Learning from Imperfect Text Guidance: Robust Long-Tail Visual Recognition with High-Noise Labels

Mengke Li
SCSE, Shenzhen University
Shenzhen, China
mengkeli@szu.edu.cn

Haiquan Ling
SCSE, Shenzhen University
Shenzhen, China
2410815024@mails.szu.edu.cn

Yiqun Zhang
SCST, Guangdong University of Technology
Guangzhou, China
yqzhang@gdut.edu.cn

Yang Lu
INFORMATICS, Xiamen University
Xiamen, China
luyang@xmu.edu.cn

Hui Huang*
SCSE, Shenzhen University
Shenzhen, China
hhzhiyan@gmail.com

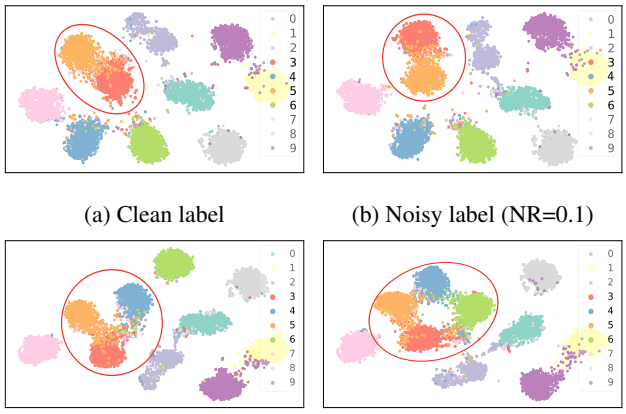
Abstract

Real-world data often exhibit long-tailed distributions with numerous noisy labels, substantially degrading the performance of deep models. While prior research has made progress in addressing this combined challenge, it overlooks the severe label-image mismatch inherent to high-noise settings, thereby limiting their effectiveness. Given that observed labels, though mismatched with images, still retain category information, we propose employing auxiliary text information from labels to address label-image inconsistencies in long-tailed noisy data. Specifically, we leverage the intrinsic cross-modal alignment in pre-trained visual-language models to correct the label-image inconsistencies. This supervisory signal, referred to as Weak Teacher Supervision (WTS), is unaffected by label noise and data distribution biases, albeit exhibits limited accuracy. Therefore, the activation of WTS is determined by evaluating the discrepancy between text-predicted labels and observed labels. Extensive experiments demonstrate the superior performance of WTS across synthetic and real-world datasets, particularly under high-noise conditions. The source code is available at https://anonymous.4open.science/r/WTS-0F3C.

Keywords: ***Noisy label learning, Long-tail learning, Pre-trained Model, CLIP***

1. Introduction

With the availability of large-scale public datasets [49, 54, 46], significant progress has been made in the field of computer vision [29, 42] and large models [46, 11].



(a) Clean label (b) Noisy label (NR=0.1)

(c) Noisy label (NR=0.5) (d) Noisy label (NR=0.6)

Figure 1: T-SNE visualization of the feature distributions on the test set, obtained by models trained on label-noisy datasets with varying noise rates[1]. (a) Classes 3 and 5 exhibit overlap. (b) Classes 3 and 5 remain partially overlapping. (c) Classes 3, 4, and 5 are not entirely separated. (d) Classes 3, 4, 5 and 6 are not fully separated.

However, real-world visual datasets typically exhibit two critical limitations: (1) severe class imbalance, where a small number of head classes dominate the sample distribution while tail classes remain substantially underrepresented [67], and (2) pervasive label noise caused by erroneous annotations [53, 61]. Creating balanced datasets with correctly labeled classes to address these challenges is expensive and unsustainable. To address these issues, the practical problem of long-tailed noisy label (LTNL) learning [40, 65] has been introduced.

*Corresponding author

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 [48, 52, 3, 56, 18]; (2) selecting clean samples based on carefully designed criteria [57, 60, 40, 25]; (3) developing improved representation learning methods [70, 64, 65, 34]. 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. To illustrate this phenomenon, we visualize test set features from models trained on LTNL datasets with varying noise ratios (NRs), as shown in Figure 1. The feature distributions under low NR closely resemble those of clean labels, as shown in Figures 1a and 1b. Consequently, long-tailed approaches are less affected in the low noise scenario, as the labels are relatively reliable. In contrast, when comparing Figures 1c and 1d, we can observe that the feature distributions under high noise ratios differ significantly from those of clean labels. 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 integrating auxiliary linguistic information into supervisory signal calibration, as textual information inherently captures the semantics of the labels themselves and is inherently robust to label noise and data distribution biases in the training set.

Considering that in 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) [21, 46] 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 these 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. The main contributions of this paper are summarized as follows:

- We empirically demonstrate that even a text encoder from a pre-trained VLM with suboptimal performance can contribute to performance improvements in LTNL learning, and provide an in-depth analysis of the underlying rationale.
- 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.

2. Related Work

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 [29, 51]. (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 [19, 68] and BBN-based methods, where BBN denotes Bilateral-Branch Network [69, 66], separate representation learning from classifier training. They first extract representations from the original long-tailed dataset, and then retrain the classifier using either class-balanced sampling data [19] or reverse sampling data [69]. Ensembling learning includes redundant ensembling [55, 24, 27, 2], which aggregates outputs from separate classifiers or networks

[1]The training set is CIFAR-10-LTN with asymmetric noise and an imbalance factor of 10. The model is Adaptformer [5], fine-tuning on CLIP [46]. The loss function employed is logit adjustment [41].

within a multi-expert framework, and complementary ensembling [69, 9], which involves the statistical selection of different data partitions. (3) Output level. Existing methods enhance model representation and refine the classifier by calibrating model logits based on specific criteria. For example, logit adjustment methods [41, 47] calibrate the predicted output distribution to achieve a balanced distribution. Re-margining methods [4, 30, 41, 31] introduce class size-based constants that assign larger margins to tail classes compared to head classes.

2.2. Noisy Label Learning

Noisy label supervision with high-noise datasets critically compromises model recognition performance. A straightforward and effective approach is to distinguish between clean and noisy samples, with methods such as MentorNet [39], Co-teaching [12] and DivideMix [26] treating samples with small training losses as clean samples. FedFixer [16] introduces the personalized model that collaborates with the global model to effectively select clean and client-specific samples, and NPN [50] directly corrects labels by accumulating model predictions. Divergence-based approaches rely on margin metrics, such as AUM [45] which measures the logit differences between a specified class and the top non-specified class, or distribution similarity employed by methods like Jo-SRC [63] and UNICON [20] via Jensen-Shannon divergence. In addition, several methods design noise-robust loss functions to mitigate the impact on noisy data, such as backward and forward loss correction [44], gold loss correction [13], MW-Net [52] and Dual-T [62]. Other effective methods focus on evaluating the noise transfer matrix [62, 6, 35] or reweighting examples for noisy label learning [48, 37].

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 [60] improves upon the small loss method [12] by identifying a subset of high-loss samples as clean. TABASCO [40] addresses a complex scenario where noisy labels can cause an intrinsic tail class to be misrepresented as a head class. To solve this issue, it proposes a bi-dimensional separation metric that effectively adapts to different cases. Another promising path is to emphasize the importance of different samples by reweighting or regularization [48, 52, 3, 56, 18]. Concurrently, improving representation learning [70, 64, 65, 34] leverages intrinsic feature-space structures to mitigate noise propagation. However, 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 utilization of feature-based strategies to address both noise and class imbalance.

3. Proposed Method: WTS - a Weak yet Effective Teacher

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) [21, 46] 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 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 2.

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 $y_i \in \mathcal{Y}$. $\mathcal{Y}$ 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 [40, 65]. 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[2], 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.

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

Logit Adjustment (LA). LA [41] is a statistically grounded approach that addresses class imbalance by modifying classifier logits based on label frequencies. Formally, given the original logit z_i and the estimated class prior π_i, the adjusted logit is computed as $\tilde{z}_i = z_i + \log \pi_i$. This adjustment acts as a relative margin that effectively penalizes majority classes during optimization, thereby enforcing consistency with the balanced error rate and improving recognition on tail categories.

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 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 [41], and LADE [14], for long-tailed learning. Meanwhile, $\mathcal{L}_T$ is the loss based on text-predicted labels. Equation (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% (additional results can be found in Section 4.3), 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 Equation (1)) into probabilities:

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

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). Regarding the temperature scale used in the KL divergence, we implemented it as a learnable parameter rather than a fixed scalar. This allows the model to adaptively adjust the sharpness of the probability distribution during training.

Supervision Switch Control. Since the weak teacher is not always accurate, and as discussed in Section 1, 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. For a mini-batch of size B, the overlap ratio is defined as $OR = \frac{1}{B} \sum_{i=1}^B \mathbf{1}\left(y_i^t = \hat{y}_i\right)$, where y_i^t and $\hat{y}_i$ denote the text-predicted label and the observed label of the i-th sample in the batch, respectively, and $\mathbf{1}(\cdot)$ is the indicator function. The overlap ratio OR measures the batch-wise agreement between text-predicted and observed labels and serves as an indicator for activating weak teacher supervision. 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 Section 4.3. We used $\alpha = 2.0$ and $\beta = 2.0$ for the Beta distribution. 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 Equation (2). Conversely, when the switch control of 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.

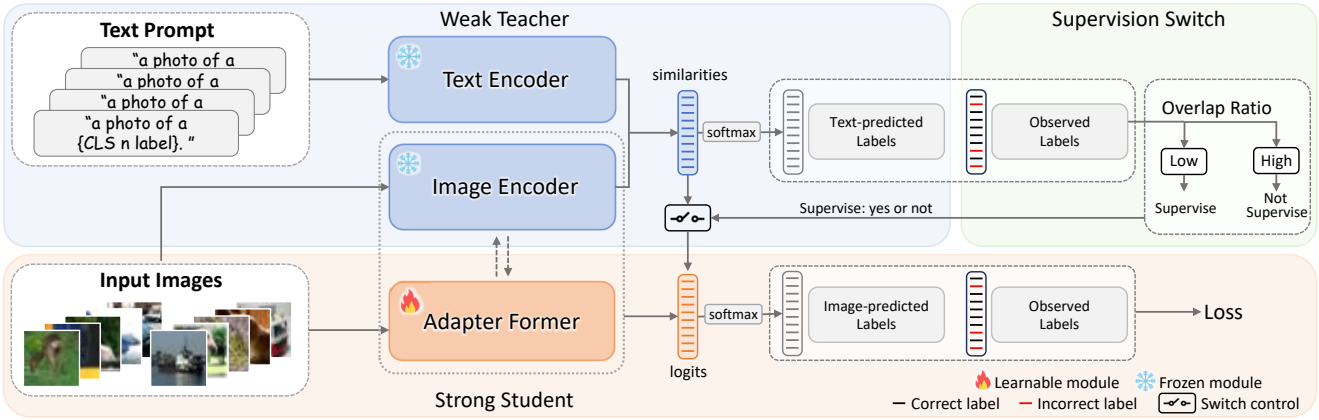


Figure 2: Overview of WTS. We leverage the text encoder in pre-trained visual-language models to obtain text-based predictions, using text-image alignment to correct label-image inconsistencies. Since this supervisory signal is not always accurate, we evaluate the discrepancy between the text-predicted and observed labels to determine when to activate it.

Algorithm 1 Training Algorithm of WTS

Require: Training set $\mathcal{D}$, pre-trained model $\mathcal{M}$;
Ensure: Fine-tuned model;
Initialize the fine-tuning module ϕ;
for $k = 1$ to K **do**
  Sample batches data $\mathcal{B}_e \sim \mathcal{D}$;
  Compute text-predicted labels $\hat{\mathcal{Y}}^t_{\mathcal{B}_e}$ by Equation (1);
  Compute OR between $\hat{\mathcal{Y}}^t_{\mathcal{B}_e}$ and $\hat{\mathcal{Y}}^o_{\mathcal{B}_e}$;
  if $OR < \tau$ **then**
    Compute the TL supervision: $\mathcal{L}_T = \mathrm{KL}(\mathcal{P}^t \| \mathcal{P}^I)$;
    Sample a by $a \sim \mathrm{Beta}(\alpha, \beta)$ and compute the final loss by $\mathcal{L} = a\mathcal{L}_O + (1-a)\mathcal{L}_T$;
  else
    Compute the final loss by $\mathcal{L} = \mathcal{L}_O$;
  end if
  Update ϕ by $\phi_{k+1} = \phi_k - \eta_k \cdot \nabla\mathcal{L}$
end for

The algorithm of WTS is summarized in Algorithm 1.

Remark 1. *The advantages of WTS can be summarized in three main aspects:* ***(1) Label noise-robust feature calibration.*** *The pre-trained text encoder in VLM, which is unaffected by noisy labels, exclusively focuses on the semantics of the labels and is inherently aligned with visual features, serves as the teacher model to rectify 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.* ***(2) Sample distribution resilient bias corrector.*** *WTS predictions are distribution-agnostic, enabling them to effectively mitigate long-tail bias in samples by propagating distribution-independent reference gradients. This approach helps mitigate potential classification errors that severely affect tail classes and improves the performance of all classes, including both head and tail.* ***(3) Highly efficient training.*** *WTS introduces minimal computational overhead, with the primary cost arising from the fine-tuning of the image encoder.*

3.3. Effectiveness Analysis 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. The efficacy of the proposed method 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 Equation (4) is:

$$\begin{aligned}\mathrm{KL}(\mathcal{P}^t \| \mathcal{P}^I) &= -\sum p_c^t \log p_c^I - \left(-\sum p_c^t \log p_c^t\right) \\ &= -\sum p_c^t \log p_c^I + H(\mathcal{P}^t), \end{aligned} \tag{6}$$

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 Equation (2), by substituting $\mathcal{L}_T$ with Equation (6) and replacing $\mathcal{L}_O$ with the expanded form of the CE-based loss, we can obtain:

$$\begin{aligned}\mathcal{L} &= a\left(-\sum_{c=1}^{C} p_c^o \log p_c^I\right) + (1-a)\left(-\sum_{c=1}^{C} p_c^t \log p_c^I\right), \\ &= -\sum_{c=1}^{C}\left(a\cdot p_c^o + (1-a)\cdot p_c^t\right)\cdot \log p_c^I, \end{aligned} \tag{7}$$

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

Proposition 1 illustrates that WTS has the following two impacts 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 $q_i = \frac{e^{z_i}}{\sum_{j=1}^C e^{z_i}}$. 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:*

$$\frac{\partial \ell}{\partial z_i} = q_i - p_i, \tag{8}$$

where p_i denotes the target probability of class i.

Proof. For a specific class, we assume, without loss of generality, that class c is considered. Then, the derivative on z_c is given by:

$$\begin{aligned}\frac{\partial \ell}{\partial z_c} =& -\left(p_c \frac{\sum_j e^{z_j}}{e^{z_c}} \cdot \frac{e^{z_c}\sum_j e^{z_j} - e^{z_c}e^{z_c}}{\left(\sum_j e^{z_j}\right)^2} \right. \\ &\left. + \sum_{i\neq c} p_i \frac{\sum_j e^{z_i}}{e^{z_c}} \cdot \frac{-e^{z_i}e^{z_c}}{\left(\sum_j e^{z_j}\right)^2}\right) \\ =& -\left(p_c \frac{\sum_j e^{z_j} - e^{z_c}}{\sum_j e^{z_j}} - \sum_{i\neq c} p_i \frac{e^{z_c}}{\sum_j e^{z_j}}\right) \\ =& q_c \sum_{i\neq c} p_i - p_c(1-q_c). \end{aligned} \tag{9, 10}$$

According to the property that the sum of a probability distribution is 1, we have $\sum_{i\neq c} p_i = 1 - p_c$. Substituting this into Equation (10), we get:

$$\begin{aligned}\frac{\partial \ell}{\partial z_c} &= q_c(1-p_c) - p_c(1-q_c) \\ &= q_c - p_c. \end{aligned} \tag{11}$$

By substituting c for i, we obtain Equation (8). □

Remark 2. *The gradient of logit adjustment methods reduces the positive signal contributions from head classes while amplifying those from tail classes.*

Proof. We compare the gradient difference d_g for the target class between the LA loss ($\mathcal{L}_{LA}$) and the CE loss ($\mathcal{L}_{CE}$) to analyze how LA adjusts the gradient during model training[6]. According to Theorem 1, the gradient differences is:

$$\begin{aligned} d_g &= \left(q_y^{LA} - 1\right) - \left(q_y^{CE} - 1\right) \\ &= \frac{e^{z_y - m_y}}{\sum_j e^{z_j - m_j}} - \frac{e^{z_y}}{\sum_j e^{z_j}} \\ &= \frac{\sum_j e^{z_j}\cdot e^{z_y - m_y} - \sum_j e^{z_j - m_j}\cdot e^{z_y}}{\sum e^{z_j - m_j}\sum e^{z_j}} \\ &= \frac{e^{z_y}\left(\sum_j e^{z_j - m_y} - e^{z_j - m_j}\right)}{\sum e^{z_j - m_j}\sum e^{z_j}}. \end{aligned} \tag{12}$$

The denominator of Equation (12) is always positive, thus the sign of d_g is determined entirely by the numerator. If y belongs to the tail class, in the extreme case where y is the smallest class, m_y exceeds that of the other classes. Therefore, $d_g < 0$, and then $\frac{\partial \mathcal{L}_{LA}}{\partial z_y} < \frac{\partial \mathcal{L}_{CE}}{\partial z_y}$. Gradient descent updates the parameters by subtracting the gradient to minimize the objective function. As a result, the positive signal $(1 - q_y \geq 0)$ for the target class in tail classes is amplified compared to the base loss (CE loss). For head classes, the opposite holds, that is, the positive signal corresponding to the target class is reduced in tail classes. □

For the notation, we define the modified probability p_c^m for class c as follows:

$$p_c^m = a\cdot p_c^o + (1-a)\cdot p_c^t. \tag{13}$$

Theorem 1 gives that the derivatives of $\mathcal{L}_O$ and $\mathcal{L}$ (Equation (7)) with respect to the logit of the target class are:

$$\frac{\partial \mathcal{L}_O}{\partial z_y} = p_y^I - 1,\ \frac{\partial \mathcal{L}}{\partial z_y} = p_y^I - p_y^m. \tag{14}$$

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 2, 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 introduces text-encoder-derived semantic constraints to attenuate class-specific positive correlations, thereby counteracting error signal amplification while preserving discriminative feature learning.

4. Experiment

4.1. Basic Settings

Datasets and Implementation Details. We evaluate the proposed WTS on both simulated and real-world noisy long-tailed datasets, following TABASCO [40] and RCAL [65]. Specifically, synthetic scenarios are created based on CIFAR-10/100 [22], real-world noise is introduced in mini-ImageNet [17] (referred to as red Mini-ImageNet, abbreviated as Img-LTN^r), and WebVision-50 [33] 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 [67], 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 [46] ViT-B/16 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 set to 128, and WTS is trained for 10 epochs across all datasets.

Definition of Noise Types. For joint noise, each element T_{ij}^{JN} in the transfer matrix T^{JN} is defined as:

$$T_{ij}^{JN} = P^{JN}(\hat{y} = j \mid y = i) = \begin{cases} 1-\gamma & \text{if } i = j \\ \frac{n_j}{N-n_i}\gamma & \text{if } i \neq j \end{cases}, \tag{15}$$

where $\hat{y}$ denotes the observed label, and y is the ground-truth label. γ denotes the noise ratio. N is the total number of training samples, and n_i is the number of training samples in class i.

For symmetric noise, the element T_{ij}^{SN} in the transfer matrix is expressed as:

$$T_{ij}^{SN} = \begin{cases} \gamma \cdot \frac{1}{C} + (1-\gamma), & \text{if } i = j \\ \gamma \cdot \frac{1}{C}, & \text{if } i \neq j \end{cases}, \tag{16}$$

where C represents the total number of classes.

For asymmetric noise, the element T_{ij}^{AN} in its transfer matrix is given by:

$$T_{ij}^{AN} = \begin{cases} 1-\gamma, & \text{if } i = j \\ \gamma \cdot P(i \to j), & \text{if } i \neq j \end{cases}, \tag{17}$$

where $P(i \to j)$ is the probability of a sample with true label i being mislabeled as j and satisfies $\sum_{j=1}^{C} P(i \to j) = 1$. Similar to previous work [52, 40], we adopt $P(i \to j)$, where only one element is 1 while all others are 0 in our experiments.

4.2. Comparison Results

Comparison Methods. We compare our method with the following three types of approaches: (1) *Long-tail (LT) learning methods*: LDAM [4], NCM [19], MiSLAS [38], logit adjustment (LA) [41], and influence-balanced loss (IB) [43]. (2) *Label-noise (LN) learning methods* : Co-teaching (CT) [12], CDR [59], Sel-CL [32] DivideMix [26], and UNICON [20]. (3) *Long-tailed noisy label (LTNL) learning*: MW-Net [52], ROLT [57], HAR [64], ULC [15], TABASCO [40], RCAL [65] and ECBS [34].

Results on CIFAR-10/100-LTN. Tables 1 and 2 compare joint versus symmetric/asymmetric noise results 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+ [32], 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.

Table 3 presents the performance comparison of various methods on CIFAR-10/100-LTN under an NR of 0.6, representing an extremely high-noise regime. In this challenging setting, observed labels are highly unreliable, leading to poor performance from all methods under asymmetric noise conditions. The experimental results demonstrate that incorporating WTS can significantly enhance the performance of all methods, especially under asymmetric and symmetric noise. For example, WTS improves the CLIP+LA method by more than 10%, achieving 74.7% compared to 63.1% on the CIFAR-10-LTN dataset with an imbalance factor of 100.

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. 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]	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 [19]	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[38]	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]	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 [59]	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+[32]	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 [57]	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[57]	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[64]	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 [65]	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 [34]	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 [34]	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 (zero-shot)	87.2	87.2	87.2	87.2	87.2	87.2	64.4	64.4	64.4	64.4	64.4	64.4
CLIP+CE	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+CE+WTS (ours)	95.2	95.1	94.5	90.1	88.9	87.8	80.8	78.7	77.8	67.7	66.1	64.9
CLIP+LA	**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+LA+WTS (ours)	**96.5**	**96.2**	**95.6**	**95.6**	**95.2**	**91.9**	**83.2**	**81.2**	**80.4**	**77.8**	**77.1**	**76.9**

Table 2: Top-1 acc. (%) on CIFAR-10/100-LTN 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]	70.5	62.0	31.3	23.1	40.1	33.3
LA [41]	70.6	54.9	29.1	23.2	39.3	28.5
IB [43]	73.2	62.6	32.4	25.8	45.0	35.3
DivdeMix [26]	82.7	80.2	54.7	45.0	58.1	42.0
UNICON [20]	84.3	82.3	52.3	45.9	56.0	44.7
MW-Net [52]	70.9	59.9	32.0	21.7	42.5	30.4
RoLT [57]	81.6	76.6	42.0	32.6	48.2	39.3
HAR [64]	77.4	63.8	38.2	26.1	48.5	33.2
ULC [15]	84.5	83.3	54.9	44.7	54.5	43.2
TABASCO [40]	85.5	84.8	56.5	46.0	59.4	50.5
ECBS [34]	86.4	83.9	56.7	48.1	60.5	52.1
CLIP (zero-shot)	87.2	87.2	64.4	64.4	64.4	64.4
CLIP+CE	94.9	94.4	78.4	74.7	78.7	62.5
CLIP+CE+WTS (ours)	95.2	95.0	78.9	75.9	79.1	**67.4**
CLIP+LA	**95.8**	**95.2**	**80.2**	**76.8**	**79.3**	67.2
CLIP+LA+WTS (ours)	**96.1**	**95.3**	**80.7**	**78.0**	**79.8**	**69.5**

Table 4: Top-1 acc. (%) on Img-LTN^r with NR of 0.4.

Imbalance Factor	10	100
CE	31.5	31.5
LDAM [4]	23.5	15.6
LA [41]	25.9	9.6
IB [43]	22.1	16.3
DivdeMix [26]	49.0	34.7
UNICON [20]	41.6	31.1
MW-Net [52]	40.3	31.1
RoLT [57]	24.2	16.9
HAR [64]	38.7	31.3
ULC [15]	47.1	34.8
TABASCO [40]	49.7	37.1
ECBS [34]	50.8	36.9
CLIP (zero shot)	77.1	77.1
CLIP+CE	82.9	80.5
CLIP+CE+WTS	**83.3**	**81.3**
CLIP+LA	81.9	79.5
CLIP+LA+WTS	**83.1**	**80.9**

Table 5: Top-1 acc. (%) on WebVision-50.

Train	WebVision-50	
Test	WV50^3	IMG12^3
CE	62.5	58.5
CT [12]	63.6	61.5
MentorNet [39]	63.0	57.8
ELR+ [36]	77.8	70.3
MoPro [28]	77.6	76.3
NGC [58]	79.2	74.4
Sel-CL+ [32]	80.0	76.8
RCAL+ [65]	79.6	76.3
ECBS [34]	80.0	76.1
CLIP (zero-shot)	74.5	78.0
CLIP+CE	83.4	83.8
CLIP+CE+WTS	83.5	83.6
CLIP+LA	**85.2**	**84.1**
CLIP+LA+WTS	**85.2**	**84.2**

Recent proposed LTNL learning methods, including RCAL [65], TABASCO [40] and ECBS [34], have shown enhanced robustness across various noise ratios. However, there remains room for further improvement, particularly on more challenging datasets. For example, 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%, highlighting the effectiveness of the CLIP-introduced prior and the text-based knowledge in WTS. These results demonstrate the generalization capability of WTS across different noise types and its robustness under high-noise conditions.

Results on Real-World Datasets. Table 4 reports the performance on the test set of Img-LTN^r. It can be observed that with the noise ratio reaching 0.4, directly applying LT learning methods can 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 a low imbalance ratio. In contrast, the LTNL learning method demonstrates a more significant improvement. For instance, when IF is 100, TABASCO [40] 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,

Table 3: Acc. (%) on CIFAR-10/100-LTN. NR is 0.6. JN, AN, and SN stand for joint, asymmetric, and symmetric noise, respectively.

Dataset	CIFAR-10-LTN						CIFAR-100-LTN					
Imbalance Factor	10			100			10			100		
Noise Type	JN	AN	SN	JN	AN	SN	JN	AN	SN	JN	AN	SN
CLIP+CE	91.2	19.6	94.4	71.4	22.4	87.1	74.6	21.6	74.7	58.4	23.5	62.0
CLIP+CE+WTS	94.1 (↑ 2.9)	38.8	95.0 (↑ 0.6)	83.7 (↑ 12.3)	52.5 (↑ 30.1)	91.5 (↑ 4.4)	75.7 (↑ 1.1)	23.4	75.9 (↑ 1.2)	61.6 (↑ 3.2)	29.0	65.1 (↑ 3.1)
CLIP+LDAM	93.0	1.9	94.8	75.8	11.4	87.5	74.8	7.2	74.7	58.5	16.5	60.0
CLIP+LDAM+WTS	94.5 (↑ 1.5)	5.1	95.0 (↑ 0.2)	83.4 (↑ 7.6)	26.5	90.3 (↑ 2.8)	75.2 (↑ 0.4)	9.9	75.7 (↑ 1.0)	60.3 (↑ 1.8)	20.3	62.6 (↑ 2.6)
CLIP+LA	93.9	56.7	95.2	85.9	63.1	91.7	78.4	47.1	76.8	73.7	45.7	66.8
CLIP+LA+WTS	94.2 (↑ 0.3)	63.8 (↑ 7.1)	95.7 (↑ 0.4)	88.2 (↑ 2.3)	74.7 (↑ 11.6)	94.2 (↑ 2.5)	79.0 (↑ 0.6)	48.8 (↑ 1.7)	78.0 (↑ 1.2)	74.3 (↑ 0.6)	49.5 (↑ 3.8)	70.5 (↑ 3.7)

Table 6: Top-1 accuracy (%) on Red mini-ImageNet dataset with real-world noise.

Imbalance Factor	10						100					
Noise Ratio	0.1	0.2	0.3	0.4	0.5	0.6	0.1	0.2	0.3	0.4	0.5	0.6
CLIP (zero-shot)	77.1	77.1	77.1	77.1	77.1	77.1	77.1	77.1	77.1	77.1	77.1	77.1
CLIP+CE	85.7	84.4	84.1	82.9	80.9	79.6	**84.0**	**82.6**	81.3	80.5	79.7	78.4
CLIP+CE+WTS	**86.1** (↑ 0.4)	**85.3** (↑ 0.9)	**85.1** (↑ 1.0)	**83.3** (↑ 0.4)	82.8 (↑ 1.9)	**82.8** (↑ 3.2)	**84.2** (↑ 0.2)	**82.7** (↑ 0.1)	**82.6** (↑ 1.3)	**81.3** (↑ 0.8)	**81.0** (↑ 1.3)	**80.6** (↑ 2.2)
CLIP+LDAM	84.8	83.4	82.7	81.2	79.8	77.9	80.2	78.3	77.8	77.3	75.5	74.2
CLIP+LDAM+WTS	84.8 (0.0)	84.1 (↑ 1.7)	82.8 (↑ 0.1)	82.4 (↑ 1.2)	81.5 (↑ 1.7)	80.4 (↑ 2.5)	80.2 (0.0)	78.9 (↑ 0.6)	78.3 (↑ 0.5)	77.6 (↑ 0.3)	77.8 (↑ 2.3)	76.1 (↑ 1.9)
CLIP+LADE	84.4	83.6	83.1	81.9	81.1	79.7	79.9	80.5	79.1	78.9	78.4	77.8
CLIP+LADE+WTS	84.8 (↑ 0.4)	84.2 (↑ 0.6)	83.4 (↑ 0.3)	83.0 (↑ 1.1)	**83.5** (↑ 2.4)	82.2 (↑ 2.5)	80.4 (↑ 0.5)	81.5 (↑ 1.0)	79.6 (↑ 0.5)	79.7 (↑ 0.8)	80.3 (↑ 1.9)	**79.3** (↑ 1.5)
CLIP+LA	85.1	84.5	83.1	81.9	81.1	79.0	81.3	81.0	79.9	79.5	78.9	77.6
CLIP+LA+WTS	**86.0** (↑ 0.9)	**85.9** (↑ 1.4)	**85.7** (↑ 2.6)	**83.1** (↑ 1.2)	**83.6** (↑ 2.5)	**82.5** (↑ 3.5)	83.4 (↑ 2.1)	82.6 (↑ 1.6)	**81.5** (↑ 1.6)	**80.9** (↑ 1.4)	**81.4** (↑ 2.5)	**80.6** (↑ 3.0)

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 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.

4.3. Further Analysis

Impact of WTS on Different Classes. We conduct experiments to demonstrate that WTS can enhance the performance of all classes. Figure 3 shows the top-1 classification accuracy across different class types. As shown in the figures, compared to the basic CE loss, LA improves tail-class performance at noise ratios of 0.4 and 0.5, with a slight trade-off in head-class performance. Building on LA, WTS further improves the performance of all classes.

The Influence of Overlap Ratio Control Threshold τ. The supervision switch control in Section 3.2 relies on a hyper-parameter τ. We conduct an ablation study on the parameter τ to examine its impact on model training. Figure 4 shows the results. When the noise ratio is low (e.g., 0.3 or 0.4), the choice of parameter τ influences model performance. In such cases, the reliability of observed labels is relatively higher It is crucial to assess whether supervision from the weak teacher may teacher introduce uncertainty, potentially impacting training. Therefore, determining whether to trust the label $\mathcal{Y}^t$, which depends on τ, becomes essential. Figure 4 shows minimal fluctuation, demonstrating that WTS remains stable and informative across different values of τ. When NR is high, the reliability of observed labels decreases, making the supervision signal from the weak teacher more critical. The observed label set $\mathcal{Y}^o$ deviates significantly from $\mathcal{Y}^t$, further emphasizing the necessity of guidance from the weak teacher Consequently, during training, the control switch remains active, ensuring that model performance is largely unaffected by τ.

[3]WV50 and IMG12 are abbreviated for WebVision-50 and ILSVRC12 [23], respectively.

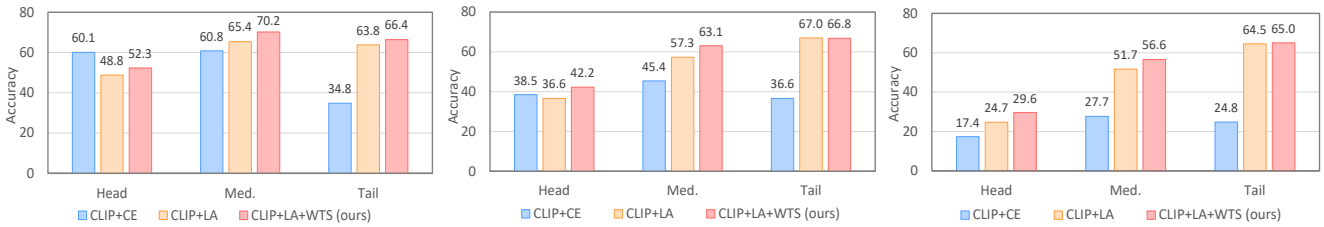
Table 7: Acc. (%) of CLIP (zero-shot) on the training set of CIFAR-100-NLT.

Imbalance Factor	Head	Med.	Tail	All
10	66.5	64.1	66.1	65.5
100	66.8	64.0	60.6	64.0

CLIP performance on training set. Table 7 presents the performance of CLIP on the training set of CIFAR-100-LTN. Since the predictions obtained by CLIP are independent of class labels, they remain unaffected by label noise and class distribution biases. However, the predicted labels generated by the text encoder of CLIP are not sufficiently accurate. We therefore characterize the textual supervision from CLIP as a weak supervisory signal that, while imperfect, provides valuable guidance for model training under noisy conditions.

5. Concluding Remarks

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



(a) NR = 0.4 (b) NR = 0.5 (c) NR = 0.6

Figure 3: Accuracy of different class types. (CIFAR100-LTN with IR of 100 and asymmetric noise)

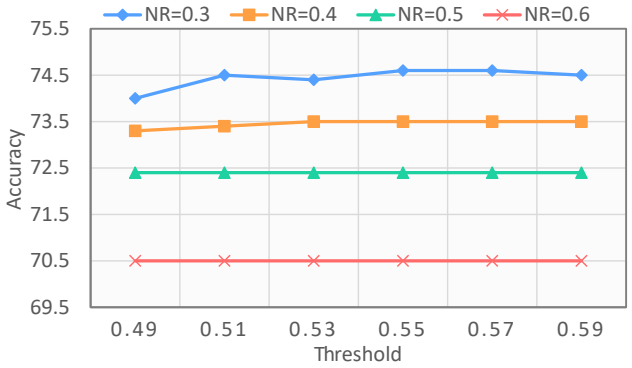


Figure 4: Ablation of τ in supervision switch. The dataset is CIFAR-100-LTN with IR=100 and symmetric noise.

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 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.

Acknowledgements

This work was supported in parts by NSFC (62306181), Guangdong Basic and Applied Basic Research Foundation (2024A1515010163), Shenzhen Science and Technology Program (RCBS20231211090659101, KJZD20240903100022028), National Key Lab of Radar Signal Processing (JKW202403), and Scientific Development Funds from Shenzhen University.

References

[1] P. Albert, D. Ortego, E. Arazo, N. E. O'Connor, and K. McGuinness. Addressing out-of-distribution label noise in webly-labelled data. In *WACV*, pages 392–401. IEEE, 2022. 9

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

[3] K. Cao, Y. Chen, J. Lu, N. Aréchiga, A. Gaidon, and T. Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *ICLR*, 2021. 2, 3

[4] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, pages 1567–1578, 2019. 3, 4, 7, 8

[5] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *NeurIPS*, volume 35, pages 16664–16678, 2022. 2, 4, 7

[6] D. Cheng, Y. Ning, N. Wang, X. Gao, H. Yang, Y. Du, B. Han, and T. Liu. Class-dependent label-noise learning with cycle-consistency regularization. In *NeurIPS*, 2022. 3

[7] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, pages 113–123, 2019. 2

[8] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, pages 3008–3017, 2020. 2

[9] J. Cui, S. Liu, Z. Tian, Z. Zhong, and J. Jia. Reslt: Residual learning for long-tailed recognition. *IEEE TPAMI*, 45(3):3695–3706, 2023. 3

[10] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019. 2

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[12] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, volume 31, 2018. 3, 7, 8

[13] D. Hendrycks, M. Mazeika, D. Wilson, and K. Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, volume 31, 2018. 3

[14] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, pages 6626–6636, June 2021. 4

[15] Y. Huang, B. Bai, S. Zhao, K. Bai, and F. Wang. Uncertainty-aware learning against label noise on imbalanced datasets. In *AAAI*, volume 36, pages 6960–6969, 2022. 7, 8

[16] X. Ji, Z. Zhu, W. Xi, O. Gadyatskaya, Z. Song, Y. Cai, and Y. Liu. Fedfixer: Mitigating heterogeneous label noise in federated learning. In *AAAI*, pages 12830–12838, 2024. 3

[17] L. Jiang, D. Huang, M. Liu, and W. Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *ICML*, volume 119, pages 4804–4815, 2020. 7

[18] S. Jiang, J. Li, Y. Wang, B. Huang, Z. Zhang, and T. Xu. Delving into sample loss curve to embrace noisy and imbalanced data. *AAAI*, 36:7024–7032, 2022. 2, 3

[19] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 2, 7, 8

[20] 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. 3, 7, 8

[21] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 2, 3

[22] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 7

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, volume 25, 2012. 9

[24] B. Li, Z. Han, H. Li, H. Fu, and C. Zhang. Trustworthy long-tailed classification. In *CVPR*, pages 6970–6979, 2022. 2

[25] H.-T. Li, T. Wei, H. Yang, K. Hu, C. Peng, L.-B. Sun, X.-L. Cai, and M.-L. Zhang. Stochastic feature averaging for learning with long-tailed noisy labels. In *IJCAI*, pages 3902–3910, 2023. 2

[26] J. Li, R. Socher, and S. C. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020. 3, 6, 7, 8

[27] J. Li, Z. Tan, J. Wan, Z. Lei, and G. Guo. Nested collaborative learning for long-tailed visual recognition. In *CVPR*, pages 6949–6958, 2022. 2

[28] J. Li, C. Xiong, and S. C. H. Hoi. Mopro: Webly supervised learning with momentum prototypes. In *ICLR*, 2021. 8

[29] M. Li. *Advances in Long-Tailed Visual Recognition*. PhD thesis, Hong Kong Baptist University, 2022. 1, 2

[30] M. Li, Y.-m. Cheung, and Z. Hu. Key point sensitive loss for long-tailed visual recognition. *IEEE TPAMI*, 45(4):4812–4825, 2023. 3

[31] M. Li, Y.-m. Cheung, and Y. Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *CVPR*, pages 6929–6938, June 2022. 3

[32] S. Li, X. Xia, S. Ge, and T. 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. 7, 8

[33] W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 7

[34] Z. Li, H. Zhao, Z. Li, T. Liu, D. Guo, and X. Wan. Extracting clean and balanced subset for noisy long-tailed classification, 2024. 2, 3, 7, 8

[35] Y. Lin, Y. Yao, and T. Liu. Learning the latent causal structure for modeling label noise. In *NeurIPS*, volume 37, pages 120549–120577, 2024. 3

[36] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, volume 33, pages 20331–20342, 2020. 8

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

[38] Y. Liu, B. Cao, and J. Fan. Improving the accuracy of learning example weights for imbalance classification. In *ICLR*, 2022. 7, 8

[39] J. Lu, Z. Zhou, T. Leung, L.-J. Li, and F.-F. Li. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pages 2304–2313, 2018. 3, 8

[40] Y. Lu, Y. Zhang, B. Han, Y.-m. Cheung, and H. Wang. Label-noise learning with intrinsically long-tailed data. In *ICCV*, pages 1369–1378, 2023. 1, 2, 3, 7, 8

[41] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 2, 3, 4, 7, 8

[42] M. Pang, B. Wang, M. Ye, Y.-M. Cheung, Y. Zhou, W. Huang, and B. Wen. Heterogeneous prototype learning from contaminated faces across domains via disentangling latent factors. *IEEE TNNLS*, 2024. 1

[43] S. Park, J. Lim, Y. Jeon, and J. Y. Choi. Influence-balanced loss for imbalanced visual classification. In *ICCV*, pages 735–744, 2021. 7, 8

[44] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 2233–2241, 2017. 3

[45] G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In *NeurIPS*, volume 33, pages 17044–17056, 2020. 3

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, pages 8748–8763, 2021. 1, 2, 3, 7

[47] J. Ren, C. Yu, s. sheng, X. Ma, H. Zhao, S. Yi, and h. Li. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, volume 33, pages 4175–4186, 2020. 3

[48] M. Ren, W. Zeng, B. Yang, and R. Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, volume 80, pages 4331–4340, 2018. 2, 3

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein,

et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015. 1

[50] M. Sheng, Z. Sun, Z. Cai, T. Chen, Y. Zhou, and Y. Yao. Adaptive integration of partial label learning and negative learning for enhanced noisy label learning. In *AAAI*, pages 4820–4828, 2024. 3

[51] J.-X. Shi, T. Wei, Z. Zhou, J.-J. Shao, X.-Y. Han, and Y.-F. Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *ICML*, 2024. 2

[52] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, pages 1917–1928, 2019. 2, 3, 7, 8

[53] H. Song, M. Kim, and J.-G. Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *ICML*, pages 5907–5915, 2019. 1

[54] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pages 843–852, 2017. 1

[55] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021. 2

[56] T. Wei, J.-X. Shi, Y.-F. Li, and M.-L. Zhang. Prototypical classifier for robust class-imbalanced learning. In *PAKDD*, pages 44–57, 2022. 2, 3

[57] T. Wei, J.-X. Shi, W.-W. Tu, and Y.-F. Li. Robust long-tailed learning under label noise. *ArXiv*, 2021. 2, 7, 8

[58] Z.-F. Wu, T. Wei, J. Jiang, C. Mao, M. Tang, and Y. Li. Ngc: A unified framework for learning with open-world noisy data. In *ICCV*, pages 62–71, 2021. 8

[59] X. Xia, T. Liu, B. Han, C. Gong, N. Wang, Z. Ge, and Y. Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2020. 7, 8

[60] X. Xia, T. Liu, B. Han, M. Gong, J. Yu, G. Niu, and M. Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. In *ICLR*, 2022. 2, 3

[61] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015. 1

[62] Y. Yao, T. Liu, B. Han, M. Gong, J. Deng, G. Niu, and M. Sugiyama. Dual t: reducing estimation error for transition matrix in label-noise learning. In *NeurIPS*, pages 7260–7271, 2020. 3

[63] Y. Yao, Z. Sun, C. Zhang, F. Shen, Q. Wu, J. Zhang, and Z. Tang. Jo-SRC: A contrastive approach for combating noisy labels. In *CVPR*, pages 5188–5197, 2021. 3

[64] X. Yi, K. Tang, X.-S. Hua, J.-H. Lim, and H. Zhang. Identifying hard noise in long-tailed sample distribution. In *ECCV*, pages 739–756, 2022. 2, 3, 7, 8

[65] M. Zhang, X. Zhao, J. Yao, C. Yuan, and W. Huang. When noisy labels meet long tail dilemmas: A representation calibration method. In *ICCV*, pages 15844–15854, 2023. 1, 2, 3, 7, 8

[66] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, pages 2361–2370, 2021. 2

[67] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng. Deep long-tailed learning: A survey. *IEEE TPAMI*, 45(9):10795–10816, 2023. 1, 7

[68] Z. Zhong, J. Cui, S. Liu, and J. Jia. Improving calibration for long-tailed recognition. In *CVPR*, pages 16489–16498, 2021. 2

[69] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 9719–9728, 2020. 2, 3

[70] X. Zhou, X. Liu, D. Zhai, J. Jiang, X. Gao, and X. Ji. Prototype-anchored learning for learning with imperfect annotations. In *ICML*, volume 162, pages 27245–27267, 2022. 2, 3