This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Local and Global Logit Adjustments for Long-Tailed Learning

Yingfan Tao[1,*], Jingna Sun[2,*], Hao Yang[2], Li Chen[2], Xu Wang[2], Wenming Yang[1,†], Daniel Du[2], Min Zheng[2]
[1]Tsinghua University, [2]ByteDance Inc

Abstract

Multi-expert ensemble models for long-tailed learning typically either learn diverse generalists from the whole dataset or aggregate specialists on different subsets. However, the former is insufficient for tail classes due to the high imbalance factor of the entire dataset, while the latter may bring ambiguity in predicting unseen classes. To address these issues, we propose a novel Local and Global Logit Adjustments (LGLA) method that learns experts with full data covering all classes and enlarges the discrepancy among them by elaborated logit adjustments. LGLA consists of two core components: a Class-aware Logit Adjustment (CLA) strategy and an Adaptive Angular Weighted (AAW) loss. The CLA strategy trains multiple experts which excel at each subset using the Local Logit Adjustment (LLA). It also trains one expert specializing in an inversely long-tailed distribution through Global Logit Adjustment (GLA). Moreover, the AAW loss adopts adaptive hard sample mining with respect to different experts to further improve accuracy. Extensive experiments on popular long-tailed benchmarks manifest the superiority of LGLA over the SOTA methods.

1. Introduction

Deep learning has brought profound improvements to various vision tasks, including classification, detection, segmentation, *etc*. The success of deep learning is undoubtedly inseparable from large-scale well-designed datasets, such as ImageNet [9], COCO [25] and Places [51], which usually exhibit approximately uniform distribution over different classes. However, constructing these artificially balanced datasets is extremely difficult: sufficient instances must be collected for the *tail classes* whose samples are few by nature. According to the inherently existing *power law* [43], most real-world data follows a *long-tailed distribution*: a few head classes occupy a large portion of samples, while most tail classes only have small portions. It is challenging to learn directly from these long-tailed data, because deep learning models tend to be dominated by the head classes that appear most during training, while resulting in poor performance on tail classes. This paper aims to mitigate such problems in model training on long-tailed data.

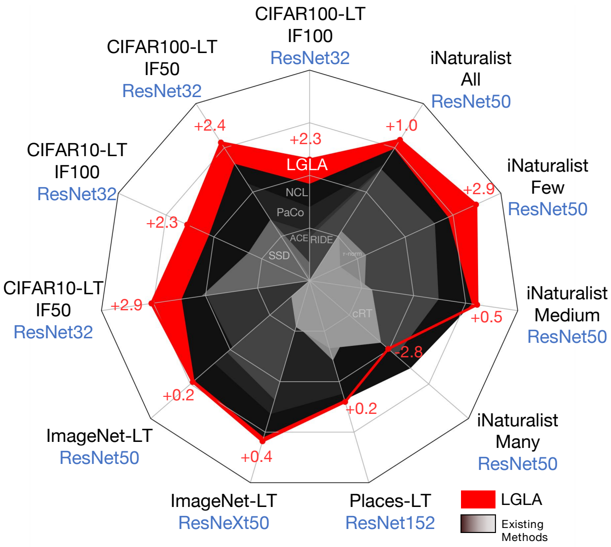
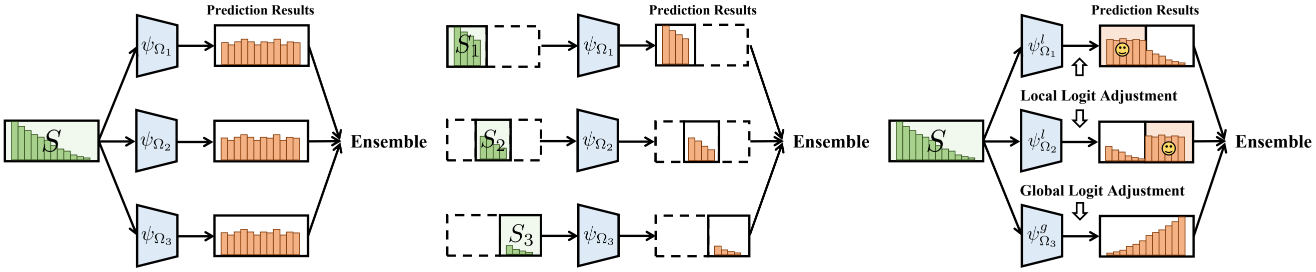


Figure 1: Our LGLA exhibits advantageous performance, i.e., Top-1 Acc (%), over existing SOTA approaches on most long-tailed benchmarks, all under the same settings for fair comparisons. On iNaturalist 2018, though surpassed on Many-shot, LGLA achieves superiority over others on Medium-/Few-shot and the overall dataset ("All").

Among the existing literature coping with the long-tail problem, some design class re-balancing strategies for training, including re-sampling [3, 12, 17, 33] or cost-sensitive learning [24, 20, 2, 40, 34]. In addition, decoupled learning proposes a two-stage training process that decouples the representation learning and the classifier learning [18, 23] to preserve the broken feature representation caused by the re-balancing methods. Most recent efforts rely on ensemble learning to achieve state-of-the-art performances on long-tail visual recognition, where multiple experts are trained in a complementary manner, then aggregated together for in-

[*] Equal contribution
[†] Corresponding author



(a) Training generalists with full data.

(b) Training specialists with partial data.

(c) Ours: LGLA with full data.

Figure 2: (a) Some ensemble methods train generalists ($\psi_{\Omega_{1,\ldots,3}}$) on the entire dataset S that is severely imbalanced. (b) Some others train specialists on individual subsets ($S_{1,\ldots,3}$) that are less imbalanced, but these specialists may suffer from limited perceptions. (c) Our method trains all experts **on the entire dataset** S, with a novel CLA strategy to ensure an adaptive local/global awareness for the experts. Assuming this is a three-expert model, the first two experts $\psi^l_{\Omega_{1,2}}$ controlled by LLA excel at different subsets (marked with smile symbols), while the last expert ($\psi^g_{\Omega_3}$) optimized by GLA will further boost the ensemble results and delivers superior performance.

ference. The ensemble process involves averaging the predictions of each expert to make the final decision. These experts are commonly trained either using the entire dataset to create generalists [39, 21] (Figure 2a) or by combining specialists trained on different subsets [1, 6, 41] (Figure 2b). However, the former approach experiences a high level of data imbalance during training, while the latter restricts the vision of each model seeing only a subset of the training data, which impairs the capabilities of each specialist as well as the overall ensemble model since they only encounter a limited number of classes and lack effective collaboration during training.

Inspired by the above insights, we propose a novel framework for long-tailed recognition: the **L**ocal and **G**lobal **L**ogit **A**djustments (**LGLA**) framework. LGLA contains two core parts: the *Class-aware Logit Adjustment* (**CLA**) strategy and the *Adaptive Angular Weighted* (**AAW**) loss. CLA has two adjustment strategies, namely *Local Logit Adjustment* (**LLA**) and *Global Logit Adjustment* (**GLA**). LLA trains multiple experts on the whole dataset but specializing in non-overlapping subsets and possessing the most comprehensive knowledge in their respective areas. Meanwhile, GLA trains one expert that achieves a global perspective and excels at handling inversely long-tailed distributions. AAW further improves classification performance by introducing adaptive hard sample mining, which enhances discriminative ability by capturing and reweighting hard samples for each expert during training.

The core insight of LGLA is simple: *each skill-diverse expert should always have access to the whole data during training*. The various skills of each expert controlled by LLA within a certain subset, as well as the capability of the last expert learned by GLA, should be obtained through an adaptive approach, instead of arbitrarily restricting their training data. Besides, the sharing of the same vision for all experts would also eliminate possible ambiguity during their ensemble.

As illustrated in Figure 2, instead of training generalists on the full dataset (Figure 2a), or training specialists on partial subsets (Figure 2b), LGLA (Figure 2c) combines the advantages of both the generalist and the specialists by training experts all on the full data, leading to diversely skilled recognition ability w.r.t both fine-grained subsets and the whole data distribution.

To sum up, LGLA inherits the strengths of generalists and specialists, resulting in improved classification performance on long-tailed data. As illustrated by Figure 1, under a fair setting ensuring the same backbone and same training data, our approach achieves superior performance over existing methods on most public benchmarks. Overall, our contributions can be summarized as follows:

- We propose a novel Local and Global Logit Adjustments (LGLA) method to boost long-tailed recognition tasks. LGLA possesses the merits of more fine-grained optimization, higher diversity among experts, and an entire feature space learning.
- We propose a Class-aware Logit Adjustment (CLA) strategy to instruct differentiated learning among experts, and an Adaptive Angular Weighted (AAW) loss for adaptive instance re-weighting, which better handles samples of different difficulties.
- Extensive experiments on popular long-tailed benchmarks, including CIFAR-10/100-LT [8], ImageNet-LT[27], iNaturalist 2018 [36] and Places-LT [51] have demonstrated the superiority of LGLA over the SOTA competitors, as shown by Figure 1.

2. Related Work

This section introduces existing methods for long-tailed learning. We focus on training strategies and intentionally exclude methods that require a large amount of extra training or pre-training data. [28, 35].

Class Re-balancing Class re-balancing handles long-tailed distribution problem by re-balancing the contribution of each class during training. This method can be divided into two types: re-sampling and cost-sensitive learning. The re-sampling methods re-balance data distribution by over-sampling the tail classes [3, 12] or under-sampling the head classes [11, 17, 33]. However, over-sampling often causes tail category overfitting, while under-sampling inevitably neglects a large amount of information from head data. On the other hand, the cost-sensitive learning methods emphasize the tail classes by assigning larger weights on them [24, 20, 2, 40] or randomly neglect the gradients from head classes [34] by assigning different weights in loss for different samples. Nevertheless, cost-sensitive learning often makes the network difficult to optimize, especially on large-scale data [30]. It also brings more performance fluctuations on tail classes [43].

Decoupled Learning Although yielding great improvement in classifying long-tailed data, re-balancing methods often suffer from the degeneration of feature representation during the representation learning stage. To cope with this, two-stage decoupled methods are proposed [18, 4, 16, 44, 50], which first train a feature extractor with standard instance-balanced sampling, then fine-tune the classifier with class-balanced sampling. Previous works focusing on classifier learning can be summarized in three ways: cRT, NCM and τ-norm [18]. Decoupled learning makes it more convenient and effective to combine re-balancing strategies, since it only trains the classifier and thus will not impair representation learning. However, two-stage learning makes the training process more redundant and complex than end-to-end learning.

Ensemble Learning Ensemble learning utilizes complementary knowledge by training and aggregating multiple experts, which can be further divided into two categories: The first category [39, 21, 46, 50] directly trains each expert on the whole dataset. For example, NCL [21] highlights the importance of cooperation by conducting distillation between every two model pairs. RIDE [39] brings about diversified experts by maximizing the KL-divergence between the predicted probabilities from diverse models. However, training on a highly imbalanced dataset inevitably causes negligence in the sample-few classes. The second category [41, 1, 6] often trains each expert on a subset of training data, which leads to a less severe data imbalance problem and results in better performance. For instance, ResLT [6] trains three experts on three subsets respectively corresponding to the all, the medium+tail, and the tail classes, and combines their outputs as the final result. However, the learned feature space of each expert in ResLT is incomplete with unseen classes, which would bring ambiguity to the ensemble results. Unlike these methods, our LGLA takes advantages of both strategies above. Through an adaptive training on the whole data for all experts, and by leveraging a novel logit adjustments strategy combining LLA and GLA, LGLA achieves advantageous performance on most public benchmarks over existing approaches.

3. Methodology

3.1. Preliminaries

As shown by Figure 3, LGLA consists of K experts with a shared backbone f_Ω. Denote the experts as $\Psi = \{\psi^l_{\Omega_1}, ..., \psi^l_{\Omega_{K-1}}, \psi^g_{\Omega_K}\}$, with $\psi^l_{\Omega_{1\leq k\leq K-1}}$ being the $K-1$ experts controlled by LLA and the last expert $\psi^g_{\Omega_K}$ learned by GLA; Ω_k is the parameter of the k-th expert. We use $W_k = \{W_k^1, W_k^2, ..., W_k^C\} \in \mathbb{R}^{d\times C}$ to represent the weight of the last FC layer within the classifier of the k-th expert (which is not shared across experts), where C is the number of classes, d denotes the dimension of features. Note that the rows $W_k^{1\leq j\leq C}$ of W_k can also be regarded as the center of the learned features w.r.t the k-th expert on the j-th class.

Let $S = \{x_i, y_i\}$ be the training data, with x_i being the i-th image and y_i the corresponding label. Given x_i, the backbone f_Ω extracts features $z_i = f_\Omega(x_i)$ from x_i, then the classifier ψ_{Ω_k} obtains the logit $v_i = \psi_{\Omega_k}(z_i)$ from z_i. It is straightforward to apply softmax onto the logit v_i as:

$$p(x_i) = \frac{\exp\left(v_i^{y_i}\right)}{\sum_{j=1}^{C}\exp\left(v_i^j\right)}, \tag{1}$$

with v_i^j being the j-th value of v_i. Losses like cross-entropy are then imposed on $p(x_i)$ for training.

3.2. Class-aware Logit Adjustment Strategy

The vanilla softmax suffers from the negligence of the discrepancy between the posterior distributions of training and test data, leading to biases under long-tailed scenarios. In order to encourage greater intra-class compactness and inter-class separability, *logit adjusting* – a strategy proposed by face recognition approaches [10, 37, 26] – also reveals its potential in some long-tailed learning methods [31, 29, 2, 48]. Existing methods employ logit adjustment with the following formula:

$$p(x_i) = \frac{\exp\left(v_i^{y_i} + T(y_i)\right)}{\sum_{j=1}^{C}\exp\left(v_i^j + T(j)\right)}, \tag{2}$$

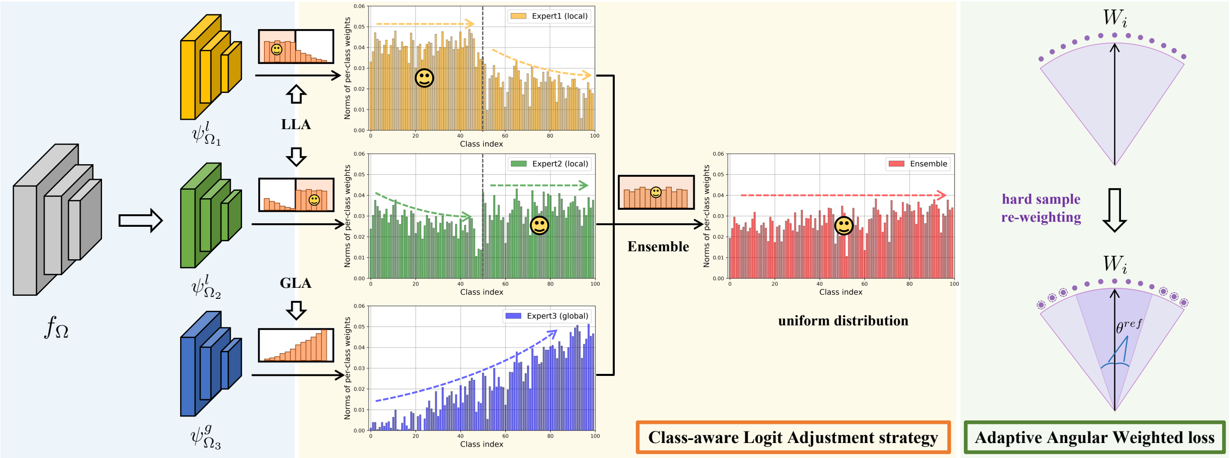


Figure 3: The framework of LGLA: (1) A shared backbone f_Ω for feature learning and a group of experts $\psi^{l|g}_{\Omega_k}$ for ensemble learning. (2) CLA strategy conbines LLA and GLA: LLA trains $K-1$ experts to be skilled w.r.t different subsets. The smile symbol marks the specialized classes, while in the non-specialized regions, experts are still affected by long-tailed data. GLA is proposed to deal with it, which learns an expert excelling at an inversely long-tailed distribution. Finally, the ensemble result presents uniform distribution. (3) AAW loss is designed to adaptively re-weight hard samples (in the light purple area) whose angles are greater than expert-associated references (θ^{ref}) during training.

where $T(\cdot)$ is a logit adjustment function conditioning only on class labels. Unlike such logit adjustment used by existing methods [31, 29, 2, 48], our proposed Class-aware Logit Adjustment (CLA) strategy cooperates with K experts presenting different adjustments for different experts, and the probability output of the k-th expert can be formulated as:

$$p(x_i, k) = \frac{\exp\left(v^{y_i}_{i,k} + T(k, y_i)\right)}{\sum_{j=1}^{C} \exp\left(v^{j}_{i,k} + T(k, j)\right)}. \tag{3}$$

Unlike the $T(j)$ in Eq. 2 which conditions only on the class label j, $T(k, j)$ in Eq. 3 is designed to reflect the awareness of the k-th expert w.r.t the j-th class. In specific, we first sort all the classes in descending order according to their cardinality; then we divide all the C classes into $K-1$ groups $C = \{C_1, C_2, ..., C_{K-1}\}$, ensuring that the number of samples $\|S_k := \{(x_i, y_i) : y_i \in C_k\}\|$ belonging to each group C_k is approximately equal, *i.e.* $\|S_p\| \approxeq \|S_q\|, 1 \leq p, q \leq K-1$. Through this way, C_1 contains mostly the head classes while C_{K-1} includes mostly the tail classes. The entire training dataset can be denoted as $S = S_1 \bigcup S_2 \bigcup ... \bigcup S_{K-1}$.

Now, we list the designed CLA function $T(k, j)$ of the k-th expert model for the class j as follows:

$$T(k, j) = \begin{cases} \log(n_j), & j \in C_k, \quad k \neq K \\ \log(n_{max}), & j \notin C_k, \quad k \neq K \\ \tau \cdot \log(n_j), & k = K \end{cases} \tag{4}$$

where n_j is the frequency of class j, n_{max} denotes the frequency of the largest class, and τ is a hyper-parameter.

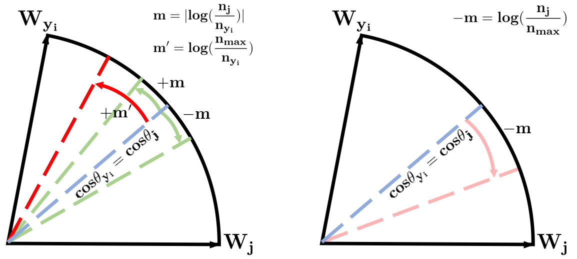
To illustrate the intuition behind our design, we substitute Eq. 4 into Eq. 3, obtaining the following equation:

$$P(x_i, k) = \frac{\exp\left(v^{y_i}_{i,k}\right)}{\sum_{j=1}^{C} \exp\left(v^{j}_{i,k} + M\left(y_i, j, k\right)\right)}, \tag{5}$$

with $M\left(y_i, j, k\right) := T(k, j) - T(k, y_i)$. $M(y_i, j, k)$ adjusts the decision boundaries between the ground truth class y_i and all the other non ground truth classes j for the k-th expert. Its formulation can be expanded as:

$$M(y_i, j, k) = \begin{cases} \log(\frac{n_j}{n_{y_i}}), & y_i \in C_k, j \in C_k, k \neq K \\ \log(\frac{n_{max}}{n_{y_i}})[> 0], & y_i \in C_k, j \notin C_k, k \neq K \\ \log(\frac{n_j}{n_{max}})[< 0], & y_i \notin C_k, j \in C_k, k \neq K \\ \log(\frac{n_{max}}{n_{max}})[= 0], & y_i \notin C_k, j \notin C_k, k \neq K \\ \tau \cdot \log(\frac{n_j}{n_{y_i}}), & y_i, j \in C, k = K \end{cases}. \tag{6}$$

Local Logit Adjustment We propose Local Logit Adjustments (LLA) to meticulously explore subset representation ability and eliminate ambiguity from unseen classes. LLA, denoted by the first four lines ($k \neq K$) in Eq. 6, trains the first $K-1$ experts with full data and restricts their specialization to the sample groups $S_1, ..., S_{K-1}$. In specific, when $y_i \in C_k$ and $j \in C_k$, indicating that both the two classes y_i and j belong to the group that the k-th expert is skilled in, then $M(y_i, j, k) = \log(n_j / n_{y_i})$ which is positive when $n_{y_i} < n_j$ while negative when $n_j < n_{y_i}$. As



$\mathbf{W_{y_i}}$
$\mathrm{m = |log(\frac{n_j}{n_{y_i}})|}$
$\mathrm{m' = log(\frac{n_{max}}{n_{y_i}})}$
+m
+m'
−m
$\cos\theta_{y_i} = \cos\theta_j$
$\mathbf{W_j}$
$\mathbf{W_{y_i}}$
$\mathrm{-m = log(\frac{n_j}{n_{max}})}$
−m
$\cos\theta_{y_i} = \cos\theta_j$
$\mathbf{W_j}$

(a) $y_i \in C_k$

(b) $y_i \notin C_k$

Figure 4: The decision boundary of first $K-1$ experts controlled by LLA in different scenarios. (a) The ground truth class y_i is in the specialized area. (b) The ground truth class y_i is not in the specialized area.

depicted by Figure 4a, this means that during the training, the decision boundary (blue line) would be pushed toward y_i with a margin of $m = |\log(n_j/n_{y_i})|$ (upper green line) if the size of class y_i is smaller, while pushed toward j with the same margin m (lower green line) if the size of class j is smaller, to ensure that the samples in the smaller classes are more emphasized. However, when $y_i \in C_k$ and $j \notin C_k$, meaning that the k-th expert specializes in y_i but does not care about j, $M = \log(n_{max}/n_{y_i}) > 0$: the boundary is pushed toward y_i by a larger margin (red line in Figure 4a) so that samples in the class y_i are more emphasized.

Figure 4b shows the other two cases when $y_i \notin C_k$. If $j \in C_K$, the decision boundary (pink line) is moved from y_i to j, highlighting the classes j in C_k that are specialized by the k-th expert. If both y_i and j are not in C_k, the classification loss degrades to the vanilla Softmax (blue line), since the expert does not care about both of them. Generally speaking, the design of the LLA always extracts more discriminative feature representations for classes in C_k (where the k-th expert is good at, $k \neq K$) by tightening their decision boundaries.

Global Logit Adjustment LLA enables the first $K-1$ experts to acquire specialized knowledge on certain subsets. However, the non-specialized subsets trained with original Softmax are still susceptible to the long-tailed distribution. To address this issue and further promote the ensemble performance, we introduce Global Logit Adjustment (GLA), presented in the last line ($k = K$) of Eq. 6. GLA trains another expert to learn an inversely long-tailed distribution, which offsets the long-tailed distribution of the non-specialized subsets in the first $K-1$ experts. The decision boundary for this expert has an offset of $\tau \cdot \log(n_j/n_{y_i})$, with a scaler $\tau > 1$ to amplify the margin and adjust the ensemble results.

Figure 3 also shows the L2 norms of per-class weights from the classifiers of a three-expert LGLA model. Higher norms in a classifier usually contribute more to the performance [50]. The first two experts excel at the first and last subsets, respectively, while the last expert is skilled in the inversely long-tailed distribution. The ensemble weights are obtained by averaging the weights of all experts, resulting in an approximately uniform among various categories, which shows the effectiveness of our proposed LGLA on a uniform test set. It is noteworthy that LGLA enlarges the variety among all the experts meanwhile ensures the integrity of each learned feature space.

3.3. Adaptive Angular Weighted Loss

Hard sample mining emphasizes hard instances during training by giving them higher weights in loss. During the process of training, tail classes iterate fewer times than head classes and gradually become hard to score, so the instance-level re-weighting approach is important and effective for long-tailed learning. However, it is strenuous to find suitable weights to fit the model. Existing approaches like OHEM [32] score hard examples with high weights, yet it ignores the optimization of easy samples; Focal loss [24] reduces the weights of simple samples, while paying more attention to hard samples, but it does not consider the nature of distribution within long-tailed learning. Different from these methods, we propose a novel Adaptive Angular Weighted (AAW) loss to address hard instances for long-tailed data and re-weight them adaptively during training.

We measure the difficulty of the i-th sample w.r.t the k-th expert using the angle between the input feature ($f_k^{y_i}$) of the last FC layer and its corresponding class centers of the k-th expert ($W_k^{y_i}$) in the cosine space, which is formulated as:

$$\theta_k^{y_i} = \arccos\left(\frac{W_k^{y_i} \cdot f_k^{y_i}}{||W_k^{y_i}|| \cdot ||f_k^{y_i}||}\right). \tag{7}$$

For each iteration, we compute $K-1$ *reference* angles using $\theta_k^{ref} = mean(\theta_k^{y_i}), y_i \in C_k$ to guide the optimization of samples belonging to different groups. Based on the fact that features with larger cosine distances are more difficult to learn, we define an instance in C_k with the angle greater than θ_k^{ref} as the hard sample and make its weight increase to $1 + \theta_k^{y_i} - \theta_k^{ref}$. Meanwhile, the samples whose angles are smaller than the reference angles will not be neglected by keeping their weights unchanged. Therefore, AAW loss can adaptively re-weight those instances away from the class center during different training phases, and the re-weighting function is formulated as:

$$g(\theta_k^{y_i}, \theta_k^{ref}) = \begin{cases} 1, & \theta_k^{y_i} \leq \theta_k^{ref} \\ 1 + \theta_k^{y_i} - \theta_k^{ref}, & \theta_k^{y_i} > \theta_k^{ref} \end{cases} \tag{8}$$

Cooperated with CLA, the overall loss of our proposed LGLA is formulated as:

$$L = -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N} g(\theta_k^{y_i}, \theta_k^{ref}) \log \frac{\exp\left(v_{i,k}^{y_i} + T(k, y_i)\right)}{\sum_{j=1}^{C} \exp\left(v_{i,k}^{j} + T(k, j)\right)} \tag{9}$$

Method	CIFAR-100-LT		CIFAR-10-LT	
	100	50	100	50
CB Focal loss [8]	38.7	46.2	74.6	79.3
LDAM+DRW [2]	42.0	45.1	77.0	79.3
LDAM+DAP [16]	44.1	49.2	80.0	82.2
LDAM+M2m [19]	43.5	–	79.1	–
BBN [50]	39.4	47.0	79.8	82.2
LFME [41]	42.3	-	-	-
CAM [47]	47.8	51.7	80.0	83.6
RIDE [39]	49.1	-	-	-
Logit Adj. [29]	43.9	-	77.7	-
LADE [15]	45.4	50.5	–	–
MiSLAS [49]	47.0	52.3	82.1	85.7
Hybrid-SC [38]	46.7	51.9	81.4	85.4
DiVE [14]	45.4	51.3	-	-
SSD [22]	46.0	50.5	-	-
ACE [1]	49.6	51.9	81.4	84.9
PaCo [7]	52.0	56.0	-	-
ResLT [6]	49.7	54.5	-	-
NCL [21]	54.2	58.2	85.5	87.3
Ours	56.5	60.6	**87.8**	**90.2**
Ours (GC)	**57.2**	**61.6**	87.5	89.8

Table 1: Top-1 accuracy (%) on CIFAR-100-LT and CIFAR-10-LT with IF=100/50. All the methods use the ResNet-32 backbone. The best and the secondary results are marked in **bold** and underline, respectively.

3.4. Advantages over Previous Methods

In general, four appealing properties of LGLA make it stand out among previous multi-expert ensemble methods: First, LLA makes experts extremely explore the representation ability in each subset. Preserving the merits of previous ensemble specialists methods [1, 41], we also shrink down the specialized area to a subset with a lower imbalance factor, which is more helpful to extract the discrimination among long-tailed classes. Different from them, all the experts in LGLA possess the same feature space, which is more reasonable to fuse the logits for obtaining stunning ensemble performance. Secondly, the number of experts in LGLA depends on the number of subsets, which can be easily expanded for a more fine-grained optimization, and experiments in Figure 7 verify that more experts can get better results. Thirdly, AAW loss can adaptively re-weight hard samples in training to further improve the recognition ability. Finally, as shown in Figure 6, the ensemble model can bias to a specific region, like many, medium, or few split, by increasing the weights of the logit from the k-th expert, which makes our method adapt to test sets with different class distributions.

4. Experiments

4.1. Datasets and Protocols

We validate the effectiveness of our proposed method on five major long-tailed datasets: CIFAR-10/100-LT [8], ImageNet-LT[27], iNaturalist 2018 [36] and Places-LT [51]. **CIFAR-10/100-LT** [8] are the different long-tailed versions of the original CIFAR datasets [20], with 10 and 100 classes, respectively. Both of them follow an exponential decay in the sample sizes across different classes, which is controlled by an Imbalance Factor (IF). IF evaluates how imbalanced a dataset is, denoted as the ratio of the number of the largest class to that of the smallest class. Two values of IF, namely 50 and 100 are used for each dataset. **ImageNet-LT** [27] is created from the ImageNet [9] dataset following the Pareto distribution with power value $\alpha = 6$. It obtains 115K images from 1,000 classes with the IF=256. **Places-LT** is a terribly imbalanced version of Places [51], containing 62.5K samples of 365 categories with the IF=996. **iNaturalist 2018** [36] is currently the largest long-tailed visual dataset, possessing 437.5K images from 8,142 classes with the IF=512.

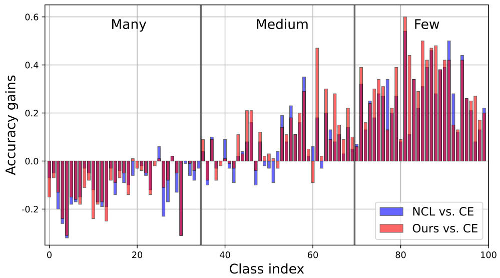


Figure 5: Accuracy gain comparisons between our method and SOTA on all the many-/medium-/few-shot. Experiments are conducted on CIFAR-100-LT dataset (IF=100).

We evaluate our proposed method on balanced test sets after long-tailed training and report the top-1 accuracy on all categories. We follow [27], to divide a validation dataset into three subsets: many (with more than 100 instances), medium (with 20 to 100 instances), and few (with less than 20 instances) splits.

4.2. Implementation Details

Following [2, 21, 39], we adopt ResNet-32 [13] as the backbone for CIFAR-10/100-LT, ResNet-50/ResNeXt-50 [42] for ImageNet, ResNet-50 for iNaturalist 2018 and pretrained ResNet-152 for Places-LT, respectively. In addition, current ensemble methods [46, 21] constitute independent convolution layers and a fully-connected layer as an expert. However, we find that group convolution provides a better alternative, which can slightly promote classification accuracy on many test sets without introducing additional parameters. The effect of group convolution will be discussed in 4.4. Moreover, following [21, 7], we use RandAugment [5] on all the benchmarks except Places-LT, and 8 NVIDIA Tesla V100 GPUs are employed for training. We exclude methods that rely on a large amount of extra training or pre-training data [28, 35] for fair comparison.

Method	ImageNet-LT		Places-LT
	ResNet-50	ResNeXt-50	ResNet-152
OLTR [27]	-	-	35.9
BALMS [31]	-	-	38.7
BBN [50]	48.3	49.3	-
NCM [18]	44.3	47.3	36.4
cRT [18]	47.3	49.6	36.7
τ-norm [18]	46.7	49.4	37.9
LWS [18]	47.7	49.9	37.6
RIDE [39]	55.4	56.8	-
DisAlign [45]	52.9	–	-
DiVE [14]	53.1	-	-
SSD [22]	-	56.0	-
ACE [1]	54.7	56.6	-
PaCo [7]	57.0	58.2	41.2
ResLT [6]	-	57.6	41.0
NCL [21]	59.5	60.5	41.8
Ours	**59.7**	60.9	**42.0**
Ours (GC)	59.6	**61.1**	-

Table 2: Top-1 accuracy (%) on ImageNet-LT and Places-LT. For Places-LT, we only report the result of "Ours", due to the usage of pretrained model following [7, 6, 21].

Method	iNaturalist 2018			
	Many	Medium	Few	All
OLTR [27]	59.0	64.1	64.9	63.9
BBN [50]	49.4	70.8	65.3	66.3
DAP [16]	-	-	-	67.6
cRT [18]	69.0	66.0	63.2	65.2
τ-norm [18]	65.6	65.3	65.9	65.6
LDAM+DRW [2]	-	-	-	68.0
Logit Adj. [29]	-	-	-	66.4
CAM [47]	-	-	-	70.9
RIDE [39]	70.9	72.4	73.1	72.6
ACE [1]	-	-	-	72.9
PaCo [7]	-	-	-	73.2
ResLT [6]	**73.0**	72.6	73.1	72.9
NCL [21]	72.7	75.6	74.5	74.9
Ours	69.9	76.1	77.4	75.9
Ours (GC)	70.1	**76.2**	**77.6**	**76.2**

Table 3: Top-1 accuracy (%) on iNaturalist 2018 with ResNet-50.

4.3. Comparing with Existing Methods

Experimental results on the major long-tailed benchmarks, including CIFAR-10/100-LT [8], ImageNet-LT[27], iNaturalist 2018 [36] and Places-LT [51] are listed in Table 1 2 3, respectively, demonstrating the superiority of our method over the state-of-the-art. To ensure the fairness of comparisons, we list the results of LGLA with both the original network and K-group convolution (denoted by GC). It is worth noting that without additional explanation, the number of experts in LGLA is set to 3 for fair comparisons with other three-expert ensemble modules [46, 21, 39, 1].

CIFAR-10/100-LT. Table 1 suggests that our proposed LGLA consistently outperforms the state-of-the-art by a large margin on CIFAR-100-LT and CIFAR-10-LT with IF=50 and IF=100. Concretely, compared with the current SOTA[21], obvious improvements are gained by 2.3% (IF=100) and 2.4% (IF=50) on CIFAR-100-LT, and 2.3% (IF=100) and 2.9% (IF=50) on CIFAR-10-LT. When using "GC", the accuracies are further improved to 57.2% (IF=100) and 61.6% (IF=50) on CIFAR-100-LT without introducing additional parameters. Class-wise accuracy gain with NCL is compared in Figure 5, which shows the effectiveness of LGLA on medium-/few-shot.

Method	w/o RandAug	w/ RandAug
Softmax	41.88	47.97
BALMS	47.97	55.24
NCL	49.22	54.42
Ours (GC)	**49.58**	**57.15**

Table 4: Top-1 accuracy (%) of distinct models on CIFAR-100-LT (IF=100) trained with or without RandAugment.

Softmax	BALMS	CLA	AAW	Acc.
✓				47.97
	✓			55.24
		✓		56.53
	✓		✓	56.51
		✓	✓	**57.15**

Table 5: Ablation studies on CIFAR-100-LT (IF=100). A three-expert BALMS [31] model is trained with RandAugment[5] to illustrate the effectiveness of the proposed CLA and AAW.

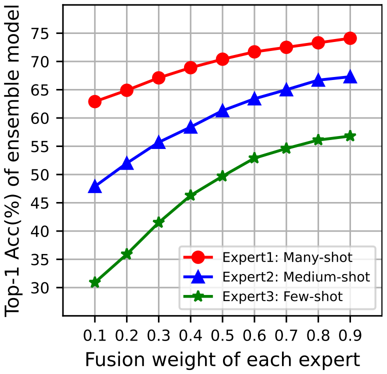


Figure 6: Effects of the fusion weights controlled by LLA.

ImageNet-LT. We report the results on ImageNet-LT with different backbones in Table 2. It can observe that the performance is further improved to 59.7% with backbone ResNet-50 and 61.1% (GC) with backbone ResNeXt-50.

iNaturalist 2018. Table 3 presents the performance comparison of LGLA and other SOTA competitors on iNaturalist 2018. We find that LGLA results in around 3% better performance on few-shot, showing the effectiveness of our method on tail classes. Finally, there is around 1.3% (GC) performance gain over SOTA in all categories.

Places-LT. Following previous work [7, 6, 21], we use a pretrained model on ImageNet and fine-tune 30 epochs on Places-LT. So that we only provide results with the original expert network. As shown in Table 2, we can observe that

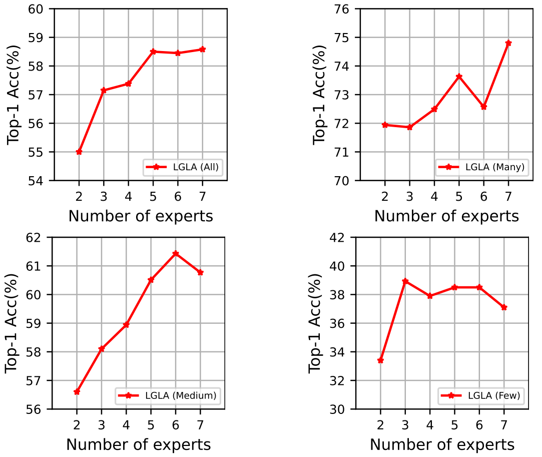


Figure 7: Comparing different expert numbers for each split (All and Many-/Medium-/Few-shot) of CIFAR-100-LT with IF=100.

our proposed LGLA still yields a state-of-the-art result on the extremely imbalanced benchmarks Places-LT at 42.0%.

4.4. Component Analysis

Data augmentation. Data augmentation techniques have proved to be useful in long-tailed learning. Many works adopt data augmentation for obtaining richer feature representations. For example, [1, 47, 49] use Mixup and [21, 7] employ RandAugment [5]. Following PaCo [7] and NCL [21], RandAugment is used in our method. In order to investigate the effect of data augmentation, we conduct experiments of different ensemble methods with and without RandAugment. From Table 4, we can observe that RandAugment brings significant improvement to classification accuracies. However, our method sees a major performance increase over the current SOTA methods irrespective of whether RandAugment is used or not.

CLA. We evaluate the effectiveness of CLA on CIFAR-100-LT. The first three rows of Table 5 show the classification accuracies of a three-expert model trained with Softmax, BALSM, and CLA respectively. It observes that compared with Softmax and BALSM, CLA considerably improves the performance by 8.56% and 1.29% respectively.

To validate that LLA is actually making experts excel at a specific data region, we train a 4-expert model on CIFAR-100-LT, with each expert specializing in Many-/Medium/Few-shot, respectively. The weight of the last expert controlled by GLA is fixed, we increase the fusion weight of each expert from 0.1 to 0.9 and evaluate the ensemble model on each split. Figure 6 shows that when we raise the fusion weight of an expert, the corresponding accuracy on the ensemble model simultaneously increases, which makes LGLA adapt to different class distributions.

AAW loss. As shown in Table 5, we apply AAW to different classification losses, namely BALMS, and CLA, to explore the effects of the proposed Adaptive Angular Weighted (AAW) loss. There can be seen obvious increases in accuracy by adding AAW to BALMS (55.24% to 56.51%), and CLA (56.53% to 57.15%). This indicates that AAW is helpful to improve performance on long-tailed data.

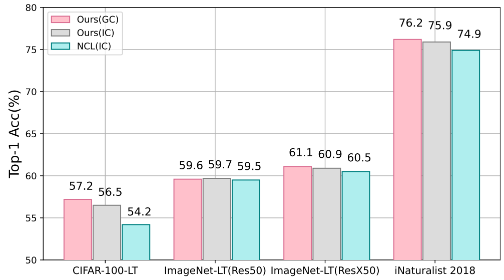


Figure 8: Comparisons of Independent Convolution (IC) and Group convolution (GC) with different methods.

Number of experts. LGLA is a multi-expert model, and we evaluate the performance under different numbers of experts. From Figure 7, we can observe that increasing the number of experts from 2 to 7 brings obvious improvements to "All" classes. In specific, The performance is greatly improved when only using a small number of experts, however, when $K > 3$, the performance gains tend to be stable.

Independent Convolution vs. Group Convolution. The current ensemble methods [39, 21, 46, 31] construct each expert with Independent Convolution (IC) layers and a fully-connected layer, but we find that using Group Convolution (GC) instead of IC can achieve better results without introducing extra parameters. When using GC, representations should be firstly concatenated K times and then K groups of convolution are utilized for representation learning. Then the obtained K groups of features are sent to K independent fully-connected layers, respectively. Figure 8 shows the results of LGLA using IC and GC with current SOTA [21] which employs IC for comparison. It observes that GC results in relatively higher accuracies and whether IC or GC is used, our method achieves better performance.

5. Conclusion

This paper presents a novel Local and Global Logit Adjustments (LGLA) method for long-tailed learning. LGLA consists of two main components: CLA and AAW. The CLA strategy, applies LLA to extract more discriminative features for each subset and applies GLA to further build superior ensemble performance. The AAW loss also boosts performance by capturing and adaptively re-weighting hard instances according to the deviations between features and class centers. Extensive experiments on popular long-tailed benchmarks highlight the effectiveness and superiority of our LGLA over SOTA methods.

References

[1] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–121, 2021.

[2] 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*, 32, 2019.

[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[4] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *European Conference on Computer Vision*, pages 694–710. Springer, 2020.

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[6] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022.

[7] 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.

[8] 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.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[11] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003.

[12] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 235–244, 2021.

[15] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021.

[16] 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.

[17] Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung. Classification of imbalanced data by combining the complementary neural network and smote algorithm. In *International Conference on Neural Information Processing*, pages 152–159. Springer, 2010.

[18] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[19] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13896–13905, 2020.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[21] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2022.

[22] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 630–639, 2021.

[23] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10991–11000, 2020.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[26] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 212–220, 2017.

[27] 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.

[28] Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. A simple long-tailed recognition baseline via vision-language model. *arXiv preprint arXiv:2111.14745*, 2021.

[29] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[31] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Proceedings of Neural Information Processing Systems(NeurIPS)*, Dec 2020.

[32] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[33] Muhammad Atif Tahir, Josef Kittler, and Fei Yan. Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, 45(10):3738–3750, 2012.

[34] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11662–11671, 2020.

[35] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022.

[36] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[37] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[38] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 943–952, 2021.

[39] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

[40] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer, 2020.

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

[42] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[43] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, pages 1–36, 2022.

[44] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020.

[45] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021.

[46] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Advances in Neural Information Processing Systems*, 2022.

[47] Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3447–3455, 2021.

[48] Yan Zhao, Weicong Chen, Xu Tan, Kai Huang, and Jihong Zhu. Adaptive logit adjustment loss for long-tailed visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3472–3480, 2022.

[49] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021.

[50] 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.

[51] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.