Label-Noise Learning with Intrinsically Long-Tailed Data

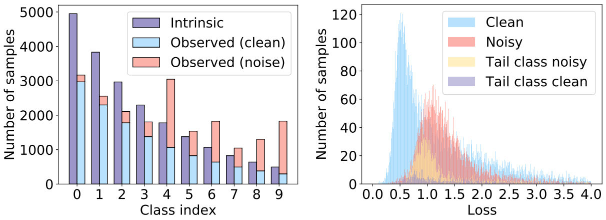
Yang Lu[1,2]* Yiliang Zhang[1,2] Bo Han[3] Yiu-ming Cheung[3] Hanzi Wang[1,2]
[1]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen, China
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, Xiamen, China
[3]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
luyang@xmu.edu.cn ylzhangcs@hotmail.com {bhanml, ymc}@comp.hkbu.edu.hk hanzi.wang@xmu.edu.cn

Abstract

Label noise is one of the key factors that lead to the poor generalization of deep learning models. Existing label-noise learning methods usually assume that the ground-truth classes of the training data are balanced. However, the real-world data is often imbalanced, leading to the inconsistency between observed and intrinsic class distribution with label noises. In this case, it is hard to distinguish clean samples from noisy samples on the intrinsic tail classes with the unknown intrinsic class distribution. In this paper, we propose a learning framework for label-noise learning with intrinsically long-tailed data. Specifically, we propose two-stage bi-dimensional sample selection (TABASCO) to better separate clean samples from noisy samples, especially for the tail classes. TABASCO consists of two new separation metrics that complement each other to compensate for the limitation of using a single metric in sample separation. Extensive experiments on benchmarks demonstrate the effectiveness of our method. Our code is available at https://github.com/Wakings/TABASCO.

1. Introduction

Under the support of a large amount of high-quality labeled data, deep neural networks have achieved great success in various fields [27, 40, 7]. However, it is expensive and difficult to obtain a large amount of high-quality labeled data in many practical applications. Instead, the commonly used large-scale training data is usually obtained from the Internet or crowdsourcing platforms like Amazon Mechanical Turk, which is unreliable and may be mislabeled [52, 43]. The models trained on this kind of unreliable data, called noisy-labeled data, often produce poor generalization performance because deep neural networks tend to overfit noisy samples due to their large model capacity. In the literature, there are some works to obtain a robust model trained on noisy-labeled data [11, 44]. Among them, the most straightforward and effective way is to differentiate between clean and noisy samples based on their differences in specific metrics, such as training loss [12, 29, 23].



(a) Class distribution (b) Loss distribution

Figure 1: (a) An example of observed class distribution with noisy labels with long-tailed intrinsic class distribution. (b) The training loss of each sample under noisy-labeled and long-tailed data.

These label-noise learning methods generally assume that the intrinsic class distribution of the training data is balanced, where each class has almost the same number of samples in terms of their unknown ground-truth labels. However, the data in real-world applications are often imbalanced, e.g., the LVIS dataset [9] and the iNaturalist dataset [17]. The class imbalance usually exhibits in the form of a long-tail distribution, where a small portion of classes possess a large number of samples, and the other classes possess a small number of samples only [38, 18]. In this case, the model training tends to the head classes and ignores the tail classes [60, 62]. When both noisy labels and long-tail distribution exist, training a robust model is even more challenging. There are two key challenges in this scenario. (1) ***Distribution inconsistency***: The observed and intrinsic distributions are likely inconsistent due to noise labels, making the model more difficult to discover and focus on the intrinsic tail classes. As illustrated in Fig. 1(a), the intrinsic class distribution of the dataset is long-tailed, while the existence of noisy labels makes the distribution more

*Yang Lu is the corresponding author: luyang@xmu.edu.cn balanced. The intrinsic tail classes, e.g., classes 6 and 9, are occupied by a large number of noisy data, making them no longer tail classes by observation. (2) ***Tail inseparability***: Even if the tail class is identified, it is more difficult than ever to distinguish between clean and noisy samples in the tail class because clean samples are overwhelmed by noises that make their values of the separation metric highly similar. As illustrated in Fig. 1(b), the training loss of clean and noisy samples in the tail class are generally inseparable compared with the ones in the other classes. Several preliminary works have studied the joint problem of label noise and long-tail distribution [21, 49, 24, 2]. These methods implicitly reduce the complexity of the problem by assuming a similar noise rate for each class. However, this assumption is too strong to apply because noisy samples from the head classes may be the majority, resulting in a higher noise rate in the tail class than in other classes.

In this paper, we propose a Two-stAge Bi-dimensionAl Sample seleCtiOn (TABASCO) strategy to address the problem of label-noise learning with intrinsically long-tailed data. In the first stage, we propose to use two new separation metrics for sample separation, i.e., weighted Jensen-Shannon divergence (WJSD) and adaptive centroid distance (ACD), which work corporately to separate clean samples from noisy samples in the tail classes. The proposed metrics are complementary, where WJSD separates the samples from the output perspective while ACD does that from the feature perspective. In the second stage, we determine the separation dimension with better separability for each class and perform sample selection. In order to evaluate the method uniformly and effectively, we introduce two benchmarks with real-world noise and intrinsically long-tailed distribution. The main contributions of our work can be summarized as follows:

- We present a more general problem of label-noise learning with intrinsically long-tailed data. The key challenges in this problem are distribution inconsistency and tail inseparability.
- We propose an effective solution called TABASCO. With the help of two new separation metrics, it is able to effectively identify and select clean samples of the intrinsic tail class.
- We introduce two benchmarks with real-world noise and intrinsically long-tailed distribution. Extensive experiments on them show the effectiveness of our method and the limitations of existing methods.

2. Related Work

2.1. Label-Noise Learning

A straightforward strategy to deal with noisy data is to reduce the proportion of noise in training samples by separating noisy samples from clean samples. Methods such as co-teaching [12] and DivideMix [29] adopt the small loss trick, while Jo-SRC [55] and UNICON [23] use Jensen-Shannon divergence instead for sample selection. In contrast to methods that separate at the label level, some methods [30, 34] attempt to separate samples at the feature space. There are also some methods to avoid over-fitting the model to noisy data by imposing regularization constraints on model parameters [50, 10] or labels [37, 58, 32]. Other methods mitigate the influence of noisy data by adjusting the loss functions, such as backward and forward loss correction [36], gold loss correction [16], MW-Net [41] and Dual-T [54].

2.2. Long-tail Learning

Re-balancing the data for long-tail distributions is a classical strategy to solve the problem of long-tail learning, such as re-sampling [4, 31, 8, 13] and data augmentation [61, 57, 5]. In addition, there are methods to improve the model generalization by introducing long-tail robust loss functions [6, 3, 45, 39]. Methods such as FTL [56], RIDE [46] and DiVE [15] try to solve the problem by using the idea of transfer learning. Recently, approaches based on decoupling [22, 61, 59] split end-to-end learning into feature learning and classifier retraining such that the obtained feature extractor is less affected by the long-tail distribution. In contrast to the above methods with the supervised learning paradigm, CReST [47], ABC [28] and DARP [25] attempt to solve the long-tail problem in the manner of semi-supervised learning.

2.3. Label-Noise Learning on Long-tailed Data

Research on this joint problem has just been explored. CNLCU [51] relaxes the constraint of the small loss trick by regarding a portion of large loss samples as clean samples to reduce the probability of misclassifying clean samples to noisy samples in the tail class. RoLT [49] uses the distance from the samples to the centroid of the current class instead of the training loss for sample selection. HAR [2] uses an adaptive approach to regularize noise and tail class samples. Karthik et al. [24] uses the idea of decoupling to fine-tune the loss function for better robustness after feature learning by the self-supervised method. ULC [19] introduces uncertainty to enhance the separation ability of noise samples. H2E [53] reduces hard noises to easy ones by learning a classifier as noise identifier invariant to the class and context distributional changes.

3. Problem Definition

Given a training set $\mathcal{D} = \{x_i, \hat{y}_i\}_{i=1}^N$, where $\hat{y}_i \in [M]$ is the observed label of the sample x_i. N is the number of training samples, and M is the number of classes. In our problem, $\mathcal{D}$ has the following properties:

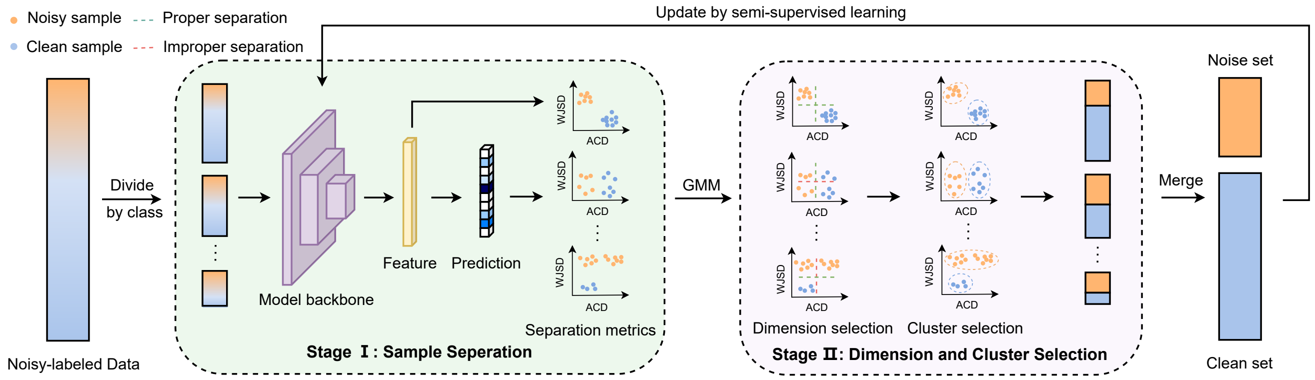


Figure 2: The proposed framework for label-noise learning with intrinsically long-tailed data.

- Noisy-labeled. There is a subset of samples $\hat{\mathcal{D}} \in \mathcal{D}$, where a sample $\{x, \hat{y}\} \in \hat{\mathcal{D}}$ has an unknown ground-truth label y different from its observed label $\hat{y}$.
- Long-tailed. The ground-truth class distribution is long-tailed. Supposing n_c is the number of the samples in class c, we have $n_1 > n_2 > \cdots > n_M$.

We call this kind of training data *intrinsically long-tailed* because the observed class distribution may not be long-tailed due to the existence of noisy labels. Therefore, our goal is to learn from noisy-labeled data with the intrinsically long-tailed class distribution. The problem is challenging from two aspects. On the one hand, directly applying long-tail learning methods [6, 3, 45] is infeasible because the observed class distribution may be inconsistent with the intrinsic class distribution. On the other hand, label-noise learning methods perform poorly in the tail classes because they usually contain more noisy samples due to insufficient data, as shown in Fig. 1(a). It results in high noise rates in the tail classes, which brings great challenges to separating clean samples and noisy samples for further training. Under this circumstance, the existing sample selection methods [29, 23] often fail to select clean samples in the tail class. The main reason is that it is difficult to distinguish the tail class data from the noisy label data by existing metrics like cross-entropy loss, as shown in Fig. 1(b).

4. Proposed Method

In this paper, we propose a two-stage bi-dimensional sample selection (TABASCO) to address the problem of label-noise learning with intrinsically long-tailed data. TABASCO decouples the sample selection process into two stages: (1) sample separation, and (2) dimension and cluster selection. First, given an initial model θ trained on the original training data $\mathcal{D}$, we propose to calculate bi-dimension metrics for each sample in each observed class based on the outputs and features of the model. Second, we determine the separation dimension with better separability for each class, and we adopt the corresponding selection strategy to select the cluster with more clean samples based on the selected metric. Last, we adopt semi-supervised learning to update the model by regarding the selected clean cluster as labeled data. The overall framework is shown in Fig. 2.

4.1. Bi-Dimensional Separation Metrics

Due to the deficiency of a single separation metric to distinguish clean samples from noisy samples in the complex situation of noisy labels with intrinsic long-tail distribution, we propose to jointly use two metrics from different perspectives: weighted Jensen-Shannon divergence (WJSD) and adaptive centroid distance (ACD). Both metrics are specifically designed for the joint problem and are complementary to cover the case when only one of them cannot separate clean samples from noisy samples well. WJSD fully utilizes the information of prediction confidence, while ACD relies on the distance in the feature space. Thus, using clustering on samples according to the values of their bi-dimensional metrics has the flexibility to separate samples with different imbalance ratios, noise ratios, and noise types.

We first reduce the separation granularity to alleviate the interference from the head class to the tail class. Specifically, the separation is performed within each observed class according to the proposed bi-dimensional metrics. We first divide training set $\mathcal{D}$ into subsets according to the observed labels $\mathcal{D}_c = \{(x, \hat{y}) \mid \hat{y} = c\}$. Given a sample $(x_i, \hat{y}_i) \in \mathcal{D}_c$, the model predictions obtained from the model θ is denoted as $\mathbf{p}_i = [p_i^1, p_i^2, ..., p_i^M]$, where p_i^j is the j's dimension of vector $\mathbf{p}_i$. The one-hot representation of the observed label $\hat{y}_i$ is denoted as $\hat{\mathbf{y}}_i$.

Weighted Jensen-Shannon Divergence The Jensen-Shannon Divergence (JSD) is a commonly used metric to separate samples by assessing the variability of model predictions [55, 23]. It is defined as:

$$JSD(x_i) = \frac{1}{2}KL\Big(\mathbf{p}_i \Big\| \frac{\mathbf{p}_i + \hat{\mathbf{y}}_i}{2}\Big) + \frac{1}{2}KL\Big(\hat{\mathbf{y}}_i \Big\| \frac{\mathbf{p}_i + \hat{\mathbf{y}}_i}{2}\Big), \tag{1}$$

where $KL(\cdot\|\cdot)$ is the Kullback-Leibler divergence. When the intrinsic distribution is balanced, the JSD values of clean samples are generally lower than the noisy samples, so the separation can be easily modeled. However, when the distribution is intrinsically long-tailed, the noisy and clean samples in a tail class tend to obtain similar prediction confidence on the observed class c, because their predictions are both towards the head classes. In this case, using JSD as a separation metric fails to separate them. The reason is given by Theorem 1, which shows the upper bound of the absolute value of the JSD difference between two samples.

Theorem 1 (Upper bound of JSD difference) *Suppose x_i and x_j are two samples in class c, p_i^c and q_i^c are the c's dimension of their prediction confidence $\mathbf{p}_i = [p_i^1, p_i^2, ..., p_i^M]$ and $\mathbf{p}_j = [p_j^1, p_j^2, ..., p_j^M]$, respectively. The upper bound of the absolute value of the difference between their JSD values is given by:*

$$|JSD(x_i) - JSD(x_j)| \leq \frac{1}{2}\log\left(\frac{p_i^c + 1}{p_i^c}\right)|p_i^c - p_j^c|.$$

Theorem 1 shows that the difference between values of JSD is only determined by the prediction confidence of the observed class, i.e., p_i^c and p_j^c. In addition, as $|p_i^c - p_j^c|$ gets smaller with a fixed value of p_i^c, $|JSD(x_i) - JSD(x_j)|$ is also smaller. This indicates that when two samples are in the same class c with close values of p_i^c, their JSD values are also close, which makes it difficult to separate them.

Nevertheless, we can utilize the prediction confidence on other classes, i.e., p_i^d for $d \neq c$, to distinguish two samples even if their values of p_i^c are close. Specifically, we propose WJSD by imposing an additional weight on JSD to further distinguish samples by inspecting their prediction confidence on other classes rather than the observed class. First, we take the maximum prediction confidence $\max(\mathbf{p}_i)$ into account because it may reflect the confidence of a noisy sample's ground-truth class y_i. Then, we calculate the ratio of the maximum prediction confidence to the prediction confidence of the observed class $\max(\mathbf{p}_i)/p_i^c$. In this manner, the additional weight is greater than one only if the class with the maximum prediction confidence is not the observed class. The larger gap between $\max(\mathbf{p}_i)$ and p_i^c makes the weight higher. To avoid exceptionally large weights by the division during normalization over all samples in class c for later clustering, we set the upper bound according to the averaged prediction confidence over all samples in class c. Finally, the additional weight can be calculated as follows:

$$W(x_i) = \min(\max(\mathbf{p}_i)/p_i^c, \max(\bar{\mathbf{p}}_c)/\bar{p}_c^c), \tag{2}$$

where $\bar{\mathbf{p}}_c = [\bar{p}_c^1, \bar{p}_c^2, ..., \bar{p}_c^M] = \frac{1}{|\mathcal{D}_c|}\sum_{i=1}^{|\mathcal{D}_c|}\mathbf{p}_i$ is the average prediction confidence of class c. Thus, WJSD can be calculated by:

$$WJSD(x_i) = W(x_i) \times JSD(x_i). \tag{3}$$

Remarks: Compared with the value of JSD that is only related to the prediction confidence on the observed class, adopting the maximum prediction confidence in WJSD can better separate clean samples from noisy samples.

Adaptive Centroid Distance. Although the sample separability of WJSD is greatly improved compared with JSD, the separation metric still relies on model prediction. When all the noisy samples labeled in a class are from another class, e.g., asymmetric noise, the model prediction for both noisy and clean samples will be highly similar because it is easy to learn a classifier that maps the features from two different classes into one. It results in low discrimination between the model predictions of clean and noisy samples in this case. Thus, solely using WJSD may not be enough to separate the clean and noisy samples when model predictions (in terms of all classes, not only about p^c) are close.

Except for separating the samples in the output space, we propose another metric calculated in the feature space to eliminate the bias from the classifier because the learned features are more robust to noise labels. Specifically, for a given sample, we can calculate the distance between its feature and the class feature centroid to evaluate how the feature of a sample deviates from its class centroid. This approach is only practical when the quality of the centroid is high. Directly calculating the centroid according to the observed label may not be accurate because it involves a certain number of features of the noisy samples. Therefore, we define purity in Definition 1 to assess the quality of the centroid of class c.

Definition 1 (Purity) *Suppose a noisy sample set of class c be $\mathcal{D}_c = \{(x, \hat{y}) \mid \hat{y} = c\}$, where the intrinsic label corresponding to the sample is y. The purity $P_{\mathcal{D}_c}$ of set $\mathcal{D}_c$ is calculated by:*

$$P_{\mathcal{D}_c} = \max_{k=1,...,M}\left\{\frac{\sum_{i=1}^{N} I(y_i = k)}{N}\right\}, \tag{4}$$

where $I(\cdot)$ is an indicator function and N is the number of samples in the set $\mathcal{D}_c$.

Purity indicates the proportion of an intrinsic class that takes the majority in the observed class c. The higher the purity of a class set used to calculate the centroid, the better the centroid helps to distinguish between noise and clean samples. RoLT [49] adopts a similar idea to use class feature centroid for noisy sample detection. However it directly calculates the feature centroid on the observed class, such that it inevitably suffers from the existence of noise features when the purity is low, especially in our problem.

In order to improve the purity of the class feature centroid for distance calculation between features, we propose ACD as the second separation metric to separate samples jointly with WJSD. The feature centroid for each class is adaptively updated by involving samples with high confidence from the observed class. The class centroid $\mathbf{o}_c$ based on a high-confidence sample set $\mathcal{D}_c^H$ is calculated by:

$$\mathbf{o}_c = \frac{1}{|\mathcal{D}_c^H|} \sum_{i=1}^{|\mathcal{D}_c^H|} \mathbf{f}_i, \tag{5}$$

$$\mathcal{D}_c^H = \{x_i | x_i \in \mathcal{D}_c,\ p_i^{t_c} > H_c\}, \tag{6}$$

where $\mathbf{f}_i$ is the feature of x_i, and $t_c = \text{argmax}_c\{\bar{p}_c^j\}$ is the class index with the largest average prediction confidence of class c. Thus, we can use the prediction confidence of class t_c for sample x_i, e.g., $p_i^{t_c}$, as the selection criteria compare with the threshold H_c. $\mathcal{D}_c^H$ is constructed by the samples in D_c whose corresponding prediction confidence of class t_c is higher than H_c. The high-confidence threshold H_c is defined as:

$$H_c = \frac{1}{D_c} \sum_{i=1}^{|D_c|} w_i \times p_i^{t_c}, \tag{7}$$

$$w_i = \max\left(1, p_i^{t_c}/\bar{p}_c^{t_c}\right). \tag{8}$$

H_c is calculated by the weighted sum of the prediction confidence of class t_c for all samples x_i. The weight w_i increases when a sample's prediction confidence of class t_c is higher than its class average. Thus, the threshold is high enough to ensure the purity of the selected samples. High confidence samples are more representative and more likely to be in the same class. It is worth noting that we choose more representative samples rather than clean samples because the clean samples in the tail classes are easily overwhelmed by the noise samples in our problem. In this case, it is difficult to select the clean samples directly because the noise samples may be more representative in the tail class. As for asymmetric noise, the choice of clean or noisy samples to obtain centroid for sample separation is equivalent.

In each round, the centroids are adaptively adjusted because the feature extractor in the model is updated, as well as the set $\mathcal{D}_c^H$. Therefore, the proposed metric ACD can be calculated by:

$$ACD(x_i) = \cos(\mathbf{f}_i, \mathbf{o}_c). \tag{9}$$

Remarks: The proposed two metrics, WJSD and ACD, are complementary for sample separation in the tail classes. When the ground-truth classes of the noisy samples are diverse, e.g. symmetric noise, WJSD plays the dominant role because the predictions can hardly be unified due to the class diversity of noisy samples. They are more likely to be predicted towards their ground-truth class, which can be captured by $\max(\mathbf{p}_i)$. In this case, ACD may not be effective in separating clean samples from noisy samples because the messy noisy samples may affect the purity of the calculated centroid. On the other hand, when the ground-truth classes of the noisy samples belong to only one class, e.g., asymmetric noise, WJSD tends to produce similar predictions for noisy and clean samples, while the adaptive centroid in ACD can help distinguish them in the feature space.

Algorithm 1: The Dimension Selection Strategy

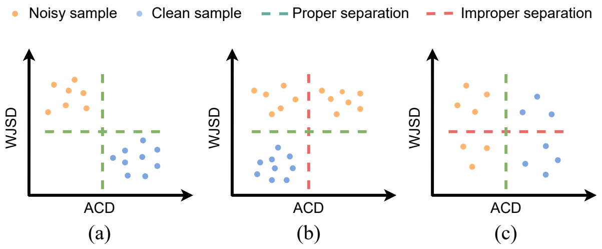
```
Input: Noise sample set $\mathcal{D}_c$ in class $c$, threshold $\eta$
Output: Cluster $\mathcal{G}_c^1$, $\mathcal{G}_c^2$
Obtain $\mathcal{G}_{wjsd}^1$, $\mathcal{G}_{wjsd}^2$ and separation threshold $d$ by
  applying GMM with values of WJSD to all
  samples in $\mathcal{D}_c$
Obtain $\mathcal{G}_{acd}^1$, $\mathcal{G}_{acd}^2$ by applying GMM with values of
  ACD to all samples in $\mathcal{D}_c$
Calculate the mean $\mu_1, \mu_2$ and variance $\sigma_1, \sigma_2$ of
  WJSD for $\mathcal{G}_{acd}^1, \mathcal{G}_{acd}^2$
if $\mu_1 < d < \mu_2$ and $\sigma_2/\sigma_1 < \eta$ then
    return $\mathcal{G}_{wjsd}^1, \mathcal{G}_{wjsd}^2$
else if $\mu_1 > d$ and $\mu_2 > d$ then
    return $\mathcal{G}_{wjsd}^1, \mathcal{G}_{wjsd}^2$
else
    return $\mathcal{G}_{acd}^1, \mathcal{G}_{acd}^2$
end
```

4.2. Bi-Dimensional Sample Selection

Once the values of the bi-dimensional metrics for all samples in class c are calculated, each sample can be represented by a point in a 2D space. For each dimension, Gaussian mixture model (GMM) can be adopted to separate the samples in class c into two clusters $\mathcal{G}_c^1, \mathcal{G}_c^2$.

Dimension Selection. Because each metric shows different separability for different noise types, we first propose a dimension selection strategy to select a proper dimension to separate clean and noisy samples. We consider three cases to select the proper dimension: (a) both of them are acceptable; (b) the separability of WJSD is better; (c) the separability of ACD is better. In short, we only need to figure out if WJSD is suitable for sample separation. If it is certain that WJSD does not show good separability, ACD will be adopted. Fig. 3 shows three cases for dimension selection.

Therefore, we use the statistics of WJSD as the criterion of dimension selection. We measure the means (μ_1, μ_2) and standard deviations (σ_1, σ_2) of WJSD for two clusters separated by GMM in terms of ACD. We have two cases that can determine when WJSD is better. We mainly compare



(a) (b) (c)

Figure 3: Examples of three cases of the optimal separation dimension correspond to sample distribution in bi-dimensional metrics.

μ_1 and μ_2 with the threshold d of GMM in WJSD. First, as shown in Fig. 3 (b), if one cluster $\mathcal{G}_{acd}^1$ of ACD has a μ_1 less than d and a σ_1 much larger than σ_2 in another cluster with a μ_2 larger than d, it means that cluster $\mathcal{G}_{acd}^1$ has more clean samples but also many noise samples. So in this case, ACD is improper and WJSD is proper. Second, if both μ_1, μ_2 are greater than d, it means that it is a representation of the previous case in the tail class because there are fewer clean samples in the cluster. So, in this case, WJSD is proper. The strategy is summarized in Algorithm 1.

Cluster Selection. Once we have determined the separation dimension, we can obtain two sample clusters $\mathcal{G}_c^1$ and $\mathcal{G}_c^2$. However, it is not clear which cluster contains more clean samples because the noisy samples may overwhelm the clean samples in some cases. Therefore, we need to further select a cluster between $\mathcal{G}_c^1$ and $\mathcal{G}_c^2$ as the clean one. For the cluster separated based on WJSD, we simply choose the one with a smaller average WJSD value as the cluster with more clean samples, denoted as $\mathcal{D}_c^{clean}$ [29, 23]. For the cluster separated based on ACD, we cannot directly select the cluster closest to the centroid as the cluster with more clean samples since we take more representative samples, which may compose of noisy samples, rather than clean samples as the centroid. Therefore, we need to determine the selection criteria based on whether the current class centroid is obtained by clean samples. Especially, the criterion is based on the similarity between centroids of the current class and other classes. Assume that the sample in $\mathcal{G}_c^1$ is near the centroid and the sample in $\mathcal{G}_c^2$ is the opposite. We determined the choice of a more clean cluster $\mathcal{D}_c^{clean}$ based on the following criteria:

$$\mathcal{D}_c^{clean} = \begin{cases} \mathcal{G}_c^2, & \text{if } |\cos(o_c, o_k) - 1| < \varepsilon \text{ and } \left|\mathcal{D}_c^H\right| < \left|\mathcal{D}_k^H\right| \\ \mathcal{G}_c^1, & \text{otherwise} \end{cases} \tag{10}$$

It means that if the centroid of another class is similar to the centroid of the current class, whichever class of high-confidence sample set has fewer samples, then the high-confidence samples in this class are the noise samples. It is because the asymmetric noise ratio cannot exceed 50% in the past assumption [29, 23]. Therefore, if the centroid is obtained from noise samples, we choose the cluster far away from the centroid as the cluster with more clean samples; otherwise, the cluster close to the centroid is selected.

4.3. Overall Training Process

After sample selection is conducted for each class, we adopt semi-supervised learning [1, 42] to train with all clean samples $\mathcal{D}^{clean}$ as labeled data and all noisy samples $\mathcal{D}^{noisy}$ as unlabeled data. The model is thus updated for the next round of training. One may also use long-tailed semi-supervised learning methods instead of normal semi-supervised learning in this stage.

5. Experiments

5.1. Experimental Setup

Datasets. In order to comprehensively evaluate the effectiveness of method, we perform experiments respectively in the scenarios with synthetic noise and realistic noise in intrinsically long-tailed distribution. We construct synthetic scenarios based on CIFAR-10 and CIFAR-100 [26]. We construct two benchmarks to imitate realistic scenarios, which is built on real-world noise datasets including Red Mini-ImageNet [20], CIFAR-10N and CIFAR-100N [48]. For all datasets, we adopt the standard paradigm of constructing a long-tailed distribution first and injecting noise later. We imitate realistic long-tailed distribution using the same setting in previous long-tailed learning works [3, 35] that long-tailed imbalance follows an exponential decay in sample sizes across different classes. The imbalance factor is denoted by the ratio between the size of the largest class and that of the smallest class.

The details for dataset construction are as follows. (1) As for CIFAR-10/100 [26], we adopt symmetric and asymmetric noise to inject synthetic noise, which is commonly used in the area of label-noise learning [41]. The noise ratio is denoted by the ratio between the number of noisy samples and the total number of samples. It should be noted that different from the previous methods [21, 49] to deal with the joint problem of noisy labeled and long-tailed data, the noise transition matrix is randomly generated only related to the noise ratio. (2) As for CIFAR-10N/100N (10N/100N) [48], we replace sample labels with human-annotated noisy labels after constructing a long-tailed distribution based on real labels. (3) As for Red mini-ImageNet (Red) [20], we inject web label noise samples after constructing a long-tailed distribution based on the original mini-ImageNet. Therefore, the observed and intrinsic distribution are likely inconsistent, especially in long-tail distribution.

Compared Methods. We compare our method with the following three types of approaches: (1) Long-tail learning methods (LT). They are LA [33], LDAM [3] and IB [35]. (2) Label-noise learning methods (NL). They are DivideMix [29] and UNICON [23]; (3) Methods aiming at dealing with noisy label and long-tail distribution (NL-LT). They are MW-Net [41], RoLT [49], HAR [2] and ULC [19].

Implementation Details. We use PreAct ResNet18 [14] as backbone for CIFAR datasets and ResNet18 as backbone for Red mini-ImageNet dataset. Both of backbones adopt SGD with an initial learning rate of 0.02, a momentum of 0.9, and a weight decay of 5×10^{-4} and a batch size of 64. For fair comparison, we train all methods for 100 epochs.

5.2. Comparative Results

CIFAR-10/100. Tab. 1 and 2 report the accuracy of different methods for intrinsically long-tailed CIFAR-10/100

Dataset		CIFAR-10		CIFAR-100	
Imbalance Factor		0.1			
Noise Ratio (**Sym.**)		0.4	0.6	0.4	0.6
Baseline	CE	71.67	61.16	34.53	23.63
LT	LA	70.56	54.92	29.07	23.21
	LDAM	70.53	61.97	31.30	23.13
	IB	73.24	62.62	32.40	25.84
NL	DivideMix	82.67	80.17	54.71	44.98
	UNICON	84.25	82.29	52.34	45.87
NL-LT	MW-Net	70.90	59.85	32.03	21.71
	RoLT	81.62	76.58	42.95	32.59
	HAR	77.44	63.75	38.17	26.09
	ULC	84.46	83.25	54.91	44.66
Our	TABASCO	**85.53**	**84.83**	**56.52**	**45.98**

Table 1: Performance comparison under symmetric noise. The best results are shown in bold.

Dataset		CIFAR-10		CIFAR-100	
Imbalance Factor		0.1			
Noise Ratio (**Asym.**)		0.2	0.4	0.2	0.4
Baseline	CE	79.90	62.88	44.45	32.05
LT	LA	71.49	59.88	39.34	28.49
	LDAM	74.58	62.29	40.06	33.26
	IB	73.49	58.36	45.02	35.25
NL	DivideMix	80.92	69.35	58.09	41.99
	UNICON	72.81	69.04	55.99	44.70
NL-LT	MW-Net	79.34	65.49	42.52	30.42
	RoLT	73.30	58.29	48.19	39.32
	HAR	82.85	69.19	48.50	33.20
	ULC	74.07	73.19	54.45	43.20
Our	TABASCO	**82.10**	**80.57**	**59.39**	**50.51**

Table 2: Performance comparison under asymmetric noise. The best results are shown in bold.

with symmetric and asymmetric noise, respectively. It can be observed three phenomena as follows: (1) The performance of existing long-tail methods is lower than baseline in most situations. It is because such methods do not have the ability to distinguish the noise samples, and the wrong attention to the noise samples leads to the further decline of the model generalization ability. (2) The advantages of existing methods aiming at both long-tailed distribution and label noise are not obvious compared with single methods. It is because existing methods do not take into account the distribution inconsistency caused by noise labels. (3) Our method significantly improved over other methods in all cases. Moreover, as the proportion of noise increases, the accuracy of our method decreases less compared to other methods. It validates that our method can effectively identify the noise in long-tail distribution.

Red mini-ImageNet and CIFAR-10N/100N. Tab. 4 reports the accuracy of different methods on real-world noise datasets with intrinsically long-tailed distribution. According to the results of label-noise learning methods on CIFAR-10N/100N, it can be observed that real-world noise is more difficult to separate than synthetic noise to some extent. It further increases the difficulty of dealing with such problem. Nevertheless, our method still performs well, which again confirms its effectiveness.

5.3. Ablation Study and Discussions

In the following experiments, we analyze each component of the proposed TABASCO on CIFAR-10/100 to verify its effectiveness. We use an imbalance factor of 0.1 and a noise ratio of 0.4 to construct the synthetic noise dataset.

Effectiveness of WJSD. To validate the effectiveness of the proposed WJSD, we compare the accuracy of models trained by different separation metrics. Specifically, we train the model by using JSD and WJSD for sample separation with a symmetric noise dataset, respectively. As for each separation metric, we simply use the small value strategy [29, 23] for sample selection. As Tab. 3 shows, the model trained by WJSD achieves better results on both symmetric noise datasets compared with JSD owing to the stronger separability of WJSD.

Dataset	Separation metric	Accuracy
CIFAR-10	JSD	83.97
	WJSD	85.57
CIFAR-100	JSD	55.80
	WJSD	56.67

Table 3: Performance comparison between JSD and WJSD.

Effectiveness of ACD. To validate the advantage of the proposed ACD, we compare it with the method which calculates the feature centroid on the observed class (hereinafter referred to as centroid distance or CD) to show their separability in CIFAR-10. Specifically, based on the same backbone, we calculate the distance from the sample to the centroid using ACD and CD in CIFAR-10 with asymmetric noise, respectively. Fig. 4 shows the distributions of samples in the tail class by CD and ACD. The purity of the tail class is 0.53. It can be observed that the CD values of noise and clean samples are highly overlapped, which leads to failure sample separation. It is because the purity of the tail class may be quite low due to the negative effects of both label-noise and long-tailed distribution. The centroid obtained by ACD has a higher purity, so there is a significant difference in ACD between them. As shown in Fig. 4(b), most clean samples are concentrated in the interval of [0.8,1], which is easily clustered by GMM.

Dataset		Red		10N	100N
Imbalance Factor		≈ 0.1		0.1	
Noise Ratio		0.2	0.4	≈ 0.4	
Baseline	CE	40.42	31.46	60.44	38.10
LT	LA	26.82	25.88	65.74	36.50
	LDAM	26.64	23.46	62.50	38.48
	IB	23.80	22.08	65.91	42.48
NL	DivideMix	48.76	48.96	67.85	44.25
	UNICON	40.18	41.64	69.54	51.93
NL-LT	MW-Net	42.66	40.26	69.73	44.20
	RoLT	22.56	24.22	75.24	46.61
	HAR	46.61	38.71	74.97	44.54
	ULC	48.12	47.06	75.71	51.72
Our	TABASCO	**50.20**	**49.68**	**80.61**	**53.83**

Table 4: Performance comparison with real-world noise and long-tail distribution. The best results are shown in bold.

Effectiveness of Bi-dimensional Sample Separation. To validate how the proposed bi-dimensional metrics complement each other, we plot the values of bi-dimensional metrics of both clean and noisy samples under different noise types in Fig. 5. For the case of symmetric noise shown in Fig. 5(a), it can be observed that ACD cannot effectively distinguish clean samples from noisy samples in the tail class. In this case, WJSD shows its advantage because the values of WJSD for most noisy samples are clustered in the top region, and the clean samples are scattered throughout the rest region. For the case of asymmetric noise shown in Fig. 5(b), it can be observed that WJSD cannot distinguish clean samples from noisy samples well for the tail class, while ACD can distinguish them well although the clean and noisy samples are closer in the tail class. The experimental observation is consistent with the previous discussion, which validates the complementarity of bi-dimensional metrics.

Effectiveness of Sample Selection. In order to validate the effectiveness of the proposed sample selection in dimension selection and cluster selection, we compare the optimal clean ratio of clusters obtained by WJSD, ACD and our selection method in CIFAR-10 with different noise types. The clean ratio of a cluster is calculated by the proportion of clean samples in the cluster to the total number of cluster samples. As shown in Fig. 6, on the tail class, the clean ratio of the cluster obtained by our selection method is aligned with the higher one between WJSD and ACD for both noise cases. It shows that we can choose the most appropriate separation dimension in different situations. Besides, the clean ratio of the cluster we selected is consistent with that of the optimal cluster in the corresponding dimension, which is sufficient to show that the method we proposed can select clean samples.

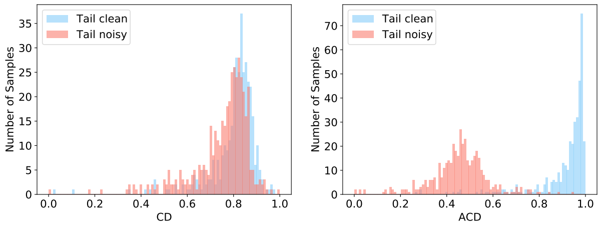
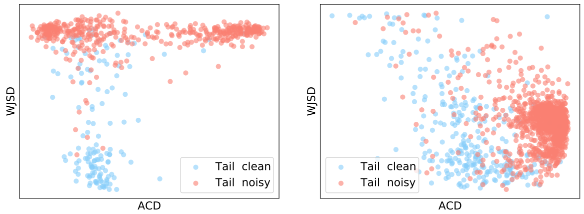
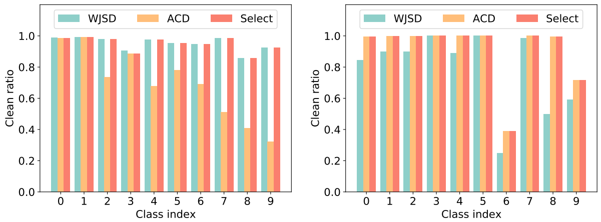


Figure 4: Comparison of the sample distributions between CD and ACD in the tail class.



(a) Symmetric noise (b) Asymmetric noise

Figure 5: Scatter plot of the values of the proposed bi-dimensional metrics with (a) symmetric noise and (b) asymmetric noise.



(a) Symmetric noise (b) Asymmetric noise

Figure 6: The performance of the proposed sample selection with (a) symmetric noise and (b) asymmetric noise.

6. Conclusion

This paper studies a more general and realistic problem of label-noise learning with intrinsically long-tailed data. The major challenge in this problem is that it is hard to distinguish clean samples from noisy samples on intrinsic tail classes. Accordingly, we propose a learning framework TABASCO for this problem. In TABASCO, two new metrics are explicitly proposed to address the problem of sample selection in tail classes. Extensive experiments on noisy-labeled datasets with long-tail distribution demonstrate its effectiveness.

Acknowledgements. We thank Dr. Shiquan Ruan for his discussion on the theory proof. This study was supported in part by the National Natural Science Foundation of China under Grants 62002302 and U21A20514; in part by the FuXiaQuan National Independent Innovation Demonstration Zone Collaborative Innovation Platform under Grant 3502ZCQXT2022008; in part by NSFC/RGC Joint Research Scheme under Grant N_HKBU214/21; and in part by the General Research Fund of RGC under Grants 12201321 and 12202622.

References

[1] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019. 6

[2] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Aréchiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *ICLR*, 2021. 2, 6

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 2, 3, 6

[4] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 2002. 2

[5] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. In *ECCV*, 2020. 2

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2, 3

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1

[8] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Comput. Intell.*, 2004. 2

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1

[10] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor W. Tsang, and Masashi Sugiyama. SIGUA: forgetting may make learning with noisy labels more robust. In *ICML*, 2020. 2

[11] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W. Tsang, James T. Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. *CoRR*, 2020. 1

[12] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 1, 2

[13] Hui Han, Wenyuan Wang, and Binghuan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *ICIC 2005*, 2005. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[15] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *ICCV*, 2021. 2

[16] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, 2018. 2

[17] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1

[18] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *CoRR*, 2017. 1

[19] Yingsong Huang, Bing Bai, Shengwei Zhao, Kun Bai, and Fei Wang. Uncertainty-aware learning against label noise on imbalanced datasets. In *AAAI*, 2022. 2, 6

[20] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *ICML*, 2020. 6

[21] Shenwang Jiang, Jianan Li, Ying Wang, Bo Huang, Zhang Zhang, and Tingfa Xu. Delving into sample loss curve to embrace noisy and imbalanced data. In *AAAI*, 2022. 2, 6

[22] 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. 2

[23] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *CVPR*, 2022. 1, 2, 3, 6, 7

[24] Shyamgopal Karthik, Jérôme Revaud, and Chidlovskii Boris. Learning from long-tailed data with noisy labels. *CoRR*, 2021. 2

[25] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *NeurIPS*, 2020. 2

[26] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 6

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 1

[28] Hyuck Lee, Seungjae Shin, and Heeyoung Kim. ABC: auxiliary balanced classifier for class-imbalanced semi-supervised learning. In *NeurIPS*, 2021. 2

[29] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020. 1, 2, 3, 6, 7

[30] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *CVPR*, 2022. 2

[31] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B*, 2009. 2

[32] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, 2020. 2

[33] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 6

[34] Diego Ortego, Eric Arazo, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *CVPR*, 2021. 2

[35] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *ICCV*, 2021. 6

[36] 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*, 2017. 2

[37] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. In *ICLR*, 2017. 2

[38] William J Reed. The pareto, zipf and other power laws. *Economics Letters*, 74(1):15–19, 2001. 1

[39] 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*, 2020. 2

[40] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[41] 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*, 2019. 2, 6

[42] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 6

[43] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: refurbishing unclean samples for robust deep learning. In *ICML*, 2019. 1

[44] Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *CoRR*, 2020. 1

[45] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *CVPR*, 2020. 2, 3

[46] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021. 2

[47] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan L. Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *CVPR*, 2021. 2

[48] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *ICLR*, 2022. 6

[49] Tong Wei, Jiang-Xin Shi, Wei-Wei Tu, and Yu-Feng Li. Robust long-tailed learning under label noise. *CoRR*, 2021. 2, 4, 6, 14

[50] 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*, 2021. 2

[51] 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. 2

[52] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 1

[53] Yi Xuanyu, Tang Kaihua, Hua Xian-Sheng, Lim Joo-Hwee, and Zhang Hanwang. Identifying hard noise in long-tailed sample distribution. In *ECCV*, 2022. 2

[54] 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*, 2020. 2

[55] 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*, 2021. 2, 3

[56] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, 2019. 2

[57] Yuhang Zang, Chen Huang, and Chen Change Loy. FASA: feature augmentation and sampling adaptation for long-tailed instance segmentation. In *ICCV*, 2021. 2

[58] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[59] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021. 2

[60] Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *AAAI*, 2021. 1

[61] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, 2021. 2

[62] Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. Cross-domain empirical risk minimization for unbiased long-tailed classification. In *AAAI*, 2022. 1