https://doi.org/10.1007/s11704-025-40860-0

RESEARCH ARTICLE

Robust long-tailed learning under label noise

Tong WEI[1,2,3*], Jiang-Xin SHI[3,4*], Min-Ling ZHANG[1,2], Yu-Feng LI[3,4]✉

1. School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
2. Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
3. National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
4. School of Artificial Intelligence, Nanjing University, Nanjing 210023, China

Received August 18, 2024; accepted March 19, 2025

E-mail: liyf@lamda.nju.edu.cn. * These authors contributed equally to this work.

Abstract

Long-tailed learning aims to enhance the generalization performance of underrepresented tail classes. However, previous methods have largely overlooked the prevalence of noisy labels in training data. In this paper, we address the challenge of noisy labels in long-tailed learning. We identify a critical issue: the commonly used small-loss noisy label detection criterion fails to perform effectively in long-tailed class distributions. This failure arises from the inherent bias of deep neural networks, which tend to misclassify tail class examples as head classes, leading to unreliable loss calculations. To mitigate this, we propose a novel small-distance criterion that leverages the robustness of learned representations, enabling more accurate identification of correctly-labeled examples across both head and tail classes. Additionally, to improve training for tail classes, we replace discrete pseudo-labels with label distributions for examples flagged as noisy, resulting in significant performance gains. Based on these contributions, we introduce the robust long-tail learning framework, designed to train models that are resilient to both class imbalance and noisy labels. Extensive experiments on benchmark and real-world datasets demonstrate that our approach outperforms previous methods, offering substantial performance improvements. Our source code is available at the website of github.com/Stomach-ache/RoLT

Keywords

long-tail learning; noisy labels; semi-supervised learning

1 Introduction

Classification problems in real-world typically exhibit a long-tailed class distribution, where most classes are associated with only a few examples, e.g., visual recognition [1–3], instance segmentation [4], and text categorization [5,6]. Due to the paucity of training examples, generalization for tail classes is challenging; moreover, naïve learning on such data is susceptible to an undesirable bias towards head classes. Recently, long-tail learning (LTL) has gained renewed interest in the community [7–15]. Two active strands of work involve normalisation of the classifier's weights, and modification of the underlying loss to account for different class penalties. Each of these strands is intuitive, and has been empirically shown to be effective [16–18].

Existing LTL methods with remarkable performance are mostly trained on *clean datasets* with high-quality human annotations. However, in real-world machine learning applications, annotating a large-scale dataset is costly and time-consuming. Some recent works resort to a large amount of web data as a source of supervision for training deep neural networks [19]. While the existing works have shown advantages in various applications [20,21], web data is naturally under long-tailed class distribution accompanied with noisy labels [22–25]. As a result, it is crucial that deep neural networks can harvest noisy and long-tailed training data. Nevertheless, deep neural networks (DNNs) are prone to overfitting noisy labels. This problem has been widely studied in the literature [26] on learning from noisy labels. Without considering noisy labels, we show that LTL methods severely degrade their performance in experiments.

In this paper, we investigate the problem of learning from noisy and long-tailed data, which is a more realistic setting but still underexplored. We provide a simple visualization of the studied problem in Fig. 1. To reduce the negative impact of noisy labels, learning with noisy labels has gained a lot of attention in recent years and a lot of approaches have been proposed [22–24,27–30]. Existing works can be roughly divided into two strands, i.e., noise transition matrix estimation [31,32] and sample selection [33–35]. Since the noise transition matrix is hard to be estimated especially when the number of classes is large, sample selection is a more promising way of handling noisy labels and is our focus in this paper. In sample

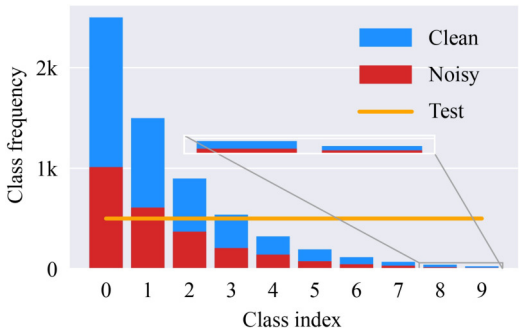

Fig. 1 Problem setting illustration

selection methods, the *small-loss* criterion is one of the most popular approaches [36–40]. It selects samples with small losses and treats these samples as correctly annotated for robust training. In recent years, the small-loss criterion has been demonstrated to be effective in many works.

However, the small-loss criterion selects possibly clean samples using a constant threshold [37,40] or mixture distribution model [38,39] for all classes, thus failing to consider different learning status and learning difficulties of different classes, which is very important in LTL. Owing to this paucity of samples for tail classes, naïve method is susceptible to an undesirable bias towards head classes. Specifically, due to the misclassification of samples with tail classes (large losses), the small-loss criterion cannot distinguish clean samples of tail classes from samples with noisy labels. Once the wrong selection is made, the inferiority of accumulated errors will arise. We further confirm this standpoint by experiments as shown in Figs. 2(a) and 2(b).

When handling noisy labels in long-tailed data, we believe it is important to keep the approach simple and free of auxiliary supervision. The benefit of doing this is that the approach can be easily absorbed by many existing frameworks for learning with long-tailed data. Guided by this belief, we propose the *class-wise small-distance* criterion. For each individual class, it selects small-distance samples as clean where distance is calculated between the sample and its class prototype in the embedding space. The intuition why the class-wise small-distance criterion can be more robust than small-loss is briefly explained as follows: 1) why is sample-distance better than small-loss? As confirmed by many previous literature [35,41–43], it is reasonable to assume that clean examples tend to be clustered around their prototypes even when training with noisy labels. 2) why can the sample selection work well in a class-wise manner? As the number of classes can be large and the population of classes varies significantly in training data, the variance of distances between samples and class prototypes becomes large. We show the distance distribution for both head and tail classes in Figs. 2(c) and 2(d). Moreover, the proposed *class-wise small-distance* is general and can be combined with semi-supervised learning to improve generalization.

Our main contributions are summarized as follows. 1) We address the underexplored problem of learning from long-tail data in the presence of noise, taking a significant step towards real-world applications. 2) We identify the limitations of the widely used small-loss criterion under long-tailed class distributions and introduce a novel criterion called class-wise small-distance to overcome this challenge. Building upon this, we propose a robust framework called RoLT, which incorporates label distributions to enhance training for tail classes. Additionally, we present an improved version, RoLT+, which can be seamlessly integrated with existing semi-supervised learning methods with minimal additional overhead. 3) Extensive experiments conducted on benchmark and real-world datasets demonstrate the superiority of our proposed method.

2 Related work

● Long-tail learning. Recently, many approaches have been proposed in LTL [12,18,44,45]. Most extant approaches can be categorized into three types by modifying: (i) the inputs to a model by re-balancing the training data [2,46,47]; (ii) the outputs of a model, for example by post-hoc adjustment of the classifier [16,48–50]; and (iii) the internals of a model by modifying the loss function [11,17,51–54]. Each of the above methods are intuitive, and have shown strong empirical performance. However, these methods assume the training examples are correctly-labeled, which is often difficult to obtain in many real-world applications. Instead, we study a realistic problem to learn from long-tailed data with label noise. Although the presence of label noise in class-imbalanced datasets has also been mentioned in HAR [55], they only consider a specific setup. In this work, we provide systematic studies concerning noisy labels in LTL.

● Label noise detection. Plenty of methods have been proposed to detect noisy examples [36,37,39,56–60]. Many works adopt the small-loss trick, which treats examples with small training losses as correctly-labeled. In particular, MentorNet [36] reweights samples according to their loss magnitude so that noisy samples contribute less to the loss. Co-teaching [37] trains two networks where each network selects small-loss samples in a mini-batch to train the other.

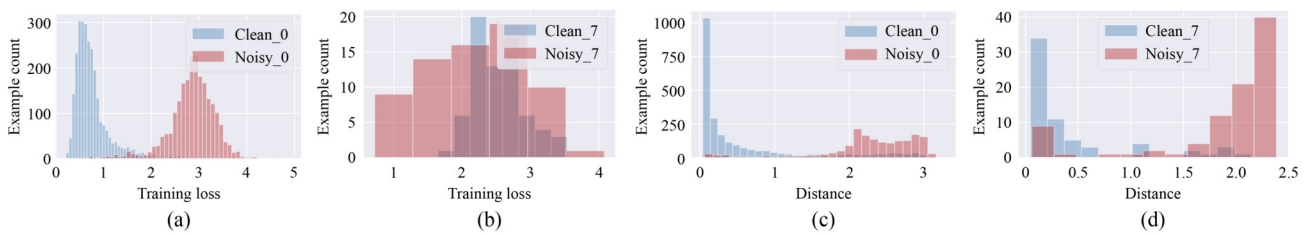

Fig. 2 (a–b) Training losses head and tail class samples. (c–d) Distance distribution between samples and their class prototype for head and tail class. Experiments are conducted on CIFAR-10 with noise level $\gamma = 50\%$ and imbalance ratio $\rho = 100$

DivideMix [39] fits a Gaussian mixture model on per-sample loss distribution to divide the training data into a clean set and a noisy set. In addition, AUM [61] introduces a margin statistic to identify noisy samples by measuring the average difference between the logit values for a sample's assigned class and its highest non-assigned class. The above methods only consider training datasets that are class-balanced and thus are not applicable for long-tailed label distribution. Recent studies [25,62] have observes that real-world datasets with label noise often exhibit class imbalanced as well. Nevertheless, they only inspect a particular case, but we provide a systematic study of noisy labels under various long-tailed scenarios and propose a novel class-wise prototypical noise detection method.

3 RoLT: Robust long-tail learning with noisy labels

3.1 Problem setting & background

Given a training dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$, where $\boldsymbol{x}_i$ is an instance feature vector and $y_i \in C = [K] = \{1, \ldots, K\}$ is the class label assigned to it. The training examples $(\boldsymbol{x}_i, y_i), 1 \leqslant i \leqslant N$ consists of two types: 1) a *correctly-labeled example* whose assigned label matches the ground-truth label, i.e., $y_i = y_i^*$, where y_i^* denotes the ground-truth label of $\boldsymbol{x}_i$, 2) a *mislabeled example* whose assigned label does not match the ground-truth label, but the input matches one of the classes in C, i.e., $y_i \neq y_i^*$ and $y_i^* \in C$. Moreover, the data follows a long-tailed class distribution where the class prior distribution $\mathbb{P}(y)$ is highly skewed, so many underrepresented classes have a very low probability of occurrence. Specifically, we denote the imbalance ratio as $\rho = \max_y \mathbb{P}(y) / \min_y \mathbb{P}(y)$ to indicate the skewness of data. Classes with high $\mathbb{P}(y)$ are referred to as *head classes*, and others are referred to as *tail classes*.

In practice, since the data distribution is unknown, Empirical Risk Minimization (ERM) uses the training data to achieve an empirical estimate of the underlying data distribution. Typically, one minimizes the softmax cross-entropy as following

$$\ell(y, f(\boldsymbol{x})) = \log\Big[\sum_{y' \in [K]} e^{f_{y'}(\boldsymbol{x})}\Big] - f_y(\boldsymbol{x}), \tag{1}$$

where $f_y(\boldsymbol{x})$ denotes the predictive probability of model f on class y. This ubiquitous approach neglects the issue of class imbalance and makes the model biased toward head classes. Moreover, it assumes training examples are correctly-labeled.

To reduce the impact of noisy labels, sample selection based on the small-loss criterion is one of the most popular approaches. It selects "easy" samples for training based on the outputs of the current networks. Specifically, given an arbitrary training example $(\boldsymbol{x}_i, y_i)$, we can obtain a loss ℓ_i, i.e., $\ell_i = \ell(f(\boldsymbol{x}_i), y_i)$. Then a certain proportion of samples with small losses are selected as probably clean samples. The selection can be achieved by a manually set hyperparameter [37] or a mixture distribution model that fits two-component mixtures on the loss distribution [38].

The small-loss criterion is demonstrated to be effective in the existing literature [36–40], but we find that it does not work well on the long-tail dataset. It is known that DNNs learn "easy" samples first, and on long-tail datasets, DNNs first learn to recognize head classes ahead of tail classes. In light of that, DNNs tend to misclassify tail class examples as head classes. This phenomenon has been widely revealed in recent works [16,47]. Therefore, the losses cannot reflect the probability that a sample is mislabeled, especially for tail classes as shown in Figs. 2(a) and 2(b).

3.2 Class-wise distance-based sample selection

We present a new criterion for selecting noisy labels, called *small-distance*. Formally, we model clean examples of class $k \in [K]$ as if they are distributed around prototype $\boldsymbol{c}_k \in \mathbb{R}^D$, and the likelihood of an example $\boldsymbol{x}$ belonging to class k decays exponentially with its distance from the prototype $\boldsymbol{c}_k$, i.e., $\mathbb{P}(\boldsymbol{x} \mid \boldsymbol{c}_k) \propto e^{-dist(\boldsymbol{c}_k, \boldsymbol{x})}$, which is a common assumption about the data distribution [63,64]. Here, $dist(\cdot, \cdot)$ represents any distance measured in the embedding space and is typically set to be the Euclidean distance. Considering the variance of data distribution of different classes, we propose to inspect the distance statistics in a *class-wise* manner.

To justify the feasibility of using distance to select clean examples, we compute the AUC based on the distance between examples and their class prototypes for each class separately and report the average value of Many (more than 100 images), Medium (20–100 images), Few (less than 20 images) and All shots in Table 1. The experiment is done on CIFAR-100 with imbalance ratio $\rho = 100$, noise level $\gamma = 20\%$ and $\gamma = 50\%$. It can be seen that the AUC is high even for tail classes, indicating the distance measure may be a useful approach to distinguish clean and noisy examples. Notice that, previous literature [35,41,42] also confirms that the representations are resilient to noisy labels.

As mentioned earlier, sample selection can be accomplished using either a thresholding hyperparameter or a mixture distribution model. In line with previous research [38,39], we adopt the latter approach. Specifically, for class k, we employ a two-component GMM [65] to model the distance distribution of clean and noisy samples, i.e., $d \sim \sum_{j=1}^2 \phi_j \mathcal{N}(\mu_j, \sigma_j^2)$, where $d = dist(\boldsymbol{c}_k, \boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{D}_k$ and ϕ_j denotes weight of the j-th component. Note that we have $\sum_{j=1}^2 \phi_j = 1$. Without loss of generality, we assume $\mu_1 < \mu_2$. Since clean examples are located around the prototype while noisy examples spread out, we flag $\boldsymbol{x}$ as clean if and only if $\mathbb{P}(d \mid \mu_1, \sigma_1) > \mathbb{P}(d \mid \mu_2, \sigma_2)$. We thus perform noise detection by estimating the Gaussians' parameters from distance statistics. In particular, for each class k, we compute its prototype as the

Table 1 Average AUC of different classes

Noise level	Many	Medium	Few	All
$\gamma = 20\%$	95.16	93.38	82.81	90.64
$\gamma = 50\%$	92.00	87.43	73.60	85.20

normalized average of the embeddings for training examples by

$$\boldsymbol{c}_k \leftarrow \text{Normalize}\left(\frac{1}{|\mathcal{D}_k|}\sum_{\boldsymbol{x}_i\in\mathcal{D}_k} f_\theta(\boldsymbol{x}_i)\right), \forall k\in[K], \tag{2}$$

where $f_\theta(\boldsymbol{x})$ denotes the extracted feature of $\boldsymbol{x}$ and $\forall \boldsymbol{x}_i \in \mathcal{D}_k$. Given $\boldsymbol{c}_k$, the distances between $\boldsymbol{c}_k$ and examples of class k are obtained by

$$dist(\boldsymbol{c}_k, \boldsymbol{x}_i) = \left\|\boldsymbol{c}_k - f_\theta(\boldsymbol{x}_i)\right\|_2^2. \tag{3}$$

We then fit a two-component Gaussian mixture model to maximize the log-likelihood value, i.e., $\max \sum_{i=1}^{|\mathcal{D}_k|} \log(\sum_{j=1}^{2} \phi_j \mathbb{P}(d_k(\boldsymbol{x}_i) \mid \mu_j, \sigma_j))$, where $d_k(\boldsymbol{x}_i) = dist(\boldsymbol{c}_k, \boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in \mathcal{D}_k$.

For simplicity, we denote the clean (noisy) data of class k as $\mathcal{X}_k$ ($\mathcal{S}_k$). Note that we have $\mathcal{D}_k = \mathcal{X}_k \bigcup \mathcal{S}_k$. Therefore, we obtain a subset of clean examples by $\mathcal{X} = \bigcup_{k=1}^{K} \mathcal{X}_k$ and noisy examples by $\mathcal{S} = \bigcup_{k=1}^{K} \mathcal{S}_k$. It is also verified that the Gaussian mixture model can be used to distinguish clean and noisy data because of its flexibility in the sharpness of distribution in previous literature [39]. We also observe that the proposed method works well on real-world datasets where the Gaussian distribution assumption is not perfectly satisfied. Recall that $\mathcal{D}_k$ may contain noisy labels, the estimate of $\boldsymbol{c}_k$ in (2) is inaccurate and the split of $\mathcal{D}_k = \mathcal{X}_k \bigcup \mathcal{S}_k$ is problematic. To remedy this, we refine class prototypes using $\mathcal{X}_k$ rather than $\mathcal{D}_k$, and acquire a new split of $\mathcal{D}_k$. By doing this, $\mathcal{X}_k$ retains most of correctly-labeled examples of class k as well as fewer mislabeled examples.

3.3 Soft pseudo-labeling with label distribution

For each example that is likely to be mislabeled, its original discrete noisy label is transformed into a continuous label distribution to incorporate the uncertainty of the ground-truth label. The benefits of using label distributions are two-fold. First, it mitigates the influence of noisy labels [66]; Second, it compensates for the learning of data scarcity tail classes [67]. To this end, the underrepresented tail classes will receive significant improvements, which is crucial in LTL. To generate label distributions, a direct approach is to leverage the prediction of the ERM (Empirical Risk Minimization) model. However, the ERM is known to be biased toward head classes [67]. Hence, refining noisy labels using the predictions of ERM may be sub-optimal for examples of tail classes. In contrast, the NCM (Nearest Class Mean) classifier can yield a balanced classification boundary [16]. Specifically, we find that the NCM classifier produces much higher recall on tail classes than the ERM in experiments. By aggregating the predictive information from the ERM and NCM classifiers, we construct diverse soft pseudo-labels for detected noisy examples. To amend the misflag of the noise detector, we also take into account the original labels as a source of soft pseudo-labels. Since the predictions of ERM and NCM classifiers may disagree, we further remedy this by the label smoothing technique [67].

Given the predictions $\hat{y}^{erm} = \arg\max_k f(\boldsymbol{x})$, $\hat{y}^{ncm} = \arg\min_k \|\boldsymbol{c}_k - f_\theta(\boldsymbol{x})\|_2$, and original noisy label y, we construct the guessing label set $\mathcal{G} = \{\hat{y}^{erm}, \hat{y}^{ncm}, y\}$ and generate the label distribution $\tilde{\boldsymbol{y}} \in \mathbb{R}^K$ for $\boldsymbol{x}$. For $k \in [K]$, we compute

$$\tilde{y}_k = \begin{cases} \frac{1}{4}\sum_{\hat{y}\in\mathcal{G}} \mathbb{I}(\hat{y}=k) + \frac{1}{4K}, & \text{if } k \in \mathcal{G}, \\ \frac{1}{4K}, & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathbb{I}(\cdot)$ is an indicator which returns 1 if the condition is true, otherwise 0. Considering the classification task with four classes (i.e., $K=4$) and $\mathcal{G}=\{1,4,2\}$, the soft pseudo-label would be $\tilde{\boldsymbol{y}} = [\frac{5}{16}, \frac{5}{16}, \frac{1}{16}, \frac{5}{16}]$. The targets $\hat{y}^{erm}$ and $\hat{y}^{ncm}$ can be set equal to the model output, but using a running average is more effective which is known as temporal ensembling [68] in semi-supervised learning. For ERM or NCM classifier, let $\boldsymbol{z}_i(t) \in \mathbb{R}^K$ be the output logits vector (pre-softmax output) for example $\boldsymbol{x}_i$ at iteration t of training, we update the momentum logits by

$$\boldsymbol{q}_i(t) = \alpha \boldsymbol{q}_i(t-1) + (1-\alpha)\boldsymbol{z}_i(t), \tag{5}$$

where $0 \leqslant \alpha < 1$ is the combination weight. For each iteration t, we then obtain $\hat{y}^{erm}$ and $\hat{y}^{ncm}$ using softmax of $\boldsymbol{q}_i(t)$. Having acquired $\mathcal{X}$, $\mathcal{S}$, and soft pseudo-labels, we compute the cross-entropy loss for clean examples using original training labels by

$$\mathcal{L}_{\mathcal{X}} = \frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}_i\in\mathcal{X}} H(\boldsymbol{y}_i, f(\boldsymbol{x}_i)), \tag{6}$$

where $\boldsymbol{y}_i$ is the one-hot label vector for $\boldsymbol{x}_i$. For noisy examples, the loss function is computed by

$$\mathcal{L}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{x}_i\in\mathcal{S}} H(\tilde{\boldsymbol{y}}_i, f(\boldsymbol{x}_i)), \tag{7}$$

where $H(\boldsymbol{q}, \boldsymbol{p}) = -\sum_{k=1}^{K} q_k \log\left(\frac{\exp p_k}{\sum_{j=1}^{K}\exp p_j}\right)$ is the cross-entropy between distributions $\boldsymbol{q}$ and $\boldsymbol{p}$. Overall, the training objective is $\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \mathcal{L}_{\mathcal{S}}$. We illustrate the whole proposed framework in Fig. 3.

3.4 Combining with semi-supervised learning

The proposed method can be further improved by using a well-established semi-supervised learning approach, where clean and noisy examples are viewed as labeled and unlabeled data, respectively. Inspired by DivideMix [39] and ELR+ [69], we use two separate neural networks, where the target of each network is computed from the output of the other network. For fair comparisons, we replace the sample selection module of DivideMix with our proposed class-wise prototypical noise detector. We call this improved method RoLT+.

3.5 Difference with prior works

Class prototypes are employed in some previous literature and we discuss the differences between this paper and some related works. In few-shot learning, prototypical networks [70] learn a metric space in which classification can be performed by computing distances to prototype representations of each class. Compared to other methods, prototypical networks reflect a simpler inductive bias that is beneficial in this limited-data regime, and achieve excellent results. In long-tailed recognition, OLTR [2] proposes to use the distances between samples and prototypes to handle open-set recognition. In self-supervised learning, PCL [71] proposes the ProtoNCE loss which encourages representations to be closer to their assigned

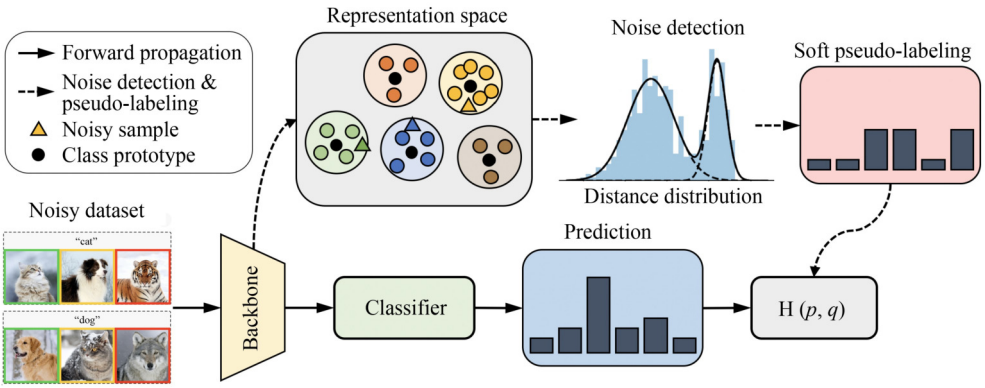

Fig. 3 The proposed framework RoLT

prototypes. However, the above-mentioned works do not consider using class prototypes to detect noisy labels in long-tailed class distribution. Moreover, the distances calculated between examples and their class prototypes are utilized in a *class-wise* manner to mitigate the influence of long-tailed class distribution.

4 Experiments

4.1 Results on simulated datasets

4.1.1 Setting

We test RoLT on CIFAR-10 and CIFAR-100 under various imbalanced ratio ρ and noise level γ. CIFAR-10 consists of 50k images across 10 classes, while CIFAR-100 contains 50k images but with 100 classes. For each dataset, we first simulate the long-tailed dataset following the same setting as LDAM [51]. The long-tailed imbalance follows an exponential decay in sample sizes across different classes. To inject noisy labels, we define the noise transition matrix or the asymmetric noise policy to construct the training set. In particular, we consider imbalance ratio $\rho \in \{10, 50, 100\}$ and noise level $\gamma \in \{20\%, 50\%\}$.

4.1.2 Results under simulated noise

Tables 2 and 3 summarize the results for CIFAR-10 and CIFAR-100.

Table 2 Comparisons of methods using a single model. Test accuracy (%) on CIFAR datasets with different imbalance ratios (i.e., 10, 50, and 100) and noise levels (i.e., 20%, 50%)

Method		CIFAR-10						CIFAR-100					
		20%			50%			20%			50%		
		10	50	100	10	50	100	10	50	100	10	50	100
CE	Best	75.61	63.60	62.17	63.25	43.38	38.71	43.27	30.27	26.21	26.92	16.97	14.23
	Last	73.86	58.60	54.15	48.39	32.73	27.05	43.06	30.04	26.08	26.60	15.59	13.47
LDAM	Best	82.37	71.34	66.26	60.30	42.95	36.66	48.14	33.43	29.70	29.62	17.51	14.19
	Last	82.07	71.11	65.88	59.10	38.33	33.38	47.89	33.30	29.50	29.38	17.29	13.24
LDAM-DRW	Best	83.73	76.41	72.28	67.93	48.88	43.23	50.44	36.60	32.27	32.24	19.48	15.21
	Last	83.67	75.67	71.08	67.68	47.38	41.45	50.29	36.16	32.05	31.72	19.23	14.75
BBN	Best	77.81	68.01	64.51	64.71	46.22	36.72	47.60	31.07	28.79	30.01	19.75	14.56
	Last	76.81	67.48	64.24	53.76	43.35	34.83	47.26	30.76	28.56	29.42	19.55	14.34
cRT	Best	76.15	65.02	59.92	64.15	43.26	36.73	42.56	30.23	26.31	25.55	17.47	16.01
	Last	75.05	64.22	58.47	62.75	41.87	34.55	41.56	30.08	26.18	23.94	17.34	15.94
MW-Net	Best	82.19	71.63	67.26	72.12	56.09	46.36	50.20	36.68	31.77	37.50	23.99	**21.24**
	Last	77.67	64.12	58.23	59.68	45.39	37.05	47.82	34.45	29.57	33.14	20.33	18.82
HAR-DRW	Best	82.43	67.44	67.88	67.39	52.35	42.80	46.24	28.86	26.29	31.30	16.75	14.78
	Last	78.44	61.08	62.73	64.75	45.06	40.07	43.04	26.11	24.71	26.96	13.87	12.42
RoLT	Best	**85.04**	**77.86**	**73.84**	**77.11**	**60.15**	**55.32**	**53.41**	**38.94**	**33.36**	**39.22**	**25.51**	20.61
	Last	**84.95**	**77.65**	**73.54**	**76.94**	**59.59**	**54.55**	**53.22**	**38.77**	**33.20**	**39.01**	**25.35**	**20.45**

Table 3 Test accuracy (%) of methods using two models on CIFAR datasets with different imbalance ratios (i.e., 10, 50, and 100) and noise levels (i.e., 20%, 50%)

Method		CIFAR-10						CIFAR-100					
		20%			50%			20%			50%		
		10	50	100	10	50	100	10	50	100	10	50	100
ELR+	Best	**88.96**	**80.21**	69.60	85.02	56.96	48.72	54.01	49.64	38.40	49.53	30.12	21.58
	Last	88.09	**79.69**	66.67	84.08	48.14	43.11	53.32	48.37	38.12	49.06	29.68	20.47
DivideMix	Best	88.79	75.34	66.90	87.54	67.92	61.81	63.79	49.64	43.91	49.35	36.52	31.82
	Last	**88.10**	73.48	63.76	86.88	65.22	59.65	63.17	48.37	42.59	48.87	35.72	31.05
RoLT+	Best	87.95	77.26	**72.31**	**88.17**	**75.11**	**64.42**	**64.22**	**51.01**	**45.35**	**53.31**	**39.78**	**35.29**
	Last	87.54	75.90	**69.12**	**87.45**	**73.92**	**61.15**	**63.31**	**49.40**	**43.16**	**52.44**	**39.27**	**34.43**

As shown in the results, previous LTL methods (i.e., LDAM, BBN [47], and cRT [16]) dreadfully degrade their performance as the noise level and imbalance ratio increase, while our methods retain robust performance. In particular, compared with CE, RoLT improves the test accuracy by 8% on average. It can be observed that the improvement becomes more significant at high noise levels, benefiting from proposed noise detection and soft pseudo-labeling. Further application of Deferred Re-Weighting (DRW) [51] enhances the performance by favoring the tail classes. This clearly demonstrates the importance of correcting noisy labels in the training data. Notice that MW-Net [52] and HAR-DRW [55] are proposed to handle label noise and class imbalance. Our method consistently outperforms them by a large margin.

We further compare RoLT+ with DivideMix [39] and ELR+ [69], which are the most popular methods for learning with noisy labels. The results are given in Table 3. First, we can see that the performance of ELR+ significantly drops as the training set becomes class-imbalanced. DivideMix is relatively robust to class imbalance than ELR+ by imposing the uniform predictions regularization in its objective. In contrast, our method RoLT+ achieves performance improvements in test accuracy by 2.57% on average. This validates the superiority of our noise detector.

4.1.3 Results under asymmetric noise

We further verify the effectiveness of the proposed detection method and the robust framework under asymmetric label noise. We conduct experiments on the long-tail CIFAR-10 dataset using the noise injection policy following previous works [39,69]. From Table 4, it can be seen that RoLT+ consistently outperforms DivideMix in all cases. It is interesting to observe that the performance gap between best and last widens as the imbalance ratio becomes large. This reflects that the model is easy to collapse under asymmetric noise and class imbalance. The proposed method can alleviate this issue substantially.

4.2 Results on real-world dataset

4.2.1 Results on WebVision dataset

We test the performance of our method on a real-world dataset. WebVision [19] contains 2.4 million images collected from Flickr and Google, characterized by real-world label noise and class imbalance. The estimated noise level is 20%. Following previous literature, we use a subset of WebVision, known as mini WebVision, which includes the first 50 classes. In Table 5, we compare RoLT+ with several previous approaches, including D2L [21], MentorNet [36], Co-teaching [37], Iterative-CV [72], HAR [55], DivideMix [39], and [69]. RoLT+ achieves superior performance than DivideMix and, particularly in terms of the top5 accuracy.

To further uncover the advantages of our method, we run experiments by controlling the imbalance ratio of WebVision dataset. The test accuracy is reported in Table 6. From the results, we can see that the superiority of our method is more significant as the imbalance ratio increases.

4.2.2 Results on ImageNet-LT dataset

We compare our method with baselines (CE and ELR) on a large long-tail benchmark, i.e., ImageNet-LT, by combining with methods

Table 4 Results on CIFAR-10 with asymmetric noise with different imbalance ratios (i.e., 10, 50, and 100) and noise levels (i.e., 20%, 40%)

Method		20%			40%		
		10	50	100	10	50	100
DivideMix	Best	83.48	73.13	66.51	78.85	67.74	59.63
	Last	78.77	56.90	44.50	74.50	46.48	32.68
RoLT+	Best	**84.05**	**75.93**	**68.53**	**79.98**	**71.44**	**65.56**
	Last	**79.81**	**67.13**	**60.38**	**74.68**	**60.07**	**54.05**

Table 5 Accuracy (%) on mini WebVision and ImageNet validation sets

Method	Metric	D2L	MentorNet	Co-teaching	Iterative-CV	HAR	DivideMix	RoLT+
WebVision	top1	62.68	63.00	63.58	65.24	75.5	77.32	**77.64**
	top5	84.00	81.40	85.20	85.34	90.7	91.64	**92.44**
ImageNet	top1	57.80	57.80	61.48	61.60	70.3	**75.20**	74.64
	top5	81.36	79.92	84.70	84.98	90.0	90.84	**92.48**

Table 6 Top1 (top5) accuracy (%) on WebVision and ImageNet

Method	Imbalance ratio $\rho = 50$		Imbalance ratio $\rho = 100$	
	WebVision	ImageNet	WebVision	ImageNet
DivideMix	64.56 (83.56)	62.68 (85.24)	55.76 (73.48)	53.92 (74.00)
w/ DRW	68.16 (84.92)	66.12 (85.40)	60.28 (74.60)	59.04 (75.68)
RoLT+	66.28 (88.68)	64.76 (89.96)	60.68 (87.84)	59.68 (88.52)
w/ DRW	70.08 (88.52)	67.28 (90.12)	65.48 (87.32)	64.80 (87.08)

Table 7 Accuracy (%) on ImageNet-LT dataset

Noise level	Method	–	w/ cRT	w/ LA
$\gamma = 20\%$	CE	28.18	34.15	34.06
	ELR	26.58	35.21	34.05
	RoLT	**29.57**	**35.76**	**35.09**
$\gamma = 50\%$	CE	17.80	21.85	22.71
	ELR	17.33	22.80	22.60
	RoLT	**21.53**	**25.61**	**25.50**

and strategies for class-imbalanced datasets, i.e., Classifier Re-training (cRT) [16] and Logit Adjustment (LA) [50]. The ImageNet-LT dataset consists of 1000 classes and its imbalance ratio is 200. We simulate various levels of label noise to assess the performance under different noise levels. From the results in Table 7, we can clearly see that our method outperforms other methods, particularly under high noise levels.

4.3 Further analysis and ablation studies

Efficacy of the noise detector. To further support our motivation, we compare the performance of the ERM and NCM classifiers in Fig. 4. It can be seen that NCM produces more balanced recall across classes, while ERM tends to predict examples as head classes, resulting in low recall for tail classes. Figure 5 shows the precision and recall of selected clean examples by our method. To better understand RoLT, we construct three groups of classes for CIFAR-100 by: many (more than 100 images), medium (20−100 images), and few (less than 20 images) shots; and CIFAR-10 by: many ($\{0,1\}$), medium ($\{2,\ldots,6\}$), and few ($\{7,8,9\}$) shots according to class indices. RoLT maintains high precision and recall, which validates the effectiveness of our method. This experiment is conducted under the imbalance ratio $\rho = 100$ and noise level $\gamma = 30\%$.

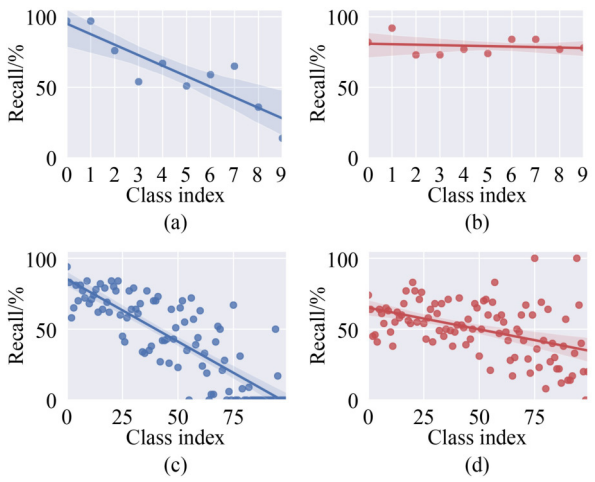

Fig. 4 Per-class recall of ERM and NCM classifiers. (a) ERM on CIFAR-10; (b) NCM on CIFAR-10; (c) ERM on CIFAR-100; (d) NCM on CIFAR-100

Efficacy of the soft pseudo-labeling. We investigate the effectiveness of soft pseudo-labeling by comparing it with three other

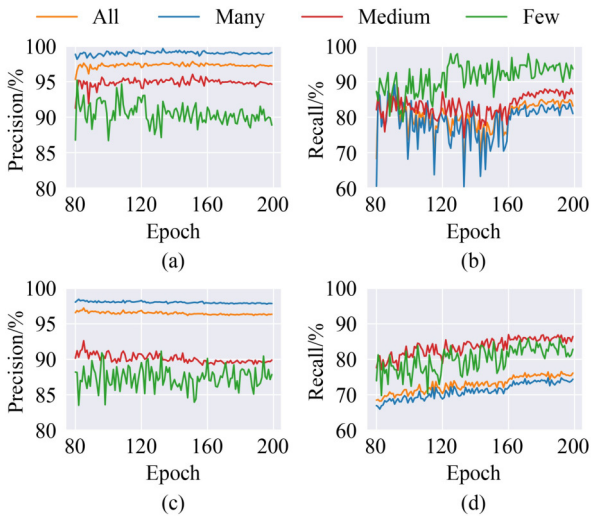

Fig. 5 Precision and Recall of selected samples. (a) CIFAR-10 Precision; (b) CIFAR-10 Recall; (c) CIFAR-100 Precision; (d) CIFAR-100 Recall

methods: (i) "Noisy" keeps the noisy labels, (ii) "ERM" rectifies labels via the ERM predictions, (iii) "Soft (w/o LS)" uses the label distribution without label smoothing (w/o LS). The "Soft (w/o LS)" method produces label distribution as following:

$$\tilde{y}_k = \begin{cases} \frac{1}{3}\sum_{\hat{y}\in\mathcal{G}} \mathbb{I}(\hat{y}=k), & \text{if } k \in \mathcal{G}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

We report the results in Table 8 for noise level $\gamma = 50\%$ and imbalance ratio $\rho = 100$. We observe that ERM and soft pseudo-labeling significantly improve the performance by over 4% in test accuracy, and the improvement is more significant under high noise levels. Moreover, the soft pseudo-labeling outperforms its ERM and "w/o LS" counterpart in most cases, demonstrating that label smoothing and label guessing can provide diverse and informative supervision under imperfect training labels.

Results on class-balanced datasets. We compare the performance of our method with DivideMix on balanced datasets with noise level $\rho \in \{20\%, 50\%\}$. The results are reported in Table 9 and our method is comparable with DivideMix. This shows that the proposed prototypical noise detector also works well on balanced datasets.

Results on clean datasets. We present the results on clean datasets in Table 10. Interestingly, RoLT consistently outperforms vanilla CE in all scenarios, highlighting the advantages of our proposed soft pseudo-labeling approach. Additionally, our method achieves performance comparable to the widely-used LDAM-DRW baseline. When compared to HAR-DRW, which also addresses class imbalance and label noise, our approach provides an average improvement of over 2%. Using an auxiliary $1k$ clean and class-balanced validation set, MW-Net achieves performance similar to RoLT-DRW in four out of six cases. However, in the remaining two cases, RoLT-DRW shows a significant performance gain. Moreover, obtaining such an auxiliary $1k$ clean and class-balanced validation set is challenging in practice, demonstrating the robustness of our method.

Table 8 Ablation studies on pseudo-labeling. Test accuracy on CIFAR-100 is reported

DRW	Method	γ=50%			
		Many	Medium	Few	All
×	Noisy	32.06	7.89	0.04	14.23
×	ERM	38.83	12.05	0.89	18.41
×	Soft (w/o LS)	36.03	15.11	2.19	18.94
×	Soft (w/ LS)	37.20	15.97	1.74	**19.56**
√	Noisy	23.77	14.53	3.41	14.76
√	ERM	32.80	17.05	2.30	18.58
√	Soft (w/o LS)	30.31	18.92	4.93	19.13
√	Soft (w/ LS)	30.94	21.32	6.22	**20.61**

Table 9 Test accuracy (%) on class-balanced datasets with different noise levels

Method		CIFAR-10		CIFAR-100	
		20%	50%	20%	50%
DivideMix	Best	**92.79**	**95.03**	77.25	73.84
	Last	**92.41**	**94.63**	77.03	**73.42**
RoLT+	Best	92.46	94.59	**78.60**	**74.11**
	Last	92.01	94.41	**78.14**	73.35

Table 10 Test accuracy (%) on clean datasets with different imbalance ratios (i.e., 10, 50, and 100)

Method	CIFAR-10			CIFAR-100		
	10	50	100	10	50	100
CE	86.75	77.38	71.83	56.31	44.15	38.88
LDAM	86.38	77.62	74.31	55.66	43.61	39.25
LDAM-DRW	87.29	81.25	78.78	57.21	**47.30**	**42.93**
BBN	87.83	81.19	78.87	58.08	45.62	40.09
cRT	86.78	77.30	71.18	56.62	43.01	39.44
MW-Net†	**87.99**	79.58	74.92	**58.66**	46.72	42.10
HAR-DRW	87.81	79.82	75.99	56.89	43.34	40.78
RoLT	**87.99**	80.50	77.70	57.47	45.38	39.35
RoLT-DRW	87.75	**83.02**	**80.57**	57.48	47.21	41.70

† MW-Net uses a 1k clean and class-balanced validation set.

5 Conclusion

In this paper, we introduced a robust learning framework designed to address both class imbalance and noisy labels. We highlighted the limitations of the widely-used small-loss sample selection criterion under long-tailed class distributions and proposed a novel small-distance criterion that more accurately identifies correctly-labeled examples across both head and tail classes. Additionally, we introduced a soft pseudo-labeling technique to further enhance training for tail classes. Through extensive experiments on benchmark and real-world datasets, we demonstrated the superiority of our framework compared with existing approaches in both long-tail learning and noisy label detection. We believe that this work will inspire further research into this underexplored yet highly practical task.

While our framework performs well in handling noisy labels, we note that, in the case of clean datasets, the noise detector may still flag certain examples as noisy due to the model's inherent bias toward a two-component Gaussian Mixture Model, leading to the loss of some accurate supervision. Addressing this challenge, by accurately estimating the noise proportion in the training data, presents an interesting direction for future research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62206049, 62225602), the Key Program of Jiangsu Science Foundation (BK20243012), and the Big Data Computing Center of Southeast University. We would like to thank anonymous reviewers for their constructive suggestions.

Competing interests

Min-Ling Zhang is an Action Editor of the journal and a co-author of this article. To minimize bias, he was excluded from all editorial decision-making related to the acceptance of this article for publication. The remaining authors declare no conflict of interest.

References

[1] Van Horn G, Aodha O M, Song Y, Cui Y, Sun C, Shepard A, Adam H, Perona P, Belongie S J. The iNaturalist species classification and detection dataset. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 8769−8778

[2] Liu Z, Miao Z, Zhan X, Wang J, Gong B, Yu S X. Large-scale long-tailed recognition in an open world. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 2532−2541

[3] Tan J, Wang C, Li B, Li Q, Ouyang W, Yin C, Yan J. Equalization loss for long-tailed object recognition. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 11659−11668

[4] Gupta A, Dollár P, Girshick R. LVIS: a dataset for large vocabulary instance segmentation. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 5351−5359

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proceedings of

the 31st International Conference on Neural Information Processing Systems. 2017, 6000−6010

[6] Wei T, Li Y F. Does tail label help for large-scale multi-label learning? IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(7): 2315−2324

[7] Cardie C, Nowe N. Improving minority class prediction using case-specific feature weights. In: Proceedings of the 14th International Conference on Machine Learning. 1997, 57−65

[8] Zhou Z H, Liu X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 63−77

[9] He H, Garcia E A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263−1284

[10] Wang Y, Ramanan D, Hebert M. Learning to model the tail. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 7032−7042

[11] Cui Y, Jia M, Lin T Y, Song Y, Belongie S. Class-balanced loss based on effective number of samples. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 9260−9269

[12] Wang X, Lian L, Miao Z, Liu Z, Yu S X. Long-tailed recognition by routing diverse distribution-aware experts. In: Proceedings of the 9th International Conference on Learning Representations. 2021

[13] Wei T, Tu W W, Li Y F, Yang G P. Towards robust prediction on tail labels. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021, 1812−1820

[14] Wei T, Gan K. Towards realistic long-tailed semi-supervised learning: consistency is all you need. In: Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023, 3469−3478

[15] Shi J X, Wei T, Zhou Z, Shao J J, Han X Y, Li Y F. Long-tail learning with foundation model: heavy fine-tuning hurts. In: Proceedings of the 41st International Conference on Machine Learning. 2024, 45014−45039

[16] Kang B, Xie S, Rohrbach M, Yan Z, Gordo A, Feng J, Kalantidis Y. Decoupling representation and classifier for long-tailed recognition. In: Proceedings of the 8th International Conference on Learning Representations. 2020

[17] Wu T, Huang Q, Liu Z, Wang Y, Lin D. Distribution-balanced loss for multi-label classification in long-tailed datasets. In: Proceedings of the 16th European Conference on Computer Vision. 2020, 162−178

[18] Wu T, Liu Z, Huang Q, Wang Y, Lin D. Adversarial robustness under long-tailed distribution. In: Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 8655−8664

[19] Li W, Wang L, Li W, Agustsson E, Van Gool L. WebVision database: visual learning and understanding from web data. 2017, arXiv preprint arXiv: 1708.02862

[20] Li W, Niu L, Xu D. Exploiting privileged information from web data for image categorization. In: Proceedings of the 13th European Conference on Computer Vision. 2014, 437−452

[21] Ma X, Wang Y, Houle M E, Zhou S, Erfani S M, Xia S T, Wijewickrema S N R, Bailey J. Dimensionality-driven learning with noisy labels. In: Proceedings of the 35th International Conference on Machine Learning. 2018, 3361−3370

[22] Xu Y, Cao P, Kong Y, Wang Y. $\mathcal{L}_{\mathrm{DMI}}$: a novel information-theoretic loss function for training deep nets robust to label noise. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019, 559

[23] Yao Y, Liu T, Han B, Gong M, Deng J, Niu G, Sugiyama M. Dual T: reducing estimation error for transition matrix in label-noise learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 609

[24] Xia X, Liu T, Han B, Gong C, Wang N, Ge Z, Chang Y. Robust early-learning: hindering the memorization of noisy labels. In: Proceedings of the 9th International Conference on Learning Representations. 2021

[25] Li J, Xiong C, Hoi S C H. MoPro: Webly supervised learning with momentum prototypes. In: Proceedings of the 9th International Conference on Learning Representations. 2021

[26] Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. In: Proceedings of the 5th International Conference on Learning Representations. 2017

[27] Natarajan N, Dhillon I S, Ravikumar P, Tewari A. Learning with noisy labels. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. 2013, 1196−1204

[28] Frenay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(5): 845−869

[29] Wang D B, Wen Y, Pan L, Zhang M L. Learning from noisy labels with complementary loss functions. In: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence. 2021, 10111−10119

[30] Song H, Kim M, Park D, Shin Y, Lee J G. Learning from noisy labels with deep neural networks: a survey. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(11): 8135−8153

[31] Liu T, Tao D. Classification with noisy labels by importance reweighting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(3): 447−461

[32] Hendrycks D, Mazeika M, Wilson D, Gimpel K. Using trusted data to train deep networks on labels corrupted by severe noise. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018, 10477−10486

[33] Wang Y, Liu W, Ma X, Bailey J, Zha H, Song L, Xia S T. Iterative learning with open-set noisy labels. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 8688−8696

[34] Yao Q, Yang H, Han B, Niu G, Kwok J T Y. Searching to exploit memorization effect in learning with noisy labels. In: Proceedings of the 37th International Conference on Machine Learning. 2020, 10789−10798

[35] Lee K, Yun S, Lee K, Lee H, Li B, Shin J. Robust inference via generative classifiers for handling noisy labels. In: Proceedings of the 36th International Conference on Machine Learning. 2019, 3763−3772

[36] Jiang L, Zhou Z, Leung T, Li L J, Li F F. MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels. In: Proceedings of the 35th International Conference on Machine Learning. 2018, 2309−2318

[37] Han B, Yao Q, Yu X, Niu G, Xu M, Hu W, Tsang I W, Sugiyama M. Co-teaching: robust training of deep neural networks with extremely noisy labels. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018, 8536−8546

[38] Arazo E, Ortego D, Albert P, O'Connor N, McGuinness K. Unsupervised label noise modeling and loss correction. In: Proceedings of the 36th International Conference on Machine Learning. 2019, 312−321

[39] Li J, Socher R, Hoi S C H. DivideMix: learning with noisy labels as semi-supervised learning. In: Proceedings of the 8th International Conference on Learning Representations. 2020

[40] Xia X, Liu T, Han B, Gong M, Yu J, Niu G, Sugiyama M. Sample selection with uncertainty of losses for learning with noisy labels. In: Proceedings of the Tenth International Conference on Learning Representations. 2022

[41] Wu P, Zheng S, Goswami M, Metaxas D, Chen C. A topological filter for learning with label noise. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 1795

[42] Wu Z F, Wei T, Jiang J, Mao C, Tang M, Li Y F. NGC: a unified framework for learning with open-world noisy data. In: Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. 2021, 62−71

[43] Tang Y, Pan Z, Hu X, Pedrycz W, Chen R. Knowledge-induced multiple kernel fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14838−14855

[44] Yang Y, Xu Z. Rethinking the value of labels for improving class-imbalanced learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 1618

[45] Zhang M L, Li Y K, Yang H, Liu X Y. Towards class-imbalance aware multi-label learning. IEEE Transactions on Cybernetics, 2022, 52(6): 4459−4471

[46] Shen L, Lin Z, Huang Q. Relay backpropagation for effective learning of deep convolutional neural networks. In: Proceedings of the 14th European Conference on Computer Vision. 2016, 467−482

[47] Zhou B, Cui Q, Wei X S, Chen Z M. BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 9716−9725

[48] Ye H J, Chen H Y, Zhan D C, Chao W L. Identifying and compensating for feature deviation in imbalanced deep learning. 2020, arXiv preprint arXiv: 2001.01385

[49] Tang K, Huang J, Zhang H. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 128

[50] Menon A K, Jayasumana S, Rawat A S, Jain H, Veit A, Kumar S. Long-tail learning via logit adjustment. In: Proceedings of the 9th International Conference on Learning Representations. 2021

[51] Cao K, Wei C, Gaidon A, Arechiga N, Ma T. Learning imbalanced datasets with label-distribution-aware margin loss. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019, 140

[52] Shu J, Xie Q, Yi L, Zhao Q, Zhou S, Xu Z, Meng D. Meta-weight-net: learning an explicit mapping for sample weighting. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019, 172

[53] Jamal M A, Brown M, Yang M H, Wang L, Gong B. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 7607−7616

[54] Ren J, Yu C, Sheng S, Ma X, Zhao H, Yi S, Li H. Balanced meta-softmax for long-tailed visual recognition. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 351

[55] Cao K, Chen Y, Lu J, Aréchiga N, Gaidon A, Ma T. Heteroskedastic and imbalanced deep learning with adaptive regularization. In: Proceedings of the 9th International Conference on Learning Representations. 2021

[56] Tanaka D, Ikami D, Yamasaki T, Aizawa K. Joint optimization framework for learning with noisy labels. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 5552−5560

[57] Wang Y, Ma X, Chen Z, Luo Y, Yi J, Bailey J. Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. 2019, 322−330

[58] Kim Y, Yim J, Yun J, Kim J. NLNL: negative learning for noisy labels. In: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. 2019, 101−110

[59] Yi K, Wu J. Probabilistic end-to-end noise correction for learning with noisy labels. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, 7010−7018

[60] Nguyen D T, Mummadi C K, Ngo T P N, Nguyen T H P, Beggel L, Brox T. SELF: learning to filter noisy labels with self-ensembling. In: Proceedings of the 8th International Conference on Learning Representations. 2020

[61] Pleiss G, Zhang T, Elenberg E R, Weinberger K Q. Identifying mislabeled data using the area under the margin ranking. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 1430

[62] Li H T, Wei T, Yang H, Hu K, Peng C, Sun L B, Cai X L, Zhang M L. Stochastic feature averaging for learning with long-tailed noisy labels. In: Proceedings of Proceedings of the 32nd International Joint Conference on Artificial Intelligence. 2023, 434

[63] Goldberger J, Roweis S, Hinton G, Salakhutdinov R. Neighbourhood components analysis. In: Proceedings of the 18th International Conference on Neural Information Processing Systems. 2004, 513−520

[64] Samuel D, Chechik G. Distributional robustness loss for long-tail learning. In: Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. 2021, 9475−9484

[65] Permuter H, Francos J, Jermyn I. A study of Gaussian mixture models of color and texture features for image classification and segmentation. Pattern Recognition, 2006, 39(4): 695−706

[66] Lukasik M, Bhojanapalli S, Menon A, Kumar S. Does label smoothing mitigate label noise? In: Proceedings of the 37th International Conference on Machine Learning. 2020, 6448−6458

[67] Zhong Z, Cui J, Liu S, Jia J. Improving calibration for long-tailed recognition. In: Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, 16484−16493

[68] Laine S, Aila T. Temporal ensembling for semi-supervised learning. In: Proceedings of the 5th International Conference on Learning Representations. 2017

[69] Liu S, Niles-Weed J, Razavian N, Fernandez-Granda C. Early-learning regularization prevents memorization of noisy labels. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 1707

[70] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 4080−4090

[71] Li J, Zhou P, Xiong C, Hoi S C H. Prototypical contrastive learning of unsupervised representations. In: Proceedings of the 9th International Conference on Learning Representations. 2021

[72] Chen P, Liao B, Chen G, Zhang S. Understanding and utilizing deep neural networks trained with noisy labels. In: Proceedings of the 36th International Conference on Machine Learning. 2019, 1062−1070

Tong WEI received his PhD degree in computer science from Nanjing University, China in 2021. He joined the Southeast University, China in 2022, and is currently an associate professor. His research interest is machine learning. Dr. Wei has served as the Area Chair of ICML'25 and IJCAI'25.

Jiang-Xin SHI received his BSc degree in 2020. He is currently working toward the PhD degree in the National Key Laboratory for Novel Software Technology at Nanjing University, China. His research interest is machine learning.

Min-Ling ZHANG received his BSc, MSc, and PhD degrees in computer science from Nanjing University, China in 2001, 2004, and 2007, respectively. Currently, he is a professor at the School of Computer Science and Engineering, Southeast University, China. His main research interests include machine learning and data mining. In recent years, Dr. Zhang has served as the General Co-Chairs of ACML'18, Program Co-Chairs of CCML'25, PAKDD'19, CCF-ICAI'19, ACML'17, CCFAI'17, PRICAI'16, Senior PC member or Area Chair of KDD 2021-2024, AAAI 2022-2025, IJCAI 2017-2024, ICML 2024, ICLR 2024, etc. He is also on the editorial board of IEEE Transactions on Pattern Analysis and Machine Intelligence, Science China Information Sciences, ACM Transactions on Intelligent Systems and Technology, Frontiers of Computer Science, Machine Intelligence Research, etc. Dr. Zhang is the Steering Committee Member of ACML and PAKDD, Vice-Chair of the CAAI (Chinese Association of Artificial Intelligence) Machine Learning Society. He is a Distinguished Member of CCF, CAAI, and Senior Member of AAAI, ACM, IEEE.

Yu-Feng LI received his BSc and PhD degrees in computer science from Nanjing University, China in 2006 and 2013, respectively. He joined the National Key Laboratory for Novel Software Technology at Nanjing University in 2013, and is currently a full professor at the School of Artificial Intelligence, Nanjing University. He is a member of the LAMDA group. His research interestes focus on robust and reliable machine learning and applications. He has published over 90 academic papers in top-tier journals and conferences in the field, with 6900 citations. The research work has been selected for the IJCAI 2021 Early-Career Spotlight Talk. He serves as an editorial board member for journals of Artificial Intelligence, Machine Learning, etc. He served as program co-chair for IEEE Big Comp 2020/CCML 2021, associated program co-chair for IJCAI 2025, and area chairs for ICML/NeurIPS/ICLR/IJCAI. He won the PAKDD Early-Career Research Award 2024.