Separating Noisy Samples From Tail Classes for Long-Tailed Image Classification With Label Noise

Chaowei Fang, Lechao Cheng, *Member, IEEE*, Yining Mao, Dingwen Zhang, *Member, IEEE*, Yixiang Fang, Guanbin Li, *Member, IEEE*, Huiyan Qi, and Licheng Jiao, *Fellow, IEEE*

***Abstract*— Most existing methods that cope with noisy labels usually assume that the classwise data distributions are well balanced. They are difficult to deal with the practical scenarios where training samples have imbalanced distributions, since they are not able to differentiate noisy samples from tail classes' clean samples. This article makes an early effort to tackle the image classification task in which the provided labels are noisy and have a long-tailed distribution. To deal with this problem, we propose a new learning paradigm which can screen out noisy samples by matching between inferences on weak and strong data augmentations. A leave-noise-out regularization (LNOR) is further introduced to eliminate the effect of the recognized noisy samples. Besides, we propose a prediction penalty based on the online classwise confidence levels to avoid the bias toward easy classes which are dominated by head classes. Extensive experiments on five datasets including CIFAR-10, CIFAR-100, MNIST, FashionMNIST, and Clothing1M demonstrate that the proposed method outperforms the existing algorithms for learning with long-tailed distribution and label noise.**

***Index Terms*— Image classification, long-tailed, noisy labels.**

Manuscript received 17 September 2022; revised 2 March 2023; accepted 20 June 2023. Date of publication 12 July 2023; date of current version 30 October 2024. This work was supported in part by the Key Scientific Technological Innovation Research Project by Ministry of Education; in part by the State Key Program and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61836009; in part by the National Natural Science Foundation of China under Grant U22B2054, Grant 62003256, Grant 62106235, Grant 62102341, Grant U21B2048, Grant 62293543, and Grant 61976250; in part by the 111 Project; in part by the Program for Cheung Kong Scholars and Innovative Research Team in University under Grant IRT 15R53; in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LQ21F020003; in part by the Basic and Applied Basic Research Fund in Guangdong Province under Grant 2022A1515010166; in part by the Exploratory Research Project of Zhejiang Laboratory under Grant 2022PG0AN01; in part by the Zhejiang Laboratory under Grant K2022KG1BB01; in part by the Shenzhen Science and Technology Program under Grant ZDSYS20211021111415025 and Grant JCYJ20220530141211024; and in part by the Open Project Program of the Key Laboratory of Artificial Intelligence for Perception and Understanding, Liaoning Province, under Grant AIPU. *(Corresponding author: Lechao Cheng.)*

Chaowei Fang is with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China, and also with the Zhejiang Laboratory, Hangzhou 310012, China.

Lechao Cheng and Yining Mao are with the Zhejiang Laboratory, Hangzhou 310012, China (e-mail: chenglc@zhejianglab.com).

Dingwen Zhang is with the Brain and Artificial Intelligence Laboratory, School of Automation, Northwestern Polytechnical University, Xi'an 710072, China.

Yixiang Fang is with the School of Data Science, Chinese University of Hong Kong, Shenzhen 518172, China.

Guanbin Li is with the School of Computer Science and Engineering, Research Institute of Sun Yat-sen University in Shenzhen, Sun Yat-sen University, Guangzhou 510275, China.

Huiyan Qi is with the School of Computer Science, Fudan University, Shanghai 200433, China.

Licheng Jiao is with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China.

Digital Object Identifier 10.1109/TNNLS.2023.3291695

I. Introduction

DEEP convolutional neural networks (CNNs) have achieved vast progress in addressing a variety of computer vision problems. Because CNNs have a large number of parameters, training them requires a massive demand for carefully labeled samples. Researchers can use inexpensive labeling systems such as coarse annotations [1], [2], [3], [4], semi-supervised pseudolabeling [5], [6], and nonexpert crowd-sourcing to reduce the cost of data annotation. However, these systems usually suffer from label noises. Relieving the drawback of noisy annotations is crucial for CNNs under such circumstance. The existing researches about learning with label noise (LLN) typically focus on splitting clean and noisy examples while neglecting the distribution imbalance. Considering real-world data usually exhibits a long-tailed distribution, learning with both long-tailed distribution and label noise as in Fig. 1 is urgent among real-world applications.

Deep CNNs have the capacity to memorize wrongly labeled samples during training but suffer from alarmingly increased generalization error [7]. Thus, LLN attracts a large amount of research interest. Arpit et al. [8] reveal that CNNs first learn the simple and general pattern of clean samples before fitting noisy examples. Based on this concept, a variety of methods are proposed to mitigate the influence of noisy samples via modeling the cross-entropy loss between predictions and given labels [9], [10], [11]. Samples with relatively smaller loss values are regarded as clean data and preserved for training models, while the remaining samples are regarded as noisy samples and excluded. However, under the long-tailed distribution, the cross-entropy losses between predictions and given labels are difficult to differentiate noisy samples from clean samples of tail classes, since the training process is dominated by head classes while ignoring tail classes. As shown in the top row of Fig. 2, we can observe this phenomenon on public datasets with synthetic label noise, especially CIFAR-10 [12], CIFAR-100, and FashionMNIST [13].

In this article, we make an early effort to tackle the learning task with long-tailed data distribution and label noise. Apart from the pure small cross-entropy loss criterion, we propose

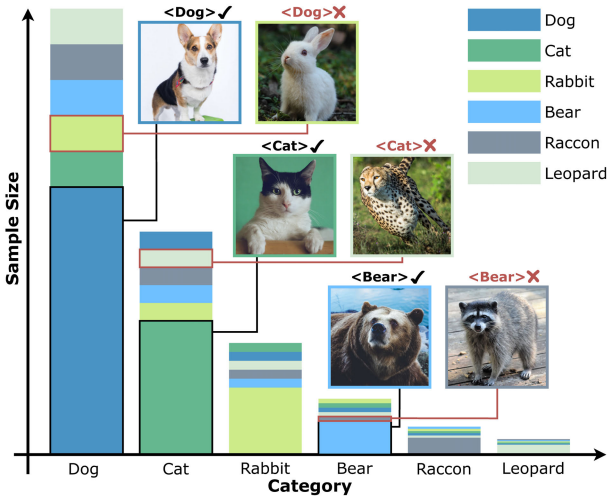


Fig. 1. This article aims at tackling the learning task with both distribution imbalance and label noise.

to model the inference sensitivity against data augmentations for increasing the discrimination between noisy samples and clean samples of tail classes. Although the CNN model is capable of memorizing all the training samples, noisy samples are more scattered than samples of tail classes, which means that conquering the former requires more effort. Thus, predictions on noisy samples are more sensitive to strong image augmentation compared with predictions on clean samples of tail classes. Hence, we devise a specific cross-augmentation matching criterion for recognizing noisy samples. Concretely, we use a shared network architecture for processing both weakly and strongly augmented images. Strongly augmented images have a feature distribution which is distinct to that of the original images. This brings shifts to the statistics of batch normalization layers. Therefore, we devise a dual-branch batch normalization (DBBN) module to process weakly and strongly augmented images with separate statistical parameters. As illustrated in the bottom row of Fig. 2, the matching between predictions on weakly and strongly augmented images significantly improves the discrimination between noise samples and clean samples of the tail class. There also exist methods [14], [15] which separate clean samples and noisy samples based on the prediction consistency between two independent networks. Since the learning paces of two networks cannot be synchronized, our method which uses a shared backbone is more effective in capturing the prediction sensitivity difference between clean and noisy samples. Furthermore, for counteracting the influence caused by noisy labels, we devise a leave-noise-out regularization (LNOR) through suppressing predictions on labels of the recognized noisy samples.

The other key point to address this problem is enhancing the learning of tail classes. Namely, we can bias the training process toward tail classes, thus balancing the loss values across classes. This can help increase the discrimination between noisy samples and clean samples of tail classes as well. To this end, we propose to penalize the prediction with classwise confidence levels which are dynamically estimated from prediction scores of the classification model. The prediction penalty achieves the effect of data rebalancing by biasing the training process toward relatively hard classes which are dominated by less-frequent classes. Compared with the existing data rebalancing strategies such as resampling [16] and re-weighting [17] which rely on class-wise sample sizes to rebalance the training data, our proposed prediction penalization strategy is robust against the existence of noisy samples in which actual classwise sample sizes are unavailable. Experiments on five public datasets including CIFAR-10 [12], CIFAR-100, MNIST [18], Fashion-MNIST [13], and Clothing1M [13] indicate that our method significantly outperforms the existing methods for learning with long-tailed distribution and label noise.

The main contributions of this article are as follows.

1) We propose a novel cross-augmentation matching criterion for detecting noisy samples under long-tailed distribution, through comparing predictions on weakly and strongly augmented images. Besides, an LNOR is introduced to counteract the influence of detected noisy samples.
2) Dual normalization branches are incorporated into the classification backbone to deal with the feature distribution shift between weakly and strongly augmented images.
3) Based on dynamically evaluated classwise confidence levels, a prediction penalization strategy is devised for rebalancing the training data.
4) Extensive experiments on five public datasets including CIFAR-10, CIFAR-100, MNIST, FashionMNIST, and Clothing1M are conducted, demonstrating the superiority of our method against the existing methods.

II. Related Work

A. Deep Learning With Long-Tailed Distribution

Training models with imbalanced training data is a critical task, considering samples usually obey the long-tailed distribution in real-world applications. Zhang et al. [20] provide a detailed recipe for guiding the choice of effective techniques in long-tailed classification, such as resampling, reweighting, and mixup [21]. Dablain et al. [22] even apply the synthetic minority oversampling technique for balancing training data. However, the resampling/reweighting algorithms are harmful to the learning of feature representations [23]. The other popular solution is formed by a two-stage pipeline, including representation learning based on uniform sampling and classifier learning based on rebalanced sampling [17], [24]. In [23], a cumulative learning algorithm is devised to combine the representation capacity of the model trained with uniform sampling and the classification capacity of the model trained with inverse proportional sampling. Zhong et al. [25] design hand-crafted functions for smoothing classes concerning their sample sizes. Wang et al. [26] attempt to transfer knowledge from head classes to tail classes, alleviating the issue of underfitting tail classes. Kang et al. [27] devise a balanced contrastive learning algorithm and a normalized classifier to take both representation and classifier learning into account. Few studies are targeted at learning models with

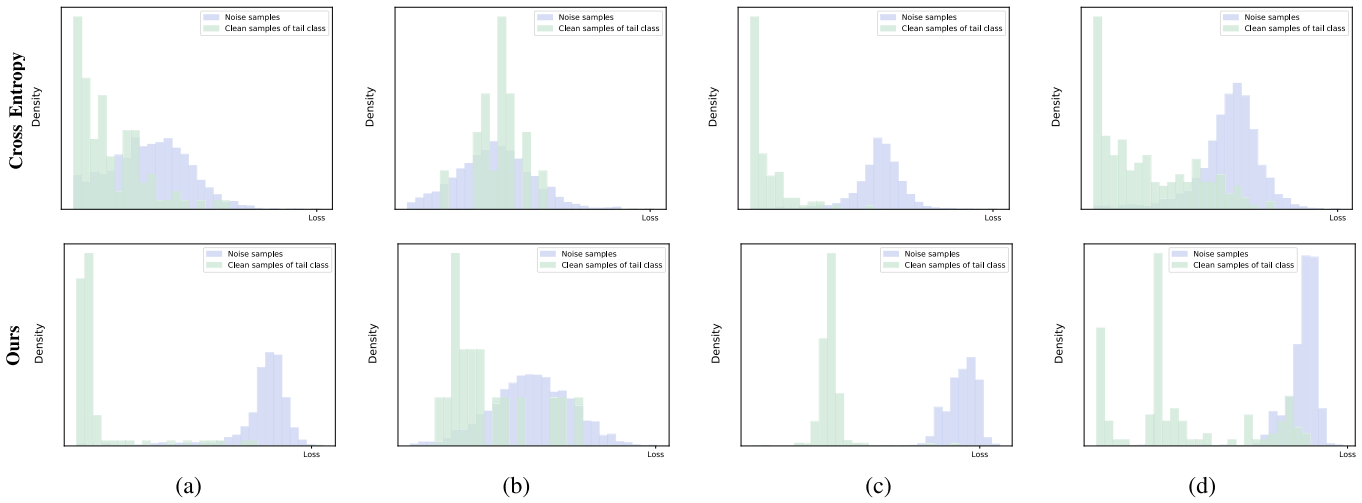


Fig. 2. Loss distributions of cross-entropy (top row) and our proposed cross-augmentation matching criterion (bottom row). The loss values are calculated with ResNet32 [19] on (a) CIFAR-10, (b) CIFAR-100, (c) MNIST, and (d) FashionMNIST. Labels of 20% training images are corrupted. The ratio between sample sizes of the minority class and the majority class is 0.01.

both long-tailed distribution and label noise. Zhong et al. [28] attempt to tackle the face recognition task under such a setting. However, this kind of study on general image classification is still underexplored. In this article, we make an early effort to tackle the general image classification task with both long-tailed distribution and label noise.

B. Deep LLN

The key to learning an image classification model with label noise is filtering out noisy samples before they are memorized by the model. One type of algorithms is devised based on the co-training mechanism, such as [9], [10], [29], and [15], which collaborates two classification models to clean noisy samples for each other. Han et al. [9] and Yu et al. [10] use two models to select small-loss samples as clean samples for each other. Li et al. [29] rely on the mixup operation to rectify the noisy labels with the help of predictions from peer networks. Instead of maintaining two diversified models, [14] attempts to encourage the agreement between two models and devise a co-regularization mechanism to select clean samples. Lu et al. [15] build up two models for weakly and strongly augmented images, respectively, and adopt the consistency loss between their predictions to preclude noisy samples. Wei et al. [30] attempt to address the label noise problem through label matrix completion. Yao et al. [31] and Yi and Huang [32] also detect noisy samples according to the difference between predictions on differently augmented images. Yao et al. [31] leverage the mean teacher model to rectify labels of noisy samples. Yi and Huang [32] use a single network to cope with weakly and strongly augmented images simultaneously. However, the feature distribution of the strongly augmented images is inconsistent with that of the original training images. This interferes with the estimation of statistical parameters in batch normalization layers. A comprehensive survey of the existing works for LLN is provided in [33]. In this article, we aim to settle the task of learning with long-tailed distribution and label noise. Inspired from [34], [35], [15], the matching between predictions on weakly and strongly augmented images is used to screen out noisy samples considering it can well separate noisy samples from clean samples of tail classes. Instead of using two networks [15], we construct only one network with the help of DBBN. Our method is more economical and capable of leveraging strongly augmented images to facilitate representation learning without affecting the feature distribution. The main differences between our method and the existing LLN methods are as follows. 1) Previous works can only conquer label noise under balanced data distribution, while our method takes the data imbalance into consideration. 2) We devise a novel method for preventing learning bias toward head classes via online prediction penalization (OPP). An LNOR is devised to counteract the negative effect caused by noisy samples. Online label smoothing [36] is targeted at tackling the overfitting problem in image classification without data imbalance. It helps prevent overconfidence on target categories and explore underlying similarity knowledge among different categories. In our method, the online classwise confidence levels are used for relieving the learning bias toward head classes via penalizing predictions on them.

III. Method

A. Preliminary

This article aims at tackling the long-tailed classification under the existence of label noise. Given a training dataset $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with N samples and K classes, we learn a deep CNN model to implement the classification task. Here, y_i ($\in \{1, 2, \ldots, K\}$) is the label of the training image x_i which is possibly incorrect. Let the number of samples in the kth class be N_k. In many real-world applications, samples are severely imbalanced with respect to classes.

B. Observation

In the task of learning long-tailed distribution and label noise, the training data are dominated by samples of head

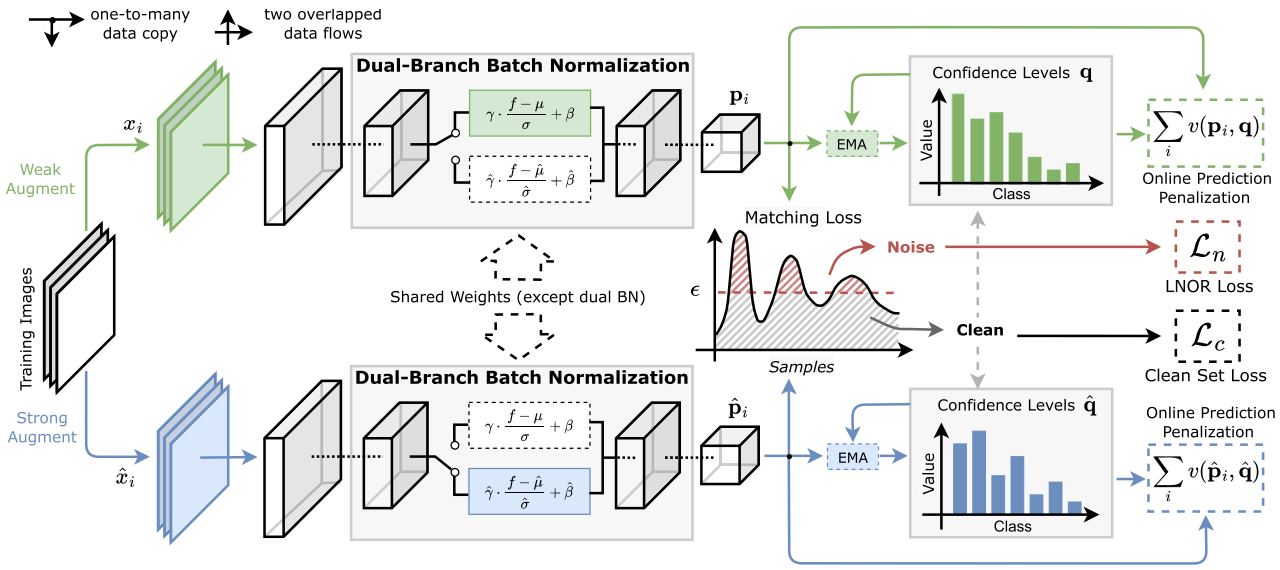


Fig. 3. Overview of our proposed method. The matching between predictions on weakly and strongly augmented images is used to identify clean and noisy samples. LNOR is used to eliminate the impact of recognized noisy samples, and an OPP scheme is devised to rebalance training data. "EMA" indicates the exponential moving average.

classes. The model merely captures patterns of head classes in the early stage. As shown in Fig. 2, only using the cross-entropy loss between predictions and given labels, which is widely applied to detect noisy samples e.g., [9] and [29], is hard to discriminate samples with wrong labels from samples of the tail class. There exists a large overlap between losses on the two kinds of samples.

C. Overview

We propose a novel algorithm to solve the image classification problem with training samples having long-tailed distribution and label noise. The CNN-based classification model is built up with the help of DBBN, which processes feature maps of weakly and strongly augmented images with separate parameters. Based on it, we use the cross-augmentation matching to detect noisy labels, which effectively tackles the issue in the above observation. When learning on the clean set of samples, an OPP strategy is devised for balancing training data. We propose to eliminate the effect caused by noisy samples through restraining predictions on given labels of these recognized noisy samples. The overall framework of our method is illustrated in Fig. 3.

D. Cross-Augmentation Matching

We use the matching between predictions of weakly and strongly augmented images for splitting wrongly labeled samples and clean samples. Suppose the prediction of the classification model on an weakly augmented image x_i is $\mathbf{p}_i = h(x_i|\theta)$. θ indicates network parameters, and $h(\cdot)$ denotes the mapping function. We define the strongly augmented counterpart of x_i as $\hat{x}_i$. The network prediction for $\hat{x}_i$ is $\hat{\mathbf{p}}_i = h(\hat{x}_i|\theta)$. The following loss function is used to judge whether x_i is correctly labeled:

$$\ell(x_i, y_i) = \ell^c(\mathbf{p}_i, y_i) + \alpha\ell^c\left(\hat{\mathbf{p}}_i, y_i'\right) \tag{1}$$

$$\ell^c(\mathbf{p}_i, y_i) = -\log\left(p_i\left[y_i\right]\right) \tag{2}$$

where α is a weighting coefficient, and $p_i[j]$ denotes the $j$the element of $\mathbf{p}_i$. $y_i' = \arg\max_k p_i[k]$, which denotes the most confident class predicted on x_i. As demonstrated in Fig. 2, the adoption of the loss function in (1) significantly improves the distinguishability between noisy samples and clean samples, in contrast to the pure cross-entropy between predictions and given labels, namely, $\ell^c(\mathbf{p}_i, y_i)$.

Samples with losses smaller than the threshold ϵ estimated via Otsu's method [37] are regarded as clean data, $\mathbb{C} = \{(x_i, y_i)|\ell(x_i, y_i) < \epsilon\}$. During training, (1) is used to calculate losses for samples in $\mathbb{C}$. The overall objective function on $\mathbb{C}$ is formulated as

$$L_c = \frac{1}{|\mathbb{C}|}\sum_{(x_i,y_i)\in\mathbb{C}} \ell(x_i, y_i). \tag{3}$$

1) Dual-Branch Batch Normalization: In the CNN-based classification model, we use DBBN to process intermediate features instead of conventional batch normalization, which is inspired from [38]. Two separate groups of parameters are adopted to cope with features of weakly and strongly augmented images, respectively. The reason is that strongly augmented images have feature distributions distinct to those of the original images caused by appearance differences between them. The DBBN is more advantageous at preventing the effect of strongly augmented images on the estimation of feature statistics of the original images. Given a weakly augmented image x_i and its strongly augmented counterpart $\hat{x}_i$, we denote the feature generated by an intermediate convolution layer as $\mathbf{f}_i$ and $\hat{\mathbf{f}}_i$, respectively. Distinct means and variances are calculated for the two types of features

$$\mu = \frac{1}{m}\sum_{i=1}^{m}\mathbf{f}_i, \quad \sigma = \sqrt{\frac{1}{m}\sum\nolimits_{i=1}^{m}(\mathbf{f}_i - \mu)^2 + \xi} \tag{4}$$

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}\hat{\mathbf{f}}_i, \quad \hat{\sigma} = \sqrt{\frac{1}{m}\sum\nolimits_{i=1}^{m}\left(\hat{\mathbf{f}}_i - \hat{\mu}\right)^2 + \xi} \tag{5}$$

where $\xi(=10^{-5})$ is a constant, and m represents the number of samples in a training batch. These statistical values are separately accumulated via exponential moving average. The outputs of the two branches in the batch normalization are denoted as $\mathbf{f}'_i = \gamma(\mathbf{f}_i - \mu)/(\sigma) + \beta$ and $\hat{\mathbf{f}}'_i = \hat{\gamma}(\hat{\mathbf{f}}_i - \hat{\mu})/(\hat{\sigma}) + \hat{\beta}$, where γ, β, $\hat{\gamma}$, and $\hat{\beta}$ are learnable affine parameters. During training, each kind of augmented images is forwarded with one group of parameters. During testing, the input images are merely processed with simple transformations, having a more similar distribution to weakly augmented images. Thus, only the parameters estimated from weakly augmented images are preserved while those estimated from strongly augmented images are deprecated.

2) Leave-Noise-Out Regularization: Except for samples in the clean set $\mathbb{C}$, we regard samples with relatively larger loss values as confident noisy samples, which are denoted as

$$\mathbb{N} = \arg\max_{\mathbb{S}} \sum_{(x_i, y_i)\in\mathbb{S}} \ell(x_i, y_i) \quad \text{s.t. } \mathbb{S}\cap\mathbb{C} = \emptyset, \quad |\mathbb{S}| \le \kappa(N - |\mathbb{C}|). \tag{6}$$

$\kappa \in (0, 1)$ is a constant which determines the percent of samples in $\mathbb{D} - \mathbb{C}$ regarded as confident noisy samples. Considering the given labels of these confident noisy samples are incorrect, they are used to regularize the network optimization via a specific LNOR. We can assume that each sample x_j in $\mathbb{N}$ belongs to certain class from $\{k = 1, \ldots, K | k \neq y_j\}$. The following regularization term is used to constrain network predictions on $\mathbb{N}$:

$$L_n = -\frac{1}{|\mathbb{N}|} \sum_{(x_j, y_j)\in\mathbb{N}} \log\big(1 - p_j[y_j]\big). \tag{7}$$

The above regularization prevents predictions of these recognized noisy samples from being positive on their labels. It benefits to mitigate the negative impact of noisy samples.

E. Online Prediction Penalization

For preventing the dominance of head classes, the existing methods focus on resampling [16], [39], [40] or reweighting [17] training data. However, these methods are dependent on classwise sample sizes which are unavailable when label noise exists. In addition, fixed classwise sample sizes cannot reflect the practical learning difficulty levels of classes. To settle the above problems, we propose to smooth the predictions on samples of head classes by penalizing their predictions with online confidence levels. Practically, we evaluate classwise confidence levels from online predictions of the network. The confidence level of the kth class in the current training iteration is estimated as follows:

$$\mathbf{q} := (1 - \tau)\mathbf{q} + \frac{\tau}{|\mathbb{C}|} \sum_{(x_i, y_i)\in\mathbb{C}} \mathbf{p}_i \tag{8}$$

where τ is a constant. $\mathbf{q}$ is initialized by the ratios of samples in K classes, namely, $q[k] = (N_k/N)$. The OPP is formulated as follows:

$$v(\mathbf{p}_i, \mathbf{q}) = -\sum_{k=1}^{K} (1 - q[k]) \log(p_i[k]). \tag{9}$$

We can acquire the following loss function after incorporating the above prediction penalization term into the cross-entropy function:

$$\ell_p(x_i, y_i, \mathbf{q}) = \ell^c(\mathbf{p}_i, y_i) + \lambda v(\mathbf{p}_i, \mathbf{q}) \tag{10}$$

where λ is a constant weighting coefficient. It can be transformed into the following formulation:

$$\ell_p(x_i, y_i, \mathbf{q}) = -(1 + \lambda K - \lambda) \sum_{k=1}^{K} s_i^k \log(p_i[k]) \tag{11}$$

$$s_i^k = \begin{cases} \dfrac{1 + \lambda(1 - q[k])}{1 + \lambda K - \lambda}, & k = y_i \\ \dfrac{\lambda(1 - q[k])}{1 + \lambda K - \lambda}, & k \neq y_i. \end{cases} \tag{12}$$

We can observe that the penalization item in (9) is able to smooth labels according to classwise confidence levels. Labels with higher confidence levels are smoothed more intensively. Meanwhile, the reduction of the target label is distributed to other labels in negative relation to their confidence levels. Considering head classes are usually "easier" than tail classes, thus leading to relatively higher confidence levels. The penalization term can help bias the optimization process toward tail classes. It is applied for predictions on both weakly and strongly augmented images, contributing to the following loss regularization term:

$$L_p = \frac{1}{|\mathbb{C}|} \sum_{(x_i, y_i)\in\mathbb{C}} \big[v(\mathbf{p}_i, \mathbf{q}) + v\big(\hat{\mathbf{p}}_i, \mathbf{q}\big)\big]. \tag{13}$$

F. Training Process

The final objective function is formed by combining (3), (7), and (13)

$$L = L_c + L_n + \lambda L_p. \tag{14}$$

The training process of our method is composed of two stages. First, we warm up the network with the help of the loss function in (10), i.e., optimizing network parameters with the following loss: $L = (1/N)\sum_{(x_i, y_i)\in\mathbb{D}} \ell_p(x_i, y_i, \mathbf{q})$. Then, we adopt the criterion in (1) to select out clean samples and confident noisy samples and optimize the network according to (14). Stochastic gradient descent (SGD) is applied for updating network parameters. The detailed training process is described in Algorithm 1.

IV. Experiments

A. Datasets

Extensive experiments are conducted on four datasets, including CIFAR-10 [12], CIFAR-100, MNIST [18], and FashionMNIST [13] with synthetic noisy labels and long-tailed distribution, and one real-word dataset Clothing1M [13], to validate the effectiveness of our method.

1) *CIFAR-10* is composed of 50 000 and 10 000 32×32 images for training and testing, respectively. The number of classes is 10.
2) *CIFAR-100* is composed of 50 000 and 10 000 32×32 images for training and testing, respectively. The number of classes is 100.

Algorithm 1 Algorithm for Training the Classification Model via Our Proposed Method. E_1 and E_2 Denote the Number of Training Epochs in the First and Second Stages, Respectively. B Means the Number of Batches. n Denotes the Batch Size

Input: Training dataset $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^{N}$.
Output: Optimized network parameters θ.

Initialize classwise confidence level vectors $\mathbf{q}$ and $\hat{\mathbf{q}}$ as: $q[k] = \hat{q}[k] = \frac{N_k}{N}$;
for $e \leftarrow 1$ to E_1 **do**
  for $b \leftarrow 1$ to B **do**
    Fetch n samples as a training batch;
    Process all samples with weak augmentation, resulting in $\{(x_i, y_i)\}_{i=1}^{n}$;
    Forward each sample x_i through the network: $\mathbf{p}_i = h(x_i|\theta)$;
    Calculate the training loss with (10): $L = \frac{1}{n}\sum_{i=1}^{n} \ell_p(x_i, y_i, \mathbf{q})$;
    Use SGD to update network parameters θ;
    Update the classwise confidence level vector: $\mathbf{q} \leftarrow (1-\eta)\mathbf{q} + \frac{\eta}{n}\sum_{i=1}^{n} \mathbf{p}_i$;
  end for
end for
for $e \leftarrow 1$ to E_2 **do**
  for $b \leftarrow 1$ to B **do**
    Fetch n training samples as a training batch;
    Process all samples with weak augmentation, resulting in $\{(x_i, y_i)\}_{i=1}^{n}$;
    Process all samples with strong augmentation, resulting in $\{\hat{x}_i\}_{i=1}^{n}$;
    Forward each sample x_i and $\hat{x}_i$ through the network: $\mathbf{p}_i = h(x_i|\theta)$, and $\hat{\mathbf{p}}_i = h(\hat{x}_i|\theta)$;
    Infer the most confident class for each sample: $y_i' = \arg\max_k p_i[k]$;
    Calculate loss value for each sample via (1): $\ell(x_i, y_i) = \ell^c(\mathbf{p}_i, y_i) + \ell^c(\hat{\mathbf{p}}_i, y_i')$;
    Select out the set of clean samples: $\mathbb{C} = \{(x_i, y_i)|\ell(x_i, y_i) < \epsilon\}$ where ϵ is a threshold calculated with Otsu's method;
    Select out the set of confident noisy samples $\mathbb{N}$, according to (6);
    Calculate overall training loss via (14);
    Use SGD to update network parameters θ;
    Update classwise confidence level vectors: $\mathbf{q} \leftarrow (1-\eta)\mathbf{q} + \frac{\eta}{|\mathbb{C}|}\sum_{(x_i,y_i)\in\mathbb{C}} \mathbf{p}_i$, $\hat{\mathbf{q}} \leftarrow (1-\eta)\hat{\mathbf{q}} + \frac{\eta}{|\mathbb{C}|}\sum_{(x_i,y_i)\in\mathbb{C}} \hat{\mathbf{p}}_i$;
  end for
end for
return Network parameters θ.

3) *MNIST* is a handwritten digit recognition dataset consisting of 60 000 and 10 000 28 × 28 grayscale images for training and testing, respectively.
4) *FashionMNIST* contains 60 000 training and 10 000 testing grayscale images with the size of 28 × 28, which belong to ten types of clothing.
5) *Clothing1M* contains 10 00 000 images with real-word label noise, which are regarded as the training set. In all, 10 000 images with clean labels are used for testing.

To synthesize distribution imbalance on CIFAR-10, CIFAR-100, MNIST, and FashionMNIST, we randomly resample training images of all the classes to make them obey a long-tailed distribution. The imbalance ratio is denoted as $\rho = (\min_k N_k)/(\max_k N_k)$. Following [9], we adopt two types of label corruption strategies to synthesize label noise: 1) Class-independent (CI) label noise is generated via randomly changing the label of every sample with a probability of η (denoted as the ratio of noisy labels). If one label is determined to be modified, we replace it with a new label randomly drawn from other labels. 2) Class-dependent (CD) label noise is synthesized through modifying the label of samples in the kth class to $(k+1)\%K$ with the probability of η. The classification accuracy is adopted for evaluation.

B. Implementation Details

Our method is implemented under PyTorch. During training, the weak augmentation is formed by random cropping and flipping. AutoAugment [41] is used to generate strongly augmented images. We adopt ResNet32 [19] as the backbone of the classification network. The learning rate of the SGD optimizer is initialized as 0.05 and decayed in the cosine annealing schedule [42]. The training epochs in two stages are both 100, and the mini-batch size is 128. Without specification, all the models are trained from scratch. Besides, we empirically set $\alpha = 2$, $\kappa = 0.2$, $\tau = 0.5$, and $\lambda = 0.1$. The 8 V100 GPUs are used to train models on Clothing1M. For other datasets, one GeForce RTX 2080Ti is used. The training process costs around two days on Clothing1M and around 2 h on other datasets.

C. Comparison With Methods for LLN

1) Existing Methods: We compare our method against the following methods for LLN.

1) *Joint Optimization* [49] proposes to alternatively optimize network parameters and update labels according to current network predictions.
2) *Co-Teaching* [9] makes usage of the divergence in the learning errors of two distinct models to remove the influence of label noise. The two models screen out noisy samples for each other.
3) *Co-Teaching+* [10] is an extended variant of *Co-Teaching*. It uses the "update by disagreement" strategy to prevent the co-teaching method degenerating into a self-training method, since the two models converge to a consensus in late training iterations.
4) *MLNT* [50] devises a beforehand meta-learning algorithm to make the network parameters tolerant to label noise. In meta-train, noisy labels are used for updating network parameters. In meta-test, the mean teacher model is used for correcting the optimization direction.

5) *MetaCleaner* [51] proposes to construct clean samples via aggregating noisy-labeled samples with estimated importance scores.
6) *PENCIL* [52] leverages the probabilistic modeling to jointly optimize network parameters and infer the latent correct labels.
7) *JoCoR* [14] adopts two networks to identify clean samples via a joint loss which is composed of individual cross-entropy losses and a co-regularized Jensen–Shannon divergence loss.
8) *DivideMix* [29] separates clean and noisy samples via a two-component Gaussian mixture model built upon the training loss. Then, the MixMatch algorithm [53] is used to explore noisy samples.
9) *ELR+* [54] reveals that semi-supervised learning regularization can neutralize the influence of noisy samples.
10) *Co-Matching* [15] matches predictions inferred from weakly and strongly augmented images to discern noisy samples. Separate models are used to infer results for weakly and strongly augmented images.
11) *MOIT* [48] proposes an interpolated contrastive loss for alleviating the side effect of label noise in representation learning. A k-nearest neighbor search algorithm is devised for detecting noisy samples, which are subsequently explored via semi-supervised learning.
12) *AUGDESC* [11] finds that choosing different augmentation policies to model the loss metric for detecting noisy samples and learning the classifier is more effective in coping with noisy labels than using the same augmentation policy. It also reveals that applying strong augmentation in the warm-up training stage encourages the network to memorize noisy data.
13) *SOA* [43] uses covariance terms to deal with instance-dependent label noises.
14) *Jo-SRC* [31] leverages the Jensen–Shannon divergence between predictions and labels to recognize noisy samples. Besides, the prediction consistency between two views of each sample is adopted for discriminating in-distribution noisy samples from out-of-distribution noisy samples, and a mean teacher model is built up for rectifying the labels of in-distribution noisy samples.
15) *JNPL* [44] introduces a joint negative and positive learning framework. For preventing the underfitting of training data, the method decreases weights of negative learning losses with respect to predictions on negative labels. Furthermore, it adopts a strict sampling strategy to select positive samples, and reweights their losses according to the predicted probabilities of the most confident classes.
16) *AutoDo* [46] devises a generalized automated dataset optimization algorithm to eliminate the distribution gap between testing data and training data with noisy labels.
17) *LFDLN* [47] rectifies labels of samples progressively, starting from samples with high prediction scores.
18) *REL* [45] separates network parameters into two groups, including critical and noncritical parameters. The latter type of parameters are decayed in the early learning strategy for eliminating the influence of noisy samples.

TABLE I
COMPARISON WITH EXISTING LLN METHODS ON CIFAR-10 WITH CI LABEL NOISE

Imbalance Factor ρ	1	0.1	0.05	0.02	0.01
CE	86.35	75.88	70.32	64.69	54.07
DivideMix [29]	84.03	69.80	60.26	54.74	36.56
Co-Teaching [9]	89.75	58.63	52.58	42.67	39.39
Co-Teaching+ [10]	88.36	42.58	35.26	33.90	34.88
Co-Matching [15]	91.21	76.14	65.33	61.83	54.05
Co-Matching+OPP	90.57	81.86	77.58	70.60	57.71
SOA [43]	90.36	79.02	67.56	58.44	51.70
JNPL [44]	91.26	75.38	66.87	51.01	44.23
REL [45]	84.71	69.00	65.71	46.73	36.31
DDLS [36]	89.88	74.59	63.45	43.33	35.14
AutoDo [46]	89.50	80.35	75.67	69.09	62.29
LFDLN [47]	89.03	80.21	62.85	54.02	43.23
JoCoR [14]	90.21	81.63	74.15	63.87	53.69
JoCoR+OPP	90.71	82.52	77.60	65.37	58.14
MOIT [48]	86.27	81.59	75.16	66.40	59.18
Jo-SRC [31]	91.01	80.43	75.21	66.11	55.97
AUGDESC [11]	91.92	86.06	55.47	35.07	30.37
Ours	**92.32**	**86.12**	**84.33**	**81.42**	**76.94**

TABLE II
COMPARISON WITH EXISTING LLN METHODS ON CIFAR-10 WITH CD LABEL NOISE

Imbalance Factor ρ	1	0.1	0.05	0.02	0.01
CE	87.55	78.82	73.79	66.10	60.41
DivideMix [29]	85.32	68.97	58.45	52.70	37.24
Co-Teaching [9]	89.14	46.34	41.43	36.65	34.99
Co-Teaching+ [10]	80.92	45.95	34.74	32.45	27.74
JoCoR [14]	89.68	57.43	49.01	41.98	41.41
JoCoR+OPP	89.91	69.80	60.39	45.10	41.09
Co-Matching [15]	82.47	58.86	50.40	43.31	35.95
Co-Matching+OPP	89.48	70.86	60.69	47.05	40.82
AUGDESC [11]	91.54	60.52	39.61	34.11	29.93
SOA [43]	87.87	78.94	64.32	60.58	49.24
JNPL [44]	91.32	72.71	59.97	46.22	45.03
LFDLN [47]	83.38	76.36	56.22	49.31	48.68
REL [45]	82.59	70.47	65.54	48.99	37.47
DDLS [36]	89.80	78.41	59.02	43.82	40.77
Jo-SRC [31]	91.01	80.43	75.21	66.11	55.97
AutoDo [46]	90.71	81.33	78.10	71.85	69.09
MOIT [48]	86.93	81.19	77.26	73.76	63.87
Ours	**92.05**	**86.13**	**82.86**	**79.44**	**73.32**

19) *DDLS* [36] presents an online label smoothing strategy, which assigns a soft probability distribution to classes based on temporally accumulated network predictions.

2) Experiments on CIFAR-10: The experimental results on CIFAR-10 with CI noisy labels and CD noisy labels are presented in Tables I and II, respectively. The noise rate η is set to 20%, and the imbalance ratio ρ varies in $\{1, 0.1, 0.05, 0.02, 0.01\}$. *CE* indicates the baseline model trained with the standard cross-entropy loss function. It can be evidently observed that the existing LLN methods are severely influenced by the data imbalance. For instance, on CIFAR-10 with CI label noise, *JoCoR*, *Co-Matching*, and *AUGDESC* achieve promising performance (90.21%, 91.21%, and 91.92%, respectively) when data are balanced ($\rho = 1$). However, when the training data are severely imbalanced ($\rho = 0.01$), their accuracy values are drastically decreased by 36.52%, 37.16%, and 61.55%, respectively. When the imbalance ratio ρ is smaller than 0.02, a few existing methods

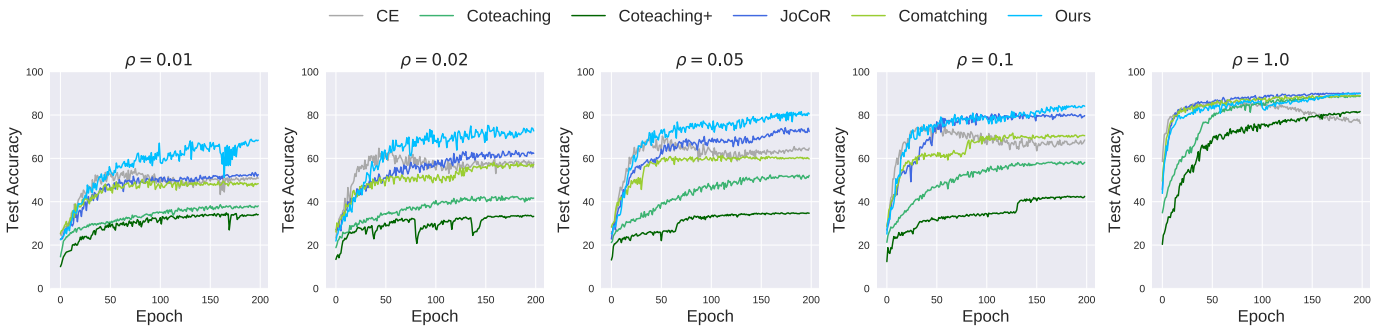


Fig. 4. Accuracy curves of methods for learning with CI label noise on CIFAR-10. From left to right, the imbalance ratio ρ is set to 0.01, 0.02, 0.05, 0.1, and 1.0, respectively.

even perform worse than *CE*. Our method is effective in coping with label noise under the imbalanced settings and derives substantially better performance than other methods. For example, on CIFAR-10 with CI label noise, the accuracy of our method (84.33%) is 8.66% higher than that of the second-best method *AutoDo* (75.67%) when $\rho = 0.05$; our method generates classification result with 14.65% higher accuracy than *AutoDo* when $\rho = 0.01$. Fig. 4 illustrates the variation curves of the test accuracy along the training epochs, indicating that our method outperforms other methods steadily in late training epochs. We also attempt to incorporate the data rebalancing strategy, namely, OPP, into the existing LLN methods *JoCoR* and *Co-Matching*, forming *JoCor + OPP* and *Co-Matching + OPP*. The adoption of OPP helps improve *JoCoR* and *Co-Matching*. However, these variants achieve inferior performance compared with the final version of our method. This indicates that our proposed cross-augmentation matching strategy is more robust in coping with label noise under long-tailed distribution.

3) Experiments on Clothing1M: The experimental results on Clothing1M are presented in Table III. For fair comparison, ResNet50 which is initialized with ImageNet pretrained parameters is used to implement our method. The results of *M-correction* [55], *Joint Optimization* [49], *MetaCleaner* [51], *MLNT* [50], *PENCIL* [52], *DivideMix* [29], *ELR+* [54], and *AUGDESC* [11] reported in their original papers are used. Our approach achieves the best performance, surpassing the second-best method *AUGDESC* by 0.38%.

Extensive experiments on four datasets (CIFAR-10, CIFAR-100, MNIST, and FashionMNIST) with various imbalance factors and noise ratios are presented in Fig. 5.

D. Comparison With Methods for LLTD

In this section, we compare our method against the existing methods capable of addressing the problem of LLTD, including *RW* (short for data reweighting) [57], *LS* [58], *Mixup* [21], *RS* (short for data resampling) [56], *LDAM-DRW* [24], and *MiSLAS* [25]. *RW* can be easily implemented via assigning relatively higher weights to samples of tail classes. *RS* usually balances training data through oversampling tail classes. Other methods are briefly introduced as below.

1) *LS* [58] (label smoothing) is a regularization technique that is effective in alleviating overfitting and overconfidence issues.

TABLE III
COMPARISON WITH LLN METHODS ON CLOTHING1M

Method	Accuracy
CE	69.21
M-correction [55]	71.00
Joint Optimization [49]	72.16
MetaCleaner [51]	72.50
MLNT [50]	73.47
PENCIL [52]	73.49
DivideMix [29]	74.76
ELR+ [54]	74.81
AUGDESC [11]	75.11
Ours	**75.49**

2) *Mixup* [21] synthesizes new samples by combining pairs of training images. Given two images and their ground-truth labels, (x_i, y_i) and (x_j, y_j), a synthetic training example $(\hat{x}, \hat{y})$ is generated as follows:

$$\hat{x} = \hat{\lambda} x_i + \left(1 - \hat{\lambda}\right) x_j \tag{15}$$

$$\hat{y} = \hat{\lambda}\,\text{onehot}(y_i) + \left(1 - \hat{\lambda}\right)\text{onehot}\left(y_j\right) \tag{16}$$

where $\hat{\lambda}$ is randomly drawn from the beta distribution, and onehot$(\cdot)$ returns the one-hot vector of the input label.

3) *LDAM-DRW* [24] proposes a label-distribution-aware margin (LDAM) loss for improving the generalization to minority classes. This work reveals that deferring the rebalancing optimization stage helps relieve the complications of rebalancing algorithms.

4) *MiSLAS* [25] uses label-aware smoothing strategies to calibrate distributions of predicted probability values. Its training phase is also composed of two stages, including a first instance-balanced stage and a second class-balanced stage. When updating the batch normalization layers in the second stage, it keeps the transformation parameters fixed and merely adjusts the mean and variance.

The results are presented in Table IV, and the variation curves of test accuracy are illustrated in Fig. 6. Our method has better generalization ability than the state-of-the-art LLTD approaches. We also attempt to replace the OPP with *RS* [56], *RW* [57], or *LS* [58] for re-balancing the clean sample set $\mathbb{C}$, producing results indicated by *Our-RS*, *Our-RW*, or *Our-LS*,

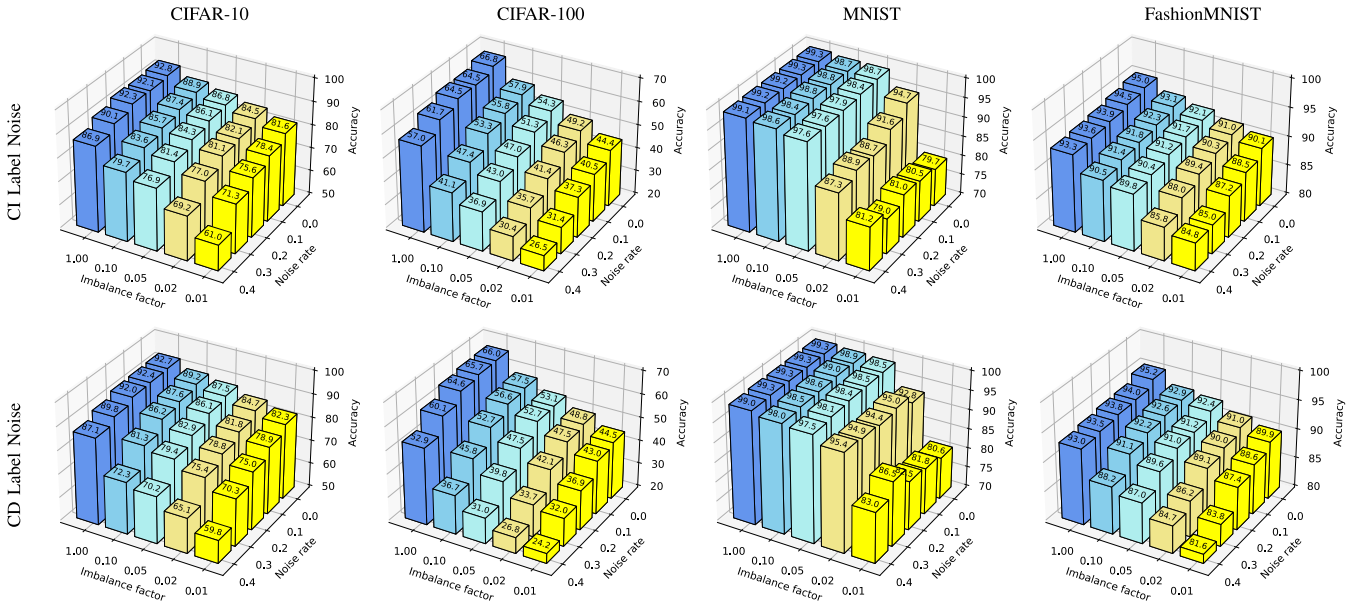


Fig. 5. Performance of our method on extensive settings for imbalance ratio and noise rate. From left to right, the experiments are conducted on CIFAR-10, CIFAR-100, MNIST, and FashionMNIST, respectively. The top row shows results of learning with CI label noise; the bottom row shows results of learning with CD label noise.

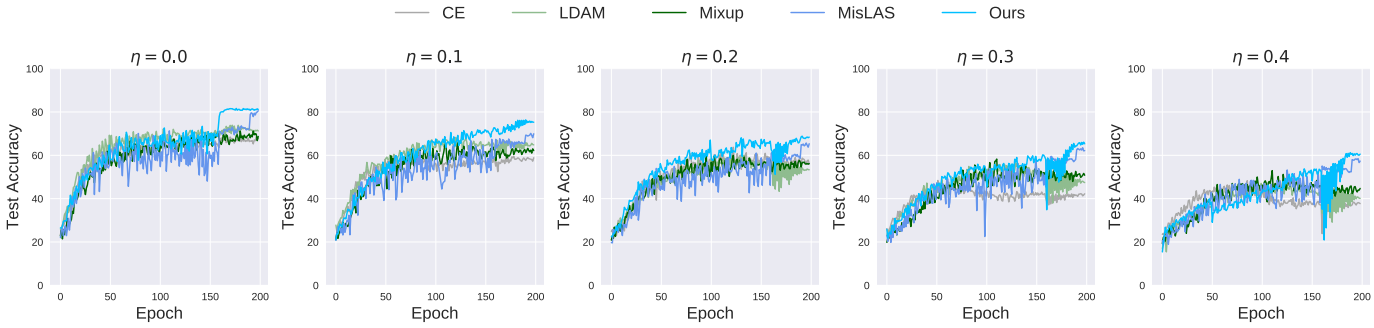


Fig. 6. Accuracy curves of methods for learning with long-tailed distribution on CIFAR-10. From left to right, the noise rate η is set to 0.0, 0.1, 0.2, 0.3, and 0.4, respectively.

TABLE IV

COMPARISON WITH EXISTING LLTD METHODS ON CIFAR-10 ($\rho = 0.01$). "RS," "RW," AND "LS" STAND FOR THE RESAMPLING STRATEGY IN [56], THE REWEIGHTING STRATEGY IN [57], AND LABEL SMOOTHING STRATEGY IN [58], RESPECTIVELY

Noise Rate η	0.0	0.1	0.2	0.3	0.4
CE	69.85	59.85	54.94	48.38	48.27
Mixup	73.06	60.90	48.11	45.94	43.92
LDAM-DRW	78.90	68.24	60.69	56.50	48.15
MiSLAS	81.91	71.32	66.23	63.11	51.34
Ours-RS	82.97	77.28	73.04	62.55	47.82
Ours-RW	77.48	76.01	72.41	66.02	51.25
Ours-LS	68.97	69.17	66.25	62.28	54.86
Ours	**86.87**	**79.66**	**76.94**	**69.23**	**61.03**

respectively. These data rebalancing strategies are inferior to OPP.

E. Comparison With Naïve Combinations of Existing LLN and LLTD Methods

We attempt to simply combine methods for LLN including *Co-Teaching* [9], *Co-teaching+* [10], *JoCoR* [14], *Co-Matching* [15], and methods for learning with long-tailed distribution including *RS* [56], *RW* [57], *LS* [58], and *LDAM* [24]. We also implement variants of our method via combining the cross-augmentation matching strategy with *RW*, *RS*, *LS*, or *LDAM*. The experimental results on CIFAR-10 with $\rho = 0.01$ and $\eta = 0.2$ are reported in Table V, and variants of our method outperform other methods consistently. For example, when incorporated with *LDAM*, our method achieves 34.44%, 29.43%, 18.54%, and 17.47% higher accuracy than *Co-Teaching*, *Co-Teaching+*, *JoCoR*, and *Co-Matching*, respectively.

TABLE V

COMPARISON OF OUR METHOD AGAINST NAÏVE COMBINATIONS OF METHODS FOR LLN AND LLTD SEPARATELY, ON CIFAR-10 WITH CI LABEL NOISES ($\rho = 100$ AND $\eta = 0.2$)

LLN \ LLTD	RW	RS	LS	LDAM
Co-Teaching	39.31	47.79	37.01	42.50
Co-Teaching+	27.33	43.06	34.61	47.51
JoCoR	57.20	60.04	23.46	58.40
Co-Matching	60.54	64.44	12.34	59.47
Ours	**72.41**	**73.04**	**66.25**	**76.94**

TABLE VI

COMPARISON OF VARIANTS OF OUR METHOD USING DIFFERENT OBJECTIVE FUNCTIONS. FOUR DATASETS ARE USED IN EXPERIMENTS. "CE;; INDICATES THE BASELINE METHOD TRAINED WITH THE CROSS-ENTROPY LOSS CALCULATED BETWEEN PREDICTIONS AND GIVEN LABELS. "CAUGMATCHING" MEANS THE MATCHING BETWEEN PREDICTIONS ON WEAKLY AND STRONGLY AUGMENTED IMAGES IS USED. "LNOR" DENOTES THE LEAVE-NOISE-OUT REGULARIZATION. "OPP" STANDS FOR THE ONLINE PREDICTION PENALIZATION

Method	CIFAR-10		CIFAR-100		MNIST		FashionMNIST	
	CI	CD	CI	CD	CI	CD	CI	CD
CE	54.07	62.57	24.64	26.01	95.83	96.77	82.36	84.40
CE + CAugMatching	61.47	65.85	30.84	31.88	96.79	97.33	85.15	85.89
CE + CAugMatching + LNOR	65.02	65.97	32.16	32.80	97.31	97.50	85.19	86.44
CE + CAugMatching + LNOR + OPP	**72.86**	**72.83**	**34.89**	**34.67**	**97.92**	**97.77**	**86.35**	**86.72**

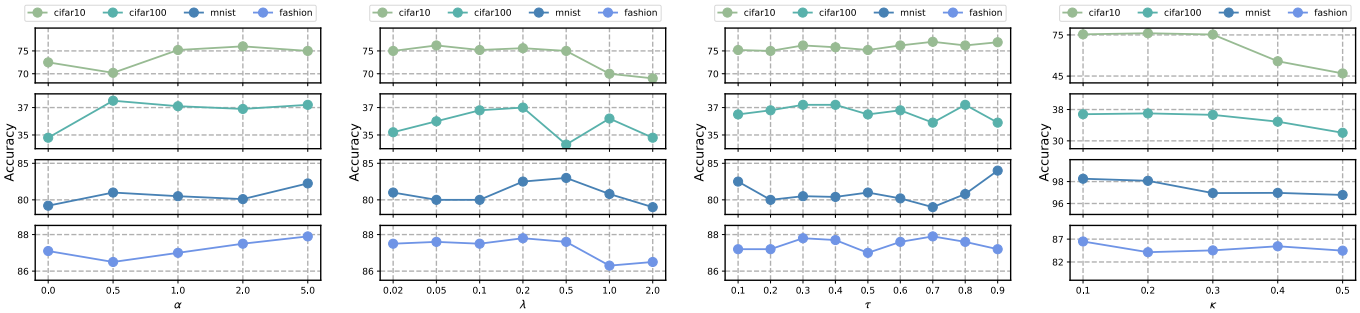


Fig. 7. Accuracy of setting hyperparameters α, λ, τ, and κ with different values on CIFAR-10, CIFAR-100, MNIST, and FashionMNIST.

F. Ablation Study

We conduct elaborate ablation study to validate core components of our method. Without specification, the CI label noise is used; η and ρ are set to 20% and 0.01, respectively.

1) Effectiveness of Loss Components: Thorough ablation study is conducted to validate loss components in our method, including cross-augmentation matching (*CAugMatching*) in (1), *LNOR* in (7), and *OPP* in (9). The experimental result is presented in Table VI. As shown in Table VI, the adoption of *CAugMatching* brings significant performance improvement, because it is advantageous in discriminating noisy samples from clean samples and enhancing the representation capacity. For example, *CAugMatching* leads to an accuracy gain of 6.20% on the CIFAR-100 dataset with CI label noise. *LNOR* helps mitigate the influence caused by noisy samples. It can also benefit the training of our model in all the cases, e.g., the accuracy is improved by 3.55% on CIFAR-10 with CI label noise. The *OPP* loss, devised for balancing training data, can bring substantial accuracy gains. For example, on CIFAR-10 and FashinMNIST with CI label noise, the accuracy is increased by 7.84% and 1.16%, respectively.

2) Robustness to Loss Criteria: We attempt to replace the cross-entropy loss in (2) with the focal loss [59] or the LDAM loss. As shown in Table VII, our method can improve baseline models in both the cases. For example, on the CIFAR-100 dataset, our method leads to the accuracy gain of 9.99% and 9.19%, when incorporated with the focal loss and the LDAM loss, respectively. This indicates that our method is not sensitive to loss functions.

3) Effect of DBBN: As shown in Table VIII, we study the impact of *DBBN* on CIFAR-10 and CIFAR-100. Thanks to the efficacy in preventing strongly augmented images from affecting the feature distribution, the DBBN can improve the classification performance consistently on both the datasets. We also implement a variant of our method (denoted as "with SN ") which uses two independent network backbones to process weakly and strongly augmented images separately. The accuracy of "with SN" is much worse the other two variants "with SN" and "with DBBN." When using two independent network backbones, their learning paces are not synchronized, and they can fit well with both weakly and strongly augmented

TABLE VII

EFFICACY OF OUR METHOD APPLIED WITH DIFFERENT LOSS FUNCTIONS. "OURS + CE," "OURS + FOCAL," AND "OURS + LDAM" DENOTE VARIANTS OF OUR METHOD TRAINED WITH CROSS-ENTROPY LOSS, FOCAL LOSS, AND LDAM LOSS, RESPECTIVELY

Method	CIFAR-10	CIFAR-100
CE	54.07	24.64
Ours + CE	**72.86**	**34.89**
Focal	56.04	25.07
Ours + Focal	**73.34**	**35.06**
LDAM	62.28	27.73
Ours + LDAM	**76.94**	**36.92**

TABLE VIII

COMPARISON OF VARIANTS OF OUR METHOD WITH OR WITHOUT DBBN. "WITHOUT DBBN" MEANS THE CONVENTIONAL BATCH NORMALIZATION WITH SHARED PARAMETERS FOR PROCESSING WEAKLY AND STRONGLY AUGMENTED IMAGES IS USED. "WITH SN" MEANS TWO SEPARATE NETWORKS ARE USED TO PROCESS WEAKLY AND STRONGLY AUGMENTED IMAGES, RESPECTIVELY

Variant	CIFAR-10	CIFAR-100	# Parameters
with SN	71.78	32.04	0.922M
without DBBN	74.11	34.22	0.461M
with DBBN	**76.94**	**36.92**	**0.464M**

TABLE IX
PERFORMANCE OF DIFFERENT METHODS IN SEPARATING CLEAN AND NOISY SAMPLES ON FASHIONMNIST

Method	CI		CD	
	η=0.2	η=0.6	η=0.2	η=0.6
Co-Teaching	65.83	37.86	56.59	16.50
Co-Teaching+	36.58	19.84	38.13	17.95
JoCoR	85.52	65.10	65.43	18.60
Co-Matching	78.21	65.53	57.51	20.56
Ours	**86.35**	**70.21**	**86.72**	**25.34**

TABLE X
IMPACT OF DATA AUGMENTATION. "WEAK" INDICATES DATA AUGMENTATION IS IMPLEMENTED WITH WEAK AUGMENTATION OPERATIONS, AND "STRONG" INDICATES DATA AUGMENTATION IS IMPLEMENTED WITH STRONG AUGMENTATION OPERATIONS. "W/O" MEANS AUGMENTATION IS NOT USED

No.	Augmentation #1	Augmentation #2	CIFAR-10	CIFAR-100
1	Weak	w/o	68.82	29.31
2	Strong	w/o	71.16	32.90
3	Weak	Weak	71.88	33.56
4	Weak	Strong	76.94	36.92
5	Strong	Strong	73.13	34.32

images. Under such a circumstance, the cross-augmentation matching has limited performance in differentiating clean samples from noisy samples. Besides, using two backbones costs twice the number of parameters in contrast to using one backbone, while the DBBN introduces few extra parameters.

4) Performance of Noise Detection: In this section, we compare our method against other LLN methods on the performance of noisy sample detection. The experiments are conducted on the FashionMNIST dataset. CI and CD label noises with $\eta = 0.2$ or 0.6 are used for performance evaluation, and ρ is set to 0.01. The accuracy metric is used for evaluation, and the experimental results are presented in Table IX. Our proposed method performs significantly better than other LLN methods.

5) Sensitivity to Hyperparameters: In this section, we discuss the sensitivity of our method against the variance of hyperparameters, including α, λ, τ, and κ. The experimental results are illustrated in Fig. 7. We can observe that the performance of our method is robust against the variance of α, λ, and τ in most cases. For κ, the performance of our method is stable when varying κ in [0.1, 0.3]. However, increasing κ means more samples are regarded as noisy samples, which is likely to mistake more clean samples as noisy ones. Hence, using a too large value for κ causes severe performance degradation.

6) Impact of Data Augmentation Strategies: In this section, we conduct experiments to validate the effectiveness of using both weak and strong data augmentations. The results are presented in Table X. First, we try to use only one augmentation to reimplement our method. In No. 1, only weak augmentation is used to generate x_i, and $\hat{x}_i$ is equivalent to x_i. In No. 2, only strong augmentation is used to generate x_i, and $\hat{x}_i$ is also equivalent to x_i. However, using the same image for x_i and $\hat{x}_i$ degrades the efficacy of the cross-augmentation matching loss in differentiating noisy samples and clean samples. No. 3 uses two different weak augmentations to generate x_i and $\hat{x}_i$, which can significantly improve the classification performance on both the datasets compared with No. 1. Using two different strong augmentations can achieve a similar effect after comparing No. 5 against No. 2. As indicated by No. 5, combining weak and strong augmentations can further improve the performance.

V. CONCLUSION

In this article, we propose an effective method to tackle the task of learning with both long-tailed distribution and label noise. We find that the integration of the matching between predictions on weakly and strongly augmented images is beneficial for improving the discrimination between noisy samples and clean samples of tail classes. It is also validated that adopting dual branches to process weakly and strongly augmented images with different parameters in the batch normalization can promote the performance since it prevents the deviation of feature distribution. We demonstrate that the LNOR on noisy samples also improves the performance since it benefits in erasing the negative impact of noisy samples. In addition, an OPP scheme is proposed for rebalancing the training data, which can achieve the efficacy of label smoothing. Extensive experiments on five datasets illustrate that the proposed method outperforms the state-of-the-art methods.

REFERENCES

[1] D. Zhang, J. Han, G. Cheng, and M. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5866–5885, Sep. 2022.
[2] P. Huang, J. Han, N. Liu, J. Ren, and D. Zhang, "Scribble-supervised video object segmentation," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 2, pp. 339–353, Feb. 2022.
[3] H. Su, Y. Ye, Z. Chen, M. Song, and L. Cheng, "Re-attention transformer for weakly supervised object localization," 2022, *arXiv:2208.01838*.
[4] D. Zhang et al., "Weakly supervised semantic segmentation via alternative self-dual teaching," 2021, *arXiv:2112.09459*.
[5] J. Huang et al., "Trash to treasure: Harvesting OOD data with cross-modal matching for open-set semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8290–8299.
[6] K. Wang et al., "Double-check soft teacher for semi-supervised object detection," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, p. 3.
[7] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," 2016, *arXiv:1611.03530*.
[8] D. Arpit et al., "A closer look at memorization in deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 233–242.
[9] B. Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," 2018, *arXiv:1804.06872*.
[10] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7164–7173.
[11] K. Nishi, Y. Ding, A. Rich, and T. Höllerer, "Augmentation strategies for learning with noisy labels," 2021, *arXiv:2103.02130*.
[12] A. Krizhevsky. (2009). *Learning Multiple Layers of Features From Tiny Images*. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
[13] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2691–2699.
[14] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13723–13732.
[15] Y. Lu, Y. Bo, and W. He, "Co-matching: Combating noisy labels by augmentation anchoring," 2021, *arXiv:2103.12814*.
[16] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," 2016, *arXiv:1608.06048*.

[17] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4109–4118.
[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[20] Y. Zhang, X.-S. Wei, B. Zhou, and J. Wu, "Bag of tricks for long-tailed visual recognition with deep convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3447–3455.
[21] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.
[22] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 27, 2022, doi: 10.1109/TNNLS.2021.3136503.
[23] B. Zhou, Q. Cui, X. Wei, and Z. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9716–9725.
[24] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," 2019, *arXiv:1906.07413*.
[25] Z. Zhong, J. Cui, S. Liu, and J. Jia, "Improving calibration for long-tailed recognition," 2021, *arXiv:2104.00466*.
[26] C. Wang et al., "Label-aware distribution calibration for long-tailed classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 25, 2022, doi: 10.1109/TNNLS.2022.3213522.
[27] N. Kang, H. Chang, B. Ma, and S. Shan, "A comprehensive framework for long-tailed learning via pretraining and normalization," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 27, 2022, doi: 10.1109/TNNLS.2022.3192475.
[28] Y. Zhong et al., "Unequal-training for deep face recognition with long-tailed noisy data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7804–7813.
[29] J. Li, R. Socher, and S. C. H. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," 2020, *arXiv:2002.07394*.
[30] Y. Wei, C. Gong, S. Chen, T. Liu, J. Yang, and D. Tao, "Harnessing side information for classification under label noise," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3178–3192, Sep. 2020.
[31] Y. Yao et al., "Jo-SRC: A contrastive approach for combating noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5188–5197.
[32] R. Yi and Y. Huang, "Transform consistency for learning with noisy labels," 2021, *arXiv:2103.13872*.
[33] H. Song, M. Kim, D. Park, Y. Shin, and J. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 7, 2022, doi: 10.1109/TNNLS.2022.3152527.
[34] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," 2019, *arXiv:1904.12848*.
[35] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," 2020, *arXiv:2001.07685*.
[36] C. Zhang et al., "Delving deep into label smoothing," *IEEE Trans. Image Process.*, vol. 30, pp. 5984–5996, 2021.
[37] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
[38] W. Chang, T. You, S. Seo, S. Kwak, and B. Han, "Domain-specific batch normalization for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7346–7354.
[39] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018.
[40] B. Kang et al., "Decoupling representation and classifier for long-tailed recognition," 2019, *arXiv:1910.09217*.
[41] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation policies from data," 2018, *arXiv:1805.09501*.
[42] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.
[43] Z. Zhu, T. Liu, and Y. Liu, "A second-order approach to learning with instance-dependent label noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10108–10118.
[44] Y. Kim, J. Yun, H. Shon, and J. Kim, "Joint negative and positive learning for noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9437–9446.
[45] X. Xia et al., "Robust early-learning: Hindering the memorization of noisy labels," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–12.
[46] D. Gudovskiy, L. Rigazio, S. Ishizaka, K. Kozuka, and S. Tsukizawa, "AutoDO: Robust AutoAugment for biased data with label noise via scalable probabilistic implicit differentiation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16596–16605.
[47] Y. Zhang, S. Zheng, P. Wu, M. Goswami, and C. Chen, "Learning with feature-dependent label noise: A progressive approach," 2021, *arXiv:2103.07756*.
[48] D. Ortego, E. Arazo, P. Albert, N. E. O'Connor, and K. McGuinness, "Multi-objective interpolation training for robustness to label noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6602–6611.
[49] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5552–5560.
[50] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Learning to learn from noisy labeled data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5046–5054.
[51] W. Zhang, Y. Wang, and Y. Qiao, "MetaCleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7365–7374.
[52] K. Yi and J. Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7010–7018.
[53] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, "MixMatch: A holistic approach to semi-supervised learning," 2019, *arXiv:1905.02249*.
[54] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-learning regularization prevents memorization of noisy labels," 2020, *arXiv:2007.00151*.
[55] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness, "Unsupervised label noise modeling and loss correction," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 312–321.
[56] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4334–4343.
[57] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[58] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[59] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

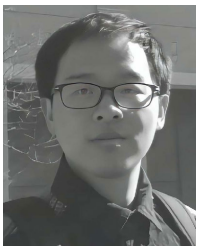
Chaowei Fang received the bachelor's degree from Xi'an Jiaotong University, Xi'an, China, in 2013, and the Ph.D. degree from the University of Hong Kong, Hong Kong, in 2019.

He is a Lecturer at the School of Artificial Intelligence, Xidian University, Xi'an, China. He has contributed as an author or a coauthor to several publications featured in prestigious journals such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON MEDICAL IMAGING (TMI), and *Medical Image Analysis* (MIA), as well as renowned conferences such as IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on Computer Vision (ICCV), AAAI Conference on Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), and ACM International Conference on Multimedia (ACM MM). Moreover, he took on the role of organizing the third International Workshop on Human-Centric Multimedia Analysis, held at ACM MM in 2022. His research interests encompass various domains, including low-level image processing, medical image analysis, and machine learning.

Lechao Cheng (Member, IEEE) received the Ph.D. degree from the College of Computer Science and Technology, Zhejiang University, Hangzhou, China, in 2019.

He is currently an Associate Professor at the Institute of Artificial Intelligence, Zhejiang Laboratory, Hangzhou, China. He has contributed more than 30 research papers to renowned academic journals and conferences such as *International Journal of Computer Vision* (IJCV), IEEE TRANSACTIONS ON MEDICAL IMAGING (TMI), IEEE TRANSACTIONS ON CYBERNETICS (TCyb), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), AAAI Conference on Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), and ACM International Conference on Multimedia (ACM MM). His research area centers around vision knowledge transfer in deep learning.

Yining Mao, photograph and biography not available at the time of publication.



Dingwen Zhang (Member, IEEE) received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2018.

From 2015 to 2017, he was a Visiting Scholar at the Robotic Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Professor with the School of Automation, Northwestern Polytechnical University. He is also an Adjunct Researcher at the Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei, China. His research interests include computer vision and multimedia processing, especially on saliency detection, video object segmentation, temporal action localization, and weakly supervised learning.

Yixiang Fang received the Ph.D. degree from the University of Hong Kong (HKU), Hong Kong, in 2017.

Currently, he is an Associate Professor at the School of Data Science, Chinese University of Hong Kong, Shenzhen, China. His research interests mainly focus on areas of data management, data mining, and artificial intelligence over big data.

Guanbin Li (Member, IEEE) received the Ph.D. degree from the University of Hong Kong, Hong Kong, in 2016.

He is currently an Associate Professor at the School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China. He has authorized and co-authorized on more than 100 papers in top-tier academic journals and conferences. His current research interests include computer vision, image processing, and deep learning.

Dr. Li was a recipient of the IEEE International Conference on Computer Vision (ICCV) 2019 Best Paper Nomination Award. He serves as an Area Chair for the conference of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP). He has been serving as a reviewer for numerous academic journals and conferences such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), *International Journal of Computer Vision* (IJCV), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE TRANSACTIONS ON CYBERNETICS (TCyb), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ICCV, European Conference on Computer Vision (ECCV), and Neural Information Processing Systems (NeurIPS).

Huiyan Qi, photograph and biography not available at the time of publication.

Licheng Jiao (Fellow, IEEE) received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University, Xi'an, where he is currently the Director of Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. His research interests include image processing, natural computation, machine learning, and intelligent information processing.

Dr. Jiao is the Foreign Member of the Academia Europaea, the Foreign Member of the Russian Academy of Natural Sciences, the Chairperson of the Awards and Recognition Committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, the Foreign Member of the Academia Europaea, the Foreign Member of the Russian Academy of Natural Sciences, the fellow of Institution of Engineering and Technology (IET)/Chinese Association of Artificial Intelligence (CAAI)/Chinese Institute of Electronics (CIE)/China Computer Federation (CCF)/Chinese Association of Automation (CAA), a Councilor of the Chinese Institute of Electronics, a Committee Member of the Chinese Committee of Neural Networks, and an expert of the Academic Degrees Committee of the State Council.