategy within each class, i.e., the M_c most confident pseudo-annotations for each class $c \in \{1, \cdots, C\}$ are selected:

$$M_c = t_c \cdot M, \qquad \sum_{c=1}^{C} t_c = 1. \tag{3}$$

Here M denotes the total number of selected pseudo-annotations within each iteration, which is a hyperparameter set by the users. t_c denotes the normalized fraction coefficient of class c, which is inversely proportional to the number of pseudo-annotations N'_c of class c among all the generated pseudo-annotations:

$$t_c \propto \frac{1}{N'_c}. \tag{4}$$

Algorithm 1 summarizes the main steps of the DASC approach. We iteratively predict the unobserved annotations and add some of them into the training data. Those pseudo-annotations with lower entropy values and contribute to rebalancing the annotation distribution are selected according to Eqs. (3) and (4). Then the learning model is retrained using the combination of observed and pseudo-annotations. This process repeats until the expected performance is reached.



(a) Weighted Doctor Net

(b) Crowdlayer and Anno-Reg

Fig. 3 The network architecture of Weighted Doctor Net (Guan et al., 2018), Crowdlayer (Filipe and Francisco, 2018), Anno-Reg (Tanno et al., 2019), where $\hat{p}^k = \hat{p}(\bar{y}^k)$

Algorithm 1 Distribution Aware Self-training based Crowdsourcing Learning

```
Input:
$D = \{(x_i, \bar{y}_i)\}_{i=1}^N$ : crowdsourcing training data where $\bar{y}_i = \{\bar{y}_i^k\}_{k=1}^K$
$\ell$: loss function
Output: classifier $f$
Initialization:
train the crowdsourcing base model, e.g, Crowdlayer on $D$
Repeat:
obtain the pseudo-annotations predictions of each worker on its unannotated instances using Eq. 5 or  8
for each pseudo-annotation, calculate its confidence score according to Eq. 2
for each class $c$, calculate the corresponding selection number $M_c$ according to Eq. 3
select the $M_c$ most confident pseudo-annotations within each class
add the selected pseudo-annotations into the training data and retrain the base model, e.g, Crowdlayer.
Until expected performance reached
```

As one primary attempt for the important but largely ignored sparse and imbalance problem, DASC is simple and universal. It can be applied to any existing crowdsourcing methods for potential significant performance improvement, without modifying their original implementations. The possible limitation is that, the performance of the self-training process depends largely on the quality of selected pseudo-annotations, which in turn depends on the base model performance and pseudo-annotation selection measure/

threshold. Thus successful application of the framework requires skills and experiences on tuning self-training process.

3.3 Deep crowdsourcing base models

With the ubiquitous success of deep neural networks (DNN), deep crowdsourcing learning has been studied by combining the strength of DNN with crowdsourcing. In this subsection, we will briefly introduce three state-of-the-art deep crowdsourcing learning models WDN (Weighted Doctor Net) Guan et al. (2018), Crowdlayer Filipe and Francisco (2018) and Anno-Reg Tanno et al. (2019), showing in experiments that DASC improves over them with a large margin.

Weighted doctor net Guan et al. (2018) Figure 3(a) shows the network architecture of Weighted Doctor Net. For each worker k, one network classifier $g(\cdot;\phi_k)$ with parameter ϕ_k is constructed. All classifiers share the same feature extractor network $f(\cdot;\theta)$. The annotation probability given by worker k for x is derived as:

$$\hat{p}(\overline{y}^k) = softmax(g(f(x;\theta);\phi_k)). \tag{5}$$

Given a specific loss function ℓ, e.g., the cross entropy loss used by WDN, the loss over all training data $\{X, \overline{Y}\}$ is defined as:

$$L_1 := \sum_{i=1}^{N} \sum_{k=1}^{K} \mathcal{I}[\overline{y}_i^k \neq 0] \ell(\hat{p}(\overline{y}_i^k), \overline{y}_i^k). \tag{6}$$

Here $\mathcal{I}$ is the indicator function. Then parameters θ and $\{\phi_k\}$ are optimized by minimizing Eq. (6) in an end-to-end manner. After $f(\cdot;\theta)$ and $\{g(\cdot;\phi_k)\}$ are learned, WDN learns the weight $\{w_k\}$ for workers through minimizing the loss between the weighted average of all workers' predictions and the mean average of the observed annotations over instances:

$$L_2 := \sum_{i=1}^{N} \ell(\frac{1}{K} \sum_{k=1}^{K} w_k \cdot \hat{p}(\overline{y}_i^k), \frac{1}{N_i} \sum_{k=1}^{K} \mathcal{I}[\overline{y}_i^k \neq 0] \overline{y}_i^k). \tag{7}$$

Here N_i is the number of collected annotations for instance x_i. After the training finish, a softmax operation is conducted over the weights. For testing, the true label prediction for some instances is taken as the weighted average of all worker classifiers' predictions.

Crowdlayer Filipe and Francisco (2018) Fig. 3b shows the network architecture of Crowdlayer approach. Using $f(x;\theta) \in [0, 1]^{1 \times C}$ to denote the softmax output of the true label classifier $f(\cdot)$ with parameter θ for instance x, the Crowdlayer model introduced K matrices $\{W_k \in \mathbb{K}^{C \times C}\}_{k=1,\cdots,K}$ to capture the annotating process of the crowds, i.e., the annotations of x given by worker k is derived as:

$$\hat{p}(\overline{y}^k) = softmax(f(x;\theta) \cdot W_k). \tag{8}$$

In Crowdlayer, although $\{W_k\}$ are real valued without any structural constraints, they are believed to be able to reflect the workers' annotating expertise, i.e., $W_k(i,j)$ can denote the process that instances belonging to class i are annotated with class label j by worker k. Larger diagonal values mean better worker expertise. Given a specific loss function ℓ, e.g., the cross entropy loss, the loss over the crowdsourcing training data is defined as:

Table 1 Statistic details of the three benchmark datasets

Dataset	Training/Testing Size	R/R_g of Annotations/ Truth	Mean workers' Accuracy	Mean workers' Labels
LabelMe	1,000/1,188	2.479/1.730	69.2% ± 18.1%	43.2
Music	700/300	1.753/1.188	73.3% ± 24.2%	66.9
CIFAR10-H	2,837/1,000	3.093/5.000	55.6% ± 19.7%	89.2

$$L := \sum_{i=1}^{N} \sum_{k=1}^{K} \mathcal{I}[\bar{y}_i^k \neq 0] \ell(\hat{p}(\bar{y}_i^k), \bar{y}_i^k). \tag{9}$$

Then regarding $\{W_k\}$ as crowdlayers after the neural network classifier $f(\cdot)$, Filipe and Francisco (2018) proposed to simultaneously optimize the classifier parameter θ and $\{W_k\}$ in an end-to-end manner by minimizing Eq. (9). After the training finished, $f(x;\theta)$ is used to predict true labels for new test instance x.

Worker regularization Tanno et al. (2019) Anno-Reg shares the same network architecture as Crowdlayer. Actually, the architecture in Fig. 3b and minimizing the loss in Eq. (9) have been the cornerstone of various deep crowdsourcing learning approaches (Tanno et al. 2019; Chu et al. 2021). They mainly differ in specific structural regularization over the expertise parameters $\{W_k\}$ with different motivations. Anno-Reg is one representative among them. As mentioned before, Crowdlayer uses real-valued $\{W_k\}$ without any constraints, which makes it hard to explain. Anno-Reg Tanno et al. (2019) addressed this by modeling each W_k as a confusion matrix, which is implemented by imposing a softmax operation over each column of real valued W_k:

$$W_{kc} \leftarrow softmax(W_{kc}). \tag{10}$$

Besides, Anno-Reg also adds an extra trace minimization term over $\{W_k\}$ to the loss function, which was theoretically shown to encourage the convergence of $\{W_k\}$ to the true workers' confusion matrices:

$$L_{norm} := \sum_{k=1}^{K} tr(W_k). \tag{11}$$

4 Experiments

4.1 Settings

Dataset We first perform experiments on three widely used read-world crowdsourcing benchmark datasets LabelMe (Filipe and Francisco, 2018), Music (Rodrigues et al. 2014) and CIFAR10-H (Peterson et al., 2019). LabelMe is an image classification dataset consisting of 2, 688 images from 8 classes. The accuracy of each worker ranges from 0 to 100%.

Music is a music genre classification dataset consisting of 1, 000 samples of songs with 30 seconds length from 10 music genres. CIFAR10-H is an image classification dataset consisting of 10, 000 images from the CIFAR-10 (Krizhevsky, 2009) test set. Originally, Peterson et al., 2019) recruited 2, 571 workers with each annotating 200 images. For our experiment, we split 1, 000 class-balanced images as the test set. To make an imbalanced training set, we adopt the long-tailed imbalance setting as in Zhou et al. (2020) which follows an exponential decay in sample sizes across different classes. The groundtruth label imbalance ratio $R_g = \frac{\max_c N_c}{\min_c N_c} = 5$. For each class c, $\hat{N}_c = \frac{\max_c \overline{N}_c}{R^{\frac{c}{10}}}$ crowd annotations are reserved. The final training set has 2, 837 images annotated by 49 workers. For LabelMe and Music, their original training/testing splitting are used.

Table 1 shows the statistic details for the three datasets. For LabelMe and Music, their imbalance ratios are relatively slight, not as heavy as that in traditional supervised imbalance learning. Thus previous crowdsourcing works didn’t take this as an issue and paid no attention. However, we will show in the following that, even the slight class distribution skewness should not be ignored for crowdsourcing learning.

Network and optimization For fair comparison, we implement the methods following the settings in Filipe and Francisco (2018). Specifically, for LabelMe, we use the pretrained CNN layers of the VGG-16 deep neural network (Simonyan and Zisserman, 2015) with one fully connected layer of 128 units, ReLU activation, one output layer on top as the classifier. Adam Kingma and Ba (2015) optimizer is adopted with a batch size of 512 and a learning rate of 1×10^{-3}. 50% random dropout and l_2 weight decay regularization with $\lambda = 5 \times 10^{-4}$ on all layers are used. In each self-training iteration, Anno-Reg and Crowdlayer are trained for 25 epochs, WDN is trained for 25 and 50 epochs in the first and second phase, 20, 000, 10, 000, 10, 000 pseudo-annotations are respectively selected without replacement for Anno-Reg, Crowdlayer and WDN.

For Music, a network of one single hidden layer of 128 units, a batch-norm layer, ReLU activation, and one output layer on top as a classifier is used. The batch size is 128. For Crowdlayer and Anno-Reg, the model is trained using Adam optimizer for 100 epochs with a learning rate of 1×10^{-3} during each iteration. For WDN, the model is trained for 100 epochs at the first training phase using Adam optimizer with a learning rate of 1×10^{-2}, and SGD optimizer for 50 epochs in the second phase with cyclic learning rate scheduler (Smith et al. 2017) between $[10^{-3}, 10^{-2}]$ and a momentum between [0.8, 0.9]. In each iteration 700, 1, 000, 1, 000 pseudo-annotations are selected without replacement for Anno-Reg, Crowdlayer and WDN.

For CIFAR10-H, a pretrained Resnet18 (He et al., 2016) with one output layer on the top as a classifier is used. The training is conducted by Adam optimizer with a batch size of 128 with learning rate 10^{-4}. l_2 weight decay regularization with $\lambda = 5 \times 10^{-5}$ on all layers are used. Crowdlayer and Anno-Reg are trained for 100 epochs in each iteration, WDN is trained for 50 epochs in both training phases. In each iteration, 2, 000 pseudo-annotations are selected without replacement for all the base models. The trace norm for Anno-Reg with $\alpha = 0.1$ is used.

Baselines To assess the performance of the proposed DASC, we conduct comparison with three groups of baselines: traditional crowdsourcing methods with MV-DL, DS-DL, imbalance learning implementations with resampling, reweighting, Focal loss, and self-training based implementations confidence, random.

Table 2 Test accuracy on LabelMe, Music and CIFAR10-H. The best performance are bold

Dataset	Baselines		Performance		
LabelMe	TraditionalCrowdsourcing Learning	MV-DL	81.3 ± 0.4		
		DS-DL	82.2 ± 0.4		
			Anno-Reg	Crowdlayer	WDN
	Imbalanced Learning	CE	82.3 ± 0.3	84.0 ± 0.3	77.1 ± 2.6
		Resampling	83.2 ± 0.5	85.6 ± 0.7	80.6 ± 1.0
		Reweighting	83.6 ± 0.4	85.6 ± 0.3	78.8 ± 1.6
		Focal	81.7 ± 0.4	84.2 ± 0.2	78.4 ± 1.5
	Self-training Methods	Confidence	82.4 ± 0.2	71.4 ± 7.3	77.7 ± 1.7
		Random	83.3 ± 1.2	77.3 ± 0.3	74.3 ± 1.2
		DASC	**85.6 ±0.3**	**87.6 ± 0.4**	**82.3 ±0.9**
Music	Traditional Crowdsourcing Learning	MV-DL	59.3 ± 1.7		
		DS-DL	65.2 ± 0.8		
			Anno-Reg	Crowdlayer	WDN
	Imbalanced Learning	CE	65.8 ± 0.7	66.3 ± 1.5	55.1 ± 4.0
		Resampling	66.3 ± 2.0	68.3 ± 1.1	51.6 ± 1.3
		Reweighting	67.1 ± 1.1	66.7 ± 0.6	56.2 ± 1.4
		Focal	63.4 ± 1.2	65.0 ± 0.6	57.5 ± 2.4
	Self-training Methods	Confidence	65.8 ± 0.9	57.6 ± 0.8	61.5 ± 1.7
		Random	66.7 ± 1.0	63.2 ± 1.5	60.1 ± 1.8
		DASC	**68.0 ± 1.1**	**68.5 ± 0.5**	**65.7 ± 1.4**
CIFAR10-H	Traditional Crowdsourcing Learning	MV-DL	44.9 ± 0.7		
		DS-DL	43.4 ± 0.8		
			Anno-Reg	Crowdlayer	WDN
	Imbalanced Learning	CE	40.9 ± 1.0	45.6 ± 0.5	17.3 ± 1.1
		Resampling	43.9 ± 1.8	48.6 ± 1.7	18.2 ± 0.5
		Reweighting	41.4 ± 1.0	47.7 ± 0.8	17.6 ± 1.3
		Focal	39.7 ± 0.4	47.6 ± 1.2	17.9 ± 1.3
	Self-training Methods	Confidence	49.4 ± 1.2	48.8 ± 1.0	31.9 ± 1.7
		Random	49.0 ± 1.5	51.2 ± 3.5	14.7 ± 0.6
		DASC	**51.0 ± 1.3**	**55.3 ± 0.6**	**36.9 ± 1.7**

Bold format means the best performance

- *MV-DL* which first aggregates over the crowd annotations, then trains the deep neural network classifier.
- *DS-DL* which first aggregates over the crowd annotations considering the workers' expertise variation using the DS model (Dawid and Skene, 1979), then trains the deep neural network classifier.
- *Resampling* which oversamples crowd annotations for minority classes such that the annotation distribution is balanced over classes.
- *Reweighting* which reweights the loss for each class c by $R_c = \frac{\max_c N_c}{N_c}$ such that the annotation losses are balanced over classes.
- *Focal* Focal loss, which trains the deep crowdsourcing base models using the imbalance focal loss function (Lin et al., 2020). The hyperparameters $\gamma = 2$ for all the experiments.

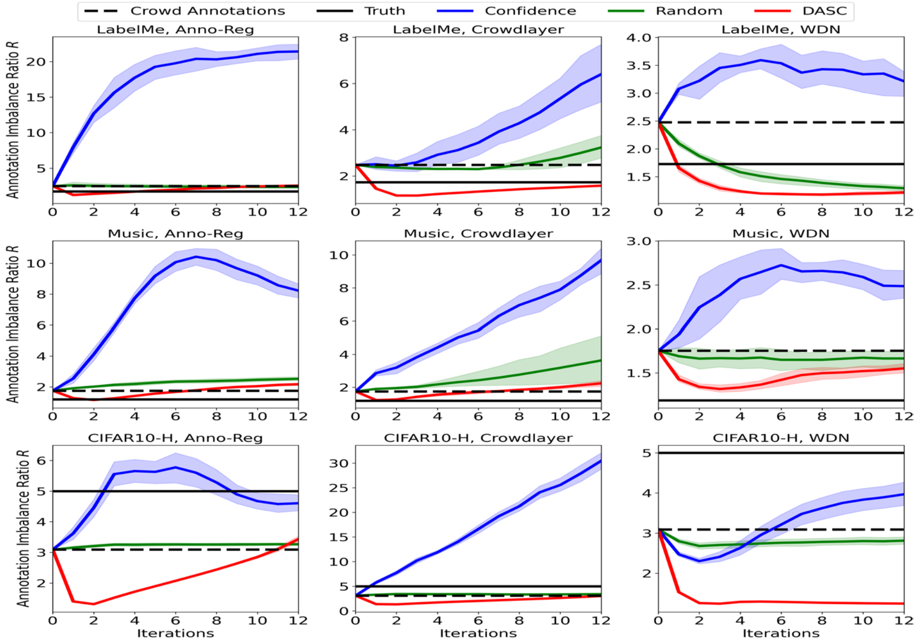


Fig. 4 The variation of imbalance ratio R for annotations during the learning procedure

- *Confidence* which trains the deep crowdsourcing base models using self-training and selects the most confident pseudo-annotations with least entropies using Eq. (2).
- *Random* which trains the deep crowdsourcing base models using self-training and randomly selects pseudo-annotations.
- *DASC* the proposed approach which trains the deep crowdsourcing base models using self-training and selects pseudo-annotations considering both confidence and the imbalance issue.

Besides, we also report the performance of the original deep crowdsourcing base models as CE (Cross Entropy), i.e., using the imbalance agnostic cross entropy loss without self-training. CE acts as the base for the imbalance and self-training implementations, and their comparisons illustrate better the effects of annotation imbalance and sparsity on crowdsourcing learning.

4.2 Results

As the test set of LabelMe and Music are slightly imbalanced (with groundtruth label imbalance ratio 1.536 and 1.461), we adopt that accuracy used by previous works is a reasonable measure. The test accuracy comparisons are shown in Table 2. The mean and standard deviation results over 5 random repetitions are reported.

As we can see, the classical aggregation methods MV-DL, DS-DL are always inferior to the deep base models with CE loss, except for WDN. For each deep base model Anno-Reg, Crowdlayer, and WDN, their DASC implementations always perform the best, improving significantly in all cases, roughly from 2 to 20% absolute test accuracy. Specially, for the inferior base model WDN, it is improved by 5.2%, 10.6% and 19.6% on LabelMe, Music and CIFAR10-H. On the most difficult benchmark CIFAR10-H, Anno-Reg, Crowdlayer, and WDN are improved by 10.1, 8.5 and 19.6%. These improvements are achieved without any sophisticated network redesign or additional information acquisition, only by a universal smarter data leverage manner.

For the imbalance learning implementations Resampling, Reweighting and Focal loss, they improve over the cross entropy loss for the three deep base models Anno-Reg, Crowdlayer and WDN on all datasets except for the case of WDN on Music using Resampling. This shows that when concerning crowdsourcing learning, the class skewness of data should be taken into account. But without dealing with annotation sparsity, the improvements are limited.

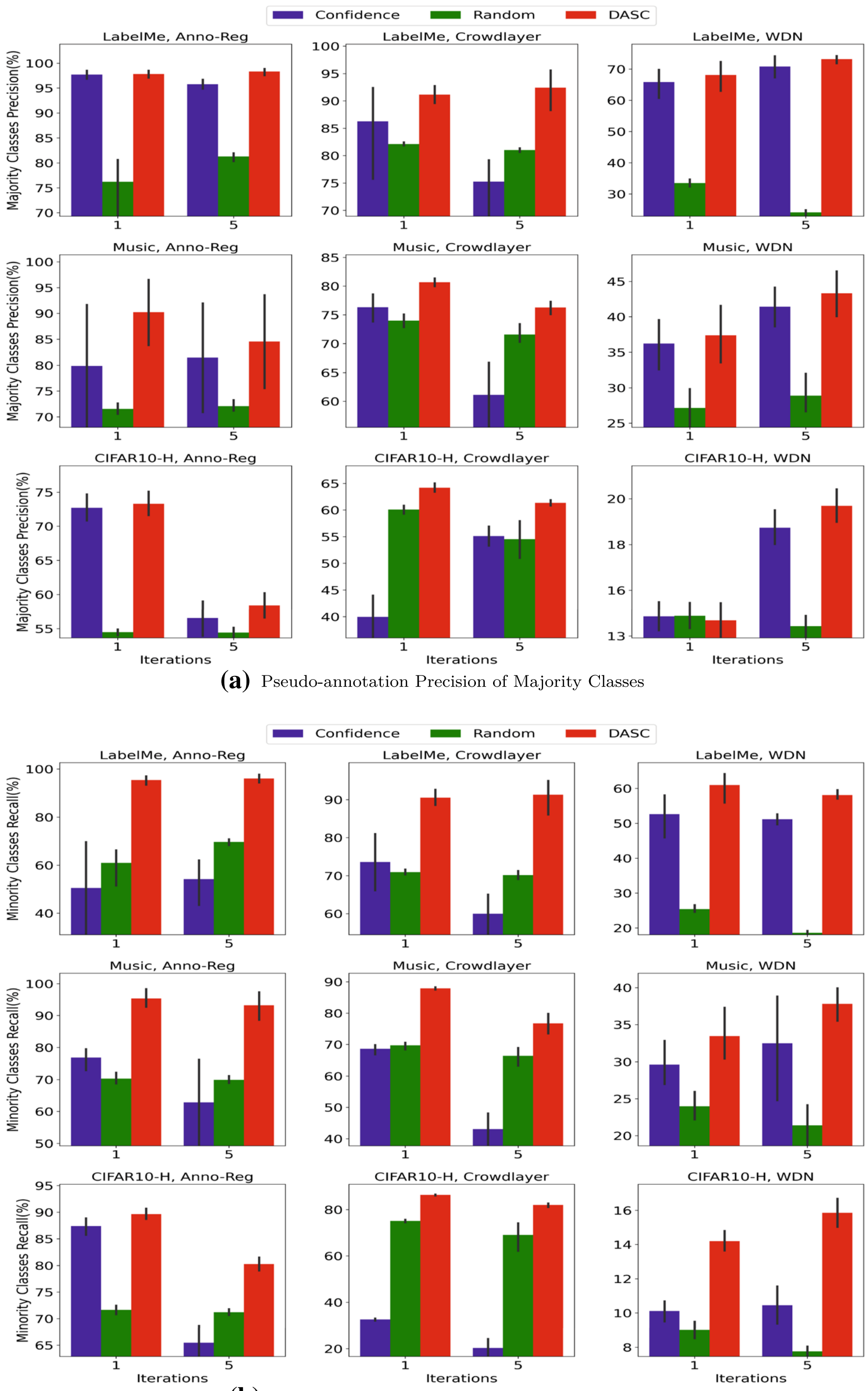
For the self-training based methods which combat annotation sparsity by supplementing the sparse annotations, the normal Confidence and Random pseudo-annotation selection strategies which don't deal with the class imbalance make positive effects sometimes but sometimes bring bad negative impacts, see Crowdlayer on LabelMe and Music, WDN with Random on LabelMe and CIFAR10-H, which are highlighted as italic. In the next, we will examine what happened during the learning procedure and explain the degeneration of Confidence and Random based self-training implementations.

Figure 4 shows the variation of the imbalance ratio R, i.e., $R = \frac{\max_c \overline{N}_c}{\min_c \overline{N}_c}$, of the combination of original and selected pseudo-annotations by Confidence, Random and DASC during the self-training procedure. For comparison convenience, we also plot the imbalance ratio of original crowd annotations and groundtruth labels. It can be seen that the R value of Confidence increases rapidly, indicating that the confidence based measure mostly selects the majority class pseudo-annotations, leading to severely imbalanced annotation distribution, which in turn can hurt the learning performance. The random selection strategy is much better than confidence based measure but still biased by the imbalance issue, see Anno-Reg and Crowdlayer models. The proposed DASC is more robust and yields much more balanced annotations, greatly alleviating the imbalance issue.

In Fig. 5, we check the pseudo-annotation quality of Confidence, Random and DASC. As shown in Fig. 4, the imbalance ratio for Confidence can be as high as 30, for which accuracy is no longer a proper quality measure. Since there tend to be lots of misleading predictions towards majority classes, the precision measure is better reflecting the non-misleading ratio. Whereas mining the minority classes is more difficult and important, the recall measure is better. Thus we divide the annotation classes into 50 vs 50% majority vs minority classes, and report the average precision for majority classes and recall for minority classes. We compute the measures by comparing them with the groundtruth labels. Results for iteration $t = 1$ and $t = 5$ are reported. We can see that DASC consistently obtains pseudo-annotations with higher precision/recall on the majority/minority classes, which explains its superior performance under the class-imbalance scenario.

4.3 Various sparsity level study

To examine the effectiveness of our approach with different sparsity levels, we remove fractions of the original training annotations for the three benchmarks LabelMe, Music and CIFAR10-H. Specifically, we remove p fractions of the observed annotations with p ranges from 20 to 80% in a uniformly random manner. Average test accuracy for 5 times repetitions are reported in Fig. 6, which demonstrates consistent results with that shown



(a) Pseudo-annotation Precision of Majority Classes

(b) Pseudo-annotation Recall of Minority Classes

Fig. 5 Pseudo-annotation quality of Confidence, Random and DASC. Results at two self-training iterations $t = 1$ and $t = 5$ are reported

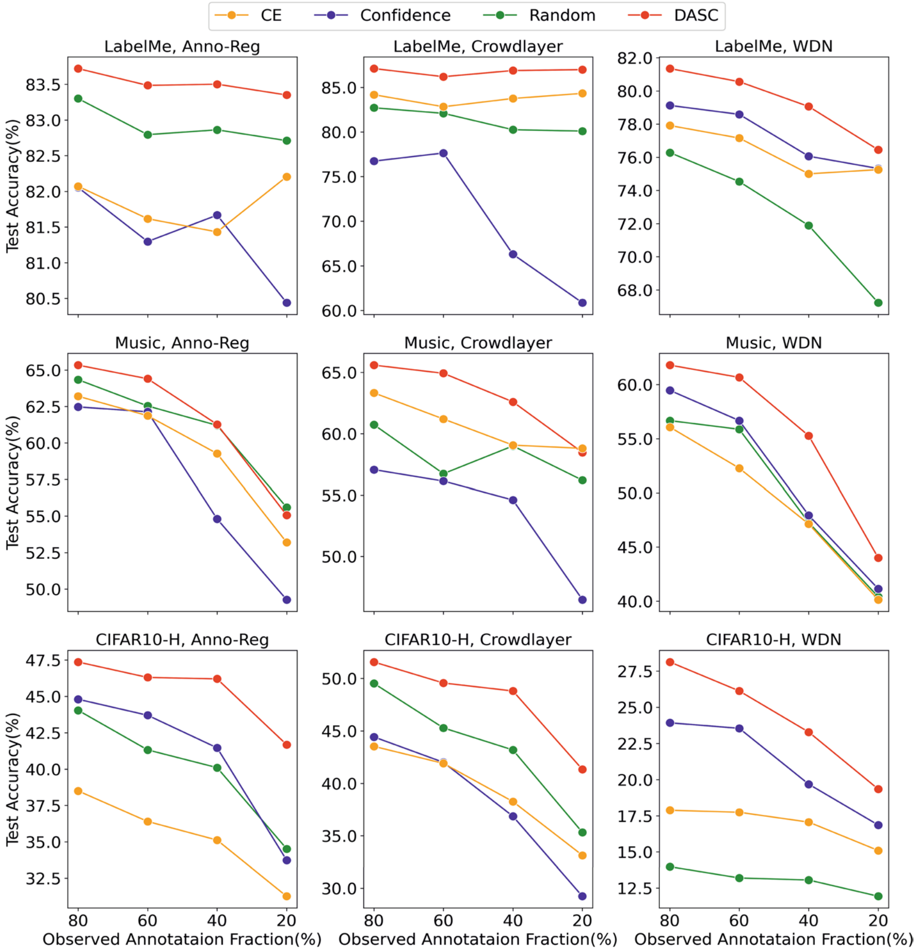
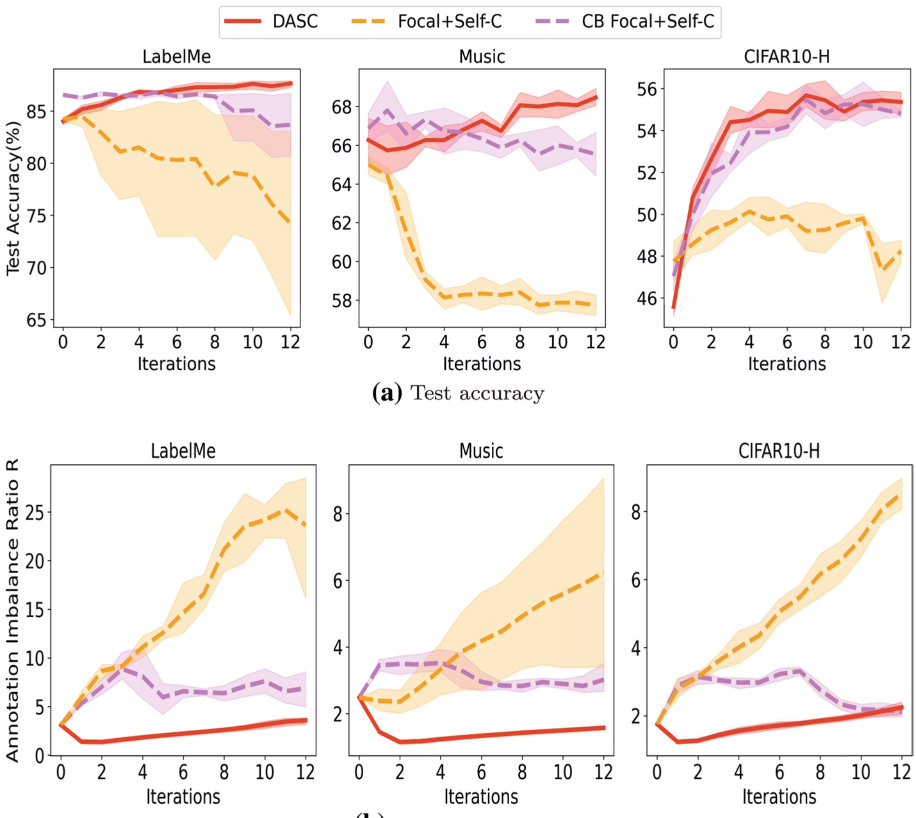


Fig. 6 Test accuracy of CE, Confidence, Random and DASC over base models Anno-Reg, Crowdlayer and WDN when the annotations are sparse on LabelMe, Music and CIFAR10-H

in Table 2. Compared with the CE base models which learn from the observed annotations without considering imbalance and sparsity, DASC stably significantly improves over them. However, the Confidence and Random self-training implementations without dealing with the imbalance issue sometimes are much worse than the CE base models.

4.4 More experiments with alternative imbalance self-training solutions

In this subsection, we further compare DASC with two alternative straightforward imbalance learning with self-training implementations: Focal+Self-C and CB Focal+Self-C, which deal with the sparse and imbalance annotations by respectively using the imbalance loss functions Focal and CB (class-balanced) Focal with confidence based self-training. CB Focal (Yin et al., 2019) improves over Focal loss by reweighting the loss of different classes using one measure named effective number rather than the nominal numbers



(a) Test accuracy

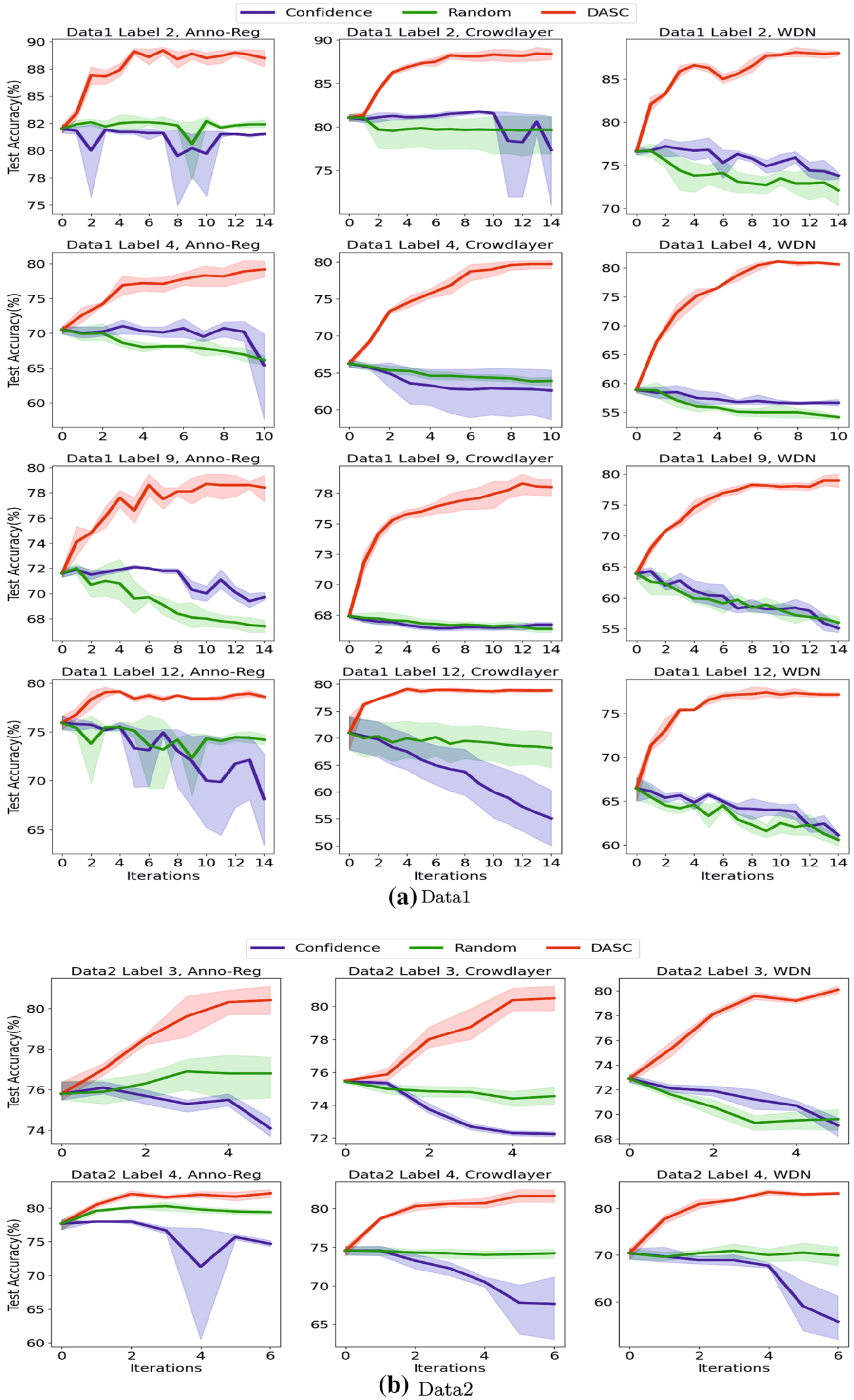
(b) Imbalance ratio R

Fig. 7 Comparison results with two alternative imbalance self-training implementations: Focal+Self-C and CB Focal+Self-C

Table 3 Statistic details of the six heavily imbalanced datasets

Dataset	Training/Testing Size	R/R_g of Annotations/ Truth	Mean workers' Accuracy	Mean workers' Labels
Data 1-label 2	1,258/200	6.530/5.989	92.7% ± 2.5%	347.8
Data 1-label 4	1,258/200	8.003/10.541	92.5% ± 2.3%	348.1
Data 1-label 9	1,258/200	7.073/7.169	91.4% ± 3.8%	347.1
Data 1-label 12	1,158/300	4.903/4.732	95.5% ± 3.0%	318.3
Data 2-label 3	500/200	3.104/3.587	89.8% ± 4.1%	188.5
Data 2-label 4	500/200	4.667/5.024	89.1% ± 4.8%	193.9

of instances. Figure 7 shows their results on the three benchmark datasets when Crowd-layer is used as base model. The two hyperparameters for Focal and CB Focal are tuned as $\beta = 0.9999$ and $\gamma = 2$. From Fig. 7a, it can be seen that they are inferior and less robust compared with DASC. We check their pseudo-annotations imbalance ratios during the



(a) Data1

(b) Data2

Fig. 8 Test accuracy of Confidence, Random and DASC over base models Anno-Reg, Crowdlayer and WDN on the heavily imbalanced crowdsourcing tasks of Data1 and Data2

self-training process in 7b, which are much higher than DASC. This indicates that the Focal and CB-Focal loss functions derive more skewed annotation class distributions, which explains their inferior performance.

4.5 More experiments on heavily imbalanced data

We further test the effectiveness of DASC on six more datasets with the heavily imbalanced class distribution. We take the two multi-label image crowdsourcing datasets Data1 and Data2 from Li et al. (2019), which respectively consist of 1, 458 images, 15 workers and 700 images, 18 workers. We use labels 3, 4, 5, 12 of Data1 and label 3, 4 of Data2 as 6 binary classification crowdsourcing tasks, each of them is highly class-imbalanced. For our experiment, we split a balanced test set for each task and use the rest as the training set. Table 3 shows the datasets statistic details.

We use a network of a hidden layer of 1000 units, ReLU activations and one output layer on top as the classifier. The training is conducted by using Adam optimizer with a batch size of 128 and a learning rate of 1×10^{-3}. For the Crowdlayer and Anno-Reg base models, we respectively train them for 100 and 200 epochs. For WDN base model, we train it for 75 epochs at both training phases.

Figure 8 shows the test accuracy of the three self-training based implementations Confidence, Random and DASC over the deep base models Anno-Reg, Crowdlayer and WDN. Here results for 14 iterations are recorded, in each iteration 500 pseudo-annotations are selected without replacement. As we can see, for the heavily imbalanced crowdsourcing data, the Confidence self-training method which supplements the sparse annotations with most confident pseudo-annotations decreases the base models rapidly, due to its selection bias towards the majority annotations, which is further intensified during the self-training process. The Random implementation ignores both the imbalance issue and pseudo-annotation quality, leading to similar decreasing trends as Confidence on Data1, slightly better on the Data2. In contrast, DASC stably improves all base models on all the tasks for a large margin, mostly with 5% to 20% absolute accuracy. This shows that significant performance gains are obtainable when the data are properly leveraged considering their imbalance and sparse characteristics.

5 Conclusion

In this paper, we propose a distribution aware self-training based method DASC to deal with the annotation sparsity and class-imbalance issues in crowdsourcing learning. Through supplementing the sparse annotations by adding confident pseudo-annotations and re-balancing the distribution, we show by extensive experiments that, significant performance gains can be obtained over existing models by using distribution aware self-training strategy. As a primary attempt, we emphasize that, for crowdsourcing learning with sparse annotations, even slight annotation imbalance should not be ignored. However, such one aspect is rarely considered in previous work. We believe this point would be worthy of more attention, and can be better solved by combining with sophisticated techniques from imbalanced learning, deep crowdsourcing learning and semi-supervised learning sotas.

Author Contributions S.-Y.L. and S.-J.H. contributed to the study conception and design. Material preparation, data collection and analysis were performed by S.-Y.L. and Y.S.. The first draft of the manuscript

was written Ye Shi and all authors commented on the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by National Natural Science Foundation of China (Grant No.61906089), National Key R&D Program of China (Grant No.2020AAA0107000),Jiangsu Province Basic Research Program (Grant No.BK20190408) and China Postdoc Science Foundation (Grant No.2019TQ0152).

Declarations

Conflict of interest Shao-Yuan Li, Ye Shi and Sheng-Jun Huang declare that they have no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material The datasets used in the study are publicly available from their corresponding authors.

Code availability The code for this study are not publicly available until the paper is published, but are available from the corresponding author Shao-Yuan Li on reasonable request.

References

Atarashi, K., Oyama, S., & Kurihara, M. (2018). Semi-supervised learning from crowds using deep generative models. In: *Proceedings of the 32nd Conference on Artificial Intelligence*, pp. 1555–1562

Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., & Raffel, C. (2020). Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In: *Proceedings of the 8th International Conference on Learning Representations*

Berthelot, D., Carlini, N., Goodfellow, I. J., Papernot, N., Oliver, A., & Raffel, C. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems, 32*, 5050–5060.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249–259.

Byrd, J., & Lipton, Z.C. (2019). What is the effect of importance weighting in deep learning? In: *Proceedings of the 36th International Conference on Machine Learning*, pp. 872–881

Chu, Z., Ma, J., & Wang, H. (2021). Learning from crowds by modeling common confusions. In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 5832–5840

Chu, Z., & Wang, H. (2021). Improve learning from crowds via generative augmentation. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 167–175

Cui, Y., Jia, M., Lin, T., Song, Y., Belongie, S.J. (2019). Class-balanced loss based on effective number of samples. In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277

Cui, Y., Jia, M., Lin, T., Song, Y., & Belongie, S.J. (2019). Class-balanced loss based on effective number of samples. In: *Proceedings of the 2019IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28*, 20–28.

Feng, J., Wang, X., & Liu, W. (2021). Deep graph cut network for weakly-supervised semantic segmentation. *Science China Information Science, 64*(3), 130105.1–130105.12

Filipe, R., & Francisco, P.C. (2018). Deep learning from crowds. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, p. 8. AAAI Press

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the 2014 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 580–587

Guan, M.Y., Gulshan, V., Dai, A.M., & Hinton, G.E. (2018). Who said what: Modeling individual labelers improves classification. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 3109–3118

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778

Horvitz, E. (2007). Reflections on challenges and promises of mixed-initiative interaction. *AI Magazine, 28*(2), 13–22.

Iscen, A., Tolias, G., Avrithis, Y., & Chum, O. (2019). Label propagation for deep semi-supervised learning. In: *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5070–5079

Jamal, M.A., Brown, M., Yang, M., Wang, L., & Gong, B. (2020). Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7607–7616

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–449.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y. (2020). Decoupling representation and classifier for long-tailed recognition. In: *Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2018). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 29*(8), 3573–3587.

Kim, H.C., Ghahramani, Z. (2012). Bayesian classifier combination. In: *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics Artificial Intelligence and Statistics*, pp. 619–627

Kingma, D.P., Ba, J. (2015). Adam: A method for stochastic optimization. In: *Proceedings of the 3rd International Conference on Learning Representations*, p. (Poster)

Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., Philbin, J., & Fei-Fei, L. (2016). The unreasonable effectiveness of noisy data for fine-grained recognition. In: *Proceedings of the 14th European Conference on Computer Vision*, pp. 301–320

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1106–1114.

Lee, D.H., et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: *Workshop on challenges in representation learning, ICML,3*, 896

Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., Fan, W., & Han, J. (2014). A confidence-aware approach for truth discovery on long-tail data. In: *Proceedings of the VLDB Endowment*, *8*, 425–436

Li, S.Y., Huang, S.J., & Chen, S. (2021). Crowdsourcing aggregation with deep bayesian learning. *Science China Information Science*, *64*(3), 130104.1–130104.11

Li, S.Y., Jiang, Y. (2018). Multi-label crowdsourcing learning with incomplete annotations. In: *Proceedings of the 15th Pacific Rim International Conference on Artificial Intelligence*, pp. 232–245

Li, S. Y., Jiang, Y., Chawla, N. V., & Zhou, Z. H. (2019). Multi-label learning from crowds. *IEEE Transactions on Knowledge and Data Engineering, 31*(7), 1369–1382.

Li, X.C., Zhan, D.C., Yang, J.Q., & Shi, Y. (2021). Deep multiple instance selection. *Science China Information Science*, *64*(3)

Li, Y., Rubinstein, B.I.P., & Cohn, T. (2019). Exploiting worker correlation for label aggregation in crowdsourcing. In: *Proceedings of the 36th International Conference on Machine Learning*, pp. 3886–3895

Lin, T., Goyal, P., Girshick, R. B., He, K., & Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Knowledge and Data Engineering, 42*(2), 318–327.

Liu, Q., Peng, J., & Ihler, A. (2012). Variational inference for crowdsourcing. *Advances in Neural Information Processing Systems, 25*, 692–700.

McLachlan, G. J. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association, 70*(350), 365–369.

Moreno, P. G., Artés-Rodríguez, A., Teh, Y. W., & Perez-Cruz, F. (2015). Bayesian nonparametric crowdsourcing. *Journal of Machine Learning Research, 16*, 1607–1627.

Nassar, I., Herath, S., Abbasnejad, E., Buntine, W.L., & Haffari, G. (2021). All labels are not created equal: Enhancing semi-supervision via label grouping and co-training. In: *Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7241–7250

Pang, N., Zhao, X., Wang, W., Xiao, W., & Guo, D. (2021). Few-shot text classification by leveraging bi-directional attention and cross-class knowledge. *Science China Information Science*, *64*(3)

Peterson, J.C., Battleday, R.M., Griffiths, T.L., & Russakovsky, O. (2019). Human uncertainty makes classification more robust. In: *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, pp. 9616–9625

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., & Raiko, T. (2015). Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems,*, *28*, 3546–3554.

Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research, 11*, 1297–1322.

Rizve, M.N., Duarte, K., Rawat, Y.S., & Shah, M. (2021). In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In: *Proceedings of the 9th International Conference on Learning Representations*

Rodrigues, F., Pereira, F.C., & Ribeiro, B. (2014). Gaussian process classification and active learning with multiple annotators. In: *Proceedings of the 31th International Conference on Machine Learning*, pp. 433–441

Shadi, A., Christoph, B., Felix, A., Vasileios, B., Stefanie, D., & Nassir, N. (2016). Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Transactions on Medical Imaging, 35*(5), 1313–1321

Shen, L., Lin, Z., & Huang, Q. (2016). Relay backpropagation for effective learning of deep convolutional neural networks. In: *Proceedings of the 14th European conference on computer vision*, pp. 467–482

Shi, W., Gong, Y., Ding, C., Ma, Z., Tao, X., & Zheng, N. (2018). Transductive semi-supervised deep learning using min-max features. In: *Proceedings of the 15th European Conference of Computer Vision*, pp. 311–327

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., & Meng, D. (2019). Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in Neural Information Processing Systems, 32*, 1917–1928.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In: *Proceedings of the 3rd International Conference on Learning Representations*

Simpson, E., Roberts, S., Psorakis, I., Smith, A. (2013). Dynamic bayesian combination of multiple imperfect classifiers. Decision making and imperfection, p. 1-35

Smith, L.N. (2017). Cyclical learning rates for training neural networks. In: *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision*, pp. 464–472

Snow, R., O'Connor, B., Jurafsky, D., Ng, A. (2008). Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 254–263. Honolulu, Hawaii

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E.D., Kurakin, A., & Li, C. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence *Advances in Neural Information Processing Systems,* 33

Su, G., Chen, W., & Xu, M. (2021). Positive-unlabeled learning from imbalanced data. In: IJCAI.

Tang, K., Huang, J., & Zhang, H. (2020). Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33

Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., & Silberman, N. (2019). Learning from noisy labels by regularized estimation of annotator confusion. In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11244–11253

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: *Proceedings of the 5th International Conference on Learning Representations*

Thierry, B., Henrik, S.J., Marcel, F.R., & Rolf, P. (2010). Crowdsourcing, open innovation and collective intelligence in the scientific method: a research agenda and operational framework. In: *Proceedings of the 20th International Conference on the Synthesis and Simulation of Living Systems*, pp. 679–686.

Venanzi, M., Guiver, J., Kazai G.and Kohli, P., & Shokouhi, M. (2014). Community-based bayesian aggregation models for crowdsourcing. In: *Proceedings of the 23rd international conference on World wide web*, pp. 155–164. Seoul, Republic of Korea.

Wei, C., Sohn, K., Mellina, C., Yuille, A.L., & Yang, F. (2021). Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: *Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10857–10866

Welinder, P., Branson, S., Belongie, S., & Perona, P. (2010). The multidimensional wisdom of crowds. *Advances in Neural Information Processing Systems, 23*, 2024–2432.

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., & Movellan, J. R. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems, 22*, 2035–2043.

Xiao, T., Xia, T., Yang, Y., Huang, C., & Wang, X. (2015). Learning from massive noisy labeled data for image classification. In: *Proceedings of the 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2691–2699.

Xie, Q., Luong, M.T., Hovy, E.H., & Le, Q.V. (2020). Self-training with noisy student improves imagenet classification. In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698.

Xu, M., & Guo, L.Z. (2021). Learning from group supervision: the impact of supervision deficiency on multi-label learning. *Science China Information Science, 64*(3), 130101.1–130101.13

Yin, C., Menglin, J., Tsung-Yi, L., Yang, S., & Serge, B. (2019). Class-balanced loss based on effective number of samples. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277

Zhang, J., Wu, X., & Sheng, V. S. (2015). Active learning with imbalanced multiple noisy labeling. *IEEE Transactions on Cybernetics, 45*(5), 1081–1093.

Zhou, B., Cui, Q., Wei, X., & Chen, Z. (2020). BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9716–9725

Zhou, D., Basu, S., Mao, Y., & Platt, J. (2012). Learning from the wisdom of crowds by minimax entropy. *Advances in Neural Information Processing Systems, 25*, 2195–2203.