Stochastic Feature Averaging for Learning with Long-Tailed Noisy Labels

Hao-Tian Li[1,2,3] , **Tong Wei**[1,2*] , **Hao Yang**[3] , **Kun Hu**[3] , **Chong Peng**[3] ,
Li-Bo Sun[3] , **Xun-Liang Cai**[3] , **Min-Ling Zhang**[1,2]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Lab. of Computer Network and Information Integration (Southeast University), MOE, China
[3]Meituan, Shanghai, China
{liht, weit, zhangml}@seu.edu.cn, {yanghao52, hukun05, pengchong, sunlibo03, caixunliang}@meituan.com

Abstract

Deep neural networks have shown promising results on a wide variety of tasks using large-scale and well-annotated training datasets. However, data collected from real-world applications can suffer from two prevalent biases, i.e., long-tailed class distribution and label noise. Previous efforts on long-tailed learning and label-noise learning can only address a single type of data bias, leading to a severe deterioration of their performance. In this paper, we propose a distance-based sample selection algorithm called Stochastic Feature Averaging (SFA), which fits a Gaussian using the exponential running average of class centroids to capture uncertainty in representation space due to label noise and data scarcity. With SFA, we detect noisy samples based on their distances to class centroids sampled from this Gaussian distribution. Based on the identified clean samples, we then propose to train an auxiliary balanced classifier to improve the generalization for the minority class and facilitate the update of Gaussian parameters. Extensive experimental results show that SFA can enhance the performance of existing methods on both simulated and real-world datasets. Further, we propose to combine SFA with the sample-selection approach, distribution-robust, and noise-robust loss functions, resulting in significant improvement in performance over the baselines. Our code is available at https://github.com/HotanLee/SFA.

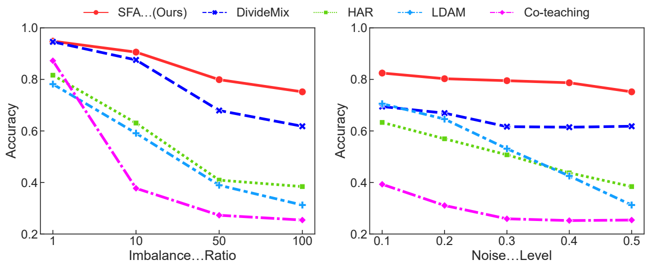


Figure 1: (left) Test accuracy of different approaches under a fixed noise level 50% but various imbalance ratios and (right) a fixed imbalance ratio 100 but various noise levels. Our proposed approach, SFA, significantly outperforms existing methods in various settings.

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success on a wide variety of tasks by leveraging large-scale and well-annotated datasets. Nevertheless, data collected from real-world applications usually follow a *long-tailed class distribution*, i.e., most classes are associated with only a small amount of training data, leading to inferior generalization on the minority class [Zhou *et al.*, 2020; Xiang *et al.*, 2020; Menon *et al.*, 2020; Wei and Li, 2020; Cui *et al.*, 2021]. On the other hand, data annotated by human labelers or web crawling are easily corrupted (*label noise*) in practice. When training with label noise, over-parameterized DNNs can achieve perfect training accuracy due to the memorization effect but cannot generalize [Arpit *et al.*, 2017; Han *et al.*, 2018; Liu *et al.*, 2020; Xia *et al.*, 2021; Wu *et al.*, 2021; Wei *et al.*, 2022b; Zhou, 2022]. Hence, it is crucial to tackle those two types of data biases, i.e., the long-tailed class distribution and label noise, to train robust DNNs.

Many methods for long-tailed learning (LTL) and label-noise learning (LNL) have been proposed to address those two common types of data biases. LTL aims to deal with the imbalanced class distribution, and LTL methods have been evolving in three main directions: (1) re-balancing the training data [Shen *et al.*, 2016; Liu *et al.*, 2019; Zhou *et al.*, 2020]; (2) adjusting the outputs of a model [Kang *et al.*, 2020; Menon *et al.*, 2020; Tang *et al.*, 2020]; and (3) designing distribution-robust loss functions [Cao *et al.*, 2019; Jamal *et al.*, 2020; Ren *et al.*, 2020]. However, existing LTL methods do not take the presence of label noise into account. On the other hand, LNL aims to deal with the problem of label noise by (1) designing noise-robust loss functions [Ghosh *et al.*, 2017; Liu *et al.*, 2020] or (2) detecting and cleaning the noisy data [Jiang *et al.*, 2018; Li *et al.*, 2020]. As one of the most commonly used criteria, a training sample is suggested as noisy data if the loss value between the prediction and its label is higher than a threshold [Han *et al.*, 2018; Arazo *et al.*, 2019; Xia *et al.*, 2021]. However, these methods can detect many false noisy data because even clean data of

[*]Corresponding author

the minority class has large losses.

To mitigate such issues for detecting label noise under long-tailed class distribution, we propose a novel and practical framework, *stochastic feature averaging* (SFA). SFA provides a high-quality splitting of clean and noisy data for both the majority and minority classes. Instead of using the loss values, SFA utilizes the distance distribution between samples and the estimated class centroids in the latent representation space. We find that the class centroid may be inaccurate due to the label noise and data scarcity; thus, we approximate the Gaussian posterior distribution over the running average of estimated class centroids. We sample class centroids from this Gaussian distribution and then compute distances. Based on the selected clean data, we add an auxiliary balance classifier to improve the generalization of the minority class and facilitate the estimate of Gaussian parameters. The proposed SFA framework significantly outperforms various existing approaches by a large margin, illustrated in Figure 1. Further, we propose to combine SFA with different existing methods: the sample-selection approach, distribution-robust, and noise-robust loss functions. The key contributions of this work are summarized as follows:

- We study an under-explored problem of learning from noisy data under long-tailed class distribution, which is more challenging yet practical.
- We propose a novel framework SFA for detecting noisy samples using the distance-based criterion and improving the minority class generalization. SFA can be an alternative to loss-based approaches to boost the performance balance across classes.
- Experimental results show that SFA significantly outperforms existing methods on various datasets and settings. Further, SFA can be used as a universal add-on for mainstream LTL and LNL methods.

2 Related Work

Long-tailed learning has drawn significant attention in recent years. Many approaches have been proposed, which can be roughly categorized into three types by modifying: (1) the inputs to a model by re-balancing the training data [Shen *et al.*, 2016; Liu *et al.*, 2019; Zhou *et al.*, 2020]; (2) the outputs of a model, for example by post-hoc adjustment of the classifier [Kang *et al.*, 2020; Menon *et al.*, 2020; Tang *et al.*, 2020] and (3) the internals of a model by modifying the loss function [Cui *et al.*, 2019; Cao *et al.*, 2019; Jamal *et al.*, 2020; Ren *et al.*, 2020]. It is worth noting that most of these LTL approaches rely on the assumption that the labels in the training set are correct. However, this is often not the case in real-world applications and label noise has been shown to deteriorate their performance severely.

Label-noise learning approaches can be broadly divided into three areas of focus: (1) noise transition matrix estimation [Patrini *et al.*, 2017; Hendrycks *et al.*, 2018; Cheng *et al.*, 2022], (2) noise-robust loss functions design [Ghosh *et al.*, 2017; Zhang and Sabuncu, 2018; Liu *et al.*, 2020], and (3) clean sample selection [Han *et al.*, 2018; Jiang *et al.*, 2018; Arazo *et al.*, 2019; Li *et al.*, 2020]. Among these approaches, the "small-loss" criterion, which treats samples with small loss as clean samples, is one of the most frequently used methods and has achieved excellent performance. However, it ignores the class imbalance problem in the training data and thus cannot generalizes to long-tailed datasets.

Learning with long-tailed noisy labels aims to tackle both the long-tailed class distribution and label noise issues. A few attempts have been made in this direction by (1) reweighting samples to put greater emphasis on clean long-tailed data [Ren *et al.*, 2018; Shu *et al.*, 2019; Wei *et al.*, 2022a; Jiang *et al.*, 2022]; (2) designing robust loss functions for handling both label noise and class imbalance [Cao *et al.*, 2020], (3) developing better representation learning methods [Zhou *et al.*, 2022; Yi *et al.*, 2022], and (4) selecting clean samples using carefully designed criteria [Wei *et al.*, 2021; Xia *et al.*, 2021]. Our approach follows the idea of sample selection based on the "small-distance" criterion, which detects noisy samples in the latent representation space. Further, we propose an auxiliary balanced classifier to improve the generalization based on selected clean samples.

3 The Proposed Approach

3.1 Problem Formulation

Consider a K class classification task, we denote the training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where the training sample $\boldsymbol{x}_i \in \mathbb{R}^d$ and its label $y_i \in [K] = \{1, 2, ...K\}$. We group training samples of class $k \in [K]$ as $\mathcal{D}_k = \{(\boldsymbol{x}_i, y_i) \mid y_i = k\}$. The training dataset $\mathcal{D}$ follows a long-tailed class distribution in our problem setting. Specifically, let $\rho = \max_k |\mathcal{D}_k| / \min_k |\mathcal{D}_k|$ as the imbalance ratio, we have $\rho \gg 1$, e.g., $\rho = 100$. With label noise, a fraction of training samples are incorrectly labeled, i.e., $\exists i \in [N], y_i \neq y_i^*$ where y_i^* denotes the ground-truth label for $\boldsymbol{x}_i$. In the rest of the paper, we denote the fraction of incorrectly labeled training samples as $\gamma \in (0, 1)$. Given $\mathcal{D}$, our goal is to learn a classifier $f : \mathbb{R}^d \to [K]$ that can generalize to unseen data.

Basic idea. To tackle the problem of learning from long-tailed noisy data, we design a novel framework termed SFA. On the one hand, SFA follows the sample selection paradigm and selects clean samples in the latent representation space, which applies to both the majority and minority classes. On the other hand, we propose to employ an auxiliary balanced classifier for training using selected yet long-tailed clean samples to improve the generalization for the minority class. Figure 2 provides an overview of our proposed framework.

3.2 A New Sample Selection Framework

The Small-Distance Criterion. To select clean samples from long-tailed and noisy data, we proposed to adopt the *small-distance* criterion. It has been verified that the small-distance criterion is more effective than the small-loss counterpart in the recent literature [Wei *et al.*, 2021]. The small-distance criterion makes an assumption on the data distribution that the likelihood of a sample $\boldsymbol{x}_i$ belonging to class k decays exponentially with its distance from its class centroid $\boldsymbol{c}_k$, i.e., $\mathbb{P}(\boldsymbol{x}_i \mid \boldsymbol{c}_k) \propto e^{-dist(\boldsymbol{c}_k, \boldsymbol{x}_i)}$, where $dist$ is the distance measure in the latent representation space and is typically set to be the Euclidean distance. The assumption allows

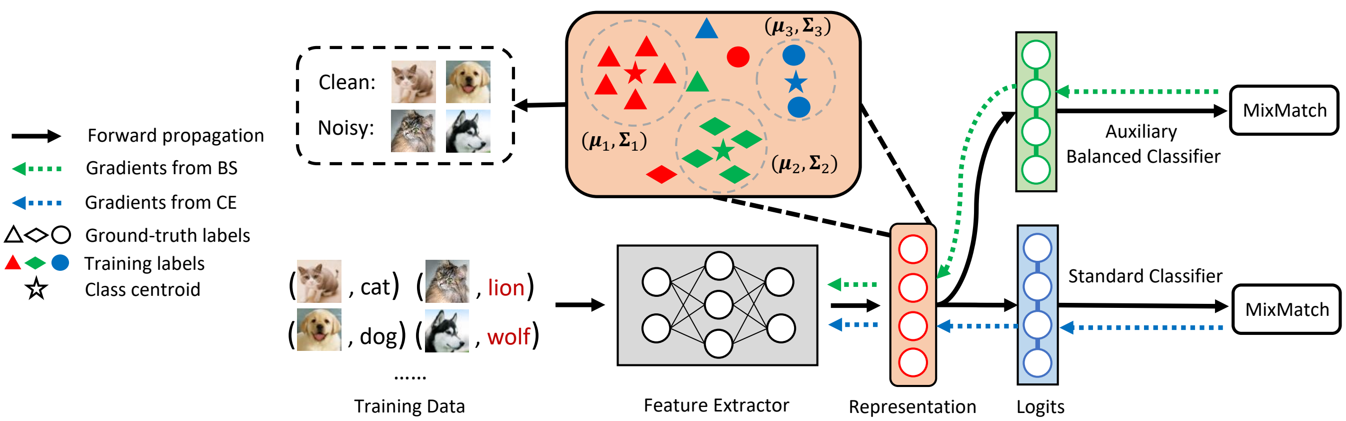


Figure 2: Illustration of the proposed SFA framework, which selects clean samples in the latent representation space. SFA consists of two branches with a shared feature extractor, i.e., the standard classifier and the auxiliary balanced classifier.

for an unrestricted number of samples, making it suitable for both the majority and minority classes. In other words, the closer the samples are to their corresponding class centroids, the more likely they are to be clean samples. The challenge of this approach is how to accurately estimate the class centroids in the presence of label noise, even when only a few samples are accessible for the minority class. In the following, this paper seeks to solve this problem by estimating the posterior distribution of the class centroids.

Instant Centroid Estimation. A direct approach to estimate the class centroids is to compute the mean of all feature vectors of each class after each training epoch referred to as the *instant centroid estimation*. However, it may be difficult to produce accurate estimates when label noise exists in data, as it uses noisy samples in the computation. To mitigate this issue, one feasible approach is to use as many clean samples as possible to compute the class centroids. We find that the majority and minority classes tend to have similar ranges of predicted confidence output by the auxiliary balanced classifier, and most clean samples have higher confidence than noisy samples. This observation motivates us to select a fraction of clean samples for each class to compute a rough estimate of class centroids using a single confidence threshold τ. Specifically, we compute the centroid of class k by:

$$\hat{\boldsymbol{c}}_k \leftarrow \text{Normalize}(\frac{1}{N'_k} \sum_{\boldsymbol{x} \in \mathcal{D}_k} \mathbb{I}(f_{abc}(\boldsymbol{x}) > \tau) g(\boldsymbol{x})), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function, N'_k is the number of selected samples, $g(\boldsymbol{x})$ returns the latent representation for $\boldsymbol{x}$, and $f_{abc}(\boldsymbol{x})$ outputs the predicted confidence by the auxiliary balanced classifier, which is introduced in Section 3.3. In practice, since more and more samples tend to receive high predicted confidence as the training proceeds, we set a dynamic threshold τ_t as an increasing function of iteration t, which is given by $\tau_t = \phi^t \hat{\tau}$, where ϕ is a constant. For example, we set $\phi = 1.005$ and $\hat{\tau} = 1/K$ in our experiments. With this instant centroid estimation, we obtain a relatively reliable centroid for selecting clean samples. However, the accuracy of the estimate can still be affected by the scarcity of the minority class samples.

Stochastic Feature Averaging. To further reduce the estimation error, inspired by [Maddox *et al.*, 2019], a Bayesian extension of the class centroid is designed to conduct Bayesian inference using a Gaussian approximation to the posterior distribution over class centroids. To start with, SFA maintains the running average of class centroids:

$$\boldsymbol{C}_{\text{SFA}} = \beta \cdot \boldsymbol{C}_{\text{SFA}} + (1 - \beta) \cdot \boldsymbol{C}_t, \tag{2}$$

where $\boldsymbol{C}_{\text{SFA}} = [\boldsymbol{c}_1; \cdots; \boldsymbol{c}_K]$ is the running estimate of class centroids, $\boldsymbol{C}_t$ denotes the instant centroid estimates computed by Eq. (1) after the t-th training epoch, and β is the smoothing factor. This averaging in the latent representation space captures the training dynamics of DNNs and provides a robust estimate of the centroid.

SFA conducts Bayesian inference using Gaussian approximation to the posterior distribution over class centroids. This paper considers a simple diagonal format for the covariance matrix. To fit a diagonal covariance approximation, we maintain a running average of the second moment for class centroids:

$$\boldsymbol{C}'_{\text{SFA}} = \beta \cdot \boldsymbol{C}'_{\text{SFA}} + (1 - \beta) \cdot \boldsymbol{C}_t^2 \tag{3}$$

$\boldsymbol{C}^2$ denotes the element-wise square. Thus, a diagonal covariance matrix can be approximated by $\boldsymbol{\Sigma}_{\text{SFA}} = \text{diag}(\boldsymbol{C}'_{\text{SFA}} - \boldsymbol{C}^2_{\text{SFA}})$ and the posterior over class centroid can be constructed as a Gaussian distribution $\mathcal{N}(\boldsymbol{C}_{\text{SFA}}, \boldsymbol{\Sigma}_{\text{SFA}})$. Note that previous work [Maddox *et al.*, 2019] also proposes a higher-rank approximation for the covariance, but we only focus on the diagonal approximation for simplicity. Then we sample several stochastic class centroids from this Gaussian distribution and perform distance averaging for sample selection. Specifically, given the sampling rates S, we first sample $\widetilde{\boldsymbol{C}} \sim \mathcal{N}(\boldsymbol{C}_{\text{SFA}}, \boldsymbol{\Sigma}_{\text{SFA}})$ and then compute the Euclidean distances between $\hat{\boldsymbol{c}}_k$ and samples of class k by:

$$dist(\tilde{\boldsymbol{c}}_k, \boldsymbol{x}_i) = ||\tilde{\boldsymbol{c}}_k - g(\boldsymbol{x}_i)||_2^2 \tag{4}$$

This procedure is repeated for S times and an averaged distance is obtained; thus, training samples are well clustered with the distance to the class centroid by fitting them into a two-component Gaussian mixture model (GMM) [Permuter *et al.*, 2006], i.e., $dist \sim \sum_{j=1}^{2} \phi_j \mathcal{N}(\mu_j, \sigma_j^2)$ where μ_j, σ_j^2

Algorithm 1: The SFA Framework

```
Input: training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)_{i=1}^N\}$, model
  parameters $\theta$, sampling rate $S$, warm-up epochs $T_0$,
  and total training epochs $T$.
// warm-up for $T_0$ epochs
for $t = 1, ..., T_0$ do
    $\mathcal{L} = \ell_{CE}(\mathcal{D}, f) + \ell_{BS}(\mathcal{D}, f_{abc})$
    $\theta_t = \text{SGD}(\mathcal{L}, \theta_{t-1})$
end
for $t = T_0 + 1, ..., T$ do
    // sample selection
    $\mathcal{D}^{\text{clean}} = \emptyset$, $\mathcal{D}^{\text{noisy}} = \emptyset$
    Compute confidence threshold $\tau_t = \gamma^t \hat{\tau}$
    for $k = 1, ..., K$ do
        Compute instant centroid by Eq. (1)
        Update the first moment $\boldsymbol{C}_{\text{SFA}}$ by Eq. (2) and
          the second moment $\boldsymbol{C}'_{\text{SFA}}$ by Eq. (3)
        for $s = 1, ..., S$ do
            Sample $\tilde{\boldsymbol{c}}_k \sim \mathcal{N}(\boldsymbol{C}_{\text{SFA}}, \boldsymbol{\Sigma}_{\text{SFA}})$
            $dist_{i,s} = ||\tilde{\boldsymbol{c}}_k - g(\boldsymbol{x}_i)||_2$, where $\boldsymbol{x}_i \in \mathcal{D}_k$
        end
        $dist_i = \frac{1}{S} \sum_{s=1}^S dist_{i,s}$
        $\mathcal{D}_k^{\text{clean}}, \mathcal{D}_k^{\text{noisy}} = \text{GMM}(dist)$
        $\mathcal{D}^{\text{clean}} = \mathcal{D}^{\text{clean}} \cup \mathcal{D}_k^{\text{clean}}$,$\mathcal{D}^{\text{noisy}} = \mathcal{D}^{\text{noisy}} \cup \mathcal{D}_k^{\text{noisy}}$
    end
    // semi-supervised learning
    $\mathcal{L}_{SSL} = \text{MixMatch}(\mathcal{D}^{\text{clean}}, \mathcal{D}^{\text{noisy}}, f)$
    $\mathcal{L}_{ABC} = \text{MixMatch}(\mathcal{D}^{\text{clean}}, \mathcal{D}^{\text{noisy}}, f_{abc})$
    $\mathcal{L} = \mathcal{L}_{SSL} + \mathcal{L}_{ABC}$
    $\theta_t = \text{SGD}(\mathcal{L}, \theta_{t-1})$
end
```

are the mean and variance of the j-th Gaussian component. Assume $\mu_1 < \mu_2$ without loss of generality, as clean samples distribute around class centroids while noisy samples spread out, we denote the clean probability of one sample as:

$$\mathbb{P}(\text{clean} \mid \boldsymbol{x}_i) = \frac{\phi_1 \mathcal{N}(\mu_1, \sigma_1^2)}{\sum_{j=1}^2 \phi_i \mathcal{N}(\mu_j, \sigma_j^2)} \tag{5}$$

We select samples with $\mathbb{P}(\text{clean} \mid x_i) > 0.5$ as clean samples and others as noisy samples. Next, the training dataset $\mathcal{D}$ is divided into the clean sample set $\mathcal{D}^{\text{clean}}$ and noisy sample set $\mathcal{D}^{\text{noisy}}$. As data in $\mathcal{D}^{\text{clean}}$ may still follow a long-tailed distribution, we propose an auxiliary balanced classifier to obtain unbiased predictions.

3.3 Auxiliary Balanced Classifier

In the literature on long-tailed learning, class re-balancing strategies are the prominent and effective methods proposed to alleviate imbalance problems, which can significantly promote classifier learning and affect the representation learning w.r.t. the original data distribution. Several existing works [Kang *et al.*, 2020; Zhou *et al.*, 2020; Lee *et al.*, 2021] suggest decoupling representation and classifier learning and demonstrate their superiority over conventional learning methods. To further boost the performance, we adopt a two-branch network to combat class imbalance during model training: we add an auxiliary classifier to the backbone of the neural network for balanced classifier learning and meanwhile maintain the original classifier for the sake of representation learning. We denote the two classifiers as f_{abc} and f, where f is trained using the standard cross-entropy loss $\ell_{CE}(\boldsymbol{x}, y) = -\log \frac{e^{z_y}}{\sum_{k=1}^K e^{z_k}}$, where $\boldsymbol{z} = f(\boldsymbol{x})$. Next, we describe obtaining a balanced classifier f_{abc}.

Balanced Softmax. To produce unbiased predictions, we propose an auxiliary balanced classifier f_{abc}, which is jointly learned with the standard classifier f by sharing the feature extractor. Specifically, f_{abc} seeks to minimize the Balanced Softmax (BS) function [Ren *et al.*, 2020]:

$$\ell_{BS}(\boldsymbol{x}, y) = -\log \frac{n_y e^{z_y}}{\sum_{k=1}^K n_k e^{z_k}}, \tag{6}$$

where $n_k = |\mathcal{D}_k^{clean}|$ is the number of samples of class k counted from $\mathcal{D}^{clean}$, and z_k stands for the k-th logit produced by $f_{abc}(\boldsymbol{x})$. The loss can penalize more heavily on samples in the majority class while placing a lower penalty on samples in the minority class. This helps to learn a balanced classifier and prevent the model from being significantly biased towards the majority class.

Incorporating with Semi-supervised Learning. To improve the utilization of noisy data, we treat samples in noisy set $\mathcal{D}^{\text{noisy}}$ as unlabeled data and incorporate our methods into a popular semi-supervised learning framework MixMatch [Berthelot *et al.*, 2019]. MixMatch is a state-of-the-art semi-supervised learning algorithm that combines data augmentation, consistency regularization, and mixup to achieve excellent performance. For unlabeled samples in $\mathcal{D}^{\text{noisy}}$, we perform label guessing using predictions from the standard classifier f. The total loss is $\mathcal{L} = \mathcal{L}_{SSL} + \mathcal{L}_{ABC}$, where $\mathcal{L}_{SSL}$ and $\mathcal{L}_{ABC}$ stands for a MixMatch loss computed from the standard classifier f and the balanced classifier f_{abc}. The pseudo-code of our SFA framework is summarized in Algorithm 1. At test time, both classifiers are utilized to return the classification results. Furthermore, we also employ two separate neural networks to combat confirmation biases following DivideMix [Li *et al.*, 2020] and RoLT+ [Wei *et al.*, 2021].

3.4 Time Complexity Analysis

The computation of distances between features and class centroids, and the fitting of GMM, are the main additional computational cost in sample selection. The time complexity of computing distances is $\mathcal{O}(SNd)$, where N is the number of samples, d is the latent feature dimension, and S is the sampling rate. We set $S = 1$ for large-scale datasets. The time complexity of fitting the two-component GMM using the EM algorithm requires $\mathcal{O}(Nd'ct)$, where d' is the dimension of distances, c is the number of components, and t is the number of iterations. In our approach, we have $d' = 1$, $c = 2$, and $t = 100$. The total time complexity is $\mathcal{O}((200 + d)N)$, which is efficient enough compared with existing approaches.

		CIFAR-10						CIFAR-100					
Noise Level		0.2			0.5			0.2			0.5		
Imbalance Ratio		10	50	100	10	50	100	10	50	100	10	50	100
CE	Best	77.86	64.38	61.79	60.72	46.50	38.43	45.97	33.41	29.85	28.70	18.49	16.24
	Last	74.00	61.38	55.69	44.29	32.69	27.78	45.75	33.12	29.58	23.70	16.56	14.19
BBN	Best	78.44	69.05	64.24	64.51	48.88	37.75	48.60	29.08	27.44	31.05	20.33	15.51
	Last	77.67	68.01	64.15	53.67	45.06	34.93	47.72	28.87	27.04	30.11	19.97	14.95
cRT	Best	77.67	68.50	60.85	62.37	42.60	35.75	43.56	31.07	24.65	26.31	19.65	15.41
	Last	75.36	67.94	58.67	60.35	41.58	33.86	42.75	30.43	23.97	25.12	19.32	14.82
ELR+	Best	88.96	80.21	69.60	85.02	56.96	48.72	54.01	49.64	38.40	49.53	30.12	21.58
	Last	88.09	79.69	66.67	84.08	48.14	43.11	53.32	48.37	38.12	49.06	29.68	20.47
DivideMix	Best	88.79	75.34	66.90	87.54	67.92	61.81	63.79	49.64	43.91	49.35	36.52	31.82
	Last	88.10	73.48	63.76	86.88	65.22	59.65	63.17	48.37	42.59	48.87	35.72	31.05
MW-Net	Best	82.19	71.63	67.26	72.12	56.09	46.36	50.20	36.68	31.77	37.50	23.99	21.24
	Last	77.67	64.12	58.23	59.68	45.39	37.05	47.82	34.45	29.57	33.14	20.33	18.82
HAR	Best	81.63	66.45	56.95	63.07	54.54	38.41	45.28	29.74	26.79	29.30	17.33	14.47
	Last	78.04	60.17	54.78	61.13	48.61	35.40	44.52	26.13	23.90	26.46	14.68	12.36
RoLT+	Best	87.95	77.26	72.31	88.17	75.11	64.42	64.22	51.01	45.35	53.31	39.78	35.29
	Last	87.54	75.90	69.12	87.45	73.92	61.15	63.31	49.40	43.16	52.44	39.27	34.43
PCL	Best	90.92	84.12	79.54	84.04	71.44	66.33	65.23	51.73	47.38	**57.65**	42.51	38.42
	Last	90.81	83.71	78.34	83.51	71.44	64.69	65.14	51.46	47.12	**57.65**	42.51	38.36
SFA (ours)	Best	**92.53**	**85.96**	**80.26**	**90.57**	**79.89**	**75.17**	**66.32**	**54.29**	**48.51**	57.41	**44.37**	**39.73**
	Last	**92.13**	**84.80**	**79.22**	**90.08**	**78.93**	**74.06**	**65.65**	**53.10**	**47.73**	57.28	**43.41**	**39.73**

Table 1: Test accuracy (%) on simulated CIFAR datasets with varying levels of noise and imbalanced ratios. The best results are in **bold**.

4 Experiments

4.1 Datasets and Implementation Details

Benchmark Datasets. We first test our approach on CIFAR-10 and CIFAR-100 datasets by simulating training data with long-tailed class distribution and label noise following prior work [Wei *et al.*, 2021]. Formally, denote the imbalance ratio as ρ and noise level as γ. We set the number of samples for the k-th class to $N_k = N/\rho^{\frac{k-1}{K-1}}$ and generate a long-tailed dataset. Next, we inject the label noise into this dataset via a noise transition matrix T defined as:

$$T_{ij} = \mathbb{P}(Y = j \mid Y^* = i) = \begin{cases} 1-\gamma & \text{if } i = j \\ \frac{N_j}{N-N_i}\gamma & \text{otherwise.} \end{cases} \quad (7)$$

We use an 18-layer PreAct ResNet [He *et al.*, 2016] and train it using SGD with a momentum of 0.9, a weight decay of 5×10^{-4}, a batch size of 128 and an initial learning rate of 0.02. The model is trained for 200 epochs with 1 NVIDIA GeForce RTX 3090. We perform sample selection after a warm-up period of 30 epochs and anneal the learning rate by a factor of 10 after 150 epochs. For all CIFAR experiments, we choose ρ from $\{10, 50, 100\}$ and γ from $\{0.2, 0.5\}$, and use the same hyperparameters $\beta = 0.99$ and $S = 5$.

Real-World Dataset. WebVision is a large-scale dataset with real-world noisy labels and long-tailed distributions. It contains 2.4 million images crawled from Flickr and Google using the 1,000 concepts in ImageNet ILSVRC12. As previously done in [Li *et al.*, 2020], we use the Inception-ResNet v2 architecture [Szegedy *et al.*, 2017] to evaluate the performance of baseline methods on the first 50 classes of the Google image subset. We set the hyperparameters to $\beta = 0.9$, $S = 1$ and train the network using SGD with a momentum of 0.9, a weight decay of 1×10^{-3}, and a batch size of 32. The initial learning rate is set to 0.01 and reduced by a factor of 10 after 50 epochs. The warm-up period is one epoch, and the model is trained for 100 epochs in total with 2 NVIDIA GeForce RTX 3090.

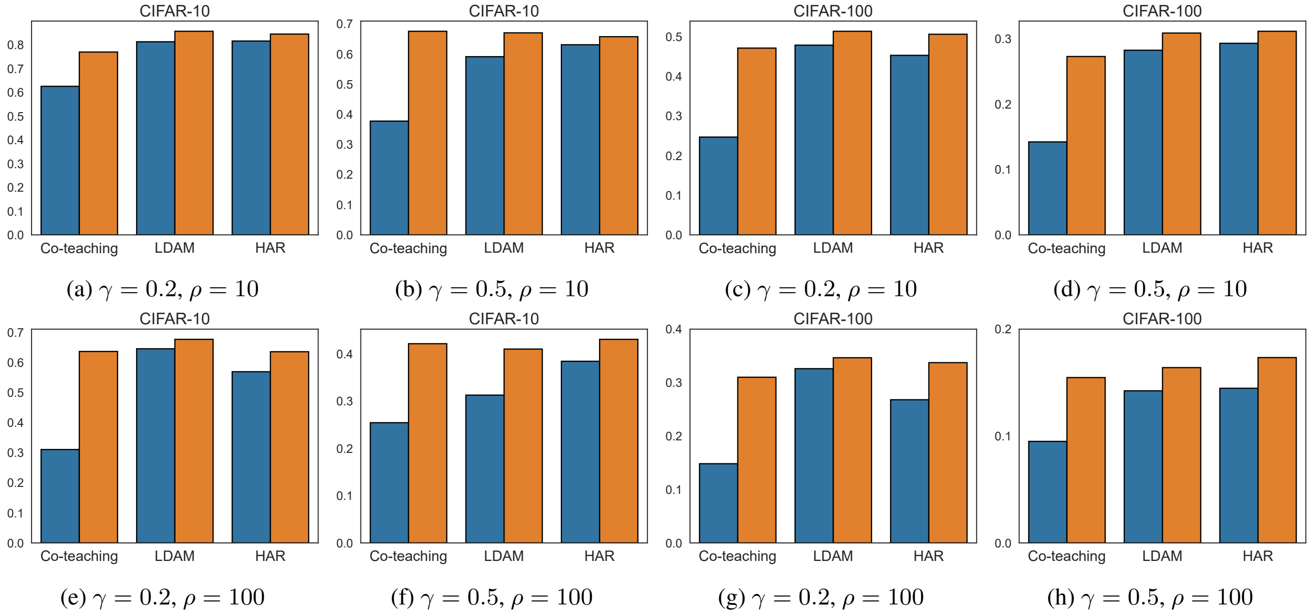
4.2 Comparison Methods

We compare the performance of our approach SFA with nine baselines with the same network architecture. In addition to the direct approach using cross-entropy loss (CE), we compare SFA with the following three groups of methods:

- **Long-tailed learning**. Methods commonly used to address the long-tailed classification include BBN [Zhou *et al.*, 2020] and cRT [Kang *et al.*, 2020].
- **Label-noise learning**. Recent state-of-the-art methods for learning with noisy labels, including ELR+ [Liu *et al.*, 2020] and DivideMix [Li *et al.*, 2020]
- **Learning with long-tailed noisy data**. Methods designed to tackle noisy labels and long-tailed distributions simultaneously, including MW-Net [Shu *et al.*, 2019], HAR[Cao *et al.*, 2020], RoLT+ [Wei *et al.*, 2021], and PCL [Wei *et al.*, 2022a].

4.3 Results on Benchmark Dataset

Table 1 summarizes the results on CIFAR-10 and CIFAR-100 datasets under different levels of label noise and imbalance



(a) $\gamma = 0.2$, $\rho = 10$ (b) $\gamma = 0.5$, $\rho = 10$ (c) $\gamma = 0.2$, $\rho = 10$ (d) $\gamma = 0.5$, $\rho = 10$

(e) $\gamma = 0.2$, $\rho = 100$ (f) $\gamma = 0.5$, $\rho = 100$ (g) $\gamma = 0.2$, $\rho = 100$ (h) $\gamma = 0.5$, $\rho = 100$

Figure 3: Test accuracy on CIFAR-10 and CIFAR-100 with varying levels of noise and imbalance ratios. Note that the blue and orange bars are results for without and with SFA, respectively.

ratios. We report the "best" test accuracy across all training iterations and the "last" test accuracy at the end of training. As shown in the results, initial model training with cross-entropy loss is susceptible to biases in the training set, leading to a significant decrease in its classification accuracy as the data biases become more pronounced. Also, previous long-tailed learning methods (i.e., BBN and cRT) dreadfully degrade their performance as the noise level increases. When the noise level is high, the last accuracy is significantly lower than the best accuracy due to the model's tendency to fit the label noise, which can negatively affect the model's ability to generalize to unseen data. The same conclusions can be drawn from other methods. However, our approach (SFA) retains the most robust performance and advances all other compared methods in almost all dataset settings. In particular, compared with the previous state-of-the-art method PCL, SFA can further improve the test accuracy on CIFAR-10 by an average of 4.04% and on CIFAR-100 by an average of 1.24%, and it can be observed that the improvement becomes more significant at high noise levels and imbalance ratios, benefiting from proposed stochastic feature averaging and balanced classifier training.

4.4 Results on Real-World Dataset

Table 2 reports the results on the WebVision dataset ($\rho \approx 6$). SFA consistently outperforms compared methods by a large margin, with an average improvement of 1.02% (0.9%) on WebVision (ImageNet) validation sets. This demonstrates the effectiveness of our approach in improving the performance of deep learning models in the presence of real-world label noise and class imbalance. To further illustrate the benefits of our approach, we also conduct experiments on the WebVision dataset by manipulating the imbalance ratios ($\rho = 50$ and $\rho = 100$) and observe a clear improvement, which suggests that the advantage of our approach is more appealing as the imbalance factor increases.

IR	Method	WebVision		ImageNet	
		top1	top5	top1	top5
$\rho \approx 6$	ELR+	77.78	91.68	70.29	89.76
	DivideMix	77.32	91.64	75.20	90.84
	HAR	75.50	90.70	70.30	90.00
	RoLT+	77.64	92.44	74.64	92.48
	PCL	77.32	92.60	75.12	91.92
	SFA (ours)	**78.96**	**93.00**	**76.16**	**92.68**
$\rho = 50$	DivideMix	64.56	83.56	62.68	85.24
	RoLT+	66.28	88.68	64.76	89.96
	PCL	68.00	88.44	65.00	86.32
	SFA (ours)	**70.64**	**89.96**	**69.04**	**90.36**
$\rho = 100$	DivideMix	55.76	73.48	53.92	74.00
	RoLT+	60.68	87.84	59.68	88.52
	PCL	62.12	85.88	59.60	84.20
	SFA (ours)	**65.68**	**88.52**	**65.08**	**88.92**

Table 2: Test accuracy (%) on mini-WebVision and ImageNet with various imbalance ratios.

4.5 Ablation Studies and Further Analyses

Insights into the Key Components. We conduct ablation studies to better understand the impact of different components of our SFA framework. Table 3 reports the results on CIFAR datasets with varying levels of noise and imbalance ratios, where w/o ICE/SCC/ABC means removing the proposed instant centroid estimation, stochastic class centroid, and auxiliary balanced classifier, respectively. The following

Noise Level		0.2			0.5		
Imbalance Ratio		10	50	100	10	50	100
SFA	Best	92.53	85.96	80.26	90.57	79.89	75.17
	Last	92.13	84.80	79.22	90.08	78.93	74.06
w/o ICE	Best	91.76	83.90	78.39	89.99	78.81	73.81
	Last	91.73	83.62	78.00	89.84	78.43	73.80
w/o SCC	Best	88.22	79.08	74.28	89.47	76.30	73.74
	Last	88.20	78.47	74.28	89.18	76.05	72.76
w/o ABC	Best	91.84	82.70	76.38	89.55	78.51	71.19
	Last	91.59	82.58	74.86	88.97	77.61	70.73

Table 3: Ablation studies on key components of our proposed SFA framework. We report the test accuracy on CIFAR-10 dataset.

analyses of results can be derived: (1) in the presence of noisy labels, instant centroid estimation based on the confidence of samples from the balanced classifier can provide a more accurate approximation of the actual class centroids compared to simply taking the mean value of all sample features. (2) The dramatic decrease in accuracy among all settings without stochastic class centroid highlights the significance of the Gaussian posterior distributions over class centroid for better sample selection. (3) the auxiliary balanced classifier can significantly enhance the model's performance, especially in cases of heavy class imbalance.

Efficacy of Sample Selection. We compare the F1-scores with existing sample selection-based methods (i.e., DivideMix and RoLT+) on two CIFAR datasets to verify the effectiveness of sample selection in our approach. Results in Figure 4 illustrate the superiority of our approach in selecting clean samples from noisy and long-tailed data. We believe this is due to the following reasons: (1) we adopt the small-distance criterion, which is more robust than small-loss when noisy labels and class imbalance co-exist in datasets; (2) sampling stochastic class centroid from the Gaussian posterior distribution, which is used in distance averaging, can further reduce the impact of label noise and data scarcity on the estimation of the actual class centroid; (3) conducting Bayesian inference of class centroid allows the model to access a diverse range of clean samples at different epochs, which helps to prevent overfitting to a particular clean subset. This can be further illustrated in Figure 5. We compare the performance of our method with RoLT+ [Wei *et al.*, 2021] in terms of the mean loss of actual clean samples on CIFAR-10 under different data settings. Both methods use the small-distance criterion, but our method utilizes Bayesian inference to estimate the class centroids. We believe a method that can achieve lower losses on the whole clean samples can more effectively identify and use these samples to improve the performance. The plot demonstrates that our approach can obtain lower losses and, therefore, better learning from clean samples than the method without stochastic class centroids.

Collaboration with Existing Approaches. To further showcase the versatility of the proposed sample selection approach, we present its three applications. We combine SFA with a variety of existing techniques, including the sample-selection approach (Co-teaching [Han *et al.*, 2018]), the distribution-robust loss (LDAM [Cao *et al.*, 2019]), and the noise-robust loss (HAR [Cao *et al.*, 2020]). Specifically, we use SFA in two ways: a replacement for the sample selection module in Co-teaching and a method for identifying clean samples to train models with LDAM and HAR. Figure 3 summarizes the performance of different approaches with and without applying SFA under various settings. Results show that SFA improves the generalization of all approaches, leading to a more robust model.

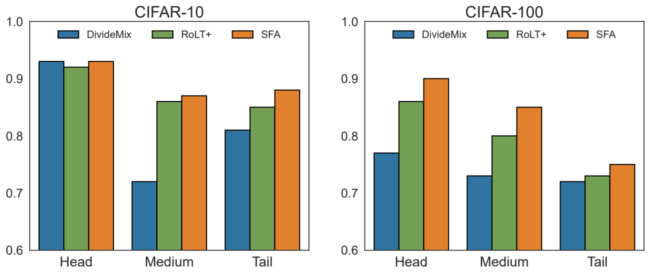
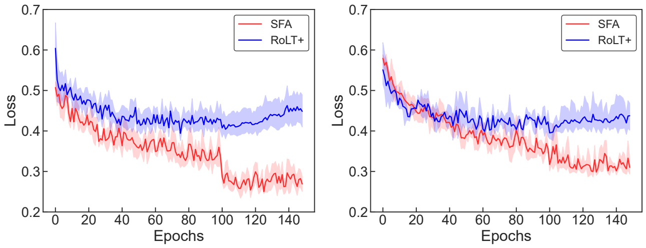


Figure 4: F1-score of the head, medium and tail classes on CIFAR-10 and CIFAR-100 datasets under $\gamma = 0.5$ and $\rho = 100$.



(a) $\gamma = 0.2, \rho = 10$ (b) $\gamma = 0.5, \rho = 100$

Figure 5: Mean losses of all clean samples on CIFAR-10 with different levels of noise and imbalance ratios. Our approach SFA can achieve lower losses compared to RoLT+.

5 Conclusion

This paper proposes a distance-based sample selection approach called stochastic feature averaging (SFA) and an auxiliary balanced classifier for handling datasets with both noisy labels and long-tailed class distribution. The SFA framework first estimates an instant centroid for each class and uses the exponential running average of the centroid to fit a Gaussian posterior distribution, which is utilized to identify noisy samples based on their distances to the centroid sampled from the Gaussian. Next, a standard and an auxiliary balanced classifier are jointly learned via a semi-supervised learning framework to improve the generalization of the minority class and facilitate the estimation of the Gaussian parameters. The proposed approach is evaluated on various datasets, including CIFAR-10, CIFAR-100, and WebVision, and is shown to outperform previous state-of-the-art consistently. Our proposed approach, SFA, can be an alternative to existing loss-based approaches for learning with long-tailed noisy labels and can be used as a universal add-on for various methods. For future work, we plan to investigate and provide a theoretical explanation for the effectiveness of stochastic class centroids.

Acknowledgments

This research was supported by the National Science Foundation of China (62206049). This research work is supported by the Big Data Computing Center of Southeast University.

References

[Arazo *et al.*, 2019] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *Proceedings of International Conference on Machine Learning*, pages 312–321, 2019.

[Arpit *et al.*, 2017] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *Proceedings of International Conference on Machine Learning*, pages 233–242, 2017.

[Berthelot *et al.*, 2019] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 2019.

[Cao *et al.*, 2019] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, pages 1565–1576, 2019.

[Cao *et al.*, 2020] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *International Conference on Learning Representations*, 2020.

[Cheng *et al.*, 2022] De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16630–16639, 2022.

[Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[Cui *et al.*, 2021] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724, 2021.

[Ghosh *et al.*, 2017] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1919–1925, 2017.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in Neural Information Processing Systems*, pages 8536–8546, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of European Conference on Computer Vision*, pages 630–645, 2016.

[Hendrycks *et al.*, 2018] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in Neural Information Processing Systems*, pages 10477–10486, 2018.

[Jamal *et al.*, 2020] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020.

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

[Jiang *et al.*, 2022] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7024–7032, 2022.

[Kang *et al.*, 2020] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

[Lee *et al.*, 2021] Hyuck Lee, Seungjae Shin, and Heeyoung Kim. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. *Advances in Neural Information Processing Systems*, pages 7082–7094, 2021.

[Li *et al.*, 2020] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

[Liu *et al.*, 2019] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

[Liu *et al.*, 2020] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in Neural Information Processing Systems*, 2020.

[Maddox *et al.*, 2019] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 2019.

[Menon *et al.*, 2020] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2020.

[Patrini *et al.*, 2017] 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 *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

[Permuter *et al.*, 2006] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, pages 695–706, 2006.

[Ren *et al.*, 2018] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of International Conference on Machine Learning*, pages 4334–4343, 2018.

[Ren *et al.*, 2020] Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, pages 4175–4186, 2020.

[Shen *et al.*, 2016] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *Proceedings of European Conference on Computer Vision*, pages 467–482, 2016.

[Shu *et al.*, 2019] 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. *Advances in Neural Information Processing Systems*, pages 1917–1928, 2019.

[Szegedy *et al.*, 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017.

[Tang *et al.*, 2020] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, pages 1513–1524, 2020.

[Wei and Li, 2020] Tong Wei and Yu-Feng Li. Does tail label help for large-scale multi-label learning? *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2315–2324, 2020.

[Wei *et al.*, 2021] Tong Wei, Jiang-Xin Shi, Wei-Wei Tu, and Yu-Feng Li. Robust long-tailed learning under label noise. *ArXiv Preprint ArXiv:2108.11569*, 2021.

[Wei *et al.*, 2022a] Tong Wei, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. Prototypical classifier for robust class-imbalanced learning. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2022.

[Wei *et al.*, 2022b] Tong Wei, Hai Wang, Weiwei Tu, and Yufeng Li. Robust model selection for positive and unlabeled learning with constraints. *Science China Information Sciences*, 65(11):1–13, 2022.

[Wu *et al.*, 2021] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: A unified framework for learning with open-world noisy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 62–71, 2021.

[Xia *et al.*, 2021] 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 *International Conference on Learning Representations*, 2021.

[Xiang *et al.*, 2020] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *Proceedings of European Conference on Computer Vision*, pages 247–263, 2020.

[Yi *et al.*, 2022] Xuanyu Yi, Kaihua Tang, Xian-Sheng Hua, Joo-Hwee Lim, and Hanwang Zhang. Identifying hard noise in long-tailed sample distribution. In *Proceedings of European Conference on Computer Vision*, pages 739–756, 2022.

[Zhang and Sabuncu, 2018] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems*, 2018.

[Zhou *et al.*, 2020] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2020.

[Zhou *et al.*, 2022] Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, and Xiangyang Ji. Prototype-anchored learning for learning with imperfect annotations. In *Proceedings of International Conference on Machine Learning*, pages 27245–27267, 2022.

[Zhou, 2022] Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 9(8):nwac123, 2022.