Weak Teacher Helps Long-Tail Learning with High-Noisy Labels

Abstract

Real-world data often exhibit a long-tailed distribution with numerous noisy labels, which significantly degrades the performance of deep models. Training on data with a high noise ratio presents a particular challenge. While previous studies have made progress in dealing with this practical joint problem, they ignore the relatively severe label-image matching bias in high-noise environments, resulting in suboptimal performance. Given that observed labels, despite mismatching the images, still contain category information, we propose leveraging auxiliary text information from the labels to correct label-image inconsistencies in the long-tailed noisy labeled data. In particular, we leverage the text encoder in pre-trained visual-language models to obtain text-based predictions, using this text-image alignment prior to correct the label-image inconsistencies. This supervisory signal, referred to as weak teacher supervision (WTS), is not always accurate. Therefore, we evaluate the discrepancy between the text-predicted and observed labels to decide when to activate WTS. By calibrating the bias between learned features and labels, WTS ultimately enhances model robustness. Extensive experiments on benchmark datasets demonstrate that the proposed method achieves significant performance gains on both simulated and real-world datasets, especially under high-noise scenarios. The source code is temperately available at https://anonymous.4open.science/r/WTS-0F3C.

Keywords: long-tail, noisy label, high noise, visual-language.

1 Introduction

With the availability of large-scale public datasets [44, 47, 51], significant progress has been made in the field of computer vision [28, 40] and large models [11, 44]. However, real-world datasets tend to suffer from several issues, particularly class imbalance, where head classes contain most samples while tail classes are significantly underrepresented [65], and image mislabeling [50], referred to as noisy labels [59]. Creating balanced datasets with correctly labeled classes to address these challenges is prohibitively expensive and unsustainable. To address these issues, the practical problem of long-tailed noisy label (LTNL) learning [38, 63] has been introduced. Recently, the challenging task of LTNL learning has garnered significant attention. Broadly, the approaches can be divided into the following categories: (1) emphasizing the importance of different samples by reweighting or regularization [3, 17, 46, 49, 53]; (2) selecting clean samples based on carefully designed criteria [24, 38, 54, 57]; (3) developing improved representation learning methods [33, 62, 63, 68]. The aforementioned methods can effectively enhance the robustness of models on long-tailed noisy labeled data. However, they overlook the impact of different noise ratios on model training, resulting in suboptimal performance in high-noise scenarios. Notably, low levels of label noise exert a relatively minimal impact on model performance. Therefore, targeted processing is necessary for high noise ratio scenarios, where unreliable labels constitute one of the primary issues in noisy label learning with long-tailed data. In such circumstances, noisy labels introduce substantial misleading supervisory signals, making it challenging to effectively distinguish noisy samples from clean ones or improve feature representation. This results in accumulated feature learning biases and amplifies the combined challenges of label noise and class imbalance. To this end, we propose leveraging auxiliary linguistic information to assist in calibrating the supervisory signals. Considering that in the long-tailed noisy labeled data, the observed labels contain category information but may be inconsistent with the corresponding images, we propose using auxiliary text information from the observed labels to correct these inconsistencies, thereby fully utilizing the label information. Specifically, we leverage the text encoder from pre-trained visual-language models (VLMs) [20, 44] to obtain text-based predictions, utilizing this text-image alignment prior to correct label-image inconsistencies. This text-image alignment prior, serving as a supervisory signal, is not always accurate. Therefore, we evaluate the discrepancy between the text-predicted labels from the text encoder and observed labels to decide whether to activate this supervision. If the predicted labels from the pre-trained text encoder deviate significantly from the observed labels, we consider this text-based predictions to be more informative and incorporate them to guide model training. This approach enables the effective application of existing long-tailed learning methods. Since text-predicted labels generally have lower accuracy than direct fine-tuning of the image encoder, we regard the text encoder as a "weak teacher" and refer to our approach as weak teacher supervision (WTS). Experiments on benchmarks with multiple types of noisy labels and intrinsically long-tailed distributions demonstrate that the proposed WTS improves the performance of the strong student, particularly in scenarios with a high noise ratio.

2 Related works

2.1 Long-tail Learning

Long-tail learning methods typically assume correct labeling within datasets [10]. These methods then apply class-wise operation, generally falling into three main categories [28, 48]. (1) Input level. Data manipulation techniques, such as re-weighting/sampling [10] and data augmentation [7, 8], are implemented to enhance classification performance. (2) Representation level. Modifications are made to the model structure to better capture the underlying characteristics of the data. Decoupling representation [18, 66] and BBN-based methods [64, 67] separate representation learning from classifier training. These methods first extract representations from the original long-tailed dataset, and then retrain the classifier using either class-balanced sampling data [18] or reverse sampling data [67]. Ensembling learning includes redundant ensembling [2, 23, 25, 52], which aggregates outputs from separate classifiers or networks within a multi-expert framework, and complementary ensembling [9, 67], which involves the statistical selection of different data partitions. (3) Output level. Existing methods enhance model representation and refine the classifier by calibrating the model logits based on specific criteria. For example, logit adjustment techniques [39, 45] calibrate the predicted output distribution to achieve a balanced distribution. Re-margining methods [4, 29, 30, 39] introduce class size-based constants that assign larger margins to tail classes compared to head classes.

2.2 Label-noise Learning

Training a model using a dataset with a large number of noisy labels inevitably suffers from mis-supervision of noisy labels, which in turn significantly reduces the recognition performance of the model. A straightforward and effective approach to this problem is to distinguish between clean and noisy samples, with many methods such as MentorNet [37], Co-teaching [12] and DivideMix [26] treating samples with small training losses as clean samples, while AUM [43] identifies noise samples by measuring the average difference between the logarithmic value of a sample's specified category and the highest logarithmic value of a non-specified category, Jo-SRC [60] and UNICON [19] use Jensen-Shannon divergence for sample selection. In addition, some methods design noise-robust loss functions to mitigate the influence of noisy data, such as backward and forward loss correction [42], gold loss correction [13], MW-Net [49] and Dual-T [61]. Other methods evaluating the noise transfer matrix [6, 58, 61] or reweighting examples for noisy label learning [35, 46].

2.3 Noisy Label Learning on Long-Tailed Data

Numerous studies have emerged to address the challenges posed by the task of joining noisy labels and unbalanced/long-tailed data. A common strategy is to distinguish between clean and noisy samples. For example, CNLCU [57] enhances the small loss method [12] by designating a subset of samples with large losses as clean. TABASCO [38] addresses a complex scenario where noisy labels may cause an intrinsic tail class to appear as a head class, proposing a bi-dimensional separation metric to adapt to different cases. Another promising path is to emphasize the importance of different samples by reweighting or regularization [3, 17, 46, 49, 53]. For example, HAR [3] introduces an adaptive regularization approach that jointly addresses noise and imbalance by assigning larger regularization to examples with high uncertainty and low density. Another line of work focuses on improving representation learning [33, 62, 63, 68]. For example, RCAL [63] uses unsupervised contrastive learning to produce noise-resistant representations, which are then used to restore class distributions and enable balanced sampling to mitigate long-tailed effects. ECBS [33] develops a pseudo-labeling method using class prototypes and distribution matching, leveraging optimal transport to address noise and imbalance. In high-noise environments, the aforementioned methods fall short because noisy labels undermine sample reliability and obscure distinctions between noisy and tail-class samples. Moreover, label noise distorts feature space, complicating the use of feature-based strategies to address both noise and class imbalance.

3 Method

Noisy labels weaken the reliability of the supervision signal from observed labels, especially at high noise ratios, leading to accumulated biases in feature learning and exacerbating the challenges of label noise and class imbalance. Fortunately, observed labels still provide category names and counts, making external label support valuable. Recent advancements in visual-language models (VLMs) [20, 44] provide a powerful tool for incorporating label information. To achieve this process, we introduce prediction probabilities from the text encoder of VLM as auxiliary text supervision during model training, referred to as weak teacher supervision (WTS). Since WTS may introduce additional errors, we use a switch to determine whether to activate it based on the overlap ratio between observed and text-predicted labels. The overall structure of the WTS is illustrated in Figure 1.

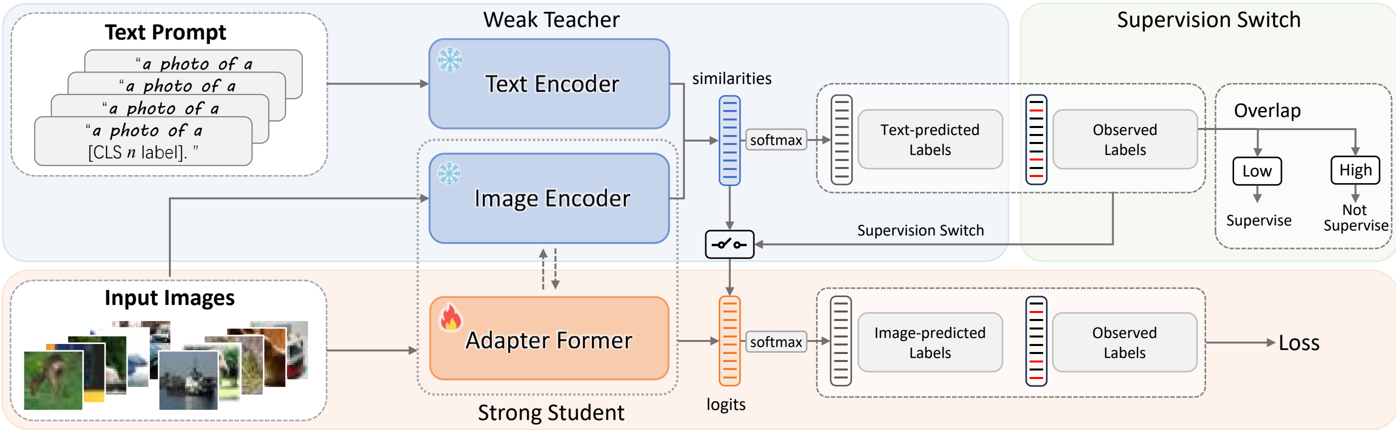

Figure 1. Overview of our proposed WTS. We leverage the text encoder in pre-trained visual-language models to obtain text-based predictions, using text-image alignment to correct label-image inconsistencies. Since this supervisory signal is not always accurate, we evaluate the discrepancy between the text-predicted and observed labels to determine when to activate it.

3.1 Preliminaries

Problem Definition. Consider a training set $\mathcal{D} = \{(x_i, \hat{y}_i)\}_{i=1}^{N}$, where each $(x_i, \hat{y}_i)$ pair represents an input and its observed label, and N is the total number of samples in $\mathcal{D}$. Suppose $\mathcal{D}$ includes C classes, with class c having n_c training samples. Then, the total number of samples is given by $N = \sum_{c=1}^{C} n_c$. The training set $\mathcal{D}$ exhibits the following properties:

1. Noisy labels. The observed label $\hat{y}_i \in \hat{\mathcal{Y}}$ may be different from the ground truth $y_i \in \mathcal{Y}$. And $\mathcal{Y}$ is unavailable.

2. Long-tailed distribution. Without loss of generality, we arrange the classes in descending order by training sample count, so that $n_1 > n_2 > \ldots > n_C$ with $n_1 \gg n_C$.

This learning task is defined as long-tailed noisy label (LTNL) learning [38,63]. Property 1 results in a distribution derived from $\hat{\mathcal{Y}}$ that is inconsistent with $\mathcal{Y}$. Existing long-tailed learning methods typically rely on precise sample counts for each class to effectively adjust logits and/or select suitable structures. As a result, these methods are inadequate for addressing Property 2, as inaccuracies in category counts can cause over-regularization in certain classes.

Basic Notation. In the following sections, scalars are represented by lowercase letters, while vectors are denoted by lowercase boldface letters. Sets or distributions are represented by uppercase script letters. The superscripts I[1], t and o are used to differentiate the outputs obtained from the fine-tuned image encoder, the pre-trained text encoder, and the observed labels, respectively.

3.2 Weak Teacher Supervision

Supervision from Linguistic Information. The text and image encoders in a pre-trained VLM can provide text embeddings ($\mathbf{t}_c$) of labels for class c and image embeddings ($\mathbf{f}_i$) for input images x_i. By comparing the

[1]To avoid confusion with the index (subscript i), the output of the fine-tuned image encoder is represented by the capital letter I.

similarity between t_c and f_i, we can derive the predicted label from the label-based text prompt:

$$s_{i,c} = \frac{\mathbf{t}_c^T \mathbf{f}_i}{\|\mathbf{t}_c\| \|\mathbf{f}_i\|}, \; y_i^t = \underset{c \in [C]}{\arg\max} \{s_{i,c}\}_{c=1}^{C}, \tag{1}$$

where y_i^t is the text-predicted label for input image x_i. Subsequently, a straightforward solution is to integrate the VLM text-predicted label (TL) with the observed label (OL) into the training to provide auxiliary supervision:

$$\mathcal{L} = a \underbrace{\mathcal{L}_O\left(x_i, \hat{y}_i\right)}_{\text{OL supervision}} + (1-a) \underbrace{\mathcal{L}_T\left(x_i, y_i^t\right)}_{\text{TL supervision}}, \tag{2}$$

where a is a hyper-parameter. $\mathcal{L}_O$ represents the loss calculated on observed labels, and can employ cross-entropy (CE) loss or existing logit adjustment methods, such as LDAM [4], LA [39], and LADE [14], for long-tailed learning. Meanwhile, $\mathcal{L}_T$ is the loss based on text-predicted labels. Eq.(2) can be utilized to fine-tune a VLM. However, the text-predicted labels $\mathcal{Y}^t = \{y_i^t\}_{i=1}^N$ also imprecise, for instance, its accuracy on the test set of CIFAR-100 is only 64.4%, indicating that $\mathcal{L}_T\left(x_i, y_{T,i}\right)$ introduces another form of label noise. To address this dilemma, we propose leveraging feature similarity between text and image to provide additional supervision. Since it can provide not only discrete one-hot optimization objectives but also insights into inter-class relationships. In detail, we utilize softmax to convert the similarity between the label text features and the input images (as shown in Eq.(1)) into probabilities:

$$p^t(x_i | y = c) = \frac{s_{i,c}}{\sum_{j=1}^{C} s_{i,j}}. \tag{3}$$

For convenience and without loss of generality, for the input x_i, we abbreviate this as p_c^t. Similarly, the probability p_c^I for the input image can be derived from the similarity between the fine-tuned image features and the classifier weights. Following, we can incorporate the text supervision information from the pre-trained VLM into the training process by minimizing the divergence between image and text prediction probability distributions. Kullback-Leibler Divergence (KL) is employed in this paper:

$$\mathcal{L}_T = \mathrm{KL}(\mathcal{P}^t \| \mathcal{P}^I), \tag{4}$$

where $\mathcal{P}^t = \{p_c^t\}_{c=1}^C$ and $\mathcal{P}^I = \{p_c^I\}_{c=1}^C$ are the probability distributions of text-prediction and fine-tuned image encoder prediction, respectively. Since the fine-tuned model has fewer training parameters and has the ability to quickly adapt to new datasets, we treat the image encoder that fine-tuned on LTNL data as a strong student, while the pre-trained VLM serves as a weak teacher and provides weak teacher supervision (WTS).

Supervision Switch Control. Since the weak teacher is not always accurate, and as discussed in Sec. **??**, observed labels can be used directly in low-noise scenarios without additional processing. The challenge, however, lies in the fact that the proportion of noisy labels cannot be determined in advance. Therefore, an indicator is needed to assess when the teacher provides effective supervision. The overlap ratio $OR = \frac{|\mathcal{Y}^t \cap \hat{\mathcal{Y}}|}{|\hat{\mathcal{Y}}|}$ between text-predicted labels and observed labels can serve as this indicator. When the overlap ratio OR is high, it indicates that the two types of labels are largely consistent, and the information provided by WTS is limited. Therefore, we opt to deactivate it. In contrast, when OR is low, the observed labels and the visual-language alignment prior are significantly different, indicating a need for auxiliary supervision. Then, the supervision switch based on OR can be calculated as:

$$a = \begin{cases} 1 & \text{if } OR \geq \tau \\ a \sim \text{Beta}(\alpha, \beta) & \text{if } OR < \tau \end{cases}, \tag{5}$$

where τ is the overlap ratio control threshold, which we will empirically analyze in detail in Sec. **??**. We estimate this value in an online manner by calculating the overlap rate between the two types of labels in each batch. When WTS needs to be turned off, $a = 1$, and only $\mathcal{L}_O$ is included in Eq.(2). Conversely, when WTS is turned on, we set a to a random number sampled from the beta distribution. In this paper, we choose Adapterformer [5] for VLM fine-tuning.

Remarks: The advantages of WTS can be summarized in three main aspects: (1) Noise-resilient feature calibration. In scenarios with high noise ratios, WTS shifts its focus from distinguishing noisy samples to addressing feature misalignments caused by noisy labels. This strategy helps mitigate potential classification errors that can severely affect tail classes and improves the performance of all classes, including both head and tail. In low-noise scenarios, WTS automatically deactivates supervision that could introduce errors, relying exclusively on observation labels, which are considered relatively more reliable. (2) Noise-robust bias corrector. The pre-trained VLM, which is unaffected by noisy labels and inherently aligned with visual-language features, serves as the teacher model to correct biases in the features obtained by the student model. Notably, we observe that fine-tuning with $\mathcal{L}_O$ can sometimes outperform text-prediction, which often has lower accuracy. Nevertheless, WTS still provides valuable guidance in correcting misalignments introduced by noisy labels. A detailed rationale for this aspect will be provided in Sec. 3.3. (3) Highly efficient training. WTS introduces minimal computational overhead, with the primary cost arising from the fine-tuning of the image encoder.

3.3 Rationale behind Effectiveness of WTS

Although the proposed WTS may seem intuitive and straightforward at first glance, it is built on a solid theoretical foundation. In this section, we explore the underlying theoretical rationale behind WTS. Its effectiveness is analyzed from the perspectives of noisy label learning and long-tail learning.

Effectiveness on Noisy Label Learning. We investigate the impact of WTS on noisy label learning by analyzing how $\mathcal{L}_T$ revise incorrectly observed labels, leading to the following proposition.

Proposition 1. *WTS corrects observed labels based on the predicted probabilities provided by the pre-trained VLM with the ratio a.*

Proof. Another form to write Eq.(4) is:

$$\text{KL}(\mathcal{P}^t \| \mathcal{P}^I) = -\sum_{c=1}^{C} p_c^t \log p_c^I - \left(-\sum p_c^t \log p_c^t\right) \tag{6}$$

$$= -\sum_{c=1}^{C} p_c^t \log p_c^I + H(\mathcal{P}^t), \tag{7}$$

where $H(\cdot)$ represents cross entropy. $\mathcal{P}^t$ is provided by the pre-trained VLM. Since the VLM parameters are not updated during training, $H(\mathcal{P}^t)$ can be considered a constant throughout the training process. Therefore, this term can be ignored when optimizing the loss function. Without loss of generality, in Eq.(2), by substituting $\mathcal{L}_T$ with Eq.(7) and replacing $\mathcal{L}_O$ with the expanded form of the CE-based loss, we can obtain:

$$\mathcal{L} = a\left(-\sum_{c=1}^{C} p_c^o \log p_c^I\right) + (1-a)\left(-\sum_{c=1}^{C} p_c^t \log p_c^I\right), \tag{8}$$

$$= -\sum_{c=1}^{C}\left(a \cdot p_c^o + (1-a)\cdot p_c^t\right)\cdot \log p_c^I, \tag{9}$$

where p_c^o is the one-hot-form probability obtained based on the observed labels. □

Proposition 1 shows that WTS has the following impact on noisy labels:

- for $\hat{y}_i = y_i$, WTS modifies the observed labels through label smoothing, enabling the preservation of all inter-class relationships, in contrast to relying solely on $\mathcal{Y}^t$;
- for $\hat{y}_i \neq y_i$, WTS prevents over-confidence arising from prediction errors [26] by decreasing the probability assigned to the incorrect target class.

Effectiveness on Long-Tail Learning. We investigate the influence of WTS on long-tail learning from a gradient-based perspective. Before proceeding, we introduce a theorem, a used symbol, and a remark in our analysis.

Theorem 1. *Let p be the base probability and q be the probability obtained from the softmax function applied to logits $\mathcal{Z} = \{z_i\}_{i=1}^C$ that $q_i = \dfrac{e^{z_i}}{\sum_{j=1}^C e^{z_i}}$. The cross-entropy loss is $\ell = -\sum_{i=1}^C p_i \log q_i$. Then, the derivative of the loss function with respect to the logits z_i is given by:*

$$\frac{\partial \ell}{\partial z_i} = q_i - p_i, \tag{10}$$

where y indicates the target class.

Remark 1. *The gradient of logit adjustment methods reduces the positive signal contributions from head classes while amplifying those from tail classes.*

Detailed proof can be found in Sec. **??** and Sec. **??**. For the notation, we define the modified probability p_c^m for class c as:

$$p_c^m = a \cdot p_c^o + (1-a)\cdot p_c^t. \tag{11}$$

Theorem 1 gives that the derivative of $\mathcal{L}_O$ and $\mathcal{L}$ (Eq.(8)) with respect to the logit for target class are:

$$\frac{\partial \mathcal{L}_O}{\partial z_y} = p_y^I - 1,\ \frac{\partial \mathcal{L}}{\partial z_y} = p_y^I - p_y^m. \tag{12}$$

On the one hand, during optimization, the gradient is updated by descending in the opposite direction of the gradient. Therefore, compared to $\frac{\partial \mathcal{L}_O}{\partial z_y}$, $\frac{\partial \mathcal{L}}{\partial z_y}$ decreases the positive signal for the target class. The extent of this reduction is determined by the text encoder, specifically p_y^m, and is independent of the training set distribution. On the other hand, if $\mathcal{L}_O$ employs the existing logit adjustment method for long-tail learning, according to Remark 1, it facilitates automatic gradient balancing. However, there are errors in the labels. Gradients that are incorrectly labeled tail classes will be erroneously amplified, leading to misleading model gradient descent. WTS reduces the positive signals of all classes based on the text encoder to mitigate the over-amplification of error signals.

4 Implementation details

4.1 Comparing with the released source codes

While the released source codes is targeted at the treatment of long-tail problems, this work has migrated to the long-tailed noisy label problem, and is building on the released source codes by adding modules to deal with the joint long-tailed and noisy label problem.

4.2 Experimental environment setup

We evaluate the proposed WTS on both simulated and real-world noisy long-tailed datasets, following Lu [38] and Zhang [63]. Specifically, synthetic scenarios are created based on CIFAR-10/100 [21], real-world noise is introduced in mini-ImageNet [16] (referred to as red Mini-ImageNet, abbreviated as Img-LTN^r), and WebVision-50 [32] is used with its inherent noisy labels. For all constructed datasets, we first subsample a long-tailed version from the original dataset following the exponential decay pattern from prior works [65], and then introduce label noise. The imbalance factor is defined as the ratio of the largest class size to the smallest. Three types of noise—joint, symmetric, and asymmetric—are applied to CIFAR datasets. To distinguish it from the original data, we appended "-LTN" to the constructed long-tailed noisy label dataset. For model training, we use CLIP [44] as the backbone and Adaptformer [5] as the fine-tuning strategy. The optimizer is SGD with an initial learning rate of 0.01, a momentum of 0.9, and a weight decay of 5×10^{-4}. The batch size is 128. WTS is trained for 10 epochs on all datasets.

4.3 Main contributions

The main contributions of this paper are summarized as follows:

- We propose a novel label calibration strategy that utilizes textual information to revise label-image mismatches for LTNL learning, without requiring additional data. This approach is robust to imbalanced distributions, maximizing the use of available label information.
- We devise a simple yet effective WTS strategy that integrates seamlessly with various existing methods. It leverages text information to predict image labels and, by evaluating the consistency between text-predicted and observed labels, selectively applies supervision to improve label reliability.
- Extensive experiments on both simulated and real-world datasets demonstrate the effectiveness of WTS, showing significant performance gains in LTNL learning, especially in challenging high-noise conditions.

5 Results and analysis

Comparison Methods. We compare our method with the following three types of approaches: (1) *Long-tail (LT) learning methods*: LDAM [4], NCM [18], MiSLAS [36], logit adjustment (LA) [39], and influence-balanced loss (IB) [41]. (2) *Label-noise (LN) learning methods* : Co-teaching [12], CDR [56], Sel-CL [31] DivideMix [26], and UNICON [19]. (3) *Long-tailed noisy label (LTNL) learning*: MW-Net [49], ROLT [54], HAR [62], ULC [15], TABASCO [38], RCAL [63] and ECBS [33].

Table 1. Top-1 acc. (%) on CIFAR-10/100-LTN with joint noise. Res32 and Res18 are abbreviations for ResNet-32 and PreAct ResNet18, respectively. Results are cited from ECBS except CLIP based methods. The best and the second-best results are shown in **underline bold** and **bold**, respectively.

Dataset	CIFAR-10-LTN						CIFAR-100-LTN					
Imbalance Factor	10			100			10			100		
Noise Ratio	0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5
CE	72.4	70.3	65.2	52.9	48.1	38.7	37.4	32.9	26.2	21.8	17.9	14.2
LDAM-DRW [4] (2019)	80.2	74.9	67.9	66.7	57.5	43.2	45.1	39.4	32.2	27.6	21.2	15.2
NCM [18] (2020)	74.8	68.4	64.8	60.9	55.5	42.6	41.3	35.4	29.3	24.7	21.8	16.8
MiSLAS [36] (2021)	83.4	76.2	72.5	67.9	62.0	54.5	50.0	46.1	40.6	32.8	27.0	21.8
Co-teaching [12] (2018)	68.7	57.1	46.8	38.0	30.8	22.9	36.1	32.1	25.3	22.0	16.2	13.5
CDR [56] (2020)	73.9	68.1	62.2	46.3	42.5	32.4	35.4	30.9	24.9	22.0	17.3	13.6
Sel-CL+ [31] (2022)	84.4	80.4	77.3	65.7	61.4	56.2	50.9	47.6	44.9	35.1	32.0	28.6
RoLT [54] (2021)	83.5	80.9	79.0	66.5	57.9	49.0	47.4	44.6	38.6	27.6	24.7	20.1
RoLT-DRW [54] (2021)	83.6	81.4	77.1	71.1	63.6	55.1	49.3	46.3	40.9	30.2	26.6	21.1
HAR-DRW [62] (2022)	80.4	77.4	67.4	48.6	54.2	42.8	41.2	37.4	31.3	22.6	19.0	14.8
RCAL [63] (2023)	84.6	83.4	80.8	72.8	69.8	65.1	51.7	48.9	44.4	36.6	33.4	30.3
ECBS-Res32 [33] (2024)	87.4	85.9	84.8	76.8	75.2	73.6	53.8	52.8	51.2	38.5	37.1	35.5
ECBS-Res18 [33] (2024)	89.1	87.7	85.6	78.0	76.5	72.9	60.1	58.2	55.3	43.0	39.9	39.1
CLIP [44]+CE (2021)	95.1	94.8	93.9	89.9	88.2	83.5	79.7	78.6	77.5	66.4	64.8	62.1
CLIP [44]+CE+WTS (ours)	95.2	95.1	94.5	90.1	88.9	87.8	79.6	78.7	77.8	66.1	66.1	63.6
CLIP [44]+LA [39]	**96.3**	**96.0**	**95.3**	**95.2**	**94.0**	**90.5**	**82.0**	**80.8**	**79.7**	**77.5**	**76.6**	**75.4**
CLIP [44]+LA+WTS (ours)	**96.3**	**96.0**	**94.7**	**95.2**	**92.3**	**89.2**	**82.0**	**81.2**	**80.4**	**77.8**	**77.1**	**76.1**

Results on CIFAR-10/100-LTN. Tab. 1 shows comparison results for joint noise, while Tab. 2 presents results for symmetric and asymmetric noise on CIFAR-10/100-LTN. We observe that directly applying the LT learning method can achieve improvement to a certain extent. The logit adjustment-based method performs slightly better. NCM requires resampling the data distribution based on class labels, but the presence of noisy labels leads to unreasonable resampling, limiting its performance improvement. Under joint noise, LN learning methods, such as Sel-CL+ [31], outperform LT learning methods. For symmetric and asymmetric noise, recently proposed noise learning techniques can achieve satisfactory performance. However, these two kinds of methods are less effective when the noise ratio is high. For example, under joint noise, when the imbalance factor (IF) of CIFAR-100-LTN is 100 and the noise ratio (NR) is 0.5, MiSLAS and Sel-CL+ achieve accuracies of 21.8% and 28.6%, respectively. While these are significantly higher than CE (14.2%), they still fall short of meeting practical requirements for usage.

Recent proposed LTNL learning methods, such as TABASCO [38], RCAL [63] and ECBS [33], have demonstrated improved performance across various noise ratios. However, there is still potential for further enhancement, especially for more challenging datasets. For instance, on CIFAR-100 with an IF of 10 and NR of 0.4, ECBS achieves accuracies of 58.2%, 56.7%, and 52.1% under three types of label noise, respectively. In comparison, the proposed WTS achieves 81.2%, 80.7%, and 69.5%, demonstrating the generalization capability of the CLIP-introduced prior and the text-based knowledge in WTS across various noise types, as well as its robustness to high noise levels.

Results on Real-World Datasets Tab. 3 presents the top-1 accuracy on the test set of Img-LTN^r. It can be observed that, with a noise ratio of 0.4, applying LT learning methods may lead to adverse effects. Similar to the results on the CIFAR-10/100-LTN datasets, the LN learning method shows effectiveness on Img-LTN^r with

Table 2. Top-1 acc. (%) on the CIFAR-10/100-LTN dataset with an imbalance factor of 10 under symmetric and asymmetric noise.

Dataset	CIFAR-10-LTN		CIFAR-100-LTN		CIFAR-100-LTN	
Noise Type	Symmetric				Asymmetric	
Noise Ratio	0.4	0.6	0.4	0.6	0.2	0.4
CE	71.7	61.2	34.5	23.6	44.5	32.1
LDAM [4] (2019)	70.5	62.0	31.3	23.1	40.1	33.3
LA [39] (2021)	70.6	54.9	29.1	23.2	39.3	28.5
IB [41] (2021)	73.2	62.6	32.4	25.8	45.0	35.3
DivdeMix [26] (2020)	82.7	80.2	54.7	45.0	58.1	42.0
UNICON [19] (2022)	84.3	82.3	52.3	45.9	56.0	44.7
MW-Net [49] (2019)	70.9	59.9	32.0	21.7	42.5	30.4
RoLT [54] (2021)	81.6	76.6	42.0	32.6	48.2	39.3
HAR [62] (2022)	77.4	63.8	38.2	26.1	48.5	33.2
ULC [15] (2022)	84.5	83.3	54.9	44.7	54.5	43.2
TABASCO [38] (2023)	85.5	84.8	56.5	46.0	59.4	50.5
ECBS [33] (2024)	86.4	83.9	56.7	48.1	60.5	52.1
CLIP [44]+CE (2021)	95.1	94.4	78.4	74.7	78.7	62.5
CLIP+CE+WTS (ours)	95.2	95.0	78.5	75.9	78.9	**67.4**
CLIP [44]+LA (2021)	**96.1**	**95.2**	**80.2**	**76.8**	**79.3**	67.2
CLIP+LA+WTS (ours)	**96.0**	**95.3**	**80.7**	**78.0**	**79.8**	**69.5**

a low imbalance ratio. In contrast, the LTNL learning method demonstrates a more significant improvement. For instance, when IF is 100, TABASCO [38] achieves a top-1 classification accuracy of 37.1%, compared to 34.7% achieved by DivideMix [26]. In comparison, WTS exceeds 80%. Under real-world noisy label conditions, LA also negatively impacts the CLIP fine-tuned model, with CLIP+LA reducing CE from 82.9% to 81.9% at an imbalance ratio of 10 for example. In contrast, WTS enables the effective use of LA, further boosting performance to 83.1%.

WebVision-50 is derived from real-world datasets with NR of 0.05, out-of-distribution ratio 0.24 [1] and

Table 3. Top-1 acc. (%) on Img-LTN^r with NR of 0.4.

Imbalance Factor	10	100
CE	31.5	31.5
LDAM [4] (2019)	23.5	15.6
LA [39] (2021)	25.9	9.6
IB [41] (2021)	22.1	16.3
DivdeMix [26] (2020)	49.0	34.7
UNICON [19] (2022)	41.6	31.1
MW-Net [49] (2019)	40.3	31.1
RoLT [54] (2021)	24.2	16.9
HAR [62] (2022)	38.7	31.3
ULC [15] (2022)	47.1	34.8
TABASCO [38] (2023)	49.7	37.1
ECBS [33] (2024)	50.8	36.9
CLIP [44]+CE (2021)	82.9	80.5
CLIP+CE+WTS (ours)	**83.3**	**81.3**
CLIP [44]+LA (2021)	81.9	79.5
CLIP+LA+WTS (ours)	**83.1**	**80.9**

Table 4. Top-1 acc. (%) on WebVision-50.

Train	WebVision-50	
Test	WV50[3]	IMG12[3]
CE	62.5	58.5
CT[2] [12] (2018)	63.6	61.5
MentorNet [37] (2018)	63.0	57.8
ELR+ [34] (2020)	77.8	70.3
MoPro [27] (2021)	77.6	76.3
NGC [55] (2021)	79.2	74.4
Sel-CL+ [31] (2022)	80.0	76.8
RCAL+ [63] (2023)	79.6	76.3
ECBS [33] (2024)	80.0	76.1
CLIP [44]+CE	83.4	83.8
CLIP+CE+WTS (ours)	83.5	83.6
CLIP [44]+LA	**85.2**	**84.1**
CLIP+LA+WTS (ours)	**85.2**	**84.2**

IF of 6.78. Since the NR is low, LA can be applied directly, and the correction effect of WTS is less evident. However, WTS does lead to a slight improvement in cross-dataset test results, demonstrating an enhancement in the generalization ability of the models.

6 Conclusion and future work

In this work, we proposed a label calibration method, WTS, to tackle the compounded challenges of noisy labels and long-tailed distributions in real-world data. WTS leverages auxiliary language information from pre-trained visual-language models to correct label misalignment. By calibrating the supervisory signal, WTS enables effective feature learning and ensures that valuable category information is preserved, even in high-noise scenarios. This approach shows significant improvements in model performance across various benchmarks, particularly under challenging noise conditions. Despite WTS being effective in most scenarios, its supervision

[2]WV50, IMG12 and CT are abbreviated for WebVision-50, ILSVRC12 [22] and Co-teaching, respectively.

activation relies on an empirically chosen hyperparameter. In simple noise environments, this dependency may occasionally lead WTS to provide misleading signals, potentially impacting model performance. Our future work will focus on developing a more reasonable parameter selection to overcome this limitation.

References

[1] Paul Albert, Diego Ortego, Eric Arazo, Noel E. O'Connor, and Kevin McGuinness. Addressing out-of-distribution label noise in webly-labelled data. In *WACV*, pages 392–401, 2022.

[2] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV*, pages 112–121, 2021.

[3] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Aréchiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *ICLR*, 2021.

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, pages 1567–1578, 2019.

[5] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *NeurIPS*, volume 35, pages 16664–16678, 2022.

[6] De Cheng, Yixiong Ning, Nannan Wang, Xinbo Gao, Heng Yang, Yuxuan Du, Bo Han, and Tongliang Liu. Class-dependent label-noise learning with cycle-consistency regularization. In *NeurIPS*, 2022.

[7] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, pages 113–123, 2019.

[8] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, pages 3008–3017, 2020.

[9] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE TPAMI*, 45(3):3695–3706, 2023.

[10] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[12] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, volume 31, 2018.

[13] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, volume 31, 2018.

[14] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, pages 6626–6636, June 2021.

[15] Yingsong Huang, Bing Bai, Shengwei Zhao, Kun Bai, and Fei Wang. Uncertainty-aware learning against label noise on imbalanced datasets. In *AAAI*, volume 36, pages 6960–6969, 2022.

[16] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *ICML*, volume 119, pages 4804–4815, 2020.

[17] Shenwang Jiang, Jianan Li, Ying Wang, Bo Huang, Zhang Zhang, and Tingfa Xu. Delving into sample loss curve to embrace noisy and imbalanced data. *AAAI*, 36:7024–7032, 2022.

[18] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

[19] N. Karim, M. Rizve, N. Rahnavard, A. Mian, and M. Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *CVPR*, pages 9666–9676, 2022.

[20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, volume 25, 2012.

[23] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *CVPR*, pages 6970–6979, 2022.

[24] Hao-Tian Li, Tong Wei, Hao Yang, Kun Hu, Chong Peng, Li-Bo Sun, Xun-Liang Cai, and Min-Ling Zhang. Stochastic feature averaging for learning with long-tailed noisy labels. In *IJCAI*, pages 3902–3910, 2023.

[25] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *CVPR*, pages 6949–6958, 2022.

[26] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020.

[27] Junnan Li, Caiming Xiong, and Steven C. H. Hoi. Mopro: Webly supervised learning with momentum prototypes. In *ICLR*, 2021.

[28] Mengke Li. *Advances in Long-Tailed Visual Recognition*. PhD thesis, Hong Kong Baptist University, 2022.

[29] Mengke Li, Yiu-ming Cheung, and Zhikai Hu. Key point sensitive loss for long-tailed visual recognition. *IEEE TPAMI*, 45(4):4812–4825, 2023.

[30] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *CVPR*, pages 6929–6938, June 2022.

[31] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 316–325, 2022.

[32] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

[33] Zhuo Li, He Zhao, Zhen Li, Tongliang Liu, Dandan Guo, and Xiang Wan. Extracting clean and balanced subset for noisy long-tailed classification, 2024.

[34] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, volume 33, pages 20331–20342, 2020.

[35] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE TPAMI*, 38(3):447–461, 2015.

[36] Yuqi Liu, Bin Cao, and Jing Fan. Improving the accuracy of learning example weights for imbalance classification. In *ICLR*, 2022.

[37] Jiang Lu, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Fei-Fei Li. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pages 2304–2313, 2018.

[38] Yang Lu, Yiliang Zhang, Bo Han, Yiu-ming Cheung, and Hanzi Wang. Label-noise learning with intrinsically long-tailed data. In *ICCV*, pages 1369–1378, 2023.

[39] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

[40] Meng Pang, Binghui Wang, Mang Ye, Yiu-Ming Cheung, Yintao Zhou, Wei Huang, and Bihan Wen. Heterogeneous prototype learning from contaminated faces across domains via disentangling latent factors. *IEEE TNNLS*, 2024.

[41] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *ICCV*, pages 735–744, 2021.

[42] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 2233–2241, 2017.

[43] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. In *NeurIPS*, volume 33, pages 17044–17056, 2020.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, pages 8748–8763, 2021.

[45] Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, volume 33, pages 4175–4186, 2020.

[46] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, volume 80, pages 4331–4340, 2018.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.

[48] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *ICML*, 2024.

[49] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, pages 1917–1928, 2019.

[50] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *ICML*, pages 5907–5915, 2019.

[51] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pages 843–852, 2017.

[52] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021.

[53] Tong Wei, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. Prototypical classifier for robust class-imbalanced learning. In *PAKDD*, pages 44–57, 2022.

[54] Tong Wei, Jiang-Xin Shi, Wei-Wei Tu, and Yu-Feng Li. Robust long-tailed learning under label noise. *ArXiv*, 2021.

[55] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yufeng Li. Ngc: A unified framework for learning with open-world noisy data. In *ICCV*, pages 62–71, 2021.

[56] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2020.

[57] Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. In *ICLR*, 2022.

[58] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *NeurIPS*, volume 33, pages 7597–7610, 2020.

[59] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015.

[60] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-SRC: A contrastive approach for combating noisy labels. In *CVPR*, pages 5188–5197, 2021.

[61] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: reducing estimation error for transition matrix in label-noise learning. In *NeurIPS*, pages 7260–7271, 2020.

[62] Xuanyu Yi, Kaihua Tang, Xian-Sheng Hua, Joo-Hwee Lim, and Hanwang Zhang. Identifying hard noise in long-tailed sample distribution. In *ECCV*, pages 739–756, 2022.

[63] Manyi Zhang, Xuyang Zhao, Jun Yao, Chun Yuan, and Weiran Huang. When noisy labels meet long tail dilemmas: A representation calibration method. In *ICCV*, pages 15844–15854, 2023.

[64] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, pages 2361–2370, 2021.

[65] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE TPAMI*, 45(9):10795–10816, 2023.

[66] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, pages 16489–16498, 2021.

[67] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 9719–9728, 2020.

[68] Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, and Xiangyang Ji. Prototype-anchored learning for learning with imperfect annotations. In *ICML*, volume 162, pages 27245–27267, 2022.