Dynamic Loss for Robust Learning

Shenwang Jiang ©, Jianan Li ©, Jizhou Zhang ©, Ying Wang ©, and Tingfa Xu ©

Abstract—Label noise and class imbalance are common challenges encountered in real-world datasets. Existing approaches for robust learning often focus on addressing either label noise or class imbalance individually, resulting in suboptimal performance when both biases are present. To bridge this gap, this work introduces a novel meta-learning-based dynamic loss that adapts the objective functions during the training process to effectively learn a classifier from long-tailed noisy data. Specifically, our dynamic loss consists of two components: a label corrector and a margin generator. The label corrector is responsible for correcting noisy labels, while the margin generator generates per-class classification margins by capturing the underlying data distribution and the learning state of the classifier. In addition, we employ a hierarchical sampling strategy that enriches a small amount of unbiased metadata with diverse and challenging samples. This allows the joint optimization of the two components in the dynamic loss through meta-learning, allowing the classifier to effectively adapt to clean and balanced test data. Extensive experiments conducted on multiple real-world and synthetic datasets with various types of data biases, including CIFAR-10/100, Animal-10N, ImageNet-LT, and Webvision, demonstrate that our method achieves state-of-the-art accuracy.

Index Terms—Robust learning, label noise, class imbalance, meta learning.

I. INTRODUCTION

Deep neural networks (DNNs) have demonstrated remarkable success attributed to the abundance of labeled data [1], [2], [3]. However, real-world datasets often exhibit long-tailed distributions and inevitably contain noisy labels [4]. The presence of such biased data distributions renders DNNs susceptible to overlooking tail classes [5] and memorizing noisy training labels [6], leading to suboptimal performance on balanced and clean test data. Consequently, addressing the challenge of robust learning from long-tailed data with noisy labels has received growing attention in recent studies [7], [8].

Recent advancements in robust learning [6], [9], [10], [11] have explored various techniques, including correcting noisy labels [12], [13] and adjusting classification margins [5], [14], to address the issues of label noise and class imbalance, respectively. However, these solutions heavily rely on manually-designed rules to distinguish between noisy and clean samples, or pre-set parameters to obtain prior knowledge of class distribution. When label noise meets class imbalance, manual interventions become impractical since tail classes make it difficult to identify label noise, while noisy labels make observed class distribution unreliable. As a result, we aim to pave a new path for robust learning from biased data in a fully self-adaptive manner.

Taking the aforementioned challenges into consideration, we propose a novel meta-learning-based dynamic loss that comprises of a learnable label corrector and a margin generator. Our approach aims to learn a classifier robustly from long-tailed noisy data. In contrast to the predefined fixed objective functions adopted by priors, our dynamic loss learns to correct per-sample labels and adjust per-class classification margins simultaneously by perceiving the underlying data distribution and the learning state of the classifier. As a result, our approach provides suitable dynamic learning objectives throughout the training process.

The main challenge in correcting noisy labels lies in identifying noisy samples and disregarding or reusing them appropriately. Previous solutions [13], [15] have used a GMM with fixed hyperparameters to differentiate between noisy and clean samples based on their loss values, which is inadequate for long-tailed noisy data. This is because some clean samples of tail classes may have higher loss values than noisy samples of head classes due to the highly skewed class distribution. While a class-specific GMM [16] is a straightforward solution, it is still prone to confusion due to the significant variation in noise rates among different classes (Fig. 1). Furthermore, previous methods have relabeled identified noisy samples by combining their assigned labels and the classifier’s predictions in a fixed manner. This approach can lead to incorrect corrections since the classifier is imprecise at the beginning of training and susceptible to overfitting to noisy samples towards the end. To address these issues, we propose a novel label corrector that can learn to jointly identify and relabel noisy samples in a fully learnable fashion. This is achieved by taking into account both the class-specific loss rank of samples and the learning state of the classifier.

In order to adjust per-class classification margins, we are inspired by the fact that classes with fewer samples are associated with larger generalization error bounds, which can be minimized by increasing the classification margin [5]. However, most prior approaches [14] pre-define a fixed margin for each class based solely on its sample number (Fig. 1). This approach suffers from...
We devise a novel hierarchical sampling strategy that efficiently generates diverse and unbiased meta data, which allows for better simulation of the true data distribution and improves the generalization ability of the model.

We achieve state-of-the-art performance on multiple synthetic and real-world datasets, demonstrating the effectiveness and versatility of our proposed approach.

### Related Works

**Long-Tailed Learning**: Previous works on long-tailed learning can be broadly categorized into three main approaches: data re-sampling [20], boundary adjustment [22], and re-weighting [9], [23], [24]. The data re-sampling approach involves balancing the class distribution by over-sampling the tail classes. However, this method is prone to overfitting on the tail classes. Methods belonging to the second category aim to enlarge the classification boundary of the tail classes while narrowing that of the head classes. This is achieved by modifying the classification threshold [25] or by adjusting the weights of the output layer through normalization [10]. The re-weighting approach aims to assign larger loss weights to the tail classes. Conventional approaches of this category [24], [26] impose weights on each training sample directly, which may cause unstable training due to sensitivity to outliers [14]. Recent works [27] modify the predicted scores in the Softmax function to achieve re-weighting, which yields more stable training and promising performance. In this work, we adapt the re-weighting strategy to more challenging long-tailed scenarios with label noise.

**Learning Under Label Noise**: There are two main categories of methods for learning under label noise: sample re-weighting and relabeling. The re-weighting strategy involves assigning lower weights to samples with larger loss values, which are considered to be noisy [28], [29]. MentorNet [6] learns data-driven curriculums for deep convolutional neural networks trained on corrupted labels. Meta-Weight-Net [9] learns an explicit weight-function directly from a small set of clean data. MetaSeg [30] directly generates weights according to the feature and given label of images. On the other hand, the relabeling strategy leverages noisy samples by refining their labels. Bootstrapping [31] integrates assigned labels and model predictions through interpolation. Some works divide clean and noisy samples based on priors learned from a manually generated noisy set [32] and then take advantage of noisy samples [33].
Long-Tailed Learning Under Label Noise: Recently, there have been several approaches proposed for addressing long-tailed learning with label noise. For instance, HAR [8] applies a data-dependent regularization technique to regularize different regions of the input space differently. CurveNet [7] learns to assign appropriate weights to different samples based on their loss curves. ROLT [34] combines DivideMix and LDAM to correct noisy labels and improve the performance of tail classes. However, unlike these methods, the approach presented in this work is the first to simultaneously correct noisy labels and adaptively adjust per-class classification margins in a learnable and adaptive manner according to the training data.

III. METHODS

In this section, we will provide a detailed description of our dynamic loss and the process of optimizing it through meta-learning.

A. Overview

Our objective is to train a classifier $f_\omega$ with learnable parameters $\omega$ using a noisy and imbalanced training set $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$, where each training example comprises an image $x_i$ and its corresponding one-hot class label $y_i$. Although the training set is beset by label noise and class imbalance, our aim is to ensure that the classifier accurately recognizes all classes. To achieve this goal, we use a balanced and clean test set.

We optimize the learnable parameters $\omega$ by minimizing the classification loss on the training set, given by

$$\omega^* = \arg\min_{\omega} \sum_{i=1}^N \ell(y_i, f_\omega(x_i)).$$

where $\ell(\cdot)$ represents the cross-entropy loss. However, due to the presence of label noise and class imbalance in the training set, optimizing using this naive cross-entropy loss suffers from two significant drawbacks. First, the labels assigned to noisy samples do not correspond to their ground-truths, leading to high loss values and causing the model to memorize the noisy labels. Second, tail classes occur much less frequently than head classes, but have the same classification margins, making them susceptible to poor generalization.

To address the aforementioned issues, we introduce a novel dynamic loss function that corrects noisy labels and adjusts the classification margins for different classes in an adaptive and learnable manner. The dynamic loss is given by:

$$\mathcal{D}L = \ell(y^*_i, f_\omega(x_i)) + q,$$

where $y^*_i$ and $q \in \mathbb{R}^C$ represent the reassigned label and the additive classification margin for $x_i$, respectively.

As depicted in Fig. 2, the dynamic loss incorporates a learnable label corrector $G_{\theta_l}$, parameterized by $\theta_l$, and a margin generator $G_{\theta_m}$, parameterized by $\theta_m$, which are responsible for correcting per-sample labels and per-class classification margins, respectively. We jointly optimize these components along with the classifier $f_\omega$ using meta-learning. In the following, we elaborate on the two components and their optimization in detail.

B. Label Corrector

The label corrector operates in a class-wise manner to detect and correct wrongly assigned labels in noisy samples. To identify such samples, it first divides all samples into $C$ groups based on their class and sorts the samples in each group individually according to their loss values. It then divides the sorted samples in each group into $R$ bins of equal size. To determine whether a bin $r \in \{1, \ldots, R\}$ is dominated by noisy or clean samples, the label corrector uses a lightweight class-wise meta network. As a result, the bin index $r_i$ of the loss value for sample $i$ can be used as a reliable indicator for identifying label noise.

In terms of label correction, the classifier can learn from the clean samples that dominate the data and transfer this knowledge to the noisy samples, provided that the classifier has not severely over-fitted on biased data. This allows the classifier’s predictions on noisy samples to be more accurate and closer to their ground-truth labels, facilitating the correction of wrongly assigned labels.
Drawing from the aforementioned observation, we present our label corrector which reassigned a ground-truth label \( y_i^* \) for a given sample \( x_i \) by calculating a weighted sum of its assigned label \( y_i \) and the prediction \( y_i' \) made by the classifier. The weights used for the summation are dependent on the loss bin index \( r_i \), given by:

\[
y_i^* = \mathcal{G}_b(y_i, y_i', r_i) = y_i + y_i' \cdot (1 - g(r_i|y_i)),
\]

(3)

The function \( g : r_i|y_i \rightarrow [0, 1] \) is a class-dependent weighting function that maps the bin index \( r_i \) to a balance weight and is learned by a small meta network. The network is constructed of a one-hot encoder and a two-layer perceptron (MLP) with a Sigmoid activation function.

In instances where sample \( i \) is regarded as noisy due to a high loss value resulting in a large bin index \( r_i \), the computed value of \( g(r_i|y_i) \) tends to approach zero. Consequently, the label corrector adjusts its label assignment by incorporating the prediction made by the classifier to rectify its initial erroneous label assignment. The reverse is also true: if the sample has a low loss value and a small bin index, the weight assigned to the classifier’s prediction is close to zero, allowing the assigned label to remain dominant in the label correction process.

C. Margin Generator

To design the margin generator, we revisit the Label-Distribution-Aware Margin Loss (LDAM) [5] from the perspective of generalization error bound. Given that tail classes often have fewer training samples, they typically exhibit larger generalization error bounds when compared with head classes. Since the generalization error bound is often negatively correlated with the magnitude of the classification margin, increasing the classification margins for the tail classes can effectively reduce their generalization error bounds.

In this regard, Balanced Meta-Softmax [14] presents an unbiased extension of standard Softmax by adjusting the classification margin for class \( j \) based on its sample number \( n_j \), and adding the margin \(-\log(n_j)\) to the confidence score \( p_j \) predicted by the classifier:

\[
\hat{\ell}(\omega|y = j) = -\log \left( \frac{e^{p_j + \log(n_j)}}{\sum_{i=1}^{C} e^{p_i + \log(n_i)}} \right).
\]

(4)

However, when dealing with long-tailed data with noisy labels, the sample number \( n_j \) may not accurately reflect the true number of samples belonging to class \( j \) due to the existence of label noise. Furthermore, manually defining the margin solely based on the sample number may not account for the distinct classification difficulties among different classes.

We propose a learnable margin generator \( \mathcal{G}_m \), which consists of a two-layer multi-layer perceptron (MLP) that can dynamically adjust the margin for each class. During classifier training, the margin generator optimizes a learnable margin vector \( \mathbf{q} \in \mathbb{R}^C \), which is initialized with an all-ones vector:

\[
\mathbf{q} = \mathcal{G}_m(1) = [q_1, \ldots, q_C].
\]

(5)

By integrating the margin vector into the standard Softmax loss, we have the modified loss function:

\[
\hat{\ell}(\omega|y = j) = -\log \left( \frac{e^{p_j + q_j}}{\sum_{i=1}^{C} e^{p_i + q_i}} \right).
\]

(6)

where \( p_j \) and \( q_j \) denote the predicted score and the learned margin for class \( j \), respectively. Since the classification margin is \(-q_j \) in our formulation, the learned margin \( q_j \) for class \( j \) should be positively correlated with its sample number.

Hence, the margin generator can automatically adjust per-class margins by adapting to the underlying true class distribution in long-tailed noisy data, as well as the classification difficulty of each class. Importantly, the margin generator can accomplish this in a learnable manner, without requiring any manual intervention or prior information.

D. Hierarchical Sampling Strategy

We combine the label corrector \( \mathcal{G}_b \) and margin generator \( \mathcal{G}_m \) into a unified meta set \( \mathcal{G}_b \) which is a crucial element of our dynamic loss. We employ meta-learning to optimize \( \mathcal{G}_b \) and facilitate the learning of the classifier \( f_\omega \) to better adjust to balanced and clean test data.

To enable meta-learning, a meta set \( \mathcal{D}_c = \{(x_i, y_i)\}_{i=1}^{M} \) containing a small number of balanced and clean data must be constructed. A simple approach is to select \( M \) samples with the lowest classification loss values from each class in \( \mathcal{D} \), as samples with lower loss values computed by \( f_\omega \) are more likely to have correctly assigned labels. However, this approach can be problematic because easier samples often have lower loss values during training, which can result in fixed easy samples being selected at each epoch and increase the risk of overfitting to such samples.

Therefore, we have devised a hierarchical sampling strategy to create \( \mathcal{D}_c \). This strategy involves a two-step process: first, we randomly select \( M_0 \) samples from each class in \( \mathcal{D} \) to construct a primary set; second, we choose \( M_1 \) samples with low loss from each class in the primary set to form the final meta set. The samples that are not selected for the meta set are included in the counterpart set \( \mathcal{D}_n \). Refer to Fig. 2 for an illustration of this process.

The advantages of incorporating the primary set are two-fold. First, since the samples in the primary set are randomly selected at each epoch, it ensures that the resulting meta set is distinct across different epochs. Second, the primary set contains fewer samples than \( \mathcal{D} \), increasing the likelihood of selecting hard yet clean samples located near the decision boundary into the meta set. This hierarchical sampling strategy guarantees the dynamism and diversity of the meta set, preventing the model from overfitting to biased data.

E. Optimization

The convergence of meta-learning poses significant challenges [17]. Previous approaches [7], [9], [15] employ a strategy of inputting all known information, particularly given labels, predicted labels and loss value, into the meta net, generating
customized weights for each sample. However, due to computational complexity, meta nets often exhibit a relatively simple network structure consisting of a few fully connected layers, limiting their capability to handle complex tasks. Notably, in scenarios involving biased data, a substantial portion of the known information, such as predicted and given labels, can be omitted from the meta net’s input. This is because only one of these labels needs to be selected as the sample’s label, and a weight can be learned to combine the two labels and obtain a new label. Directly incorporating these known information into the meta net not only increases the optimization difficulty but also diminishes the utilization of the information. This is due to the implicit nature of using these information, requiring the meta net to extract it from the input, as expressed by:

\[ y^*_i = g_{0_i}(y_i, y'_i, r_i) = g(y_i, y'_i, r_i). \]  

To enhance the utilization of known information, we opt for explicit utilization through our Label Corrector, as illustrated in Fig. 3 and governed by (3).

The customization of weights for individual samples diminishes the impact of diverse input information on the meta net. However, input information often encompasses errors, such as varying sample loss values during training and overlapping loss values between noisy and clean samples. These factors make it challenging to discern between them solely based on loss values. To overcome this challenge, we employ a grouping strategy that involves sorting by loss value for noisy data and given labels for imbalanced data. This strategy brings stability to the input information, reducing its instability and enhancing the influence of each input on the meta net. By explicitly utilizing known information and implementing the grouping strategy, we simplify the task of the meta net, promoting easier convergence and the assignment of accurate weights and margins to each sample group.

IV. EXPERIMENTS ON LONG-TAILED NOISY DATA

A. Experiments on CIFAR-N-LT

Dataset and Implement Details: We assess the performance of our method on CIFAR-N-LT dataset, which comprises CIFAR-10 and CIFAR-100 [18]. These datasets contain 60,000 RGB images, out of which 50,000 images are used for training and 10,000 images for testing. The images are evenly distributed across 10 and 100 categories, respectively. Additionally, the datasets are subjected to simulated label noise and class imbalance.

To simulate a long-tailed dataset, we follow the exponential profile proposed in [5], where the imbalance ratio \( \rho \) leads to an exponential decay in the sample number across different classes. We then inject label noise into the long-tailed dataset to create the training set, with each sample’s label independently changed to class \( j \) with probability \( \frac{N_j}{N} \lambda \), where \( N \) is the total number of training samples, \( N_j \) is the frequency of class \( j \), and \( \lambda \) represents the noise rate.

Following ROLT [34], we consider imbalance ratios of \( \rho \in \{10, 100\} \) and noise rates of \( \lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5\} \). The ResNet-32 [2] is adopted as the classifier, and we train all classifiers using balanced-softmax for 300 epochs with a batch size of 512. The learning rate is initialized as 0.1 and controlled by a cosine annealing learning scheduler [37]. We train all classifiers using the same SGD optimizer with a momentum of 0.9 and a weight decay of 5e-4. Further training details can be found in Table I.

Main Results: Tables II and III present the average accuracy on CIFAR-10-N-LT and CIFAR-100-N-LT with varying imbalance ratios and noise rates. Our method exhibits consistently high accuracy across a wide range of biases, whereas previous methods suffer rapid degradation. Specifically, our dynamic loss improves the average last accuracy by 3.21% and 5.79% compared to HAR on CIFAR-10-N-LT and CIFAR-100-N-LT, respectively. Additionally, our method significantly outperforms the baseline model that simply combines strategies from DivideMix and Balanced-Softmax.

Furthermore, the performance of our last model is generally very close to that of our best model despite varying bias settings. In contrast, the last models of DivideMix and Balanced-Softmax degrade significantly compared to their corresponding best models, especially on CIFAR-10-N-LT with severe imbalance and noise (e.g., \( \rho = 100 \) and \( \lambda = 0.5 \)). These results suggest that our method is much more resistant to overfitting on biased data than the aforementioned priors.

It is worth noting that previous methods require carefully tuned hyperparameters based on unobservable noise rate [13] or perform two-stage training to obtain prior information on class distribution [8]. In comparison, our dynamic loss employs a fixed set of hyperparameters and requires only one-round end-to-end training without manual interventions on the same dataset.

B. Experiments on Webvision

Dataset and Implement Details: We also evaluate our dynamic loss on the WebVision dataset [4], which is a large-scale real-world dataset that suffers from label noise and class imbalance. It comprises 2.4 million images, of which approximately 20% are mislabeled [41]. To construct the miniWebVision dataset, we follow the methodology proposed in MentorNet [6] by selecting the top 50 classes, resulting in an observed imbalance ratio of around 6.78. Following priors [13], we train the
Inception-ResNet V2 [42] for 150 epochs using the SGD optimizer with momentum 0.9 and weight decay of 1e-4. For ResNet-50, the training details is following INOLML [40]. During the warm-up stage, which lasts for one epoch, we use a batch size of 64 and an initial learning rate of 0.02. The learning rate is then adjusted using a cosine annealing learning scheduler. We evaluate the model’s performance on the validation sets of both WebVision and ImageNet [43]. Additional training details are available in Table I.

**Main Results:** Table IV displays the performance of our method on the WebVision and ImageNet validation sets. Our approach outperforms other state-of-the-art methods even though most of them utilize additional model co-training and ensembling techniques. Notably, our method surpasses HAR and ROLT+, which are specifically designed to handle long-tailed noisy data, by at least 2.48% in terms of accuracy on WebVision, demonstrating its superiority.

The rapid advancements in the field of self-supervised learning have highlighted the increasing importance of adaptive methods for learning from biased data. Particularly, methods that can effectively adapt to models with varying initial parameters are of great significance. In this study, we performed experiments on ResNet-50 that was pre-trained using self-supervised techniques, such as CLIP [44]. The results, presented in the last row of Table IV, demonstrate a significant improvement of 3.92% in accuracy, achieved through our proposed approach, compared to the model trained using cross entropy with the same initial parameters pretrained by CLIP. These findings provide evidence of the dynamic adaptability of our method to classifiers with different initial parameters.
TABLE III
ACCURACY (%) ON CIFAR-100-N-LT WITH VARYING IMBALANCE RATIO AND NOISE RATE

| Imbalance Ratio | Noise Rate | 0.1 | 0.3 | 0.4 | 0.5 | 100 | Average |
|-----------------|------------|-----|-----|-----|-----|-----|---------|
| Cross Entropy   | Last       | 43.48| 37.25| 31.34| 25.53| 19.45| 29.92 |
|                 | Best       | 43.95| 37.64| 32.18| 29.44| 23.87| 30.52 |
| DivideMix [13]  | Last       | 54.17| 51.92| 50.44| 45.02| 43.43| 36.31 |
|                 | Best       | 54.94| 53.35| 50.93| 45.36| 43.44| 36.99 |
| ELR+ [35]       | Last       | 52.48| 51.30| 46.24| 39.98| 34.91| 33.01 |
|                 | Best       | 53.91| 51.90| 47.88| 42.61| 37.35| 33.81 |
| MOIT+ [36]      | Last       | 44.66| 40.12| NC   | NC   | NC   | NC     |
|                 | Best       | 47.70| 44.94| 42.10| 39.12| 35.50| 32.66 |
| Balanced-Softmax [14] | Last       | 58.38| 54.59| 50.49| 44.83| 40.45| 43.17 |
|                 | Best       | 58.62| 54.73| 50.66| 45.63| 40.56| 43.30 |
| DivideMix +Balanced-Softmax | Last       | 56.37| 54.80| 54.83| 51.29| 48.89| 43.26 |
|                 | Best       | 56.85| 56.06| 55.64| 52.30| 50.01| 43.67 |
| FaMUS [15]      | Last       | 46.07| 51.59| 46.07| 46.93| 43.83| 29.33 |
|                 | Best       | 47.03| 52.05| 46.41| 47.88| 44.30| 29.66 |
| CurveNet [7]    | Last       | 50.41| 47.14| 43.18| 41.23| 34.85| 22.10 |
|                 | Best       | 52.73| 51.93| 47.56| 44.08| 39.74| 25.26 |
| HAR [8]         | Last       | 58.68| 55.43| 52.57| 46.01| 43.96| 42.67 |
|                 | Best       | 59.52| 58.50| 53.44| 46.75| 44.61| 44.45 |
| Dynamic Loss (Ours) | Last       | 59.24| 57.57| 56.85| 52.07| 50.74| 47.23 |
|                 | Best       | 59.52| 57.85| 57.32| 52.66| 51.26| 47.55 |

NC: not converging, NA: not available.

TABLE IV
ACCURACY (%) ON WEBVISION AND IMAGENET VALIDATION SETS

| Methods                  | Backbone | Webvision | ImageNet |
|--------------------------|----------|-----------|----------|
|                          |          | top 1     | top 5    | top 1    | top 5    |
| Cross Entropy            | IRV2     | 72.48     | 88.48    | 65.08    | 87.88    |
| Cross Entropy            | ResNet-50| 71.72     | 87.84    | 65.96    | 86.88    |
| Cross Entropy            | ResNet-50†| 78.04    | 93.04    | 72.84    | 91.68    |
| HAR [8]                  | IRV2     | 75.50     | 90.70    | 70.30    | 90.00    |
| DivideMix [13]           | IRV2     | 77.32     | 91.64    | 75.20    | 90.84    |
| ROLT+ [34]               | IRV2     | 77.64     | 92.44    | 74.64    | 92.48    |
| ELR+ [35]                | IRV2     | 77.78     | 91.68    | 70.29    | 89.76    |
| CMW-Net-SL* [38]         | IRV2     | 78.08     | 92.96    | 75.72    | 92.52    |
| FaMUS [15]               | IRV2     | 79.40     | 92.80    | 77.00    | 92.76    |
| MOIT+ [36]               | IRV2     | 87.86     | -        | -        | -        |
| NCR [39]                 | ResNet-50| 80.50     | -        | -        | -        |
| INOLML* [40]             | IRV2     | 81.70     | 93.80    | 78.10    | 92.90    |
| Dynamic Loss             | IRV2     | 80.12     | 93.64    | 74.76    | 93.08    |
| Dynamic Loss             | ResNet-50| 78.56     | 92.52    | 72.08    | 91.36    |
| Dynamic Loss             | ResNet-50†| 81.96    | 94.48    | 77.43    | 93.40    |

Denotes the use of model training or ensembling. † Indicates the backbone is pretrained using self-supervised techniques (CLIP). IRV2: Inception-ResNet V2.

V. EXPERIMENTS ON NOISY DATA

A. Experiments on CIFAR-N

Dataset and Implement Details: CIFAR-N is a synthetic noisy dataset derived from CIFAR. It includes two common types of simulated label noise: symmetric and asymmetric. Symmetric noise is introduced by randomly changing the labels with all possible labels based on a fixed probability of $\lambda$ (also known as noise rate). Asymmetric noise, on the other hand, is designed manually to mimic real-world label noise, where labels are only altered with those in similar classes, such as deer $\rightarrow$ horse and dog $\leftrightarrow$ cat. We evaluate the performance of our dynamic loss in handling noisy data using CIFAR-10-N and CIFAR-100-N with symmetric noise rates of 0.2, 0.4, 0.6 and asymmetric noise rates of 0.2, 0.4.

Results on CIFAR-N: Table V illustrates that our proposed method outperforms previous methods specifically designed for learning on noisy data, achieving the highest average accuracy. Unlike DivideMix, which requires manual tuning of hyperparameters under different noise types and rates, our dynamic loss adapts well to various noisy scenarios in a fully self-adaptive manner without any manual intervention.

B. Experiments on Animal-10N

Dataset and Implement Details: We also evaluate our method on the real-world noisy dataset ANIMAL-10N [19], which contains a total of 55,000 images (50,000 for training and 5,000 for testing) of 5 pairs of confusing animals. All image categories are equally represented in the dataset. The images are collected from websites using the search keywords as labels, resulting in significant label noise with an estimated rate of 8%. To ensure a fair comparison with prior work [19], we train the VGG10-BN [1] for 100 epochs using the SGD optimizer with a momentum of 0.9 and weight decay of 5e-4. The warm-up stage lasts for 5 epochs, and the batch size is set to 128. We initialize the learning rate to 0.02 and control it with a cosine annealing learning scheduler. Further training details are presented in Table I.
Main Results: Table VI illustrates that our dynamic loss achieves state-of-the-art performance compared to all previous methods on the ANIMAL-10N dataset. Specifically, our method outperforms the classifier trained with cross entropy by a large margin of 7.14%, demonstrating its superior ability to handle real-world noisy data.

VI. EXPERIMENTS ON LONG-TAILED DATA

A. Experiments on CIFAR-LT

Dataset and Implement Details: The CIFAR-LT is a simulated long-tailed dataset that is derived from CIFAR by reducing the number of training samples per class according to an exponential function $n_i = n_i \mu^i$, where $n_i$, $i$, and $n$ denote the number of samples in the $i$-th class, the class index, and the maximum number of samples across all classes, respectively. We evaluate the efficacy of our dynamic loss approach in dealing with long-tailed data on clean CIFAR datasets that have varying imbalance ratios ($\rho \in \{10, 20, 50, 100\}$). For more detailed information on the training process, please refer to Table I.

Results: Table VII presents the superior performance of our proposed method compared to previous approaches specifically designed for learning on long-tailed data. Notably, our method surpasses LDAM, which adjusts classification margins based solely on the sample number, by a significant margin of 4.03%.
and 5.75% on CIFAR-10-LT and CIFAR-100-LT, respectively. This result demonstrates the effectiveness of our dynamic loss in accurately perceiving the classification difficulty of different classes and adaptively adjusting their margins accordingly. Furthermore, our approach seamlessly integrates with Logit Adjustment and Balanced Softmax techniques by utilizing them as base margins while learning the residual margin through our margin generator. As shown in Table VII, our proposed approach with Balanced-softmax and logit adjustment yields a notable improvement in final accuracy, namely 0.19% and 3.05% on CIFAR-10-LT, compared to the initial Balanced-softmax and logit adjustment techniques. However, the performance level remains comparable to that of our original dynamic loss. These findings indicate that our original dynamic loss is sufficient for accurately identifying the classification complexity of diverse categories and accounting for the number of samples per class.

B. Experiments on ImageNet-LT

Dataset and Implement Details: The ImageNet-LT dataset consists of 115.8K images, categorized into 1,000 classes based on the Pareto distribution. Consequently, the number of images per class varies between 5 and 1,280. Following previous studies [11], [55], we employ several architectures for training, including ResNet-10, ResNet-50, ResNet-152, ResNeXt-50 [56], and self-supervised ResNet-50. These models are trained for either 90 or 400 epochs, utilizing the SGD optimizer with a momentum of 0.9 and weight decay of 1e-4. No warm-up stage is conducted, and a batch size of 128 is employed. The learning rate is initialized at 0.1 and controlled using a cosine annealing learning scheduler. Further details regarding the training procedure are provided in Table I.

Main Results Table VIII presents the experimental findings, showcasing the superior accuracy achieved by our proposed method when employed in conjunction with various classifiers, namely ResNet-50, ResNet-152, ResNeXt-50, and self-supervised ResNet-50. Remarkably, even when utilizing the ResNet-10 classifier trained from scratch, our method surpasses the performance of the PaCo framework [55] equipped with self-supervised techniques by 2.12%. These results underscore the robustness and efficacy of our proposed method in effectively handling the challenges posed by long-tailed datasets.

Additionally, we conducted supplementary experiments by incorporating our method with the pretrained ResNet50 using self-supervised techniques. The results reveal a significant performance improvement of 2.42% compared to the PaCo framework. This further highlights the compatibility of our proposed method with self-supervised approaches, leading to enhanced performance gains.

VII. Qualitative Analysis

A. Label Corrector

Behavior of Label Corrector: We examine the behavior of the label corrector on the balanced CIFAR-10-N with asymmetric noise, which is designed to emulate the structure of real-world label noise by assigning distinct noise rates to various classes. Fig. 4 illustrates the learned weight $g(r|y)$ by the label corrector and the percentage of noisy labels corresponding to an increasing loss bin index $r$ for each class. We observe that for the classes that contain noisy labels, clean samples mainly appear in the top-ranked (low-loss) bins while noisy samples occupy the bottom-ranked (high-loss) bins. This finding supports our hypothesis that the loss bin index $r$ can be used as a reliable input indicator for the label corrector to differentiate between noisy and clean samples. Accordingly, the generated weight $g(r|y)$ remains at 1 and suddenly drops to 0 at around bin 60, indicating that the label corrector preserves the assigned ground-truth label for clean samples and resorts to the predicted label that is more likely to be the ground-truth for noisy samples. On the other hand, for the classes without noisy labels, $g(r|y)$ remains at 1. As a result, our label corrector consistently outputs the correct labels for both noisy and clean samples across different classes.

We also visualize the learned label weights of the real-world datasets miniWebvision and Animal-10n in Figs. 5 and 6, respectively. For miniWebvision, we select 2 categories at intervals of 25 to visualize their learned label weights considering the large number of categories. As shown in Fig. 5, the learned label weights vary with different classes, suggesting that the noise rates of different classes are different, which is consistent with real-world datasets. In Animal-10N, two categories were selected at intervals of 5 to visualize the learned label weights. As depicted in Fig. 6, the learned label weights remain 1 and drop to 0 at around bin 92 (red dotted line), indicating that the noise rate estimated by the label corrector is about 8%, which is consistent with the well-recognized estimated noise rate on Animal-10N [19].

Corrected Label Accuracy: Fig. 7 illustrates the accuracy of the corrected labels on balanced CIFAR-10-N, measured as the proportion of samples with ground-truth labels after the label correction process. The label accuracy gradually increases as the training progresses, and the classifier becomes more reliable. Eventually, the label accuracy reaches over 90% on CIFAR-10-N with noise rates of 0.2 and 0.4. Remarkably, the accuracy has also significantly improved by 35% under a noise rate of 0.6. The high accuracy of the corrected labels validates our design choices.

| Methods                      | Epoch | RN-10 | RN-50 | RN-152 | RNeXt-50 |
|------------------------------|-------|-------|-------|--------|----------|
| Cross Entropy                | 90    | 34.01 | 44.60 | 46.20  | 42.78    |
| Focal Loss [23]              | 90    | 32.64 | 41.61 | 44.36  | 41.58    |
| LDAM-DRW [5]                 | 90    | 36.03 | 48.80 | 51.83  | 51.43    |
| CDDB-S [57]                  | 90    | 37.70 | 41.80 | 46.40  | 45.10    |
| Logit adjustment [53]        | 90    | 38.43 | 48.89 | 47.86  | 51.85    |
| Balanced-Softmax[14]         | 90    | 38.21 | 50.96 | 53.93  | 51.73    |
| MSLAS [52]                   | 400   | 44.36 | 53.05 | 48.82  | 51.88    |
| PaCo* [55]                   | 400   | 42.89 | 57.00 | 58.55  | 58.20    |

* Indicates the use of self-supervised techniques. RN: ResNet. RNeXt: ResNeXt.
Learnable Weight Varied Trend in Training: Fig. 8 illustrates the evolution of the learnable weight across training epochs. At the beginning, the label corrector mainly relies on the given labels to train the classifier, while gradually shifting towards trusting the predicted labels for samples with higher rank bins, as indicated by the decreasing values of the learnable weight at epoch 45. Furthermore, the label corrector accurately estimates the noise rate to be approximately 35% (for a noise rate of 40%, there are actually 35% noisy samples). The plot, together with Fig. 7, reveals that the epoch at which the label corrector begins to trust the classifier is delayed as the noise rate increases. This suggests that the label corrector considers the classifier to require more training epochs to produce more reliable predicted labels as the noise rate of the training set increases. Therefore, our results demonstrate that the label corrector is capable of dynamically adjusting the relabelling of noisy labels based on the status of the classifier and the training set.

B. Margin Generator

Behavior of Margin Generator: We conducted an analysis of the behavior of the margin generator on clean CIFAR-10-LT with imbalance factor $\rho = 20$. The left subfigure of Fig. 9 shows the generated margins for different classes. We observed that as the class index increases and the sample number decreases, the learned margin also decreases as expected. This suggests that the margin generator has the ability to automatically discern the sample numbers of different classes and adaptively adjust the margins for each class accordingly.

Interestingly, we also observed some irregularly larger margins on class 9 and 10. To explain this observation, we visualized the feature distribution of the meta set using T-SNE [58] as shown in the right subfigure of Fig. 9. We found that the feature distribution of these two classes correspond to the two rightmost clusters, indicating that they are easier to be distinguished from the other classes. This evidence supports the claim that the margin generator takes into account not only the sample number but also the classification difficulty of each class when generating comprehensively adaptive margins during classifier training.

from two perspectives. First, this finding supports our assumption that the classifier primarily focuses on fitting in the dominant clean samples, and can transfer the acquired knowledge to the noisy samples for predicting their ground-truth labels. Second, the label corrector can precisely identify the noisy samples and rectify their labels with the predicted correct ones.
Learnable Weight Varies With Imbalance Ratios: Fig. 12 depicts the learned margins generated by the margin generator under different imbalance ratios. We can observe that as the class index increases, corresponding to decreasing sample size, the generated margin consistently decreases, irrespective of the varying imbalance ratios. Additionally, the variation in learned margins across different classes tends to increase as the imbalance ratio becomes more severe. Both quantitative and qualitative analyses provide evidence that the margin generator effectively respects and adapts to various class distributions by automatically learning to assign appropriate margins.

Learnable Weight Varies in Training: In Fig. 13, the variations of classification margins and feature distributions of meta set are depicted across different training epochs. It can be inferred that the margin generator continually adapts the classification margins based on the feature distributions of meta set during the training process. For instance, in the case of class 10, a relatively small classification margin is assigned to it since it is challenging to recognize at epoch 50. However, as it becomes easier to recognize at epoch 250, the margin generator assigns a larger classification margin to it. This observation supports the claim that the margin generator can dynamically regulate the classification margin based on the classification difficulty.

We present visualizations of the learned label weights for the real-world miniWebvision and simulated ImageNet-LT datasets in Figs. 10 and 11, respectively. In Fig. 10, the generated margins over different classes in miniWebvision accord with the complex variation of sample size, which demonstrates the adaptability of our method to handle complex real-world biased data. For the simulated ImageNet-LT dataset, we visualize the learned margins for 333 categories in intervals of 3 in Fig. 11. As shown in the figure, the learned margins consistently vary with the sample number of different classes, suggesting that the margin generator can generate proper margins for different classes. This finding provides evidence that our method can handle datasets with numerous categories.
C. Behavior of Hierarchical Sampling.

We examine the efficacy of our hierarchical sampling strategy in enhancing the construction of meta sets. Fig. 14 visually presents the feature distribution of the meta set generated using our hierarchical sampling method (left) and the meta set generated using a naive sampling method (right). The feature distribution of samples selected using hierarchical sampling demonstrates greater dispersion within distinct clusters compared to naive sampling. This indicates that the creation of a primary set prior to random sampling enables a more diverse meta data selection, encompassing both easy and challenging samples. As a result, this approach helps to mitigate biased learning on easier samples.

VIII. ABLATION STUDIES.

Effect of Label Corrector: We construct a model variant wherein the label corrector component is removed to assess its efficacy. Table IX(a) illustrates that the average final accuracy of this variant decreases by 8.22\% compared to the complete dynamic loss configuration. This notable decrease in performance provides compelling evidence of the effectiveness of the label corrector.

Effect of Margin Generator: As indicated in Table IX(a), to assess the efficacy of the margin generator, we initially utilize only the label corrector, which yields an average final accuracy of 71.08\%. Subsequently, we incorporate Balanced-Softmax to address class imbalance, which only results in a marginal improvement of 1.56\% in accuracy. Finally, by implementing our margin generator, we are able to further improve the accuracy to 79.73\%. These results serve as evidence supporting the importance of utilizing a dynamic margin in managing long-tailed noisy data.

Effect of Hierarchical Sampling: Table IX(b) demonstrates that the substitution of hierarchical sampling with naive random sampling leads to a reduction of up to 1.20\% in average final accuracy. This observation suggests that the meta set created using hierarchical sampling possesses a more comparable distribution to the test set.

Effect of Class-Specific Label Corrector: To validate the class-specific design of our label corrector, we constructed a class-agnostic variant and evaluated its performance on CIFAR-10-N, a dataset containing 40\% asymmetric noise with varying noise rates across different classes. Our results demonstrate that the class-specific label corrector significantly outperforms its class-agnostic counterpart by a large margin of 3.95\% in terms of accuracy (94.51\% versus 90.56\%), thus providing clear evidence for the effectiveness of our class-specific design.

Effect of Explicit Utilizing Given and Predicted labels: To assess the impact of explicitly utilizing known information, particularly given labels and predicted labels, we conducted a comparative analysis between the implicit and explicit approaches in utilizing such information. The results, as presented in Table IX(c), demonstrate that the explicit utilization of known information resulted in a remarkable 22.52\% increase in final accuracy and a substantial 6.13\% improvement in best accuracy compared to the implicit approach. These findings indicate that simplifying the task of the meta net through the explicit utilization of known information can significantly enhance the performance of the classifier. Furthermore, the notable discrepancy
between the final accuracy and the best accuracy underscores the contribution of explicit utilization of known information in facilitating the convergence of the meta net.

Effect of Group-Wise Optimization: To validate the efficacy of the group optimization approach employed in our method, we conducted experiments comparing it with sample-wise optimization. The results, presented in Table IX(d), reveal that utilizing a sample-wise approach with both the label corrector and margin generator hindered classifier convergence, leading to significantly lower final accuracy compared to the best achieved accuracy. Similarly, employing the margin generator in a group-wise manner alone failed to yield convergence. However, when extending the group-wise approach to the label corrector, a final accuracy of 79.73% was attained. These findings highlight the significance of optimizing the meta net task, as it enhances learning efficiency and improves the classifier’s robustness.

Effect of Meta Net Architecture: In order to validate the architecture design of the meta net, we simplified the label corrector and margin generator by replacing them with a $R$-length and $C$-length learnable vector, respectively. As shown in Table IX(f), this modification resulted in a noticeable performance drop of 0.23%. Our experimental observation suggests that this is due to the fact that MLPs are able to quickly learn appropriate label weights and per-class margins, whereas the learnable vectors suffer from slow convergence.

The average accuracy is reported to evaluate model performance. BS represents balanced softmax.

### TABLE IX

**Ablation Studies Conducted on CIFAR-10-N-LT, Considering Varying Noise Rates (0.1-0.5) and Imbalance Ratios (10, 100)**

| (a) Effect of key components. | (b) Ablations on sampling strategy. | (c) Ablation on utilization of inputs. |
|------------------------------|------------------------------------|--------------------------------------|
| $G_{l_c}$ | $G_{l_m}$ | BS | Last | Best | Sampling | Last | Best | Methods | Last | Best |
| ✗ | ✗ | ✗ | 52.73 | 60.43 | Naive | 78.53 | 79.58 | Cross Entropy | 52.73 | 60.43 |
| ✗ | ✗ | ✓ | 71.51 | 76.84 | Hierarchical | 79.73 | 80.55 | Implicit | 57.21 | 74.42 |
| ✓ | ✗ | ✗ | 71.08 | 77.11 | ✓ | 72.64 | 74.96 | Explicit | 79.73 | 80.55 |
| ✓ | ✓ | ✓ | 79.73 | 80.55 | ✓ | 72.64 | 74.96 | |

(d) Ablation on optimization approach.

| $G_{l_c}$ | $G_{l_m}$ | Last | Best |
|-----------|-----------|------|------|
| ✗ | ✗ | 52.73 | 60.43 |
| Sample-wise | Sample-wise | 44.63 | 74.42 |
| Sample-wise | Group-wise | 40.76 | 68.08 |
| Group-wise | Group-wise | 79.73 | 80.55 |

(e) Ablation on noisy label identifier.

| Loss rank | Uncertainty | Last | Best |
|-----------|-------------|------|------|
| ✗ | ✗ | 52.73 | 60.43 |
| ✗ | ✓ | 72.91 | 77.18 |
| ✓ | ✗ | 79.73 | 80.55 |
| ✓ | ✓ | 80.62 | 81.23 |

(f) Ablations on meta net architecture.

### TABLE X

**Accuracy (%) on CIFAR10-N-LT With Varying Imbalance Ratio**

| Noise Rate | 0.2 | 0.4 | Avg. |
|------------|-----|-----|------|
| Imbalance Ratio | 10 | 50 | 200 | 10 | 50 | 200 | 54.22 |
| Cross Entropy | 78.03 | 65.53 | 43.06 | 63.04 | 47.56 | 29.09 | 54.22 |
| HAR-DRW [8] | 88.81 | 82.74 | 73.98 | 84.03 | 75.36 | 63.95 | 78.15 |
| FSR [59] | 85.70 | 77.40 | 65.30 | 81.60 | 69.80 | 49.50 | 71.58 |
| INOLML [40] | 90.10 | 80.10 | 66.60 | 89.10 | 78.10 | 61.60 | 77.60 |
| Dynamic Loss | 92.40 | 83.56 | 76.72 | 90.46 | 78.54 | 66.47 | 81.36 |

WRN-28-10 used as the classifier.

### IX. CONCLUSION

This work introduces a novel dynamic loss for robust learning from long-tailed data with noisy labels. The dynamic loss consists of a learnable label corrector and margin generator,
which jointly correct noisy labels and adjust classification margins to guide classifier learning. The meta net and classifier are co-optimized through meta-learning using a hierarchical sampling strategy that provides unbiased yet diverse meta data. Extensive evaluations on both synthetic and real-world data demonstrate the effectiveness of our dynamic loss, which exhibits high adaptability and robustness to various types of data biases.

REFERENCES

[1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Representations, 2015, pp. 211–252.
[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[3] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, “RepVGG: Making VGG-style ConvNets great again,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 13738–13747.
[4] W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool, “Webvision database: Visual learning and understanding from web data,” 2017, arXiv:1708.01826.
[5] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, “Learning imbalanced datasets with label-distribution-aware margin loss,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2020, pp. 1567–1578.
[6] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, “MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 2304–2313.
[7] S. Jiang, J. Li, Y. Wang, B. Huang, Z. Zhang, and T. Xu, “Delving into sample loss curve to embrace noisy and imbalanced data,” in Proc. Conf. Assoc. Adv. Intell. Artif., 2022, pp. 7024–7032.
[8] K. Cao, Y. Chen, J. Liu, N. Arechiga, A. Gaidon, and T. Ma, “Heteroskedastic and imbalanced deep learning with adaptive regularization,” in Proc. Int. Conf. Learn. Representations, 2020, pp. 1–16.
[9] J. Shu et al., “Meta-Weight-Net: Learning an explicit mapping for sample weighting,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2019, pp. 1919–1930.
[10] B. Kang et al., “Decoupling representation and classifier for large-tailed recognition,” in Proc. Int. Conf. Learn. Representations, 2019, pp. 1–16.
[11] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, “Large-scale long-tailed recognition in an open world,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2537–2546.
[12] G. Zheng, A. H. Awadallah, and S. Dumas, “Meta label correction for noisy learning,” in Proc. Conf. Assoc. Adv. Intell. Artif., 2021, pp. 11053–11061.
[13] J. Li, R. Socher, and S. C. Hoi, “DivideMix: Learning with noisy labels as semi-supervised learning,” in Proc. Int. Conf. Learn. Representations, 2020, pp. 1–14.
[14] J. Ren et al., “Balanced meta-softmax for long-tailed visual recognition,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2020, pp. 4175–4186.
[15] Y. Xu, L. Zhu, L. Jiang, and Y. Yang, “Fast meta update strategy for face recognition and attribute prediction,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 11, pp. 2781–2794, Nov. 2020.
[16] J. Tan et al., “Equalization loss for long-tailed object recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 11659–11668.
[17] M. P. Kumar, B. Packer, and D. Koller, “Self-paced learning for latent variable models,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2010, pp. 1149–1157.
[18] J. Huang, L. Qu, R. Jia, and B. Zhao, “O2U-Net: A simple noisy label detection approach for deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3325–3333.
[19] S. Jiang et al., “MetaSeg: Content-aware meta-net for semi-supervised semantic segmentation,” IEEE Trans. Neural Netw. Learn. Syst., early access, Apr. 10, 2023, doi: 10.1109/TNNLS.2023.3263335.
[20] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, “Training deep neural networks on noisy labels with bootstrapping,” in Proc. Int. Conf. Learn. Representations (Workshop), 2015, pp. 1–11.
[21] W. Chen, C. Zhu, and Y. Chen, “Sample prior guided robust model learning to suppress noisy labels,” 2022, arXiv:2112.01197.
[22] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel, “MixMatch: A holistic approach to semi-supervised learning,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2019, pp. 5049–5059.
[23] T. Wei, J.-X. Shi, W.-W. Tu, and Y.-F. Li, “Robust long-tailed learning under label noise,” 2021, arXiv:2108.11569.
[24] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, “Early-learning regularization prevents memorization of noisy labels,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2020, pp. 10331–10342.
[25] 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 Conf. Comput. Vis. Pattern Recognit., 2021, pp. 6602–6611.
[26] J. Tan et al., “Deep learning against label noise on imbalanced datasets,” in Proc. Int. Conf. Learn. Representations, 2017, pp. 1–16.
[27] J. Shu, X. Yuan, D. Meng, and Z. Xu, “CMW-Net: Learning a class-aware sample weighting mapping for robust deep learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 10, pp. 11521–11539, 2023.
[28] A. Iscen, J. Valmadre, A. Arnb, and C. Schmid, “Learning with neighbor consistency for noisy labels,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 4662–4671.
[29] D. A. Hoang et al., “Maximizing the utility of validation sets for imbalanced noisy label meta-learning,” 2022, arXiv:2208.08132.
[30] M. Chen, H. Cheng, Y. Du, M. Xu, W. Jiang, and C. Wang, “Two wrongs don’t make a right: Combating confirmation bias in learning with label noise,” in Proc. AAAI Conf. Artif. Intell., 2023, pp. 14765–14773.
[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-V4, Inception-ResNet and the impact of residual connections on learning,” in Proc. Conf. Assoc. Adv. Intell. Artif., 2017, pp. 4278–4284.
[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.
[33] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 8748–8763.
[34] Y. Cheng, W. Zhang, P. M. Goswami, and C. Chen, “Learning with feature-dependent label noise: A progressive approach,” in Proc. Int. Conf. Learn. Representations, 2021, pp. 1–13.
[35] Y. Chen, S. X. Hu, X. Chen, S. Ai, and J. A. K. Suykens, “Compressing features for learning with noisy labels,” IEEE Trans. Neural Netw. Learn. Syst., early access, Jul. 08, 2022, doi: 10.1109/TNNLS.2022.3189030.
[36] B. Huang et al., “Co-teaching: Robust training of deep neural networks with extremely noisy labels,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2018, pp. 8536–8546.
[37] E. Englesson and H. Azizpour, “Generalized Jensen-Shannon divergence loss for learning with noisy labels,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2021, pp. 30284–30293.
[38] Y. Cui, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 630–645.
[39] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in convolutional neural networks,” in Proc. Conf. Neural Inf. Process. Syst., 2017, pp. 2999–3007.
[51] S. Alshammari, Y.-X. Wang, D. Ramanan, and S. Kong, “Long-tailed recognition via weight balancing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 6887–6897.

[52] Z. Zhong, J. Cui, S. Liu, and J. Jia, “Improving calibration for long-tailed recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 16484–16493.

[53] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, “Long-tail learning via logit adjustment,” in Proc. Int. Conf. Learn. Representations, 2021, pp. 1–24.

[54] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi, “The majority can help the minority: Context-rich minority oversampling for long-tailed classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 6887–6896.

[55] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia, “Parametric contrastive learning,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 695–704.

[56] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5987–5995.

[57] S. Sinha, H. Ohashi, and K. Nakamura, “Class-difficulty based methods for long-tailed visual recognition,” Int. J. Comput. Vis., vol. 130, no. 10, pp. 2517–2531, 2022.

[58] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, no. 11, pp. 2579–2605, 2008.

[59] Z. Zhang and T. Pfister, “Learning fast sample re-weighting without reward data,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 725–734.

[60] S. Zagoruyko and N. Komodakis, “Wide residual networks,” in Proc. Brit. Mach. Vis. Conf., 2016, pp. 1–13.

Shenwang Jiang received the BE and PhD degrees from the School of Optoelectronics, Beijing Institute of Technology, China, in 2014 and 2020, respectively. He is currently a postdoc researcher with the School of Optoelectronics, Beijing Institute of Technology, China. His research interests include image classification, object detection, and image processing.

Jianan Li received the BS and PhD degrees from the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China, in 2013 and 2019, respectively. He is currently an assistant professor with the School of Optoelectronics, Beijing Institute of Technology, China. He is mainly working on image classification and object detection.

Jizhou Zhang received the PhD degree from the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China, in 2020. He is currently working toward the postdoctoral degree with the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China. His research interests include computer vision, hyperspectral imaging, computational imaging and machine learning.

Ying Wang received the BE degree in optical engineering from the Beijing Institute of Technology, China, in 2019. She is currently working toward the PhD degree with the Beijing Institute of Technology, China. Her research interests include single object tracking and object detection.

Tingfa Xu received the PhD degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Changchun, China, in 2004. He is currently a professor with the School of Optoelectronics, Beijing Institute of Technology, Beijing, China. His research interests mainly include computer vision and optoelectronic imaging.