

BAPE: Learning an Explicit Bayes Classifier for Long-tailed Visual Recognition

Chaoqun Du Yulin Wang Shiji Song Gao Huang

Department of Automation
Tsinghua University
Beijing, China

dcq20@mails.tsinghua.edu.cn gaohuang@tsinghua.edu.cn

Abstract

Bayesian decision theory advocates the Bayes classifier as the optimal approach for minimizing the risk in machine learning problems. Current deep learning algorithms usually solve for the optimal classifier by *implicitly* estimating the posterior probabilities, *e.g.*, by minimizing the Softmax cross-entropy loss. This simple methodology has been proven effective for meticulously balanced academic benchmark datasets. However, it is not applicable to the long-tailed data distributions in the real world, where it leads to the gradient imbalance issue and fails to ensure the Bayes optimal decision rule. To address these challenges, this paper presents a novel approach (BAPE) that provides a more precise theoretical estimation of the data distributions by *explicitly* modeling the parameters of the posterior probabilities and solving them with point estimation. Consequently, our method directly learns the Bayes classifier without gradient descent based on Bayes' theorem, simultaneously alleviating the gradient imbalance and ensuring the Bayes optimal decision rule. Furthermore, we propose a straightforward yet effective *distribution adjustment* technique. This method enables the Bayes classifier trained from the long-tailed training set to effectively adapt to the test data distribution with an arbitrary imbalance factor, thereby enhancing performance without incurring additional computational costs. In addition, we demonstrate the gains of our method are orthogonal to existing learning approaches for long-tailed scenarios, as they are mostly designed under the principle of *implicitly* estimating the posterior probabilities. Extensive empirical evaluations on CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, and iNaturalist demonstrate that our method significantly improves the generalization performance of popular deep networks, despite its simplicity.

1 Introduction

As grounded in Bayesian decision theory, the Bayes classifier is usually identified as the optimal classifier that minimizes the risk in machine learning tasks [60, 61, 47]. However, the complex and often intractable nature of real data distributions presents significant challenges to the direct computation of the posterior distribution, and hence hinders attaining Bayes-optimal decision-making. In most cases, popular methodologies are designed to *implicitly* estimate posterior probabilities based on training data. This typically involves the optimization of a Softmax classifier through the application of cross-entropy loss and gradient descent [18, 34, 25, 29, 12, 21, 23, 22, 20]. This approach has consistently demonstrated its efficacy across a broad range of benchmark datasets [10, 33, 50].

However, real-world application scenarios are generally not as ideal as these academic benchmarks [15, 13, 24, 16, 14, 28]. As a notable difference, real data often follows a long-tailed distribution,

characterized by a significant drop in the number of samples per class from the head (high-frequency classes) to the tail (low-frequency classes) [59]. Given such an imbalance, the standard Softmax cross-entropy algorithm suffers from two major issues. Through the lens of optimization, it leads to a *minority collapse* phenomenon [17] caused by gradient imbalance, *i.e.*, classifiers for minority classes tend to become closer to each other due to the progressively suppressed gradients as the level of imbalance escalates. From the perspective of Bayesian decision-making, the Bayesian optimal decision rule is usually not ensured [46]. To alleviate these two problems, some methods have been proposed, such as re-sampling [35, 62, 5, 35, 1, 2, 42, 30], re-weighting [27, 45, 9, 40, 65, 67, 54], re-margining [4, 31, 3] and logit adjustment [26, 46, 48, 64, 49]. However, importantly, they continue to adopt the paradigm that *implicitly* estimates the posterior distribution. Such implicit estimation is developed and mostly effective for balanced training data. We argue that this design may result in sub-optimal algorithms in the context of realistic long-tailed training data.

In this paper, we seek to explore whether a better machine learning algorithm can be acquired under the principle of *explicitly* modeling the Bayesian decision process. We achieve this by proposing a BAPE approach (Bayes classifier by Point Estimation). In specific, BAPE offers a more precise theoretical estimation of data distribution by *explicitly* modeling the data distribution and employing point estimation for distribution parameters. As a consequence, it directly learns the Bayes classifier through point estimation without relying on gradient descent, effectively mitigating the issue of gradient imbalance. Furthermore, the Bayes classifier is learned explicitly based on Bayes' theorem, adhering to the optimal decision rule.

Our primary focus lies in explicitly modeling the data distribution and estimating its parameters. However, translating this idea into practice is not straightforward, as the methodologies for modeling real data distributions often involve complexities, such as the necessity of training deep generative models [19, 52]. We propose a more elegant and manageable solution to this challenge. Instead of modeling the data distribution in its original form, we propose to do so in the feature space, which is generally simpler and more tractable. Specifically, inspired by the Simplex-Encoded-Labels Interpolation (SELI) [56] that the features tend to collapse towards their corresponding class means in imbalanced learning, we adopt the von Mises-Fisher (vMF) distribution on the unit sphere to model the feature distribution. Building on this vMF distribution assumption, we can estimate the parameters using point estimation method. Crucially, this only requires the first sample moment, which can be efficiently computed across various batches during the training process. Thus, we eliminate the need for gradient descent, making our approach computationally efficient and straightforward.

Moreover, a crucial aspect we consider is the disparity in feature distributions between the training and testing set, a phenomenon often arising due to the limited size of available data for estimation. Such discrepancy implies that a Bayes classifier trained on the training set might not deliver optimal performance when applied to the test set. To counter this, leveraging our explicit estimation of data distribution, we can modify the parameters to better align with the test set which exhibits an arbitrary imbalance factor. Notably, our approach incurs no additional computational costs, making it an efficient solution. Empirical results demonstrate that it considerably enhances the performance.

The primary contributions of this study are outlined as follows: 1) we propose a novel method for explicitly learning a Bayes classifier using point estimation, which ensures the Bayesian optimal decision rule and alleviates the problem of gradient imbalance; 2) we introduce a straightforward yet effective method to adjust the distribution, which boosts performance without incurring any additional costs; 3) empirical evaluations on image classification tasks with CIFAR-10/100-LT, iNaturalist 2018, and ImageNet-LT demonstrate that the proposed BAPE algorithm consistently improves the generalization performance of existing long-tailed recognition methods.

2 Related Works

Re-sampling. The re-sampling methods aim to rectify the uneven distribution of training data by either downsampling the high-frequency classes [35, 62] or upsampling the low-frequency classes [5, 1, 2], thereby facilitating the acquisition of knowledge of the tail classes. Square-root sampling [42] is a modified version of class-balanced sampling, in which the sampling probability for each class is determined by the square root of the sample size within that specific class. Progressively-balanced sampling [30] gradually transitions between random sampling and class-balanced sampling. Empirical

findings from Decoupling [30] demonstrate that both square-root sampling and progressively-balanced sampling are superior strategies for training standard models in long-tailed recognition.

Re-weighting. In order to mitigate the impact of class imbalance, re-weighting techniques strive to adjust the training loss values associated with various classes by multiplying distinct weight factors [27, 9, 40, 65, 67]. Following this methodology, Class-balanced loss (CB) [9] introduces an effective number term to approximate the expected sample count for distinct classes and then incorporates a re-weighting term that balances classes by inversely scaling it with the effective number. Some recent studies [54, 65, 67] also seek to address the negative gradient over-suppression issue of tail classes by re-weighting. Equalization loss [54] simply reduces the impact of tail-class samples when they act as negative labels for head-class samples.

Re-margining. In tackling class imbalance, re-margining techniques attempt to modify losses by subtracting distinct margin factors for different classes. Following this idea, LDAM [4] incorporates class-specific margin factors determined by the training label frequencies, encouraging larger margins for tail classes. Recent studies further explored adaptive re-margining methods. Uncertainty-based margin learning (UML) [31] utilizes estimated class-level uncertainty to adjust loss margins. A subsequent work introduces a frequency indicator based on the inter-class feature compactness [3].

Logit Adjustment. Logit adjustment techniques [26, 46, 48, 64, 49] aim to address the class imbalance by modifying the prediction logits of a class-biased model. In a recent study [46], a comprehensive analysis of logit adjustment in the context of long-tailed recognition was conducted, proposing a logit adjustment (LA) method from the Bayesian perspective, which ensures the Bayesian optimal decision rule. Most recently, a vMF classifier [64] is introduced, which performs adjustments via inter-class overlap coefficients. While this approach shares similarities with our method in terms of utilizing the vMF distribution and employing a post-hoc adjustment technique, it involves complex adjustments to enhance model performance, albeit at the expense of simplicity. This method fundamentally differs from our approach in that it still relies on gradient descent for learning distribution parameters and is susceptible to gradient imbalance.

3 Method

In this section, we first introduce the theoretical motivation behind our approach based on Bayesian optimal decision-making. Next, we introduce the assumption and its rationale concerning the modeling of the distribution, which allows us to efficiently estimate the classifier’s parameters by point estimation. Subsequently, an algorithm is presented for efficiently estimating parameters using maximum a posteriori probability estimation during the training process. Moreover, we propose a distribution adjustment method for adapting the Bayes classifier to the testing process.

3.1 Theoretical Motivation

Preliminaries. We start by presenting the problem setting, laying the basis for introducing our method. Given the training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)_{i=1}^N\}$, the model is trained to map the images from the space \mathcal{X} into the classes from the space $\mathcal{Y} = \{1, 2, \dots, K\}$. Typically, the mapping function φ is modeled as a neural network, which consists of a backbone feature extractor $F: \mathcal{X} \rightarrow \mathcal{Z}$ and a linear classifier $G: \mathcal{Z} \rightarrow \mathcal{Y}: z \mapsto \arg \max(\mathbf{W}^T \mathbf{z} + \mathbf{b})$. The standard Softmax cross-entropy loss for a sample $\{\mathbf{x}, y\}$ in training set can be expressed as:

$$\mathcal{L}_{\text{Softmax}}(\mathbf{x}) = -\log p(y|\mathbf{x}), \quad p(y|\mathbf{x}) = \frac{\exp(\mathbf{w}_y^T \mathbf{z} + b_y)}{\sum_{y'} \exp(\mathbf{w}_{y'}^T \mathbf{z} + b_{y'})}, \quad (1)$$

where \mathbf{w}_y and b_y are the weight and bias of the linear classifier for class y , respectively. It can be observed from Eq. (1) that the model *implicitly* estimates the posterior probability of the class.

The Bayes Classifier is the optimal classifier that minimizes the probability of misclassification, which is defined based on the Bayes’ theorem:

$$p(y|\mathbf{x}) = \frac{p(y)p(\mathbf{x}|y)}{\sum_{y'} p(y')p(\mathbf{x}|y')}. \quad (2)$$

However, capturing the true distribution of real-world data proves challenging due to its inherent complexity. As a result, most existing approaches resort to approximating the Bayes classifier by the

model’s output, as illustrated in Eq. (1). Nevertheless, in long-tail recognition tasks, the test sets are typically balanced, resulting in a disparity in the prior distribution of $p(y)$ between the training and test sets. The long-tail recognition approaches for balancing the gradients overlook this difference, thereby failing to ensure the model learns the Bayes optimal classifier. To tackle this issue, Logit Adjustment [46] introduces a prior distribution over the class labels:

$$\mathcal{L}_{LA}(\mathbf{x}) = -\log p(y|\mathbf{x}), \quad p(y|\mathbf{x}) = \frac{\pi_y \exp(\mathbf{w}_y^T \mathbf{z} + b_y)}{\sum_{y'} \pi_{y'} \exp(\mathbf{w}_{y'}^T \mathbf{z} + b_{y'})}, \quad (3)$$

where π_y is the class frequency in the training or test set. Nonetheless, these methods all implicitly estimate the posterior probability of the classes using gradient descent, without leveraging information regarding the data distribution.

3.2 Distribution Assumption

In this section, we first introduce the distribution assumption underlying our method and explain its rationality. Based on this assumption, we present the specific form of the Bayes classifier.

As mentioned earlier, the complexity of real-world data makes it difficult to directly model the data distribution. Consequently, we choose to model the data distribution in the feature space, which is generally more manageable. Our assumption is motivated by the Simplex-Encoded-Labels Interpolation (SELI) [56] used to characterize the neural collapse phenomenon in imbalanced learning. It can be inferred that the features tend to collapse towards the mean values of their corresponding classes. Thus, we can assume that the feature norms of each class sample are equal and employ the von Mises-Fisher (vMF) distribution [43] on the unit sphere to represent the feature distribution.

The vMF Distribution is a fundamental probability distribution on the unit hyper-sphere \mathbb{S}^{p-1} in \mathbb{R}^p . Its probability density function for a random p -dimensional unit vector \mathbf{z} is given by:

$$f_p(\mathbf{z}|\boldsymbol{\mu}, \kappa) = \frac{1}{C_p(\kappa)} \exp(\kappa \boldsymbol{\mu}^T \mathbf{z}), \quad C_p(\kappa) = \frac{(2\pi)^{p/2} I_{(p/2-1)}(\kappa)}{\kappa^{p/2-1}}, \quad (4)$$

where \mathbf{z} is a p -dimensional unit vector, $\kappa \geq 0$, $\|\boldsymbol{\mu}\|_2 = 1$ and $I_{(p/2-1)}$ denotes the modified Bessel function of the first kind at order $p/2 - 1$, which is defined as:

$$I_{(p/2-1)}(\kappa) = \sum_{i=0}^{\infty} \frac{1}{i! \Gamma(p/2 - 1 + i + 1)} \left(\frac{\kappa}{2}\right)^{2i+p/2-1}. \quad (5)$$

The parameters $\boldsymbol{\mu}$ and κ are referred to as the mean direction and concentration parameters, respectively. A higher concentration around the mean direction $\boldsymbol{\mu}$ is observed with greater κ , and the distribution becomes uniform on the sphere when $\kappa = 0$.

Based on the above assumption and cross-entropy loss, we can obtain the optimization target and the Bayes classifier as:

$$\mathcal{L}_{BAPE}(\mathbf{x}) = -\log p(y|\mathbf{z}), \quad p(y|\mathbf{z}) = \frac{p(y)p(\mathbf{z}|y)}{\sum_{y'} p(y')p(\mathbf{z}|y')} = \frac{\pi_y \frac{1}{C_p(\kappa_y)} \exp(\kappa_y \boldsymbol{\mu}_y^T \mathbf{z})}{\sum_{y'} \pi_{y'} \frac{1}{C_p(\kappa_{y'})} \exp(\kappa_{y'} \boldsymbol{\mu}_{y'}^T \mathbf{z})}, \quad (6)$$

where \mathbf{z} is the corresponding feature embedding of input \mathbf{x} , π_y is the class frequency in the training or test set, κ_y and $\boldsymbol{\mu}_y$ are the parameters of the vMF distribution for class y . From Eq. (6), it is evident that the classifier is a linear classifier within the feature space. However, its fundamental distinction from existing methods lies in the explicit construction of the classifier based on the Bayes’ theorem and parameter estimation through maximum a posteriori estimation rather than implicit estimation through gradient descent. Based on our distribution adjustment method, the empirical analysis is depicted in Fig. 1 and Tab. 5.

3.3 Maximum A Posteriori Estimation of the vMF Distribution

In the following section, we will present a method for estimating the parameters κ_y and $\boldsymbol{\mu}_y$ in Eq. (6) using a point estimation approach during the training process. Under the assumption of

the vMF distribution, the parameters can be estimated by maximum likelihood estimation (MLE). However, during the early stages of training, the random distribution of features will lead to unstable optimization of the classifier. To tackle this concern, we adopt the Maximum A Posteriori (MAP) estimation method to incorporate a prior distribution for parameter estimation. This approach can be viewed as a regularized MLE.

Conjugate Prior. Suppose that a series of N vectors $\{(\mathbf{z}_i)_i^N\}$ on the unit hyper-sphere \mathbb{S}^{p-1} are independent and identically distributed (i.i.d.) observations from a vMF distribution. The conjugate prior can be defined as:

$$p(\boldsymbol{\mu}, \kappa) = \frac{1}{C} \frac{1}{C_p^{\alpha_0}(\kappa)} \exp(\beta_0 \kappa \mathbf{m}_0^T \boldsymbol{\mu}), \quad (7)$$

where $\alpha_0 \geq 0, \beta_0 \geq 0, \mathbf{m}_0 \in \mathbb{S}^{p-1}$ are the parameters of the prior distribution, and C is an unknown normalization constant.

Conjugate Posterior. Given $\mathbf{Z} = \{(\mathbf{z}_i)_i^N\}$, the posterior distribution of $\boldsymbol{\mu}$ and κ takes the form:

$$p(\boldsymbol{\mu}, \kappa | \mathbf{Z}) = \frac{1}{C} \frac{1}{C_p^\alpha(\kappa)} \exp(\beta \kappa \mathbf{m}^T \boldsymbol{\mu}), \quad (8)$$

where $\alpha = \alpha_0 + N, \beta = \|\beta_0 \mathbf{m}_0 + \sum_{i=1}^N \mathbf{z}_i\|_2$ and $\mathbf{m} = (\beta_0 \mathbf{m}_0 + \sum_{i=1}^N \mathbf{z}_i) / \beta$.

Proposition 1 (MAP Estimation). *Suppose that a series of N vectors $\{(\mathbf{z}_i)_i^N\}$ on the unit hyper-sphere \mathbb{S}^{p-1} are independent and identically distributed (i.i.d.) observations from a vMF distribution. The maximum a posteriori (MAP) estimates of the mean direction $\boldsymbol{\mu}$ and concentration parameter κ satisfy the following equations:*

$$\boldsymbol{\mu} = \mathbf{m}, \quad \frac{I_{p/2}(\kappa)}{I_{p/2-1}(\kappa)} = \frac{\beta}{\alpha}. \quad (9)$$

Derived from the MAP estimates of the vMF distribution, the parameters of conjugate prior can be interpreted in terms of pseudo-observations. Specifically, \mathbf{m}_0 and β_0 represent the direction and length of the pseudo-observations, respectively. Additionally, α_0 denotes the number of pseudo-observations. This can help choose reasonable hyperparameters for the prior distribution.

Based on the aforementioned MAP estimation, we can efficiently estimate the parameters of the Bayes classifier during the training process. A simple approximation [53] to κ is:

$$\hat{\kappa} = \frac{p\beta\alpha}{\alpha^2 - \beta^2}. \quad (10)$$

Furthermore, the sample mean of each class is estimated in an online manner by aggregating statistics from the current mini-batch:

$$\bar{\mathbf{z}}_j^{(t)} = \frac{n_j^{(t-1)} \bar{\mathbf{z}}_j^{(t-1)} + s_j^{(t)} \bar{\mathbf{z}}_j'^{(t)}}{n_j^{(t-1)} + s_j^{(t)}}, \quad (11)$$

where $\bar{\mathbf{z}}_j^{(t)}$ is the estimated sample mean of class j at step t and $\bar{\mathbf{z}}_j'^{(t)}$ is the sample mean of class j in current mini-batch. $n_j^{(t-1)}$ and $s_j^{(t)}$ are the sample numbers in the previous mini-batches and the current mini-batch, respectively.

Prior Parameter. We now discuss the parameter settings for the prior distribution. For \mathbf{m}_0 , we set \mathbf{m}_0^y for each class y to form a simplex equiangular tight frame (ETF) following ETF classifier [69]. Therefore, we construct an ETF and obtain the respective \mathbf{m}_0^y , as shown below:

$$\mathbf{M} = \sqrt{\frac{K}{K-1}} \mathbf{U} \left(\mathbf{I}_K - \frac{1}{K} \mathbf{1}_K \mathbf{1}_K^T \right), \quad (12)$$

where $\mathbf{M} = [\mathbf{m}_0^1, \dots, \mathbf{m}_0^K] \in \mathbb{R}^{p \times K}$, feature dimension $p \geq (K-1)$, $\mathbf{U} \in \mathbb{R}^{p \times K}$ is a partial orthogonal matrix, \mathbf{I}_K is the $K \times K$ identity matrix, and $\mathbf{1}_K$ is the K -dimensional vector of ones. When $p < (K-1)$, \mathbf{M} are calculated following [39].

Based on the preceding discussion, in order to ensure stability during the initial stages of training, we utilize gradient updates for the parameters \mathbf{m}_0 of the prior distribution. It is worth noting that as

training progresses, the influence of the prior distribution will gradually diminish. Furthermore, as the strength of the prior distribution tends towards infinity, our approach will degenerate into a cosine classifier. The experimental results are reported in Tab. 5.

For parameters α_0 and β_0 , which represent the number and length of the pseudo-observations, respectively, a reasonable approach is to set them in proportion to the number of samples per class N_y . Following this idea, we define new hyperparameters $\hat{\alpha}_0 = \alpha_0^y/N_y, \hat{\beta}_0 = \beta_0^y/N_y, y = 1, \dots, K$. We calculate all α_0^y and β_0^y after selecting appropriate values for $\hat{\alpha}_0$ and $\hat{\beta}_0$.

3.4 Overall Objective of BAPE

As aforementioned in Sec. 3.3, our method performs a computational estimation, which undergoes unstable optimization in the early training stages. To this end, we integrate an LA classifier (a linear classifier trained with logit adjustment method) with a BAPE classifier, both of which share a common backbone. The introduction of the LA classifier facilitates stable training for BAPE. In particular, we also employ an ensemble approach for prediction. Furthermore, to reduce the coupling between the two classifiers during optimization, we employ a projection head specifically for the BAPE classifier, and generate one view and two views of an input image for LA and BAPE classifier respectively. Finally, the loss functions are weighted and summed up as the overall loss function:

$$\mathcal{L} = \mathcal{L}_{\text{BAPE}} + \eta \mathcal{L}_{\text{LA}}, \quad (13)$$

where η is the weight of the LA classifier.

Importantly, \mathcal{L}_{LA} can be replaced with any off-the-shelf long-tailed learning algorithms. In fact, we have observed that BAPE is compatible with most existing methods, *i.e.*, these methods can contribute to a stable convergence at earlier learning stages, while BAPE considerably improves the final generalization performance. In other words, the gains of BAPE in terms of *explicitly* modeling the Bayesian decision process are orthogonal to existing approaches, which are mainly designed under the principle of *implicitly* estimating the Bayesian posterior probabilities. In this paper, we mainly report the improvements on top of LA due to its state-of-the-art performance.

4 Experiment

In this section, we conduct extensive experiments on multiple long-tail visual recognition benchmarks to validate the advantages of our method.

4.1 Dataset and Evaluation Protocol

We investigate the performance of our models on four prevalent long-tailed image classification datasets: CIFAR-10/100-LT, ImageNet-LT, and iNaturalist. We adopt the partition strategy proposed in [41, 30], grouping the categories into three subsets based on the number of training samples: Many-shot categories with over 100 images, Medium-shot categories with 20 – 100 images, and Few-shot categories with fewer than 20 images. For evaluation, the top-1 accuracy on the corresponding balanced validation or test sets is reported.

CIFAR-10/100-LT. CIFAR-10-LT and CIFAR-100-LT datasets are derived from their original counterparts, CIFAR-10 and CIFAR-100 [33], by employing a sampling technique [4, 9]. Specifically, we adopt an exponential function $N_j = N \times \lambda^j$, where $\lambda \in (0, 1)$. N is the size of the original training set, and N_j is the sample number in the j -th class. The balanced validation sets from the original datasets are used for testing. The imbalance degree in the datasets is measured by the imbalance factor γ , defined as $\gamma = \max(N_j)/\min(N_j)$. In our experiments, we set γ to typical values of 10, 50, 100.

ImageNet-LT. ImageNet-LT is a long-tailed version of the ImageNet dataset, constructed by sampling a subset of the original dataset following the Pareto distribution with power value $\alpha_p = 6$ [41]. The dataset contains 115.8K images with 1,000 classes, where each class has a varying number of images, ranging from 5 to 1,280. We used the standard setup for evaluation.

iNaturalist 2018. iNaturalist 2018 [59] is a large-scale dataset that contains 437.5K images from 8142 different species. The dataset is highly imbalanced, with an imbalance factor $\gamma = 500$. This makes iNaturalist an ideal dataset for evaluating long-tailed learning methods. We used iNaturalist to test the effectiveness of our method on real-world, complex datasets.

Table 1: Top-1 accuracy of ResNet-32 on CIFAR-100-LT and CIFAR-10-LT. * denotes results borrowed from [74]. † denotes our implementation. We report the results of 200 epochs.

Dataset	CIFAR-100-LT			CIFAR-10-LT		
Imbalance Factor	100	50	10	100	50	10
CB-Focal [9]	39.6	45.2	58.0	74.6	79.3	87.5
LDAM-DRW* [4]	42.0	46.6	58.7	77.0	81.0	88.2
BBN [74]	42.6	47.0	59.1	79.8	81.2	88.3
SSP [70]	43.4	47.1	58.9	77.8	82.1	88.5
VS [32]	43.5	-	-	80.8	-	-
TSC [36]	43.8	47.4	59.0	79.7	82.9	88.7
Casual model [55]	44.1	50.3	59.6	80.6	83.6	88.5
CDT [71]	44.3	-	58.9	79.4	-	89.4
ETF Classifier [69]	45.3	50.4	-	76.5	81.0	-
LADE [26]	45.4	50.5	61.7	-	-	-
MetaSAug-LDAM [37]	48.0	52.3	61.3	80.7	84.3	89.7
GCL [44]	48.7	53.6	-	82.7	85.5	-
Logit Adj.† [46]	50.5	54.9	64.0	84.3	87.1	90.9
BAPE†	52.5	57.3	66.1	85.4	88.4	92.2

Table 2: Results on ImageNet-LT and iNaturalist 2018 (Res50/ResX50: ResNet-50/ResNeXt-50, 90-epoch training). † denotes our implementation..

Method	ImageNet-LT		iNaturalist 2018
	Res50	ResX50	Res50
τ -norm[30]	46.7	49.4	65.6
MetaSAug [37]	47.4	-	68.8
SSP [70]	51.3	-	68.1
ALA [73]	52.4	53.3	70.7
DisAlign [72]	52.9	53.4	69.5
vMF classifier [64]	-	53.7	-
SSD [38]	-	53.8	69.3
ResLT [8]	-	56.1	70.2
Logit Adj.† [46]	55.1	56.5	71.0
BAPE†	56.7	57.6	72.3

Table 3: Top-1 accuracy on ImageNet-LT. We report ResNet-50 results with 90-epoch and 180-epoch training. † denotes our implementation.

Method	Many	Medium	Few	All
<i>90 epochs</i>				
τ -norm [30]	56.6	44.2	27.4	46.7
DisAlign [72]	61.3	52.2	31.4	52.9
DRO-LT [51]	64.0	49.8	33.1	53.5
RIDE [66]	66.2	51.7	34.9	54.9
Logit Adj.† [46]	65.5	53.2	32.3	55.1
BAPE†	66.3	54.3	38.1	56.7
<i>180 epochs</i>				
Logit Adj.† [46]	68.0	52.5	34.2	56.0
BAPE†	67.5	55.3	37.9	57.6

4.2 Implementation Details

The training of all models involves the utilization of an SGD optimizer with a momentum of 0.9.

CIFAR-10/100-LT. For long-tailed CIFAR-10 and CIFAR-100, we utilize ResNet-32 [25] as the backbone network. For the BAPE classifier, we employ a projection head with a hidden layer dimension of 512 and an output dimension of 128. We apply AutoAug [6] and Cutout [11] as data augmentation strategies for the LA classifier, while SimAug [57] is utilized for the BAPE classifier. The loss weight is assigned equally ($\eta = 1$) to both classifiers. We train the network for 200 epochs with a batch size of 256 and a weight decay of $4e-4$. The prior parameters $\hat{\alpha}_0$ and $\hat{\beta}_0$ are set to 40 and 8, respectively. We adopt a cosine schedule to regulate the learning rate. This approach entails gradually ramping up the learning rate to 0.3 during the first 5 epochs, followed by a smooth factor applied that varies between 0 and 1 according to a cosine function. Unless specified, our ablation study and analysis employ these training settings. Additionally, we train the model for 400 epochs with a similar learning rate schedule to enable a more thorough comparison.

ImageNet-LT & iNaturalist 2018. We adopt ResNet-50 [25] as the backbone network for both ImageNet-LT and iNaturalist 2018. The BAPE classifier is comprised of a projection head with an output dimension of 1024 and a hidden layer dimension of 2048, while the LA classifier is employed as a cosine classifier [63]. For data augmentation, we use RandAug [7] and SimAug for the LA and BAPE classifiers, respectively. The prior parameters $\hat{\alpha}_0$ and $\hat{\beta}_0$ are set as 20 and 0.6 for ImageNet-LT, while 10 and 0.3 are set for iNaturalist 2018. We also assign equal loss weight ($\eta = 1$). The model is trained for 90 epochs with a batch size of 256 and a cosine learning rate schedule. For ImageNet-LT, the initial learning rate is set to 0.1 and the weight decay is set to $5e-4$. We also train the model for

Table 5: Analysis of distribution adjustment. ✓denotes we perform adjustment during this stage.

Adjusted Classifier	Training	Testing	Many	Medium	Few	All
w/o Adjustment			69.5	51.9	28.0	50.9
LA [46]	✓	✓	66.2 67.5	51.1 48.6	28.9 25.7	49.7 48.3
BAPE	✓	✓	68.9 68.7	53.6 53.2	28.5 32.9	51.3 52.5

90 epochs with ResNeXt-50-32x4d [68], and for 180 epochs with ResNet-50. For iNaturalist 2018, the initial learning rate is set to 0.2 and the weight decay is set to $1e-4$.

4.3 Main Results

CIFAR-10/100-LT. The comparison between BAPE and existing methods on CIFAR-100-LT and CIFAR-10-LT are summarized in Tab. 1. Our BAPE significantly outperforms the competitors, demonstrating its efficacy in long-tailed classification. We also present extended and comprehensive results in Tab. 4, ensuring the preservation of an imbalance factor of 100. Specifically, compared to Logit Adj., our approach demonstrates a notable enhancement of 3.4% for the tail classes with 200 epochs. Furthermore, our method achieves a consistent improvement of approximately 2.0% across all subsets with 400 epochs.

Table 4: Top-1 accuracy of ResNet-32 on CIFAR-100-LT (imbalance factor: 100). † denotes our implementation.

Method	Many	Medium	Few	All
<i>200 epochs</i>				
DRO-LT [51]	64.7	50.0	23.8	47.3
RIDE [66]	68.1	49.2	23.9	48.0
Logit Adj. † [46]	67.2	51.9	29.5	50.5
BAPE †	68.7	53.2	32.9	52.5
<i>400 epochs</i>				
Logit Adj. † [46]	68.1	53.0	32.4	52.1
BAPE †	70.2	55.0	34.1	54.1

ImageNet-LT. We present comprehensive evaluation results of our approach on ImageNet-LT in Tab. 2. Leveraging the ResNet-50 and ResNeXt-50 backbones with 90 epochs training, BAPE surpasses Logit Adj. by a margin of 1.6% and 1.1% respectively. Furthermore, Tab. 3 lists detailed results on more training settings for ImageNet-LT dataset. Notably, BAPE significantly outperforms the Logit Adj. in tail classes, exhibiting improvements of 5.7%. With 180 epochs, BAPE exhibits a marginal decline of 0.5% in Top-1 accuracy for head classes, whereas it demonstrates an improvement of 3.7% for tail classes, resulting in an overall accuracy enhancement of 1.6%.

iNaturalist 2018. Tab. 2 also presents the experimental results obtained from implementing our BAPE on the iNaturalist 2018 dataset. Due to its highly imbalanced nature, iNaturalist 2018 serves as an exemplary platform to investigate the influence of imbalanced datasets on the performance of machine learning models. Our BAPE outperforms Logit Adj. by 1.3% under the same setting.

4.4 Effectiveness of BAPE

Ablation Study. In this section, we conduct experiments to validate different design choices of BAPE as reported in Tab. 6. We first implement an initial version of BAPE without incorporating prior samples or distribution adjustment (DA), which yields a result of 47.3%. Subsequently, DA is applied to address the discrepancies in conditional distributions between the training and test sets. The adjusted κ leads to a performance increase of 0.8%. We then investigate the effectiveness of prior distribution. As expounded in Sec. 3.3, we generate a group of prior parameters, which guarantees more stable training, resulting in a significant

Table 6: Ablation study on CIFAR-100-LT. DA denotes the Distribution Adjustment.

Ablation	Many	Medium	Few	All
w/o Prior & DA	66.1	48.2	24.2	47.3
w/o Prior	64.2	49.9	27.2	48.1
w/o DA	69.5	51.9	28.0	50.9
Ours	68.7	53.2	32.9	52.5

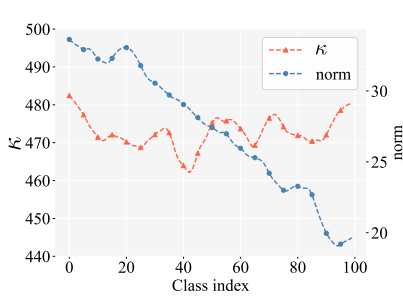


Figure 1: κ (BAPE) and the norm (products of weight norm and feature norm in LA) of different classes.

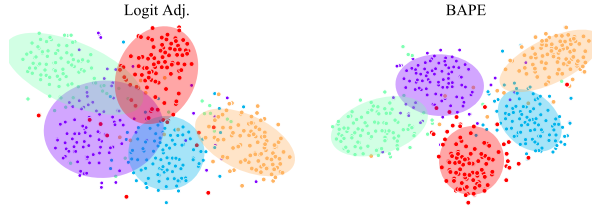


Figure 2: Visualization of the feature embedding via t-SNE. We take the mean of each category as the center and calculate the axis length and angle by utilizing the eigenvalues and eigenvectors obtained from the feature covariance matrix. This enables us to draw an ellipse capable of encompassing the majority of feature points within each category.

enhancement of 3.6%. Moreover, when combined with each other, DA and Prior exhibit substantial improvements in performance, with a respective increase of 1.6% and 4.4%. This finding demonstrates the seamless integration and synergistic effect of these two components, resulting in a further enhancement of overall performance.

Analysis of Distribution Adjustment. In this section, we embark on an in-depth exploration of the Distribution Adjustment (DA) element of our framework. The DA operates in a similar vein to the idea of post-hoc normalization on both weight and feature, albeit with a distinction that we explicitly model the norms as a parameter κ , which signifies the degree of concentration in the vMF distribution. In light of this observation, we conduct a series of experiments on multiple variants of BAPE reported in Tab. 5, employing diverse adjustment (normalization on both weight and feature for LA) configurations within the LA classifier or the BAPE classifier. All results are on CIFAR-100-LT.

To evaluate the performance of different settings, we employ the unadjusted version of BAPE as the reference standard. It is worth noting that applying adjustment to the LA classifier results in a weakened behavior, while employing adjustment to the BAPE classifier yields an improvement of accuracy. This observation provides valuable insight into the superiority of our method, which *explicitly* estimates the classifier parameters. For a clear illustration, Fig. 1 presents a visualized comparison between the norms (product of weight norm and feature norm) computed from the LA classifier and the κ parameters within the BAPE classifier of an optimized model. We can observe that the LA classifier is biased and yields norms strongly correlated with class frequency, while the BAPE classifier effectively overcomes that imbalance and focuses on learning the essential, rather than the frequency of each class, therefore learning frequency independent κ .

We also apply adjustment to LA during both stages and adopt an appropriate temperature parameter, which is equivalent to employing a cosine classifier [63]. Moreover, performing adjustment to BAPE during both stages leads to a slight decrease in accuracy, especially for tail classes. This emphasizes the necessity to preserve κ during the training stage.

Visualization of Feature Embeddings. Our approach enables the explicit estimation of the parameters of the Bayesian classifier without employing gradient descent. As a result, it effectively mitigates the issue of imbalance gradient. To depict this, we illustrate a t-SNE [58] visualization of the feature embeddings of five tail classes for optimized Logit Adj. and BAPE in Fig. 2. It can be observed that BAPE achieves a more distinct separation of different tail features compared to the Logit Adj., enabling the classifier to discern them correctly.

5 Conclusion

In this study, we introduce an innovative methodology, termed BAPE (Bayes Classifier by Point Estimation), which is devised for the *explicit* learning of a Bayes classifier. In contrast to prevailing implicit estimation methods, BAPE ensures a superior theoretical approximation of the data distribution by explicitly modeling the parameters and employing point estimation. Notably, BAPE offers dual benefits: it learns the Bayes classifier directly using Bayes’ theorem and it circumvents the need for gradient descent. This two-pronged strategy guarantees an optimal decision rule while also

addressing the issue of gradient imbalance. Furthermore, we introduce a straightforward yet potent method to adjust the distribution. This method enables our Bayes classifier, derived from an imbalanced training set, to adapt effectively to a test set with an arbitrary imbalance factor. Importantly, this adjustment technique enhances performance without accruing additional computational costs. Furthermore, we demonstrate that the advantages of our approach are independent of existing learning methodologies tailored for long-tailed scenarios, as the majority of these approaches are primarily constructed based on the principle of implicitly estimating posterior probabilities. Comprehensive experiments were conducted on popular benchmarks including CIFAR-LT-10/100, ImageNet-LT, and iNaturalist2018. The results offer strong evidence of the effectiveness and superiority of BAPE.

References

- [1] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 2018.
- [2] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *ICML*, 2019.
- [3] Dong Cao, Xiangyu Zhu, Xingyu Huang, Jianzhu Guo, and Zhen Lei. Domain balancing: Face recognition on long-tailed domains. In *CVPR*, 2020.
- [4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.
- [5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *JAIR*, 2002.
- [6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. AutoAugment: Learning augmentation strategies from data. In *CVPR*, 2019.
- [7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. RandAugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, 2020.
- [8] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. ResLT: Residual learning for long-tailed recognition. *TPAMI*, 2022.
- [9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.
- [10] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 2012.
- [11] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint*, 2017.
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [13] Chaoqun Du, Jiayi Guo, Yulin Wang, and Gao Huang. Switta: Switching domain experts and aggregating contextual features towards realistic test-time adaptation. In *ICML Workshop*, 2025.
- [14] Chaoqun Du, Yizeng Han, and Gao Huang. Simpro: A simple probabilistic framework towards realistic long-tailed semi-supervised learning. *arXiv preprint arXiv:2402.13505*, 2024.
- [15] Chaoqun Du, Yulin Wang, Jiayi Guo, Yizeng Han, Jie Zhou, and Gao Huang. Unitta: Unified benchmark and versatile framework towards realistic test-time adaptation. *arXiv preprint arXiv:2407.20080*, 2024.
- [16] Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

- [17] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *the National Academy of Sciences*, 2021.
- [18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
- [19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, 2014.
- [20] Jiayi Guo, Chaoqun Du, Jiangshan Wang, Huijuan Huang, Pengfei Wan, and Gao Huang. Assessing a single image in reference-guided image synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 753–761, 2022.
- [21] Jiayi Guo, Hayk Manukyan, Chenyu Yang, Chaofei Wang, Levon Khachatryan, Shant Navasardyan, Shiji Song, Humphrey Shi, and Gao Huang. Faceclip: Facial image-to-video translation via a brief text description. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4270–4284, 2023.
- [22] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Shiji Song, Humphrey Shi, and Gao Huang. Zero-shot generative model adaptation via image-specific prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11494–11503, 2023.
- [23] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7548–7558, 2024.
- [24] Jiayi Guo, Junhao Zhao, Chaoqun Du, Yulin Wang, Chunjiang Ge, Zanlin Ni, Shiji Song, Humphrey Shi, and Gao Huang. Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30503–30513, 2025.
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [26] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, 2021.
- [27] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016.
- [28] Gao Huang and Chaoqun Du. The high separation probability assumption for semi-supervised learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(12):7561–7573, 2022.
- [29] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.
- [30] 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.
- [31] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *CVPR*, 2019.
- [32] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *NeurIPS*, 2021.
- [33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

- [34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2017.
- [35] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, 1997.
- [36] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *CVPR*, 2022.
- [37] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. MetaSAug: Meta semantic augmentation for long-tailed visual recognition. In *CVPR*, 2021.
- [38] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *ICCV*, 2021.
- [39] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *CVPR*, 2022.
- [40] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.
- [41] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Open long-tailed recognition in a dynamic world. *TPAMI*, 2022.
- [42] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.
- [43] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional statistics*. Wiley Online Library, 2000.
- [44] Yang Lu Mengke Li, Yiu-ming Cheung. Long-tailed visual recognition via gaussian clouded logit adjustment. In *CVPR*, 2022.
- [45] Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *ICML*, 2013.
- [46] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.
- [47] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [48] Foster Provost. Machine learning from imbalanced data sets 101. In *AAAI'2000 workshop*, 2000.
- [49] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, and Shuai Yi. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, 2020.
- [50] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.
- [51] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *ICCV*, 2021.
- [52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.
- [53] Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of $i_s(x)$. *Computational Statistics*, 2012.
- [54] 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.

- [55] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020.
- [56] Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. In *NeurIPS*, 2022.
- [57] Chen Ting, Kornblith Simon, Norouzi Mohammad, and Hinton Geoffrey. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.
- [58] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, 2008.
- [59] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.
- [60] Vladimir Vapnik. Principles of risk minimization for learning theory. *NeurIPS*, 1991.
- [61] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.
- [62] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. Class imbalance, redux. In *ICDM*, 2011.
- [63] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *CVPR*, 2018.
- [64] Hualiang Wang, Siming Fu, Xiaoxuan He, Hangxiang Fang, Zuozhu Liu, and Haoji Hu. Towards calibrated hyper-sphere representation via distribution overlap coefficient for long-tailed learning. In *ECCV*, 2022.
- [65] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *CVPR*, 2021.
- [66] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2020.
- [67] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *ECCV*, 2020.
- [68] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.
- [69] Yibo Yang, Liang Xie, Shixiang Chen, Xiangtai Li, Zhouchen Lin, and Dacheng Tao. Do we really need a learnable classifier at the end of deep neural network? *NeurIPS*, 2022.
- [70] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *NeurIPS*, 2020.
- [71] Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint*, 2020.
- [72] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021.
- [73] Yan Zhao, Weicong Chen, Xu Tan, Kai Huang, and Jihong Zhu. Adaptive logit adjustment loss for long-tailed visual recognition. In *AAAI*, 2022.
- [74] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 2020.