applied sciences

MDPI

Article

BWLM: A Balanced Weight Learning Mechanism for Long-Tailed Image Recognition

Baoyu Fan, Han Ma, Yue Liu * and Xiaochen Yuan

Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China; baoyu.fan@mpu.edu.mo (B.F.); han.ma@mpu.edu.mo (H.M.); xcyuan@mpu.edu.mo (X.Y.)
* Correspondence: yue.liu@mpu.edu.mo

Abstract: With the growth of data in the real world, datasets often encounter the problem of long-tailed distribution of class sample sizes. In long-tailed image recognition, existing solutions usually adopt a class rebalancing strategy, such as reweighting based on the effective sample size of each class, which leans towards common classes in terms of higher accuracy. However, increasing the accuracy of rare classes while maintaining the accuracy of common classes is the key to solving the problem of long-tailed image recognition. This research explores a direction that balances the accuracy of both common and rare classes simultaneously. Firstly, a two-stage training is adopted, motivated by the use of transfer learning to balance features of common and rare classes. Secondly, a balanced weight function called Balanced Focal Softmax (BFS) loss is proposed, which combines balanced softmax loss focusing on common classes with balanced focal loss focusing on rare classes to achieve dual balance in long-tailed image recognition. Subsequently, a Balanced Weight Learning Mechanism (BWLM) to further utilize the feature of weight decay is proposed, where the weight decay as the weight balancing technique for the BFS loss tends to make the model learn smaller balanced weights by punishing the larger weights. Through extensive experiments on five long-tailed image datasets, it proves that transferring the weights from the first stage to the second stage can alleviate the bias of the naive models toward common classes. The proposed BWLM not only balances the weights of common and rare classes, but also greatly improves the accuracy of long-tailed image recognition and outperforms many state-of-the-art algorithms.

Keywords: long-tailedimage recognition; balanced weight; weight decay; Balanced Weight Learning Mechanism; reweighting loss

Citation: Fan, B.; Ma, H.; Liu, Y.; Yuan, X. BWLM: A Balanced Weight Learning Mechanism for Long-Tailed Image Recognition. *Appl. Sci.* **2024**, *14*, 454. https://doi.org/10.3390/app14010454

Academic Editors: Hyeonjoon Moon and Lien Minh Dang

Received: 23 November 2023
Revised: 21 December 2023
Accepted: 24 December 2023
Published: 4 January 2024

1. Introduction

At present, the application of deep neural networks [1–4] has been popularized in various fields [5–10], but the achievements of deep neural networks greatly rely on the availability of real-world annotated datasets [11–14]. Such data often follow a long-tailed distribution [15–18], where the quantity or cardinality of each type of data is severely imbalanced [14,19,20]. In fact, rare classes are equally important in recognition tasks. For example, in autonomous driving, vehicles and pedestrians are the common classes with more available samples, but animals crossing the road also need to be detected for safety reasons. [14,21]. Another sample is that the common and rare species should have the same importance in animal recognition. [22]. This has sparked in-depth research in the field of deep neural networks on long-tailed image recognition, with the aim of achieving high accuracy averaging across all classes [17].

Due to having much more training data for common classes than rare classes, the training loss is dominated by common classes [17]. Therefore, in the naive models without any strategy, the accuracy of common classes is always higher than rare classes. To alleviate the challenges posed by the training data in the long-tailed problem and balance various classes [23,24], existing research has mainly employed three types of strategies: resampling,

loss function engineering, and stage-wise training. In resampling, many methods have been proposed to balance the data distribution of each class by upsampling rare classes, downsampling common classes, or both [25–27]. In loss function engineering, the loss function is influenced by assigning low weights to common classes or high weights to rare classes, which has been widely adopted before [28,29]. However, multiple research studies have shown that artificially dividing weights can lead to poor training performance [30]. Therefore, there have been methods for implicitly defining common and rare classes, focusing on effective samples, sample difficulty, etc. [31,32]. In stage-wise training, some methods use transfer learning to learn rare classifiers based on the features of common classes [33–36]. Research has shown that the decoupling of feature learning and classifier learning is superior to traditional integrated training models [2].

The estimation of the long-tailed data using ordinary loss functions may lead to bias toward common classes. The current balanced weighted loss function either focuses on common classes or rare classes, which prompts us to explore balancing loss functions that simultaneously reduce the bias of common classes and increase the bias of rare classes. In addition, the naive training model has a high weight bias towards common classes, prompting us to explore further balancing network weights between classes.

To better balance common classes and rare classes, we train the model in two stages, as shown in Figure 1. Considering the effect of feature learning and decoupling training of classifier learning on long-tailed image recognition, we decide to transfer the feature learning from the first stage to the second stage, which enables the second stage to inherit the weights from the first stage. In the second stage, we propose the Balanced Focal Softmax (BFS) loss function that considers both common and rare class weights by combining a loss function that reduces redundancy of the common classes with a loss function that increases attention to rare classes. In order to better balance network weights between classes, we also propose a Balanced Weight Learning Mechanism (BWLM) that balances network weights in the norm. We adopt the weight decay for "punishing large weights and learning small weights" to further balance the weights of common and rare classes in the BFS loss function. We summarize the main contributions as follows:

- We propose the BFS loss function that combines the weight balance characteristics of common and rare classes. By adjusting the weights between the loss functions, the balance proportion of common and rare classes can be adjusted, which correspondingly affects the accuracy of common and rare classes.
- We propose BWLM to balance network weights, and investigate the weight balancing effect of weight decay on multiple balanced loss functions. We prove that weight decay also learns smaller balanced weights to make the model more balanced and improve the accuracy of multiple balanced loss functions, especially for BFS; BWLM performs better when the Imbalance Factor (i.e., the ratio of common classes with the most instances to rare classes with the least instances) is large. We use five long-tailed distributions including the large-scale dataset ImageNet-LT with Imbalance Factor (IF) = 256 and the dataset CIFAR100-LT with IF = [10, 50, 100, 200] to prove that our method achieves optimal results with increasing IF. Therefore, it proves that our method BWLM can handle more complex long-tailed image recognition problems.
- We adopt two-stage learning, which can narrow the weight norms gap between common and rare classes in the native model.

The remainder of this paper is structured as follows: Section 2 introduces and analyzes existing methods. Section 3 defines the problem and introduces our proposed method. Section 4 introduces the experimental setup, dataset, and analysis of experimental results. Section 5 concludes the study and introduces future research.

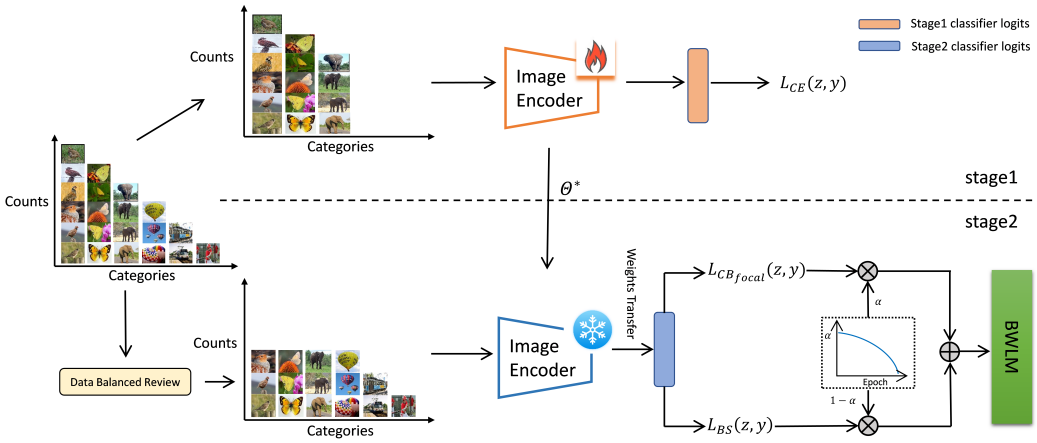


Figure 1. Overview of our two-stage training framework. In the first stage, we use common class samples to learn the stage 1 model, which uses a common cost function of cross-entropy loss L_{CE}. denotes training is in progress, while denotes that the transfer weights are frozen and cannot be trained. In the second stage, after obtaining the weights Θ^* of stage 1, we combine class balanced loss L_{CB} with a balanced softmax loss L_{BS} to obtain BWLM, which is embedded in stage 2.

2. Related Work

Data Re-Balancing. Most previous research on long-tailed imbalanced data have focused on re-balancing during the training. The re-sampling strategy includes over-sampling the rare image classes [15,37,38] and under-sampling the common image classes [15,39,40]. The purpose is to restore imbalanced training data to a regular and balanced distribution state. Usually, over-sampling has a significant effect on large datasets, but for small datasets, rare classes often exhibit over-fitting. Because the majority of the data in the common classes are discarded in cases of under-sampling, the deep model's capacity for generalization declines. When the IF is too large, the under-sampling strategy is unreasonable. The core idea of reweighting [29,35,41,42] is to assign different weights to the losses of the classes. However, such a strategy often leads to unstable training when the IF is too large. When the assigned weight is too large, the gradient will also be abnormally large, leading to a deviation in the accuracy of the class.

Loss Function Engineering. For class-level reweighting methods, it is difficult to find suitable and effective weights by directly setting weights based on the number of classes. Thus, some methods do not clearly define the common and rare classes, shifting attention to the effective sample size, the sample difficulty, etc. By assessing each class's degree of learning, Sinha et al. [43] determined loss weights and suggested a method called Class Wise Differentially Balanced Loss (CDB loss). Wang et al. [44] considered all classes as rare classes and proposed the loss function, which adaptively balances the negative gradient between different classes to improve the classifier's ability to distinguish rare classes. Seesaw loss [45] dynamically adjusts the penalty for classes with fewer samples based on the ratio of the number of samples in rare classes to that in common classes as a mitigation factor. Li et al. [46] utilized the degree of imbalance between positive and negative samples of different classes to rebalance the contribution of their losses, and extended the idea of equalization loss to the object detector [47]. Cui et al. [31] proposed the concept of effective sample size, believing that there is information overlap between data, especially in common classes with large amounts of data. This reduces the marginal benefits extracted by the model from the data. Hence, they proposed class-balanced loss (CB), which adds the effective sample size as a weight to the loss function.Tan et al. [48] found that, because the positive samples of each class are negative samples for other classes, the gradient from the negative samples for rare classes is greater than that of positive samples. Therefore, the negative gradient of common classes suppresses the learning of rare classes. Therefore, they proposed equalization loss to ignore the gradient of rare classes

in the softmax function to solve the problem of long-tailed rare classes. In Ren et al. [32], based on the Bayesian theorem, the label distribution bias leads to Softmax regression, which results in a classifier that leans more towards common classes. Therefore, they considered the changes in label distribution and proposed a balanced softmax loss.

Stage-wise Training. Dividing the training task into multiple stages to train the deep networks [8,49,50] is also a common strategy. The LFME [51] framework is a multi-stage knowledge distillation method that divides a long-tailed dataset into multiple balanced subsets and distills the knowledge in the trained expert models into the student models to guide the learning of the student model. Decoupling feature learning with the usual cross-entropy loss and classifier learning with the class balance loss, and treating them as two separate stages, was the proposal put forth by Kang et al. [2]. This study proves the important role of multi-stage training in long-tailed image recognition, as well as the importance of class balance loss in the second stage. On the basis of decoupling feature learning and classifier learning, Shaden et al. [52] used fine-tuning to balance the norm through parameter regularization in the second stage to achieve good results in long-tailed image recognition. In our paper, we drew inspiration from this stage-wise training concept, and the focus is to balance the network weights of long-tailed image recognition for different class-balanced loss functions.

3. Method

We provide a thorough explanation of our framework in this section. In Section 3.1, we first give a summary of the particular problems with long-tailed picture identification. Then, we introduce how to obtain a relatively balanced dataset and transfer the weights of stage 1 to stage 2 in Section 3.2. Finally, we propose the BFS loss function and BWLM in Section 3.3.

3.1. Preliminaries

For training datasets with long-tailed data distribution, long-tailed image recognition attempts to learn high-performance classification models. Formally, we define $D = \{x_i, y_i\}, i \in \{1, 2, \ldots, n\}$ as the long-tailed training set, where data sample x_i is labeled as $y_i \in [1, 2, \ldots, C]$, C is the total number of classes, and n_j is the number of samples in class j , where $\sum_{j=1}^{C} n_j = n$. IF $= \frac{max(n_j)}{min(n_j)}$ measures the degree of imbalance in the long-tailed training set. For long-tailed image recognition, IF $\gg 1$. The estimated class probabilities of the model are $P = [p_1, p_2, \ldots, p_C]^\top$, where $p_i \in [0, 1] \forall i$. We assume learning a classification network $f(\cdot; \Theta)$, where $\Theta = \{\theta_{l,k}\}$ is the k^{th} filter weights at layer-l. We denote θ_j as the classifier weights of class-j and optimize Θ by minimizing the loss L throughout the entire training set D to train the network:

$$\Theta^* = \underset{\Theta}{argmin} F(\Theta; D) \equiv \sum_{i=1}^{n} L(f(x_i; \Theta), y_i) \tag{1}$$

The naive classifier is influenced by the long-tailed distribution of D, and the classifier weights θ_j leans more towards common classes. Therefore, our motivation is to learn a balanced classifier by transferring the weights of common classes to rare classes, balancing the performance of the classifier by combining the advantages of two losses and further balancing classifier weights of common and rare classes by regularizing classifier weights. The specific implementation process can be seen in Figure 2.

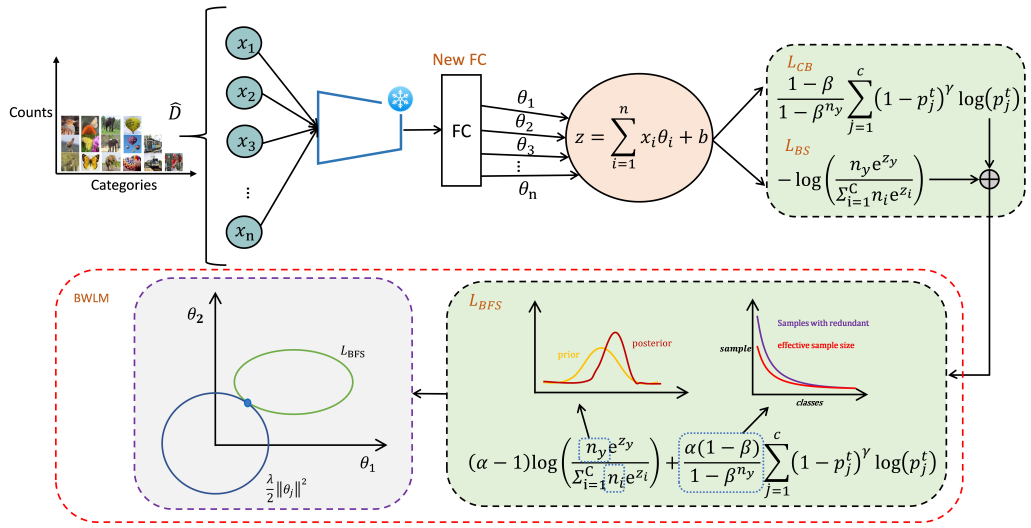


Figure 2. Detailed introduction of the second stage. We freeze the model parameters of stage 1 and add a new Fully Connected (FC) layer to achieve the weight transfer of the model. In BWLM, L_{BFS} balances the focus on common and rare classes by combining L_{CB} and L_{BS} as the loss function. L_{BS} correctly directs attention of loss to the rare class through a posterior distribution, while L_{CB} reduces the redundancy of the common sample through effective samples. Afterward, weight decay is adopted to balance the classifier weights further.

3.2. Weights Transfer

We divide the training into two stages and transfer the weights from the first stage to the second stage, and then use the proposed balanced loss function to balance the long-tailed problem.

Data Balanced Review. In order to obtain a relatively balanced dataset, we propose the data balanced review method in [53]. For class-j, D_j is the set of all its training samples. We denote the middle number of classes as $m = C/2$ and obtain the first-stage training data $D_{stage1} = \sum_{j=1}^{m} D_j$. In the second stage, we define the relatively balanced training set $\hat{D} = \{\hat{x}_i, \hat{y}_i\}, \hat{y}_i \in [1, 2, \ldots, C]$ as:

$$\hat{D} = \sum_{j=1}^{m} D_j[sam(n_{m+1})] + \sum_{j=m+1}^{C} D_j \tag{2}$$

where $sam(\cdot)$ is a random sampling method [54]. Figure 1 provides a more intuitive display.

Two-stage Transfer. Knowledge transfer currently has a wide range of research and applications, and its main idea is to transfer knowledge-rich upstream tasks to knowledge-scarce downstream tasks [55]. In the imbalanced long-tailed training set, which includes classes with sufficient knowledge and classes with insufficient knowledge, we use the Data Balanced Review [53] method to balance the dataset of downstream tasks based on the idea of head-to-tail transfer. Therefore, we use two stages to train our framework: According to Figure 1, we denote the training network for stage 1 as:

$$\Theta^* = \underset{\Theta}{argmin} F\left(\Theta; D_{stage1}\right) \tag{3}$$

At this stage, we use cross-entropy (CE) loss as the cost function denoted as L_{CE}. Based on the θ^* from stage 1 of training, we freeze the model parameters of stage 1 and add a new Fully Connected (FC) layer, and optimize FC layer weights Θ_{fc} by minimizing the loss L throughout the entire training set $\hat{D}$ to train the stage 2 network:

$$\hat{\Theta} = \underset{\Theta_{fc}}{argmin} F\left(\Theta_{fc}; \hat{D}\right)_{\Theta^*} \tag{4}$$

Through our exploration, the weight transfer method can improve the accuracy of knowledge-deficient classes, but the accuracy of knowledge-sufficient classes relatively decreases.

3.3. BWLM

We explain how BWLM works in our second stage starting by introducing the two important elements in BWLM: BFS loss function and weight decay.

BFS Loss aims to solve the problem of training from long-tailed data by focusing on a reweighting strategy that simultaneously focuses on common and rare classes.

We choose Class-Balanced (CB) [31] loss to focus on common classes. This method introduces a weight factor

$$(1-\beta)/(1-\beta^{n_j}) \tag{5}$$

We add it to the loss function to design a balanced loss function to solve the problem of data imbalance in the training. It is inversely proportional to the number of effective samples. As a hyperparameter $\beta \in [0,1)$, n_j represents the total sample size contained in class j and the balanced focal loss is defined as:

$$L_{CB}(z,y) = -\frac{1-\beta}{1-\beta^{n_y}} L_{FL(z,y)} \tag{6}$$

The focal loss $L_{FL(z,y)}$ [47] can reduce the weight of well-classified samples and make the model focus on difficult samples during training. $p_j^t = 1/(1+exp(-z_j^t))$ is the estimated probability of class j, where t means the target. The focal loss is

$$L_{FL}(z,y) = \sum_{j=1}^{C}(1-p_j^t)^{\gamma} log(p_j^t) \tag{7}$$

where γ is the focusing parameter that can reduce the weight of simple samples. The balanced focal loss is expanded as:

$$L_{CB}(z,y) = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{j=1}^{C}(1-p_j^t)^{\gamma} log(p_j^t) \tag{8}$$

Overall, the idea can be summarized as follows: By estimating the number of effective samples and reweighting the focal loss, it captures the marginal decreasing effect of the data and reduces redundant information in the common samples.

To remove the discrepancy between the test set's and the training set's posterior distributions [56], we use a balanced softmax [32]

$$\phi_j = \frac{n_j e^{z_j}}{\sum_{i=1}^{C} n_i e^{z_i}} \tag{9}$$

In order to avoid bias in testing and training, it considers label distribution shifts and derives softmax suitable for the long-tailed problem using the exponential family form of Multinomial. This tends to focus on rare classes. Similarly, we obtain the Balanced Softmax (BS) loss:

$$L_{BS}(z,y) = -log(\phi_y) = -log\left(\frac{n_y e^{z_y}}{\sum_{i=1}^{C} n_i e^{z_i}}\right) \tag{10}$$

Based on the characteristics of class-balanced focal loss and balanced softmax loss, it can be analyzed that class-balanced focal loss pays more attention to selecting effective samples for the common classes. Balanced softmax loss reduces the preference of the classifier towards the common classes caused by the difference in posterior distributions

between the training and testing sets. Therefore, we propose a balanced loss function BFS loss that can address both common and rare classes. BFS is denoted as:

$$L_{BFS} = \alpha \frac{1-\beta}{1-\beta^{n_y}} L_{FL(z,y)} + [-(1-\alpha)log(\phi_y)] \tag{11}$$

$$= \alpha \frac{1-\beta}{1-\beta^{n_y}} L_{FL(z,y)} + (\alpha-1)log(\phi_y)] \tag{12}$$

$$= (\alpha-1)log(\frac{n_y e^{z_y}}{\sum_{i=1}^{C} n_i e^{z_i}}) - \frac{\alpha(1-\beta)}{1-\beta^{n_y}} \sum_{j=1}^{C} (1-p_j^t)^{\gamma} log(p_j^t) \tag{13}$$

We discuss the selection of α as a weight and hyperparameter in Section 4.3.

Weight Decay is a relatively mature research [57,58] that constrains the networks by limiting the growth of the network weights [52], which can reduce network complexity, reduce over-fitting, and improve the generalization ability of the network. Weight decay [52] applies the L2-norm to the network weights:

$$\hat{\Theta} = \underset{\Theta_{fc}}{argmin} F\left(\Theta_{fc}; \hat{D}\right)_{\Theta^*} + \lambda \sum_j \|\theta_j\|_2^2 \tag{14}$$

We select the value of hyperparameter λ based on [52]. Weight decay increases the penalty for larger weights to prevent them from growing larger while emphasizing the learning of smaller weights. However, we find that using weight decay after Weights Transfer can improve the overall performance of long-tailed image recognition, which is beneficial for the accuracy of rare classes. To further balance the weight of common and rare classes, we adopt weight decay to adjust the impact of the model complexity on the balance loss function. Adopting weight decay can narrow the weight gap between common and rare classes, which can be reflected in Section 4.3 . Therefore, we attempt to further balance the weights for L_{CE}, L_{BS}, and our proposed L_{BFS}:

(1) By combining the encouragement of learning weight decay with L_{CB}, we obtain:

$$L_{CBW}(z,y) = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{j=1}^{C} (1-p_j^t)^{\gamma} log(p_j^t) + \lambda \left\|\theta_j\right\|^2 \tag{15}$$

which can prevent over-fitting while balancing the weight difference between the common and rare classes.

(2) By combining the encouragement of learning weight decay with L_{BS}, we obtain:

$$L_{BSW}(z,y) = -log(\phi_y) = -log\left(\frac{n_y e^{z_y}}{\sum_{i=1}^{C} n_i e^{z_i}}\right) + \lambda \left\|\theta_j\right\|^2 \tag{16}$$

(3) Similarly, our proposed balanced weight learning of L_{BFS} is denoted as:

$$L_{BWLM} = (\alpha-1)log(\frac{n_y e^{z_y}}{\sum_{i=1}^{C} n_i e^{z_i}}) - \frac{\alpha(1-\beta)}{1-\beta^{n_y}} \sum_{j=1}^{C} (1-p_j^t)^{\gamma} log(p_j^t) + \lambda \left\|\theta_j\right\|^2 \tag{17}$$

which balances the attention to common and rare classes appropriately, and punishes larger weights more heavily, allowing for learning smaller balance weights, thus facilitating long-tailed image recognition tasks.

4. Experiments

In this part, we test the proposed BWLM using CIFAR100-LT and ImageNet-LT, two popular long-tailed benchmarks. In addition, we carry out several ablation studies to assess the significance of every BWLM component in detail.

4.1. Datasets

We obtain five long-tailed image recognition datasets based on CIFAR100 and ImageNet for training (Table 1). The testing sets used for evaluation are all balanced.

CIFAR100-LT. Following [59], we use the CIFAR100 [60] dataset by downsampling the training samples to generate five long-tailed datasets of CIFAR100-LT with IF $\in$ [200, 100, 50, 10]. CIFAR100-LT includes 100 classes with the class frequency distributions of four IFs (Figure 3—left).

ImageNet-LT. The authors of [34] artificially truncate the balanced ImageNet [11] into a long-tailed version. Imagenet-LT has 1000 classes with a maximum of 1280 and a minimum of five samples of training data (Figure 3—right).

Figure 3 summarizes the class frequency distribution of these datasets. The detailed information of all the datasets is shown in Table 1. ImageNet-LT is the long-tailed distribution data with the largest IF and data volume, while CIFAR100-LT contains multiple long-tailed distributions with different IFs and relatively small data volumes.

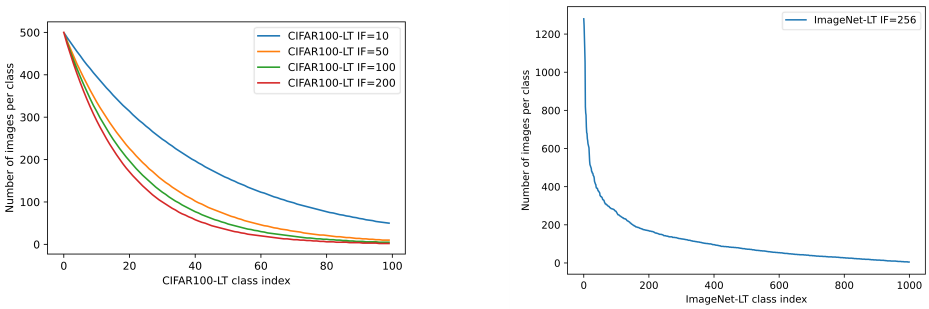


Figure 3. The class frequency distributions of five datasets. (**Left**): it indicates that as IF increases, the larger the class index, the less the class sample. (**Right**): we use the large-scale dataset ImageNet-LT, which has IF = 256.

Table 1. The specific information of five long-tailed image recognition datasets.

Dataset	IF	Training Samples	Testing Samples	Classes	Max Class Size	Min Class Size
ImageNet-LT [34]	256	115,846	50,000	1000	1280	5
CIFAR100-LT [59]	200	9502	10,000	100	500	2
	100	10,847	10,000	100	500	5
	50	12,608	10,000	100	500	10
	10	19,573	10,000	100	500	50

4.2. Implementation

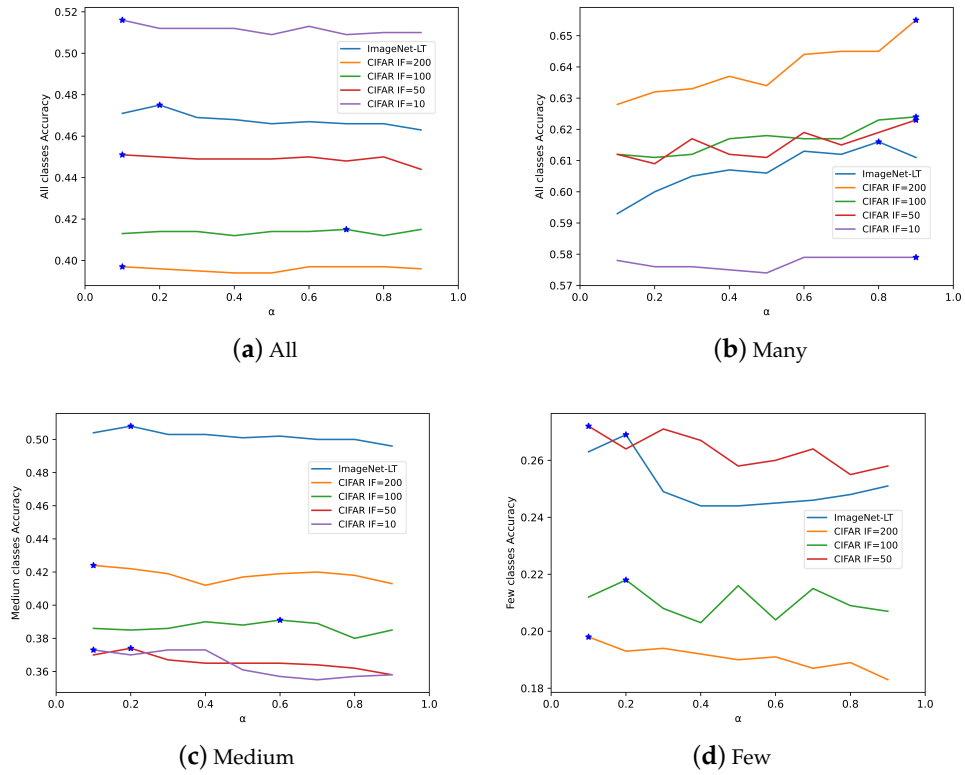
We train the models by using a long-tailed training set and evaluate them by using a balanced test set. Following [34], we divide the training set into three types based on the sample size of each category to further report accuracy: Many (greater than 100), Medium (Between 20 and 100), and Few (less than 20).

We use the PyTorch toolbox [61] to train our models. We use ResNet32 [8] on CIFAR100-LT and ResNeXt50 [62] on ImageNet-LT. The models are trained with a batch size of 64 on a single NVIDIA GeForce RTX 3090 GPU for 320 epochs on CIFAR100-LT and 500 epochs on ImageNet-LT. The momentum of the SGD optimizer is 0.9. According to the findings of [31], when the data is imbalanced, loss and validation errors gradually increase after the learning rate drops. We use a cosine learning rate regulator [63] to gradually decay the learning rate from 0.01 to 0.

4.3. Ablation Study

Choose α of BWLM. In Formula (17), we set the hyperparameter α to adjust the usage weights of CB and BS. α represents the usage weight of CB, and $1 - \alpha$ represents the

usage weight of BS. Figure 4 shows the accuracy of various classes on the five datasets, and analyzes the impact of α on All, Many, Medium, and Few classes of five datasets with different IFs. From the impact of α on the accuracy of Many classes in Figure 4b, it can be seen that almost all datasets can achieve the optimal results when α reaches its maximum value. It can be proven that CB reduces the redundancy of Many classes and focuses more on the accuracy of Many classes. In Figure 4c,d, Medium and Few classes can achieve optimal results when the value of α is small. Therefore, BS focuses more on classes with less data volume, which can also be proven. For All classes in Figure 4a, as α changes, the accuracy of all classes shows a slow downward trend in five datasets with different IFs. Since the number of Few classes is much more than that of Many classes, the overall performance of All classes is aligned with Few classes, that is, the optimal performance can be reached when α is set to be small. Although the optimal value of IF = 100 is obtained when $\alpha = 0.7$, the accuracy difference is only 0.01 when $\alpha = 0.1$ and $\alpha = 0.7$. Therefore, we suggest using BWLM by first setting $\alpha = 0.1$ by default. In the comparison experiments, we select α based on the optimal value of BWLM on the different datasets in Figure 4a.



(**a**) All (**b**) Many

(**c**) Medium (**d**) Few

Figure 4. The impact of parameter α on the accuracy of classes for datasets with different IFs. The blue star denotes the highest accuracy. (**a**) The impact of α on the accuracy of All classes. (**b**) The impact of α on the accuracy of Many classes. (**c**) The impact of α on the accuracy of Medium classes. (**d**) When IF = 10, all classes are with more than 20 samples so Few classes do not exist and the accuracy of other IFs shows an irregular downward trend.

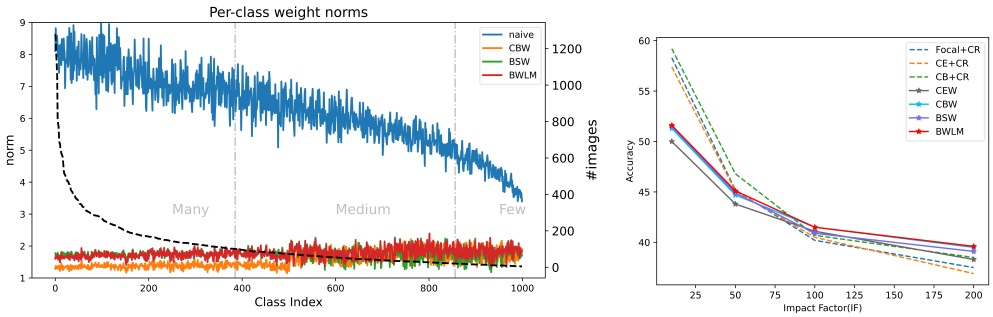
The effect of each component. To make our proposed balancing technique more suitable for application in the second stage, we also include weight decay in the first stage. This effective non-learning technique can perform post hoc scales on the classifier learned in the first stage. Our results are based on the ImageNet LT results in Table 2, where the accuracy of different classes is listed for all algorithms. We divide it into two categories for comparison: CE, CB, BS, and BFS is the reweighting loss function; CEW, CBW, BSW, and BWLM is the method after adding weight decay to the reweighting loss function. Firstly, as the weight transfer CE for the first stage increases from 39.6% of the naive model to 44.6%, it indicates the importance of weight transfer. Secondly, it is also very important

to use the balance loss function to learn classifiers based on weight transfer. Then we find that using only the equilibrium loss function without weight decay slightly improves the accuracy for only Few classes. Our method BFS balances out the accuracy of Many and Few of CB and BS, indicating that our idea is effective. This compensates for the lack of CB for Few classes and BS for Many classes. But when weight decay is used to further balance weights, the accuracy can be improved greatly. In particular, CEW reaches 46.4%. BWLM learns the weights of Few classes from BSW and Many classes from CBW, increasing the accuracy to 47.5%.

Table 2. Ablation study on ImageNet-LT with respect to accuracy (%). "naive": single model with cross-entropy loss; "+": after weight transfer in the second stage; "CE": cross-entropy; "CB": Class-Balanced focal loss; "BS": balanced softmax; "BFS": Balanced focal and softmax; "CEW": CE with weight decay; "CBW": CB with weight decay; "BSW": BS with weight decay. The best and the second-best results are shown in **underline bold** and **bold**, respectively.

	ImageNet-LT			
	Many	**Medium**	**Few**	**All**
naive	57.6	32.5	13.4	39.6
weight transfer				
+CE	55.5	48.5	19.4	44.6
+CB	51.6	45.2	18.0	41.5
+BS	52.9	47.7	25.3	44.6
+BFS	52.8	47.7	24.3	44.5
+CEW	**62.8**	**51.6**	13.1	46.4
+CBW	**61.7**	50.2	26.3	47.0
+BSW	58.2	50.1	**29.7**	**47.3**
+BWLM	60.0	**50.8**	**26.9**	**47.5**

Classifier's weight norms. Figure 5a, using the classifier weight norms of the ImageNet-LT validation set, further proves the effectiveness of weight learning. The classifier weight norms of the naive model have a relatively large numerical span from Many classes to Few classes and show a "top-heavy" downward trend. The addition of weight decay to other models has greatly shortened the weight gap between the Many and Few classes. Especially for our proposed BWLM, the classifier weight norm shows a small upward trend, which gives the Medium and Few classes a "fighting power" in the image recognition.



(**a**) Weight Norms (**b**) Descending Speed of Accuracy

Figure 5. (**a**) The classifier weight norms of the ImageNet-LT validation set, when the classes are sorted in descending n_j order (black dashed line). Blue line: a single model to learn the classifier weights. There is a significant difference between Many and Few classes. Orange line: weights after using weight decay for L_{CB}. Green line: weights after using weight decay for L_{BS}. Red line: weights after using weight decay for L_{BFS}. (**b**) Descending speed of All classes accuracy between compared methods in CIFAR100-LT different IFs. The dashed line represents the comparison method in [64]. "Focal+CR": focal loss with curvature regularization; "CE+CR": cross entropy with curvature regularization; "CB+CR": CB with curvature regularization.

4.4. Long-Tailed Image Recognition

Comparison Methods. Considering the rapid development of long-tailed image recognition, we have compared it with multiple related methods, including the method of modifying the loss function in one stage and the transfer strategy method in two stages. For better comparison, the method we chose and our commonality is that we did not add feature enhancement methods such as Mixup. For example, BS [32] for loss reweighting, NCM [2], τ-Norm [2], and cRT [2] split the training process into two stages: classifier and representation learning, while LFME [51] used transfer learning, etc. Tables 3 and 4, respectively, list the test results for the large-scale ImageNet-LT dataset and the different IFs distributions CIFAR100-LT datasets.

Table 3. The accuracy (%) on ImageNet-LT. In terms of results, our method has achieved competitive results compared to existing loss methods and two-stage transfer methods. Although not as advanced as some recent methods, these training methods added additional techniques such as Mixup and data augmentation. The best and the second-best results are shown in **underline bold** and **bold**, respectively.

Method	One-stage	Many	Medium	Few	All
CE [2]	✓	**65.9**	37.5	7.7	44.4
CE+CB [31]	✓	39.6	32.7	16.8	33.2
KD [65]	✓	58.8	26.6	3.4	35.8
Focal [31]	✓	36.4	29.9	16.0	30.5
OLTR [34]	✗	43.2	35.1	18.5	35.6
LFME [51]	✗	47.1	35.0	17.5	37.2
Range Loss [30]	✓	35.8	30.3	17.6	30.7
Equalization Loss [48]	✓	-	-	-	36.4
BS [32]	✓	50.3	39.5	25.3	41.8
NCM [2]	✗	53.1	42.3	26.5	44.3
τ-Norm [2]	✗	56.6	44.2	26.4	46.7
cRT [2]	✗	58.8	44.0	26.1	47.3
CE+CR [64]	✓	**65.1**	40.7	19.5	47.3
Our methods					
naive	✓	57.6	32.5	13.4	39.6
Weights Transfer					
+CE	✗	55.5	48.5	19.4	44.6
+CEW	✗	62.8	**51.6**	13.1	46.4
+CBW	✗	61.7	50.2	26.3	47
+BSW	✗	58.2	50.1	**29.7**	**47.3**
+BWLM	✗	60.0	**50.8**	**26.9**	**47.5**

Results. Dividing the long-tailed image recognition task into two stages and transferring the model weights from the first stage to the second stage can effectively improve the performance of long-tailed image recognition and outperform many existing methods. This conclusion can be drawn from ImageNet-LT in Table 3. Our weights transfer achieved 44.6% using the naive model with normal CE loss, which is superior to the naive model with an accuracy of 39.6%. It even outperforms other multi-stage learning methods including OLTR (35.6%) [34] and NCM (44.3%) [2]. By learning the CE classifier to regularize with weight decay, we improve accuracy from 44.6% to 46.4%. Using the weight decay to regularize the balanced loss classifier, the accuracy of the Few classes is improved by 13–16%. The accuracy of the Many classes in CBW is better than that in BSW, but the accuracy of the Few classes is lower than that in BSW. Therefore, the BWLM proposes to combine the two balanced loss functions to better balance the accuracy of Many classes and Few classes, and then reach the highest overall accuracy. For some balanced losses, such as Focal [31], Range Loss [30], Equalization Loss [48], BS [32], etc., they often choose to focus on common or rare classes, resulting in addressing one side and ignoring the other. BWLM addresses both common and rare classes, providing a more comprehensive retention of

its classification ability for Many classes, in the meantime improving its classification performance for Medium and Few classes.

Table 4. The accuracy (%) on CIFAR100-LT with different IFs. We can combine Figure 5b to analyze the rate of accuracy decline.The best results are shown in **underline bold**.

Imbalance Factor (IF)	10	50	100	200
CE [31]	55.71	43.85	38.32	34.84
CE+CB [31]	57.99	45.32	39.60	36.23
KD [65]	**59.22**	45.49	40.32	-
Focal [47]	55.78	44.32	38.41	-
LDAM [1]	56.91	-	39.60	-
Mixup [66]	58.02	44.99	39.54	-
Focal+CB [31]	57.99	45.17	39.60	35.62
CE+DRW [1]	58.10	46.50	41.00	36.90
CE+CR [64]	57.40	45.10	40.50	36.90
Focal+CR [64]	58.30	45.20	40.20	37.50
CB+CR [64]	59.20	**46.80**	40.70	38.50
Our methods				
Weights Transfer				
+CE	38.8	40.5	37.8	34.8
+CEW	50.0	43.8	41.1	38.3
+CBW	51.3	44.7	41.5	39.5
+BSW	51.5	44.9	40.9	39.1
+BWLM	51.6	45.1	**41.5**	**39.6**

In Table 4, this conclusion is also valid on other long-tailed distribution datasets with different IFs. On the four small CIFAR100-LT with different IFs, our model performs best at IF = 100 and IF = 200. We find that as IF increases, the accuracy of existing methods decreases faster than our BWLM. When the IF = [10, 100], the accuracy decreases by 17–25%, while our method controls the accuracy decrease to around 10%. In order to visually observe the descent speed, we have selected several comparison methods in Figure 5b to compare with our method. We can intuitively see that the accuracy of CE+CR [64], Focal+CR [64], and CB+CR [64] decreases significantly with the increase of IF. When IF < 100, they still have an advantage. When IF $\geq$ 100, our model surpasses other comparison methods, and our BWLM performs best among all IFs compared to other methods that add weight decay. To sum up, our method performs better on long-tailed datasets with large IF, which also proves that our model can solve more extreme long-tailed image recognition problems. Training curves on ImageNet-LT and CIFAR100-LT are shown in Figure 6. It can be seen that BWLM has a significant advantage in big data and has been proven through 150 generations of training.

Figure 7 shows the per-class accuracy for CIFAR100-LT (IF = 200). The impact of the long-tailed problem on the naive model is evident, as there is a significant decline in performance as the class IDs go from small to large. In contrast, the BWLM mitigates the steepness of this decline. It is visually apparent that BWLM substantially improves the classification results for Medium and Few classes without compromising the classification performance of Many classes. This improvement is particularly pronounced for Few classes, providing compelling evidence of the significant balancing capability of BWLM.

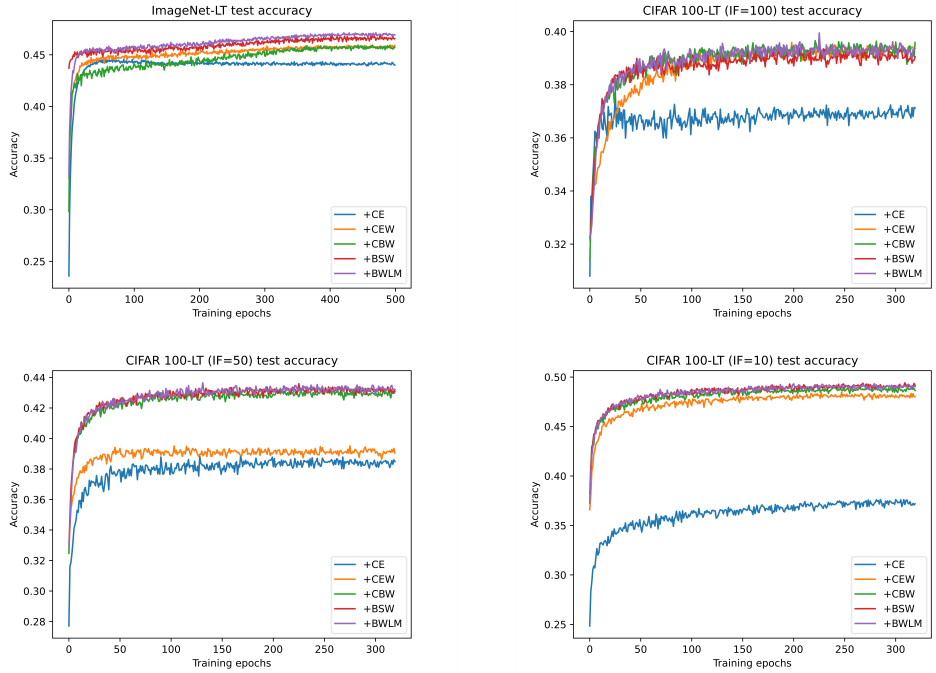


Figure 6. Training curves of ResNext-50 on ImageNet-LT and ResNet-32 on CIFAR100-LT, where IF $\in [256, 100, 50, 10]$. α of BWLM is set to be the one achieving the highest value in the corresponding dataset in Figure 4.

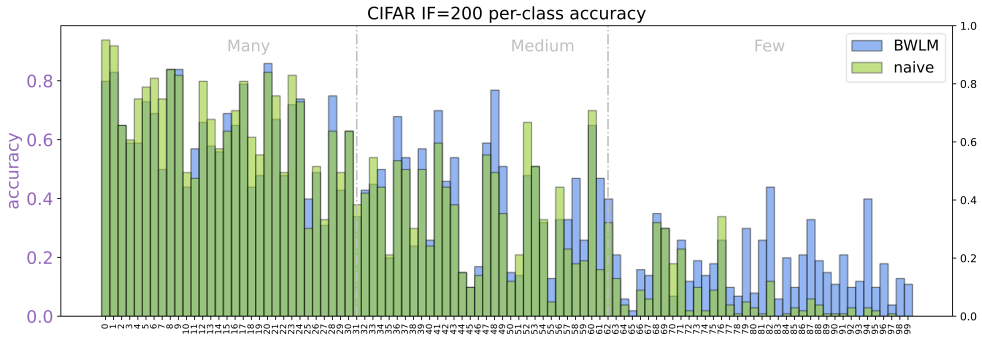


Figure 7. The per-class accuracy of CIFAR100-LT (IF = 200). The X-axis uses the numbers 0–99 to represent the classes sorted by the number of classes.

5. Conclusions

Long-tailed image recognition is a necessary challenge for the imbalanced data that often occurs in the real world. Based on our observations, the weight allocation strategy for common and rare classes on long-tailed data is divided into paying attention to common class weights or paying attention to rare class weights. We suggest focusing on both common and rare classes in the same balanced loss function by combining the balanced loss functions that focus on common class weights with the balanced loss functions that focus on rare class weights and learning the balanced weights through weight decay. Extensive experiments have shown that the correct use of regularization and balanced loss functions can greatly improve the performance of long-tailed image recognition. We introduce a relatively simple and easy-to-implement method that outperforms various existing balanced loss functions and is validated on multiple long-tailed datasets. Since our BWLM performs the best in large-scale long-tailed datasets, we believe that it can contribute to solving real-world problems related to large-scale long-tailed distributions.

Author Contributions: Conceptualization, B.F.; Methodology, B.F.; Validation, B.F.; Formal analysis, B.F., Y.L. and X.Y.; Investigation, B.F. and H.M.; Writing—original draft, B.F.; Writing—review & editing, B.F., Y.L. and X.Y.; Supervision, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Macao Polytechnic University grant number RP/ESCA-06/2021.

Data Availability Statement: The data presented in this study are openly available and can be downloaded from [34] https://doi.org/10.48550/arXiv.1904.05160.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019.
2. Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. *arXiv* **2019**, arXiv:1910.09217.
3. Yang, Y.; Xu, Z. Rethinking the value of labels for improving class-imbalanced learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19290–19301.
4. Shafiq, M.; Gu, Z. Deep residual learning for image recognition: A survey. *Appl. Sci.* **2022**, *12*, 8972. [CrossRef]
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, NV, USA, 3–6 December 2012.
6. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
7. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,Boston, MA, USA, 7–12 June 2015; pp. 1–9.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
9. Chouhan, V.; Singh, S.K.; Khamparia, A.; Gupta, D.; Tiwari, P.; Moreira, C.; Damaševičius, R.; De Albuquerque, V.H.C. A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Appl. Sci.* **2020**, *10*, 559. [CrossRef]
10. Alzubaidi, L.; Fadhel, M.A.; Al-Shamma, O.; Zhang, J.; Santamaría, J.; Duan, Y.; Oleiwi, S.R. Towards a better understanding of transfer learning for medical imaging: A case study. *Appl. Sci.* **2020**, *10*, 4523. [CrossRef]
11. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
12. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
13. Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1452–1464. [CrossRef]
14. Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; Belongie, S. The inaturalist species classification and detection dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8769–8778.
15. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef]
16. Reed, W.J. The Pareto, Zipf and other power laws. *Econ. Lett.* **2001**, *74*, 15–19. [CrossRef]
17. Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; Feng, J. Deep long-tailed learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10795–10816. [CrossRef]
18. Zhao, Y.; Kong, S.; Fowlkes, C. Camera pose matters: Improving depth prediction by mitigating pose distribution bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15759–15768.
19. Gupta, A.; Dollar, P.; Girshick, R. Lvis: A dataset for large vocabulary instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5356–5364.
20. Bansal, M.A.; Sharma, D.R.; Kathuria, D.M. A systematic review on data scarcity problem in deep learning: Solution and applications. *ACM Comput. Surv. (CSUR)* **2022**, *54*, 208. [CrossRef]
21. Kong, S.; Ramanan, D. Opengan: Open-set recognition via open data generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 813–822.

22. Romero, I.C.; Kong, S.; Fowlkes, C.C.; Jaramillo, C.; Urban, M.A.; Oboh-Ikuenobe, F.; D'Apolito, C.; Punyasena, S.W. Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 28496–28505. [CrossRef] [PubMed]
23. Ouyang, W.; Wang, X.; Zhang, C.; Yang, X. Factors in finetuning deep model for object detection with long-tail distribution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 864–873.
24. Yin, X.; Yu, X.; Sohn, K.; Liu, X.; Chandraker, M. Feature transfer learning for deep face recognition with long-tail data. *arXiv* **2018**, arXiv:1803.09014.
25. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
26. Feng, C.; Zhong, Y.; Huang, W. Exploring classification equilibrium in long-tailed object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3417–3426.
27. Estabrooks, A.; Jo, T.; Japkowicz, N. A multiple resampling method for learning from imbalanced data sets. *Comput. Intell.* **2004**, *20*, 18–36. [CrossRef]
28. Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; Van Der Maaten, L. Exploring the limits of weakly supervised pretraining. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 181–196.
29. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems 26, Lake Tahoe, NV, USA, 5–10 December 2013 .
30. Zhang, X.; Fang, Z.; Wen, Y.; Li, Z.; Qiao, Y. Range loss for deep face recognition with long-tailed training data. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5409–5418.
31. Cui, Y.; Jia, M.; Lin, T.Y.; Song, Y.; Belongie, S. Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9268–9277.
32. Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S. Balanced meta-softmax for long-tailed visual recognition. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4175–4186.
33. Jamal, M.A.; Brown, M.; Yang, M.H.; Wang, L.; Gong, B. 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, Seattle, WA, USA, 13–19 June 2020; pp. 7610–7619.
34. Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; Yu, S.X. Large-scale long-tailed recognition in an open world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2537–2546.
35. Wang, Y.X.; Ramanan, D.; Hebert, M. Learning to model the tail. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
36. Zhong, Y.; Deng, W.; Wang, M.; Hu, J.; Peng, J.; Tao, X.; Huang, Y. Unequal-training for deep face recognition with long-tailed noisy data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7812–7821.
37. Shen, L.; Lin, Z.; Huang, Q. Relay backpropagation for effective learning of deep convolutional neural networks. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VII 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 467–482.
38. Zhong, Q.; Li, C.; Zhang, Y.; Sun, H.; Yang, S.; Xie, D.; Pu, S. Towards good practices for recognition & detection. In Proceedings of the CVPR Workshops, Las Vegas, NV, USA, 27–30 June 2016; Volume 1, p. 3.
39. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
40. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
41. Huang, C.; Li, Y.; Loy, C.C.; Tang, X. Learning deep representation for imbalanced classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5375–5384.
42. Huang, C.; Li, Y.; Loy, C.C.; Tang, X. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2781–2794. [CrossRef]
43. Sinha, S.; Ohashi, H.; Nakamura, K. Class-wise difficulty-balanced loss for solving class-imbalance. In Proceedings of the Asian Conference on Computer Vision, 2020, Kyoto, Japan, 30 November–4 December 2020.
44. Wang, T.; Zhu, Y.; Zhao, C.; Zeng, W.; Wang, J.; Tang, M. Adaptive class suppression loss for long-tail object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3103–3112.
45. Wang, J.; Zhang, W.; Zang, Y.; Cao, Y.; Pang, J.; Gong, T.; Chen, K.; Liu, Z.; Loy, C.C.; Lin, D. Seesaw loss for long-tailed instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9695–9704.
46. Li, B.; Yao, Y.; Tan, J.; Zhang, G.; Yu, F.; Lu, J.; Luo, Y. Equalized focal loss for dense long-tailed object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6990–6999.
47. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.

48. Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; Yan, J. Equalization loss for long-tailed object recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11662–11671.
49. Yuan, Z.; Yan, Y.; Jin, R.; Yang, T. Stagewise training accelerates convergence of testing error over sgd. In Proceedings of the Advances in Neural Information Processing Systems 32, Vancouver, BC, Canada, 8–14 December 2019.
50. Zhao, Y.; Kong, S.; Shin, D.; Fowlkes, C. Domain decluttering: Simplifying images to mitigate synthetic-real domain shift and improve depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3330–3340.
51. Xiang, L.; Ding, G.; Han, J. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part V 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 247–263.
52. Alshammari, S.; Wang, Y.X.; Ramanan, D.; Kong, S. Long-tailed recognition via weight balancing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6897–6907.
53. Fan, B.; Liu, Y.; Cuthbert, L. Improvement of DGA Long Tail Problem Based on Transfer Learning. In Proceedings of the International Conference on Computer and Information Science, Zhuhai, China, 26–28 June 2022 ; Springer: Berlin/Heidelberg, Germany, 2022; pp. 139–152.
54. Olken, F. Random Sampling from Databases. Ph.D. Thesis, University of California, Berkeley, CA, USA, 1993.
55. Liu, B.; Li, H.; Kang, H.; Hua, G.; Vasconcelos, N. Gistnet: A geometric structure transfer network for long-tailed recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8209–8218.
56. Berger, J. The case for objective Bayesian analysis. *Bayesian Anal.* **2006**, *1*, 385–402. [CrossRef]
57. Krogh, A.; Hertz, J. A simple weight decay can improve generalization. In Proceedings of the Advances in Neural Information Processing Systems 4, Denver, CO, USA, 2–5 December 1991.
58. Moody, J.E. Note on generalization, regularization and architecture selection in nonlinear learning systems. In Proceedings of the Neural Networks for Signal Processing Proceedings of the 1991 IEEE Workshop, Princeton, NJ, USA, 30 September–2 October 1991; pp. 1–10.
59. Yue, C.; Long, M.; Wang, J.; Han, Z.; Wen, Q. Deep quantization network for efficient image retrieval. In Proceedings of the 13th Association for the Advancement of Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 3457–3463.
60. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. 2009. Available online: http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf (accessed on 8 April 2009).
61. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in Pytorch. 2017. Available online: https://note.wcoder.com/files/ml/automatic_differentiation_in_pytorch.pdf (28 October 2017).
62. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
63. Loshchilov, I.; Hutter, F.S. Stochastic Gradient Descent with Warm Restarts. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017; pp. 1–16.
64. Ma, Y.; Jiao, L.; Liu, F.; Yang, S.; Liu, X.; Li, L. Curvature-Balanced Feature Manifold Learning for Long-Tailed Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, Vancouver, BC, Canada, 17–24 June 2023; pp. 15824–15835.
65. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
66. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.