Difficulty-aware Balancing Margin Loss for Long-tailed Recognition

Minseok Son*, **Inyong Koo***, **Jinyoung Park**, **Changick Kim**

Korea Advanced Institute of Science and Technology
{ksos104, iykoo010, jinyoungpark, changick}@kaist.ac.kr

Abstract

When trained with severely imbalanced data, deep neural networks often struggle to accurately recognize classes with only a few samples. Previous studies in long-tailed recognition have attempted to rebalance biased learning using known sample distributions, primarily addressing different classification difficulties at the class level. However, these approaches often overlook the instance difficulty variation within each class. In this paper, we propose a difficulty-aware balancing margin (DBM) loss, which considers both class imbalance and instance difficulty. DBM loss comprises two components: a class-wise margin to mitigate learning bias caused by imbalanced class frequencies, and an instance-wise margin assigned to hard positive samples based on their individual difficulty. DBM loss improves class discriminativity by assigning larger margins to more difficult samples. Our method seamlessly combines with existing approaches and consistently improves performance across various long-tailed recognition benchmarks.

Code — https://github.com/quotation2520/dbm_ltr

Introduction

In recent decades, deep neural networks have demonstrated remarkable success in image recognition tasks (Simonyan and Zisserman 2014; He et al. 2016; Szegedy et al. 2015), largely due to the availability of large-scale datasets like ImageNet (Deng et al. 2009). However, real-world datasets often exhibit an imbalanced distribution, known as a long-tailed distribution, wherein a few 'head' classes contain a large number of samples, while numerous other classes, referred to as 'tail' classes, contain significantly fewer samples. This imbalance presents significant challenges: deep learning models, predominantly trained on the abundant majority classes, struggle to effectively learn features for the minority classes. As a result, models tend to underperform on these underrepresented classes, compromising their overall accuracy.

Addressing class imbalance has been a focal point in long-tailed recognition (LTR) research. Existing methods have employed various strategies to rebalance the influence of different classes. Re-sampling techniques, such as oversampling (Byrd and Lipton 2019) and undersampling (Drummond, Holte et al. 2003), adjust the occurrence of class samples to create a more balanced training set. Re-weighting approaches (Cui et al. 2019; Menon et al. 2021; Ren et al. 2020) modify class weights or logit values to emphasize learning from difficult minority classes. For instance, the label-distribution-aware margin (LDAM) loss (Cao et al. 2019) introduces larger margins for minority classes to counteract the bias towards majority classes. Despite these advances, many methods focus primarily on class-level imbalance and often overlook variations in difficulty among individual samples within each class. This oversight can lead to suboptimal performance on challenging instances, even within well-represented classes.

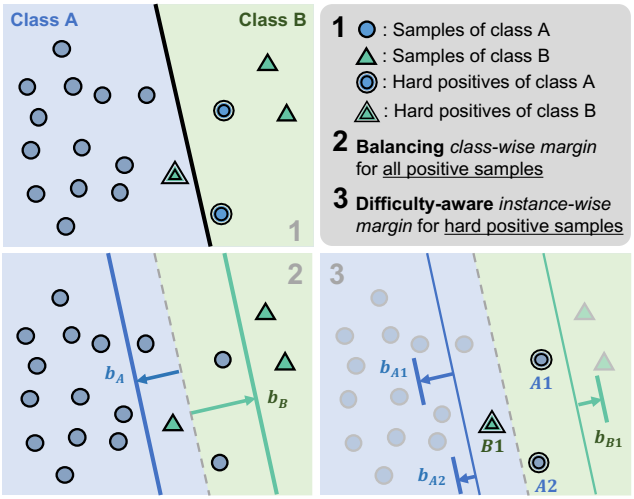


Figure 1: Overview of our method. The model is trained to align samples within decision boundaries defined by adaptive margins. (1) Hard positive samples. Misclassified samples identified during training are labeled as hard positive samples. (2) Class-wise margins. Larger margins are assigned to minority classes to ensure sufficient separation from majority classes. (3) Instance-wise Margins. We propose to apply adaptive margins to hard positive samples, considering both class frequency and sample difficulty.

*These authors contributed equally.

To address this gap, we propose a novel Difficulty-aware Balancing Margin (DBM) loss that considers both instance-level difficulty and class imbalance. Unlike previous methods that primarily address class-level bias, DBM loss incorporates two key components: a class-wise margin to mitigate imbalance in class frequencies and an instance-wise margin that adapts to the difficulty of individual samples. By assigning additional margins to hard positive samples, our approach enhances class discriminability even more.

Figure 1 illustrates the overview of our method. Here, we consider a binary classification problem where class A has more samples than class B. The decision boundary determined by the classifier is denoted by the black line in Fig. 1(1). The misclassified samples, indicated by a two-line border, are identified as hard positive samples. Our margin loss assigns a tighter decision boundary to each sample, aiming to bring sample features closer to their class centers. First, a class-wise margin of varying sizes is applied to each sample based on its class frequency. As a result, different decision boundaries, b_A and b_B, are defined as shown in Fig. 1(2), with the minority class experiencing a larger displacement in its decision boundary compared to the majority class. For hard positive samples, an additional instance-wise margin is applied. The final decision boundaries for the hard positive samples $A1$, $A2$, and $B1$ in Fig. 1(3) are denoted as b_{A1}, b_{A2} and b_{B1}, respectively. Given that $A1$ exhibits a greater angular distance from the class center compared to $A2$, $A1$ is assigned a larger instance-wise margin, leading to a more shift in b_{A1} relative to b_{A2}. This leads to a higher loss value for difficult samples, encouraging a denser feature distribution within each class.

Our method integrates seamlessly with existing LTR techniques with negligible computational impact and demonstrates consistent performance improvements across multiple benchmarks, including CIFAR-10-LT, CIFAR-100-LT (Cao et al. 2019; Kang et al. 2020b), ImageNet-LT (Liu et al. 2019c), and iNaturalist2018 (Van Horn et al. 2018). Extensive experiments validate our design choices and showcase the effectiveness and robustness of our method.

The main contributions of this paper are summarized as follows:

- We propose the difficulty-aware balancing margin (DBM) loss, which effectively balances learning bias due to class imbalance and sample-level difficulty variation within a class.
- Our DBM loss is compatible with various existing long-tailed recognition techniques, and incurs no significant additional computational overhead.
- When combined with state-of-the-art methods, our approach demonstrates competitive performance on major long-tailed recognition benchmarks.

Related Work

Long-tailed Recognition

Long-tailed recognition (LTR) has been extensively explored through multiple perspectives. Conventional approaches focus on rebalancing the bias introduced by imbalanced class influence during training, aiming to mitigate performance degradation for minority classes. Re-sampling methods (Buda, Maki, and Mazurowski 2018; He and Garcia 2009) address the class imbalance by either undersampling majority classes (Drummond, Holte et al. 2003; Tahir, Kittler, and Yan 2012) or oversampling minority classes (Byrd and Lipton 2019; Ando and Huang 2017). Re-weighting methods (Cui et al. 2019; Cao et al. 2019; Ren et al. 2020) propose class-discriminative losses to emphasize the relative contribution of minority classes. Logit compensation methods (Menon et al. 2021; Li, Cheung, and Lu 2022; Ren et al. 2020; Wang et al. 2023, 2024) adaptively adjust logit values based on prior knowledge of the sample distribution for balancing.

Another line of LTR research focuses on enhancing the robustness of representation learning to reduce model bias. Cao et al. (2019) demonstrated that applying class rebalancing methods in the later stages of training can be more effective than conventional one-stage methods. Kang et al. (2020b) proposed decoupling the training of the feature extractor from the classifier, which inspired later two-stage approaches (Zhou et al. 2020; Zhong et al. 2021). Augmentation-based methods (Li et al. 2021; Park et al. 2022; Ahn, Ko, and Yun 2023) aim to improve the sample diversity for tail classes. Inspired by the robust feature representation learned through self-supervision (He et al. 2020; Chen et al. 2020), variants of supervised contrastive learning (Khosla et al. 2020) methods have been introduced to LTR (Wang et al. 2021a; Kang et al. 2020a; Li et al. 2022b; Cui et al. 2021; Zhu et al. 2022). Suh and Seo (2023) integrated contrastive learning with logit compensation by introducing a Gaussian mixture likelihood loss, aiming to maximize mutual information between latent features and the ground truth labels. They employed a teacher-student strategy to generate contrast samples using a pre-trained teacher encoder. Ensemble-based methods (Wang et al. 2021b; Cai, Wang, and Hwang 2021; Li et al. 2022a; Tao et al. 2023) exploit the complementary knowledge from multiple experts through various incorporation methods, such as routing (Wang et al. 2021b) and distillation (Li et al. 2022a).

Most LTR studies assume that the tail classes are inherently more difficult to learn and therefore assign more weights to less frequent classes. However, some recent works (Zhao et al. 2022; Sinha and Ohashi 2023) observed that actual class-specific performance does not always correlate with class frequency. In response, they tried to consider classification difficulty in addition to sample distribution for re-weighting. We share a similar motivation and introduce an adaptive margin loss that makes instance-level adjustments based on the angular distance between the positive class center and the sample feature.

Margin Loss

Large-margin softmax loss (L-Softmax) (Liu et al. 2016) was introduced to enhance feature discrimination by encouraging intra-class compactness and inter-class separability in the embedding space. In the domain of facial recognition, margin losses have been further explored in angular space, utilizing a cosine classifier (Liu et al. 2017; Wang et al. 2018; Deng et al. 2019). These approaches aim to im-

prove discriminativity by optimizing the angular separation between class centers.

Challenges arising from class imbalance have also been addressed within margin-based frameworks. For example, face recognition methods such as fair loss (Liu et al. 2019a) and AdaptiveFace (Liu et al. 2019b), and label-distribution-aware margin (LDAM) loss (Cao et al. 2019) for LTR adaptively adjust class-wise margin values or sampling frequencies to mitigate bias. LDAM loss assigns larger margins to minority classes by explicitly incorporating class distribution priors, which helps counteract the imbalance. However, LDAM loss applies a uniform margin to all samples within a class, without accounting for variations in sample difficulty. In contrast, we propose a difficulty-aware balancing margin (DBM) loss, which introduces the consideration of instance difficulty to assign even larger margins to challenging samples. By adapting the margin based on the angular distance between the positive class center and the sample feature, DBM loss provides a more refined approach to margin adjustment, effectively addressing both class imbalance and individual sample difficulty.

Proposed Method

Preliminaries

Loss functions for Long-tailed Recognition. The cross-entropy loss with softmax function is defined as:

$$L_{\text{CE}} = -\log \frac{e^{\psi_y(x)}}{\sum_i e^{\psi_i(x)}}. \tag{1}$$

Here, $\psi_i(x)$ represents the logit function of the i-th class for sample x, which belongs to the class of index y. For models that utilize a linear classifier, the logit function is given by $\psi_i(x) = W_i^\top f(x) + b_i$, where $f(x)$ denotes the feature representation of sample x, and W_i and b_i represent the weight and bias of the linear classifier for the i-th class, respectively. Alternatively, a cosine classifier embeds features and class centers in an L2-normalized space, with logits determined by the angular distance between sample features and class centers. Specifically,

$$\psi_i(x) = s\frac{W_i^\top f(x)}{\|W_i\|\|f(x)\|} = s\cos\theta_i, \tag{2}$$

where s is the scaling factor and θ_i denotes the angular distance between W_i and $f(x)$.

In long-tailed recognition (LTR), re-weighting methods address class imbalance by incorporating class frequency n_i into the loss functions. Variants of cross-entropy loss include the class-balanced (CB) loss (Cui et al. 2019) and balanced softmax (BS) (Ren et al. 2020). The class balanced loss L_{CB} is formulated as:

$$L_{\text{CB}} = -\frac{1-\beta}{1-\beta^{n_y}} \log \frac{e^{\psi_y(x)}}{\sum_i e^{\psi_i(x)}}, \tag{3}$$

introducing a class-wise weight determined by the effective number of samples given a hyperparameter β. The balanced softmax loss L_{BS} is:

$$L_{\text{BS}} = -\log \frac{e^{\psi_y(x)+\log p_y}}{\sum_i e^{\psi_i(x)+\log p_i}}, \tag{4}$$

Methods	$\psi_y^m(\theta_y)$
SphereFace (Liu et al. 2017)	$\cos(m\theta_y)$
CosFace (Wang et al. 2018)	$\cos\theta_y - m$
ArcFace (Deng et al. 2019)	$\cos(\theta_y + m)$
LDAM (Cao et al. 2019)	$\cos\theta_y - mn_y^{-1/4}$

Table 1: $\psi_y^m(\theta_y)$ used in different margin losses.

where p_i represents the sample proportion of the i-th class over all classes, i.e., $p_i = n_i / \sum_j n_j$. The balanced softmax loss is widely adopted in later LTR studies, such as balanced contrastive learning (BCL) (Zhu et al. 2022) and nested collaborative learning (NCL) (Li et al. 2022a).

Margin-based Variants of Cross-entropy Loss. Margin losses introduce a specialized logit function associated with a margin for the positive class. A margin-based cross-entropy loss L_m can be generally formulated as:

$$L_m = -\log \frac{e^{s\psi_y^m(\theta_y)}}{e^{s\psi_y^m(\theta_y)} + \sum_{i\neq y} e^{s\cos\theta_i}}, \tag{5}$$

where $s\psi_y^m(\theta_y)$ denotes the logit function for the positive class incorporating the margin. If $\psi_y^m(\theta_y)$ adopts no margin, i.e., $\psi_y^m(\theta_y) = \cos\theta_y$, L_m is equivalent to L_{CE}.

Table 1 provides a summary of various margin-based loss functions and their respective logit formulations. Traditional margin losses (Liu et al. 2017; Wang et al. 2018; Deng et al. 2019) apply a constant margin for all classes. CosFace (Wang et al. 2018) applies a margin to the measured cosine similarity, while ArcFace (Deng et al. 2019) directly adjusts the angular distance. LDAM loss (Cao et al. 2019) follows a similar formulation to CosFace, subtracting a margin that varies with class frequency from the cosine similarity to address the class imbalance problem.

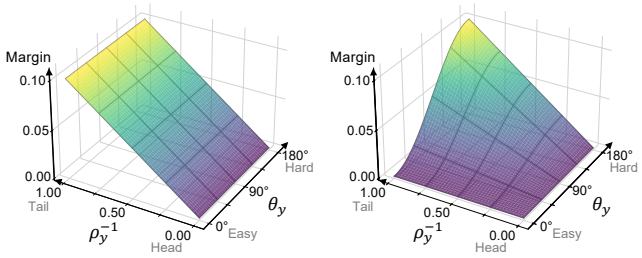
Difficulty-aware Balancing Margin Loss

Our difficulty-aware balancing margin (DBM) loss comprises two components: a class-wise margin and an instance-wise margin. By integrating these two elements, we address both the bias from class imbalance and the variation in instance difficulty within a class. Following prior works (Xiao et al. 2022; Li et al. 2024), we apply the instance-wise margin specifically to hard positive samples. Figure 2 illustrates the margins determined by class frequency and angular distance. Detailed mathematical descriptions of each component are provided below.

Class-wise Margin. The class-wise margin m_C is defined as:

$$m_C = K\rho_y^{-\tau}. \tag{6}$$

Here, $\rho_y = n_y/n_{\min}$ represents the ratio of the number of samples in class y to the number in the least frequent class. The parameter τ controls the extent of the margin difference across classes, while K scales the margin. As illustrated in Fig. 2a, m_C is solely based on the class frequency



(a) Class-wise margin. (b) Instance-wise margin.

Figure 2: Margins for $K = 0.1$ and $\tau = 1$. Less frequent classes have larger class-wise margin, and more difficult samples have larger instance-wise margin.

ratio ρ_y. By scaling inversely with ρ_y, minority classes receive a larger margin compared to majority classes, ensuring the least frequent class obtains the maximum margin of K. This helps mitigate performance degradation for minority classes. We have found that setting $\tau = 1$ is effective for our approach.

Instance-wise Margin. The instance-wise margin addresses varying sample-level difficulties. Samples with lower positive logit values are more prone to misclassification. For our cosine classifier, difficult samples are those whose feature representations are farther from the positive class center in the hypersphere. We quantify the instance difficulty d_I via following equation:

$$d_I = \frac{1 - \cos\theta_y}{2}. \tag{7}$$

Here, d_I is determined by the angular distance between the feature representation of the sample and the positive class center θ_y. A sample with its feature representation exactly at the class center has $d_I = 0$, while a sample with the feature representation at the maximum distance ($\theta_y = \pi$) has $d_I = 1$.

The instance-wise margin m_I is given by:

$$m_I = m_C \cdot d_I. \tag{8}$$

As illustrated in Fig. 2b, this margin is determined by both ρ_y and θ_y, encouraging difficult and less-frequent samples to move more aggressively towards the positive class center.

Loss formulation. Our DBM loss modifies the angular distance by incorporating both margins, similar to the ArcFace approach. Specifically, our logit function for the positive class is formulated as:

$$s\psi_y^{dbm}(\theta_y) = s\cos(\theta_y + m_C + \mathbb{1}[\operatorname*{argmin}_i(\{\theta_i\}_{i=1}^N) \neq y]m_I), \tag{9}$$

where $\mathbb{1}[\cdot]$ is an indicator function for applying the instance-wise margin only to hard positive samples. By substituting this logit function into Eq. (5), we derive the difficulty-aware balancing margin cross-entropy (DBM-CE) loss.

The DBM loss can be easily integrated with various existing LTR methods. For example, it can be combined with the class-balanced loss introduced in Eq. (3) as follows:

$$L_{\text{DBM-CB}} = -\frac{1-\beta}{1-\beta^{n_y}} \log \frac{e^{s\psi_y^{dbm}(\theta_y)}}{e^{s\psi_y^{dbm}(\theta_y)} + \sum_{i \neq y} e^{s\cos\theta_i}}. \tag{10}$$

Similarly, DBM-BS can be derived as:

$$L_{\text{DBM-BS}} = -\log \frac{e^{s\psi_y^{dbm}(\theta_y) + \log p_y}}{e^{s\psi_y^{dbm}(\theta_y) + \log p_y} + \sum_{i \neq y} e^{s\cos\theta_i + \log p_i}}, \tag{11}$$

reformulating the original balanced softmax loss described in Eq. (4). Note that our method requires adjusting the classifier from a linear to a cosine classifier.

Moreover, Our method is highly versatile and can be incorporated with a range of other LTR techniques. We demonstrate this versatility with various configurations of our method, including DBM-DRW, DBM-BCL, DBM-GML, and DBM-NCL. DRW, or deferred re-weighting (Cao et al. 2019), integrates class-balanced loss into the training process at a later stage, allowing DBM-DRW to be implemented by applying $L_{\text{DBM-CE}}$ and $L_{\text{DBM-CB}}$ sequentially according to the scheduling policy. Similarly, methods like BCL (Zhu et al. 2022), GML (Suh and Seo 2023) and NCL (Li et al. 2022a), which originally use balanced softmax loss, can incorporate our approach by substituting the classification loss with $L_{\text{DBM-BS}}$.

The integration of DBM loss into existing models does not incur significant additional computational complexity. The class-wise margin m_C is determined in advance based on the known sample distribution, ensuring that this computation does not affect the training time. The instance-wise margin m_I is computed during the logit calculation, leveraging the angular distance θ_y that is already part of the model's forward pass. This design ensures that DBM can be incorporated into existing frameworks without introducing substantial overhead.

Experiments

Datasets

To evaluate the performance of our proposed method, we conducted experiments on four benchmark long-tailed datasets. The imbalance factor ρ of each dataset is defined as the ratio of training instances between the largest and smallest classes, i.e., $\rho = n_{\max}/n_{min}$, following previous works (Cao et al. 2019; Kang et al. 2020b).

Long-tailed CIFAR-10 and CIFAR-100. We sampled long-tailed CIFAR datasets from the original CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton et al. 2009) datasets with imbalance factors of 10, 50, and 100 using an exponential down-sampling profile outlined in (Cao et al. 2019; Cui et al. 2019). Evaluations were performed on the original balanced test sets.

ImageNet-LT. ImageNet-LT (Liu et al. 2019c) is a long-tailed version of ImageNet-1K (Deng et al. 2009), sampled from a Pareto distribution with $\alpha = 6$. It comprises 1,000 categories and 115.8K training images, with an imbalanced factor of $\rho = 1280/5$.

Method	CIFAR-10-LT			CIFAR-100-LT					
	Imb. Factor			Imb. Factor			Statistics (IF 100)		
	100	50	10	100	50	10	Many	Med.	Few
CE	78.48	82.73	89.91	44.60	48.75	61.98	73.03	45.37	10.53
LDAM (Cao et al. 2019)	79.92	83.84	90.54	45.25	50.16	62.86	**75.31**	44.00	11.63
DBM-CE	**80.84**	**84.12**	**90.95**	**46.53**	**51.13**	**63.18**	73.89	**46.17**	**15.03**
CE-DRW (Cao et al. 2019)	82.24	85.05	90.94	48.28	53.89	64.25	65.89	50.74	24.87
LDAM-DRW (Cao et al. 2019)	82.60	85.36	91.22	48.99	54.27	64.58	**66.09**	50.83	26.90
DBM-DRW	**82.82**	**85.83**	**91.55**	**49.41**	**54.69**	**64.75**	63.23	**52.66**	**29.50**
BS (Ren et al. 2020)	83.57	86.45	91.26	49.35	54.79	63.93	65.77	50.14	29.27
DBM-BS	**84.60**	**87.06**	**91.42**	**51.30**	**55.84**	**65.22**	**67.29**	**50.80**	**33.23**
BCL (Zhu et al. 2022)	82.95	86.76	91.57	50.23	55.35	64.98	67.14	51.31	29.23
DBM-BCL	**84.60**	**87.16**	**91.69**	**51.66**	**55.98**	**65.25**	**67.91**	**51.91**	**32.40**
GML (Suh and Seo 2023)	85.19	88.07	92.11	53.12	58.17	66.93	71.60	54.57	28.20
DBM-GML	**85.30**	**88.35**	**92.59**	**53.70**	**58.41**	**67.15**	**72.34**	**54.89**	**30.57**
NCL (Li et al. 2022a)	87.37	89.89	93.15	56.68	61.65	69.46	**73.94**	56.97	36.20
DBM-NCL	**87.53**	**89.90**	**93.19**	**57.48**	**62.01**	**69.75**	71.49	**59.06**	**39.30**

Table 2: Top-1 accuracy (%) of ResNet-32 on CIFAR-10-LT and CIFAR-100-LT with the imbalance factor (IF) of 100, 50, and 10.

iNaturalist2018. The iNaturalist2018 dataset (Van Horn et al. 2018) is a large-scale real-world dataset that features a highly long-tailed distribution with an imbalance factor of $\rho = 1000/2$. It includes approximately 437K training images and 24.4K validation images gathered from 8,142 fine-grained species classes in the wild.

Implementation Details

For the CIFAR-10-LT and CIFAR-100-LT datasets, we integrated our method with several existing approaches including:

(1) vanilla cross-entropy (CE)

(2) CE-DRW (Cao et al. 2019), a two-stage training method applying CB loss (Cui et al. 2019).

(3) BS (Ren et al. 2020), a re-weighting method.

(4) BCL (Zhu et al. 2022), a supervised contrastive learning-based method.

(5) GML (Suh and Seo 2023), a mutual information maximization method.

(6) NCL (Li et al. 2022a), an ensemble-based method.

We ensured a fair comparison by evaluating our models under identical experimental conditions. All models utilized ResNet-32 (He et al. 2016) as the backbone network, while ResNet56 was employed as the teacher network for GML. The SGD optimizer with a momentum of 0.9 and weight decay of 2×10^{-4} was employed, along with a learning rate warm-up for the first five epochs and a cosine annealing scheduler for gradual decay. Data augmentation strategies included Cutout (DeVries and Taylor 2017) and AutoAugment (Cubuk et al. 2019). For BCL, we used an initial learning rate of 0.15 and a batch size of 256. For all other methods, we used an initial learning rate of 0.1 and a batch size of 64. Training was conducted for 200 epochs for most methods, except for NCL, which was trained for 400 epochs. In the case of DRW, class-balanced loss is introduced after 160 epochs. We used a scaling factor $s = 32$ for all our experiments, and tuned the hyperparameter for margin scaling K within the range 0.1 to 0.3, adjusting it based on datasets and baselines.

For larger datasets, our method was integrated into BCL and GML. NCL was excluded from this comparison due to its extensive training requirements of 400 epochs. For ImageNet-LT, we utilized ResNet-50 and ResNeXt-50 (Xie et al. 2017) as backbones and trained them for 90 epochs. For iNaturalist2018, we employed ResNet-50 and trained for 100 epochs. In both benchmarks, we set the scaling factor s to 30 and the margin scaling hyperparameter K to 0.1. Further details are in the supplementary materials.

Experimental Results

Long-tailed CIFAR. Table 2 presents the experimental results for CIFAR-10-LT and CIFAR-100-LT. For CIFAR-100-LT with an imbalance factor of 100, we report the accuracy across three groups of classes: 'Many ($>$ 100 shots),' 'Medium (20$\sim$100 shots),' and 'Few ($<$ 20 shots).' To ensure fairness, we have reproduced the performance of each previous method and provided these results in the corresponding cells. Methods incorporating DBM loss are highlighted in gray.

The results demonstrate that DBM consistently provides a significant performance improvement over baseline methods. When applied to CE and CE-DRW, our method achieves superior enhancement compared to LDAM and LDAM-DRW, which solely introduces a class-wise margin.

Method	R50	RX50
CE†	41.6	44.4
τ-norm (Kang et al. 2020b)	46.7	49.4
cRT (Kang et al. 2020b)	47.3	49.6
LWS (Kang et al. 2020b)	47.7	49.9
LDAM-DRW‡ (Cao et al. 2019)	49.8	–
CE-DRW‡ (Cao et al. 2019)	50.1	–
BS‡ (Ren et al. 2020)	50.9	–
ALA Loss (Zhao et al. 2022)	52.4	53.3
DisAlign (Zhang et al. 2021)	52.9	53.4
Difficulty-Net (Sinha and Ohashi 2023)	54.0	–
RIDE (3 experts) (Wang et al. 2021b)	54.9	56.4
BCL (Zhu et al. 2022)	56.0	56.7
GML (Suh and Seo 2023)	–	58.3
DBM-BCL	56.3	57.4
DBM-GML	**57.4**	**58.6**

Table 3: Top-1 accuracy (%) of ResNet-50 and ResNeXt-50 on ImageNet-LT. † and ‡ denotes results borrowed from Kang et al. (2020b) and Park et al. (2022), respectively.

Notably, DBM-BS surpasses BCL, indicating a substantial performance boost without the additional complexity introduced by BCL's contrastive learning branch. Although some algorithms show a slight decrease in accuracy for the 'Many' group compared to the baseline, our method achieves a notable increase in accuracy for the 'Medium' and 'Few' groups, demonstrating its effectiveness in mitigating performance bias.

ImageNet-LT and iNaturalist2018. Table 3 shows the performance of DBM-BCL and DBM-GML compared to the existing methods on the ImageNet-LT dataset. We report overall accuracy using ResNet-50 and ResNeXt-50 backbones. For a fair comparison, we evaluated our method against existing works that reported the performance after 90 epochs of training. DBM-BCL outperforms the baseline BCL, with an overall accuracy improvements of 0.3%p and 0.7%p for the ResNet-50 and ResNeXt-50 backbones, respectively. Although GML did not report results for the ResNet-50 model, DBM-GML demonstrates an improved performance of 0.3%p for the ResNeXt-50 backbone.

Table 4 displays the performance comparisons on the iNaturalist2018 dataset. We report overall accuracy and the accuracy of 'Many,' 'Medium,' and 'Few' groups in our experiment. To ensure a fair comparison, we excluded methods that involve extensive additional training (Cui et al. 2021; Li et al. 2022a). Since BCL and GML did not report accuracy for each group, we re-implemented their results using their official code. DBM-BCL and DBM-GML achieve improvements of 0.9%p and 0.8%p in overall accuracy, respectively, surpassing the performances of existing methods.

Analysis

In this section, we analyze the components of the DBM loss to evaluate their contributions to performance improvement. We also investigate the impact of different hyperparameters on the method's effectiveness. Additionally, we illustrate how the introduced margin enhances intra-class compactness and inter-class separability, thus improving classification performance. All experiments for analysis were conducted on CIFAR-100-LT with an imbalance factor of 100.

Methods	Many	Med.	Few	All
CE†	**73.9**	63.5	55.5	61.0
τ-norm (Kang et al. 2020b)	65.6	65.3	65.9	65.6
cRT (Kang et al. 2020b)	69.0	66.0	63.2	65.2
LWS (Kang et al. 2020b)	65.0	66.3	65.5	65.9
LDAM-DRW† (Cao et al. 2019)	–	–	–	66.1
CE-DRW‡ (Cao et al. 2019)	68.2	67.3	66.4	67.0
BS‡ (Ren et al. 2020)	65.5	67.5	67.5	67.2
DisAlign (Zhang et al. 2021)	61.6	70.8	69.9	69.5
RIDE (Wang et al. 2021b)	70.2	71.3	71.7	71.4
BCL⋆ (Zhu et al. 2022)	68.2	71.3	71.3	71.0
GML⋆ (Suh and Seo 2023)	70.7	**72.3**	71.5	71.2
DBM-BCL	65.6	71.8	73.8	71.9
DBM-GML	66.9	71.9	**73.6**	**72.0**

Table 4: Top-1 accuracy (%) of ResNet-50 on iNaturalist2018. † and ‡ denotes results borrowed from Zhou et al. (2020) and Ahn, Ko, and Yun (2023), respectively. ⋆ denotes reproduced results with the official code. RIDE (2 experts) (Wang et al. 2021b) was trained for 100 epochs.

Cosine	m_C	m_I	CE	BS
			44.60	49.35
✓			44.29	49.84
✓	✓		45.85	50.61
✓	✓	P	46.38	50.93
✓	✓	HP(ours)	**46.53**	**51.30**

Table 5: Ablation study for the components of DBM loss. 'Cosine' denotes replacing the linear classifier with cosine classifier. m_C and m_I denote class-wise and instance-wise margin, respectively. P and HP represent the cases where instance-wise margin is applied to all positive samples and hard positive samples, respectively.

Component Analysis. Table 5 presents the results of our ablation study, which examines the impact of class-wise and instance-wise margins. We integrated these components into two baseline methods: CE and BS (Ren et al. 2020). Our findings reveal that using a cosine classifier alone does not significantly improve performance. However, incorporating a class-wise margin leads to notable gains. Adding the instance-wise margin results in an additional performance increase of approximately 0.7%p for both loss functions.

We also observed differences in performance based on the way the instance-wise margin is applied. Specifically, the 'hard positive (HP)' strategy, where the margin is applied only to misclassified positive samples, yields better results compared to the 'positive (P)' strategy, which applies the margin to all positive samples. This indicates that focusing

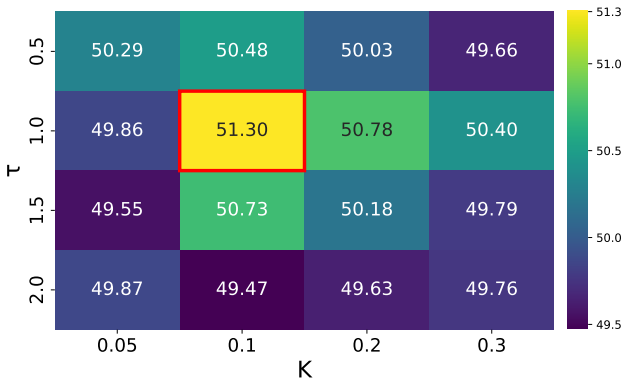


Figure 3: Analysis for effects of hyperparameters τ and K. For all cases, DBM-BS outperforms the baseline BS (Ren et al. 2020) performance of 49.35%.

on the difficulty of hard positive samples only rather than all positive samples improves performance more effectively.

Impacts of Hyperparameters. Figure 3 illustrates the effects of various hyperparameters on the performance of DBM-BS. We analyze the impact of τ and K, which are critical parameters in our method. The results show that while variations in these hyperparameters cause slight performance differences, DBM consistently outperforms the baseline across all settings.

Based on our observation, we fixed τ at 1.0 throughout all experiments on the long-tailed benchmarks. The optimal value for K varies depending on the method, but setting $K = 0.1$ generally yields satisfactory results.

Intra-class Compactness and Inter-class Separability. We apply instance-wise margins to bring hard positive samples closer to their respective class centers, aiming to enhance intra-class compactness. Figure 4 compares the distribution of angular distances between sample features and their positive class centers for the 'Many', 'Medium', and 'Few' groups in BS and DBM-BS. DBM-BS shows a reduction in the mean angular distance of approximately 10° across all groups, indicating enhanced intra-class compactness. This suggests that DBM improves the alignment of sample features with their respective class centers, which may contribute to better performance in classification tasks.

To evaluate inter-class separability, we use the Fisher criterion from Fisher's linear discriminant analysis (LDA) (Fisher 1936) as a metric to measure the distance between feature distributions of different classes. LDA aims to find a projection vector W that maximizes the separation between classes by projecting the data onto a new axis where the classes are most distinguishable. The Fisher criterion is used to determine the optimal W that maximizes the ratio of the between-class variance to the within-class variance.

The Fisher criterion is defined as:

$$J(W_{ij}) = \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2}, \tag{12}$$

where W is the projection vector, and μ_k and σ_k^2 denote the

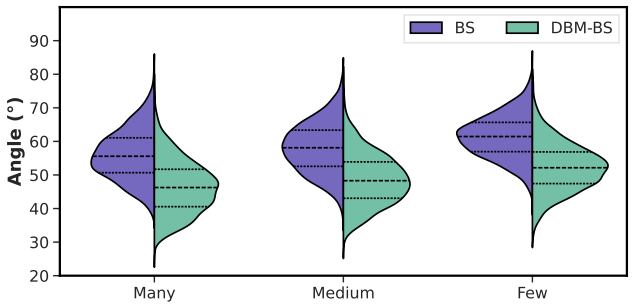


Figure 4: Comparison of BS (Ren et al. 2020) and DBM-BS of the distribution of angular distance between sample features and their positive class centers for 'Many', 'Medium', and 'Few' groups. Dashed horizontal lines denote the quartiles.

Method	Many	Med.	Few	All
BS (Ren et al. 2020)	6.02	5.98	5.65	5.89
DBM-BS	**6.21**	**6.26**	**5.95**	**6.15**

Table 6: Analysis for inter-class separability. A larger value indicates better separability.

mean and variance of the projected feature distribution for the k-th class, respectively. The objective is to find W_{ij} such that the means of the projected classes μ_i and μ_j are as far apart as possible while the variances σ_i^2 and σ_j^2 are minimized. A higher value of the Fisher criterion $J(W_{ij})$ indicates greater separability between the two classes.

After calculating the optimal projection vectors for all class pairs, we define the separability of a class S_i as:

$$S_i = \frac{1}{C-1} \sum_{j=1, j \neq i}^{C} J(W_{ij}) \tag{13}$$

where C is the number of classes. Table 6 presents the separability for 'Many,' 'Medium,' 'Few,' and 'All' groups. Our observations confirm that DBM enhances inter-class separability across all groups, leading to improved overall classification performance.

Conclusion

In this work, we propose a difficulty-aware balancing margin (DBM) loss, a novel approach designed to address class-level imbalance and instance-level difficulty variations in long-tailed datasets. The DBM loss incorporates a class-wise margin to mitigate the performance degradation caused by class imbalance and a instance-wise margin to enhance class discriminability by more effectively aligning misclassified samples with their corresponding class centers. Our method integrates effortlessly with existing long-tailed recognition techniques and consistently improves performance across benchmarks. We comprehensively evaluated our method on the long-tailed CIFAR, ImageNet-LT, and iNaturalist2018 datasets, and demonstrated its effectiveness through extensive experiments.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2018R1A5A7025409).

References

Ahn, S.; Ko, J.; and Yun, S.-Y. 2023. CUDA: Curriculum of Data Augmentation for Long-tailed Recognition. In *The Eleventh International Conference on Learning Representations*.

Ando, S.; and Huang, C. Y. 2017. Deep over-sampling framework for classifying imbalanced data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, 770–785. Springer.

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106: 249–259.

Byrd, J.; and Lipton, Z. 2019. What is the effect of importance weighting in deep learning? In *International conference on machine learning*, 872–881. PMLR.

Cai, J.; Wang, Y.; and Hwang, J.-N. 2021. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 112–121.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 113–123.

Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; and Jia, J. 2021. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 715–724.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

DeVries, T.; and Taylor, G. W. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*.

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

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188.

He, H.; and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9): 1263–1284.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kang, B.; Li, Y.; Xie, S.; Yuan, Z.; and Feng, J. 2020a. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020b. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, J.; Tan, Z.; Wan, J.; Lei, Z.; and Guo, G. 2022a. Nested Collaborative Learning for Long-Tailed Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6949–6958.

Li, J.; Xiao, D.; Lu, T.; Wei, Y.; Li, J.; and Yang, L. 2024. HAMFace: Hardness adaptive margin loss for face recognition with various intra-class variations. *Expert Systems with Applications*, 240: 122384.

Li, M.; Cheung, Y.-m.; and Lu, Y. 2022. Long-tailed visual recognition via gaussian clouded logit adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6929–6938.

Li, S.; Gong, K.; Liu, C. H.; Wang, Y.; Qiao, F.; and Cheng, X. 2021. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5212–5221.

Li, T.; Cao, P.; Yuan, Y.; Fan, L.; Yang, Y.; Feris, R. S.; Indyk, P.; and Katabi, D. 2022b. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6918–6928.

Liu, B.; Deng, W.; Zhong, Y.; Wang, M.; Hu, J.; Tao, X.; and Huang, Y. 2019a. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10052–10061.

Liu, H.; Zhu, X.; Lei, Z.; and Li, S. Z. 2019b. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11947–11956.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, W.; Wen, Y.; Yu, Z.; and Yang, M. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks. In *Proceedings of The 33rd International Conference on Machine Learning*, 507–516.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019c. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2537–2546.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*.

Park, S.; Hong, Y.; Heo, B.; Yun, S.; and Choi, J. Y. 2022. The Majority Can Help The Minority: Context-rich Minority Oversampling for Long-tailed Classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Ren, J.; Yu, C.; Sheng, S.; Ma, X.; Zhao, H.; Yi, S.; and Li, H. 2020. Balanced Meta-Softmax for Long-Tailed Visual Recognition. In *Proceedings of Neural Information Processing Systems(NeurIPS)*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sinha, S.; and Ohashi, H. 2023. Difficulty-Net: Learning to Predict Difficulty for Long-Tailed Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6444–6453.

Suh, M.-K.; and Seo, S.-W. 2023. Long-tailed recognition by mutual information maximization between latent features and ground-truth labels. In *International Conference on Machine Learning*, 32770–32782. PMLR.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Tahir, M. A.; Kittler, J.; and Yan, F. 2012. Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Pattern Recognition*, 45(10): 3738–3750.

Tao, Y.; Sun, J.; Yang, H.; Chen, L.; Wang, X.; Yang, W.; Du, D.; and Zheng, M. 2023. Local and global logit adjustments for long-tailed learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11783–11792.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wang, P.; Han, K.; Wei, X.-S.; Zhang, L.; and Wang, L. 2021a. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 943–952.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. 2021b. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. In *International Conference on Learning Representations*.

Wang, Y.; Zhang, B.; Hou, W.; Wu, Z.; Wang, J.; and Shinozaki, T. 2023. Margin calibration for long-tailed visual recognition. In *Asian Conference on Machine Learning*, 1101–1116. PMLR.

Wang, Z.; Xu, Q.; Yang, Z.; He, Y.; Cao, X.; and Huang, Q. 2024. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. *Advances in Neural Information Processing Systems*, 36.

Xiao, D.; Li, J.; Li, J.; Dong, S.; and Lu, T. 2022. IHEM loss: Intra-class hard example mining loss for robust face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7821–7831.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Zhang, S.; Li, Z.; Yan, S.; He, X.; and Sun, J. 2021. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2361–2370.

Zhao, Y.; Chen, W.; Tan, X.; Huang, K.; and Zhu, J. 2022. Adaptive logit adjustment loss for long-tailed visual recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 3472–3480.

Zhong, Z.; Cui, J.; Liu, S.; and Jia, J. 2021. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16489–16498.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. 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*, 9719–9728.

Zhu, J.; Wang, Z.; Chen, J.; Chen, Y.-P. P.; and Jiang, Y.-G. 2022. Balanced Contrastive Learning for Long-Tailed Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6908–6917.