PeerJ Computer Science

A progressive decay Mixup data augmentation for long-tailed image classification

Hongyin Li[1], Jiahao Li[2], Cheng Qian[1], Zilin Xiang[1] and Rongdong Chen[2]

[1] School of Engineering, Guangzhou College of Technology and Business, Guangzhou, China
[2] School of Electronics and Communication Engineering, Guangzhou University, Guangzhou, China

ABSTRACT

Image classification with long-tailed distribution (*i.e.*, a few classes occupy most of the training samples, while most classes have only a small number of representations) is a challenging visual task at present. In this article, we propose a novelty Mixup-based data augmentation technique, which uses two sampling mechanisms (*i.e.*, instance-based sampling and class-based sampling) to independently sample and mix the training samples. Specifically, we obtain a conventional sample distribution from instance-based sampling and a balanced sample distribution from class-based sampling. The two sets of samples are then mixed to provide a more balanced sample distribution for model training and effectively reduce the under-fitting of tail classes. In addition, we also design a decay mixing strategy to dynamically adjust the sample mixing weights during the training process. By doing so, we can gradually guide the model training and prevent over-fitting. We validate the proposed method on the long-tailed version of datasets created by CIFAR-10, CIFAR-100, CINIC-10, Tiny-ImageNet, Fashion-Mnist, and ImageNet-LT. Experimental results demonstrate that our method achieves better classification performance than other advanced Mixup-based techniques in data imbalanced scenarios.

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Computer Vision, Optimization Theory and Computation, Neural Networks
Keywords Long-tailed distribution, Mixup, Data augmentation, Over-sampling

Submitted 6 November 2024
Accepted 12 September 2025
Published 14 October 2025

Corresponding author
Jiahao Li, jiahaoli@e.gzhu.edu.cn

Academic editor
Bilal Alatas

Additional Information and Declarations can be found on page 18

DOI **10.7717/peerj-cs.3261**

INTRODUCTION

Deep convolutional neural networks (CNNs) have advanced performance in image classification (*He et al., 2016*; *Krizhevsky, Sutskever & Hinton, 2012*). However, the classification performance may decline when the data presents a long-tailed distribution (*Buda, Maki & Mazurowski, 2018*; *Park et al., 2021*; *Yang & Xu, 2020*; *Zhang et al., 2023*). Due to the dominant role of the sample sufficient classes (*i.e.*, head classes), the model's fitting to the sample lacking classes (*i.e.*, tail classes) is inadequate, which impairs the classification performance of the model.

To mitigate the adverse effects caused by long-tailed distribution, many solutions have been developed in relevant works (*Huang et al., 2016*; *Pang et al., 2023*; *Wang et al., 2023*; *Zhang et al., 2021*, *2022a*). Among these methods, re-sampling and re-weighting are the

basic techniques to solve the long-tailed distribution problem. Re-sampling technique is mainly to preprocess the training samples and achieve the balance of training sample distribution through over-sampling or under-sampling (*Ando & Huang, 2017*; *Morais & Vasconcelos, 2019*; *Zhu, Liu & Zhu, 2022*; *Zhao et al., 2023*). These methods improve the representation ability of tail classes by synthesizing the tail class samples. Re-weighting technique usually provides greater weight compensation for tail classes to alleviate the shortage of tail classes learning (*Ren et al., 2018*; *Santiago et al., 2021*; *Shu et al., 2019*).

Although the above techniques have certain advantages on the long-tailed distribution problem, some studies have found that the rebalancing methods (*i.e.*, re-sampling and re-weighting) can impair the representation learning and inhibit the performance of the model (*Cao et al., 2019*; *Zhou et al., 2020*). Therefore, some related works (*Kang et al., 2019*; *Zhou et al., 2020*) decouple the training process and coordinate the combination of conventional training and rebalancing training to achieve better performance. The typical method is bilateral-branch network (BBN) (*Zhou et al., 2020*), which designs a two-branch network to consider both representation learning and rebalancing learning, so as to effectively avoid the damage of rebalancing learning to features. Besides, some related works use data augmentation to increase the diversity of tail class samples (*Li et al., 2021*; *Mullick, Datta & Das, 2019*), like adaptive synthetic (ADASYN) (*He et al., 2008*) or Synthetic Minority Over-sampling Technique (SMOTE) (*Chawla et al., 2002*). These methods improve the representation ability of tail classes by synthesizing the tail class samples.

Our approach is the improvement method of the well-known Mixup (*Zhang et al., 2017*) data augmentation. Mixup is a common data augmentation that enhances the generalization performance of the model through the weighted mixing of training sample pairs. However, Mixup randomly combines training samples regardless of their classes, and mixing in a long-tailed distribution will introduce noise and reduce the performance of the model. Some approaches involving Mixup and long-tailed distribution have been explored recently (*Baik, Yoon & Choi, 2024*; *Chou et al., 2020*; *Galdran et al., 2021*; *Kabra et al., 2020*; *Zhang et al., 2022b*). Remix (*Chou et al., 2020*) retains part of the tail class labels to alleviate the lack of tail classes in the process of mixing. Balanced Mixup (*Galdran et al., 2021*) performs both instance-based sampling and class-based sampling during mixing to create a balanced distribution of synthetic samples. Label-Occurrence-Balanced Mixup (*Zhang et al., 2022a*) combines Mixup with over-sampling so that the occurrence of labels in each class remains statistically balanced. In contrast, Curriculum of Data Augmentation (CUDA) (*Ahn, Ko & Yun, 2023*) mitigates class imbalance by adjusting augmentation strength based on class difficulty.

In this article, we proposed a novel mixing mechanism based on the Mixup data augmentation named progressive decay Mixup. First, the mechanism simultaneously performs instance-based sampling and class-based sampling to provide a balanced sample distribution for mixing. Figure 1 shows the sample distribution diagram of the above sampling strategies. Second, we propose a progressive decay strategy which uses a linear decay factor instead of the random mixing factor in Mixup. By doing this, the proposed strategy can adaptively adjust the mixing weights of the two sampling batches and avoid

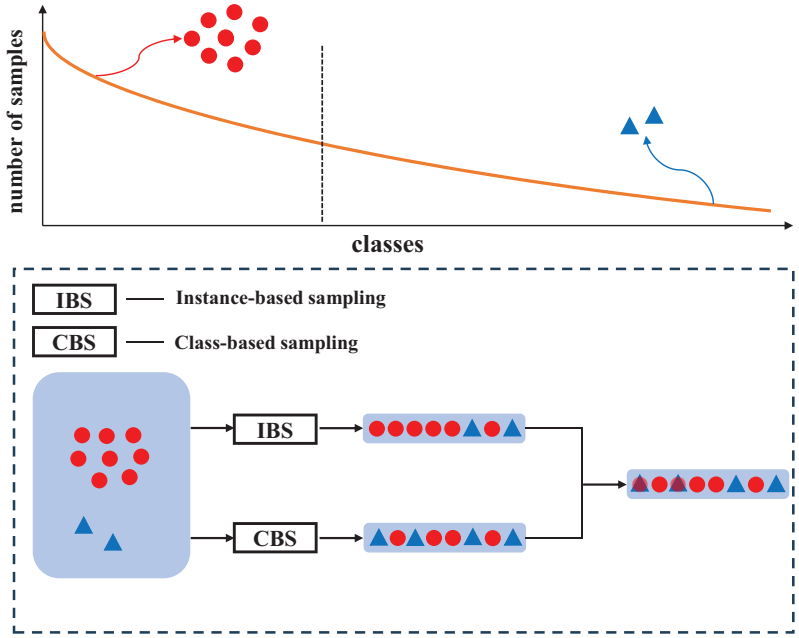


Figure 1 Distribution diagram of mixed samples obtained from two sampling strategies. By taking advantage of class-balanced sampling, the proposed mechanism can effectively increase the proportion of tail class labels in mixed samples, to alleviate the label imbalance.

Full-size DOI: 10.7717/peerj-cs.3261/fig-1

model overfitting by guiding the model learning gradually. Third, compared with the classical bilateral-branch network (BBN) (*Zhou et al., 2020*), the proposed method can availably guide the model training by assigning label distribution without introducing additional models (*Zeng et al., 2024*). The main contributions of our work are:

(a) We provide a more complete form of the Mixup algorithm for long-tailed distribution scenario, and the modified Mixup successfully integrates oversampling with Mixup organically.

(b) For the process of sample mixing, we propose a novel progressive decay strategy which enables the model to focus on tail classes progressively.

(c) By evaluating our method in various long-tailed versions of datasets, the proposed method outperforms other Mixup-based techniques in the long-tailed distribution.

RELATED WORK

Re-sampling

Re-sampling aims to pre-process imbalanced data, which mainly includes over-sampling and under-sampling (*Buda, Maki & Mazurowski, 2018*). For over-sampling, the usual practice is to randomly repeat the tail class samples and add them to the original samples. However, the appearance of repeated tail class samples may lead to overfitting of tail classes. Therefore, some data augmentation techniques have been proposed to further improve the diversity of tail class samples, and the classical methods include SMOTE

(*Chawla et al., 2002*) and ADASYN (*He et al., 2008*). Recently, some newfangled over-sampling techniques have been presented. *Li et al. (2021)* produced diverse tail class samples by translating deep features along many semantically meaningful directions. *Park et al. (2022)* proposed an over-sampling method that generated tail class samples by using the rich background of head class samples. *Zhu, Liu & Zhu (2022)* proposed an interpolating over-sampling method which generated minority samples by identifying clean tail class subregions. For under-sampling, it achieves the relative balance of sample distribution by randomly eliminating part of the head class samples (*Wan, Zhou & Wang, 2024*). However, this approach has an obvious drawback, which may inevitably lose important features (*Wang et al., 2019*), thus affecting the performance of the network. *Niu et al. (2022)* addressed the class imbalance problem by introducing a reliability-based approach grounded in Dempster-Shafer theory, aiming to alleviate the dominance of majority classes. *Zhang et al. (2025)* employed belief function theory in evidential clustering to better represent data uncertainty and imprecision. To improve classification on imbalanced datasets, *Tian et al. (2022)* developed an evidential ensemble method that combines hybrid-sampling (incorporating over-sampling and under-sampling) at the decision level.

Re-weighting

Re-weighting mainly assigns greater weight compensation to tail classes (*Ren et al., 2018*; *Tan et al., 2020*), thereby narrowing the gap between various classes in the classifier. The basic method is weighted by the reciprocal of class frequency or weighted by the reciprocal of the square root of class frequency (*Tan et al., 2020*). *Lin et al. (2017)* proposed a focal loss, whose principle was to introduce a weight coefficient to reduce the weight of negative samples in the total loss, so as to give higher weight to hard-classified tail classes. *Cui et al. (2019)* proposed a class-balanced loss, and the weight was assigned by considering the overlap of each sample. *Ren et al. (2020)* proposed a balanced-meta softmax, where predicted values were nuancedly weighted to counteract the misclassification bias arising from class imbalance. *Zhao et al. (2023)* conducted collaborative training of loss re-weighting and sample re-sampling, which significantly improved the performance of the tail classes.

Representation and classifier learning based on decoupling

The method based on decoupling learning (*Kang et al., 2019*; *Zhou et al., 2020*) is a research field that has attracted much attention in recent years, and such methods have achieved excellent performance in long-tailed distribution. Decoupling learning is to decouple the training process, which tends to learn features in the early stage of training and then tends to re-balance in the later stage of training. *Kang et al. (2019)* proposed a two-stage training strategy in which the original unbalanced sample distribution was used in the first stage of training for the representation learning, and re-sampling was used in the second stage to enhance the attention of tail classes gradually. *Zhou et al. (2020)* designed an end-to-end bilateral-branch network (BBN) to consider both classifier learning and representative learning, and its accumulative learning strategy completed the modeling of the tail classes during the training process. BBN received extensive attention in recent years because of its

excellent performance in image recognition. For example, *Baik, Yoon & Choi (2024)* added data augmentation in two-branch network architecture to strengthen the representation learning of tail classes. *Guo & Wang (2021)* proposed a cross-branch training strategy to improve the synergy of the two branches and consider the performance improvement of both head classes and tail classes.

Mixup-based data augmentation

Mixup (*Zhang et al., 2017*) is a classical data augmentation that imposes linear constraints on the model through linear interpolation between samples to improve the generalization performance of the model. However, mixing under a long-tailed distribution may create noise samples and lead to an imbalance of mixed labels. To solve this problem, many Mixup-based improvements have been presented (*Chou et al., 2020*; *Galdran et al., 2021*; *Zhang et al., 2022a*). *Chou et al. (2020)* proposed a Rebalanced Mixup which retained part of the tail class labels during the mixing process to address the imbalanced problem. *Galdran et al. (2021)* proposed a Balanced Mixup which performs class-based sampling to create a balanced distribution of synthetic samples. *Kabra et al. (2020)* proposed a synthetic over-sampling method that uses Mixup to supplement the mixed samples into the original samples. *Zhang et al. (2022a)* proposed a Label-Occurrence-Balanced Mixup which combined Mixup with over-sampling to keep each class label statistically balanced. *Ahn, Ko & Yun (2023)* introduced CUDA, a curriculum-based data augmentation approach that dynamically determines both the classes to be augmented and the corresponding augmentation strength for each class. *Baik, Yoon & Choi (2024)* proposed an effective augmentation strategy, referred to as Bilateral Mixup Augmentation, which enhances long-tailed visual recognition by mixing head and tail class samples in a balanced manner. *Pan et al. (2024)* presented Contrastive CutMix, which constructs semantically consistent augmented samples to further improve performance in long-tailed recognition tasks. *Zhao, He & Zhao (2025)* developed a dual progressive augmentation framework that progressively balances the learning process to mitigate class imbalance in long-tailed classification scenarios.

PROPOSED METHODS

In this section, we first introduce the concepts of the Mixup data augmentation and the traditional over-sampling strategy, and then we explain the details of the proposed progressive decay Mixup algorithm.

Review of Mixup technique

Mixup (*Zhang et al., 2017*) is a common data augmentation, which creates mixed samples by linear weighting between samples to improve the linear representation of the model. Specifically, assuming any training sample x and its label y, Mixup generates mixed sample $(\hat{x}, \hat{y})$ by linearly weighting two random training samples (x_a, y_a) and (x_b, y_b). The expression is as follows:

$$\begin{aligned} \hat{x} &= \lambda x_a + (1-\lambda) x_b \\ \hat{y} &= \lambda y_a + (1-\lambda) y_b. \end{aligned} \tag{1}$$

The mixing factor $\lambda \in (0, 1)$ is sampled from the $beta(\alpha, \alpha)$, where α is the hyperparameter that controls the beta distribution. Due to the randomness of data sampling, Mixup technique may synthesize noise samples when the data distribution is extremely class-imbalanced, and it also leads to the reduction of classification performance.

Data sampling strategies

When the training data presents a long-tailed distribution, tail classes are usually suppressed by head classes during the training process, which causes the underfitting of tail classes, and even tail classes are completely ignored by the model. Data-level based sampling strategies can be used to relieve this effect, such as over-sampling, although repeatedly showing the same tail class samples to the model may result in overfitting of tail classes. The general sampling strategy can be described as follows:

$$p_j = \frac{(n_j)^q}{\sum_{k=1}^{K} (n_j)^k} \tag{2}$$

where j refers to the index of classes. p_j represents the sampling frequency from the class j. n_j is the total sample number of class j. K denotes the number of classes. q is a parameter setting.

When $q = 1$, the sampling frequency is equal to the frequency of the samples on the training set (*i.e.*, instance-based sampling). In this case, the classes with more samples have a higher sampling frequency, and it obtains an imbalanced sample batch B_I.

When $q = 0$, the sampling probability $p_j = \frac{1}{K}$ remains the same for each class (*i.e.*, class-based sampling), and class-based sampling strategy receives a balanced sample batch B_C.

Proposed double sampler mixup

The two sets of data batches obtained from the above sampling strategies are then linearly weighted. By doing so, we leverage class-based sampling to effectively enhance the proportion of tail classes, while incorporating the linear mechanism of Mixup to prevent overfitting. Meanwhile, we hope to achieve the guiding effect of the model by dynamically adjusting the weights of the two sample groups. The specific mixing process is shown as follows:

$$\begin{aligned} \tilde{x} &= \rho x_I + (1-\rho) x_c \\ \tilde{y} &= \rho y_I + (1-\rho) y_c \end{aligned} \tag{3}$$

where $(\tilde{x}, \tilde{y})$ is the generated new sample. (x_I, y_I) denotes the sample from the imbalanced batch B_I, and (x_c, y_c) represents the sample from the balanced batch B_C. ρ is the proposed linear decay factor.

Proposed decay mixing strategy

The mixing factor in the original Mixup is acquired from the beta distribution $beta(\alpha, \alpha)$, but it still relies on a hyperparameter α. Motivated from cumulative learning of

bilateral-branch network (*Zhou et al., 2020*), we propose a decay mixing strategy which designs a linear decay coefficient instead of the random mixing factor. The specific mixing factor is as follows:

$$\rho = 1 - \frac{T_{epoch}}{T_{max}} \tag{4}$$

where ρ is the proposed linear decay factor, and T_{epoch} refers to the specific number of training epochs. T_{max} is the total number of training epochs. As the training epochs increased, the value of ρ is gradually decreasing. The training samples of generated batches are fed into the network for training. For each class $j \in \{1, 2, \ldots, K\}$, the probability of output *via* the softmax is as follows:

$$\hat{p} = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{5}$$

where $\hat{p}$ denotes the probability of output. z is the predicted score. The weighted cross entropy loss is formulated follows:

$$L = \rho E(\hat{p}, y_I) + (1-\rho) E(\hat{p}, y_c) \tag{6}$$

where $E(\hat{p}, y)$ is expressed as cross entropy loss. The specific process of the proposed method is shown in Fig. 2.

The proposed decay mixing strategy abandons the random mixing factor and thus eliminates the control of hyperparameter α. More importantly, the proposed strategy can adaptively adjust the weights according to the specific training epochs. During the training process, the two sampling branches coordinate with each other to create mixed samples with different label distributions by adjusting the mixing weights of each training epoch. At the early stage of training, imbalanced sample batches occupy the main weights, and it helps the model quickly establish the basic representation. As the mixing factor decays, the samples from the balanced batches dominate the mixing weights in the later training period, which can strengthen the learning of tail classes.

EXPERIMENTS

In this section, we introduce the image recognition datasets used in this experiment, which are CIFAR-10/100 (*Krizhevsky & Hinton, 2009*), CINIC-10 (*Darlow et al., 2018*), Fashion-Mnist (*Xiao, Rasul & Vollgraf, 2017*), Tiny-ImageNet (*Li, Karpathy & Johnson, 2017*), and ImageNet-LT (*Olga et al., 2015*). Then we illustrate the partitioning method of the long-tailed distribution and construct the long-tailed version of the above datasets.

Datasets introduction and long-tailed settings

CIFAR-10/100 (*Krizhevsky & Hinton, 2009*): The CIFAR dataset is a classic image classification dataset, which can be divided into CIFAR-10 and CIFAR-100 subsets according to the different numbers of categories. Both CIFAR-10 and CIFAR-100 include 50,000 training samples and 10,000 testing samples, and the size of images is 32×32.

CINIC-10 (*Darlow et al., 2018*): the CINIC-10 dataset consists of 270,000 images, which is 4.5 times the size of the CIFAR-10 dataset. It is built on two datasets, ImageNet

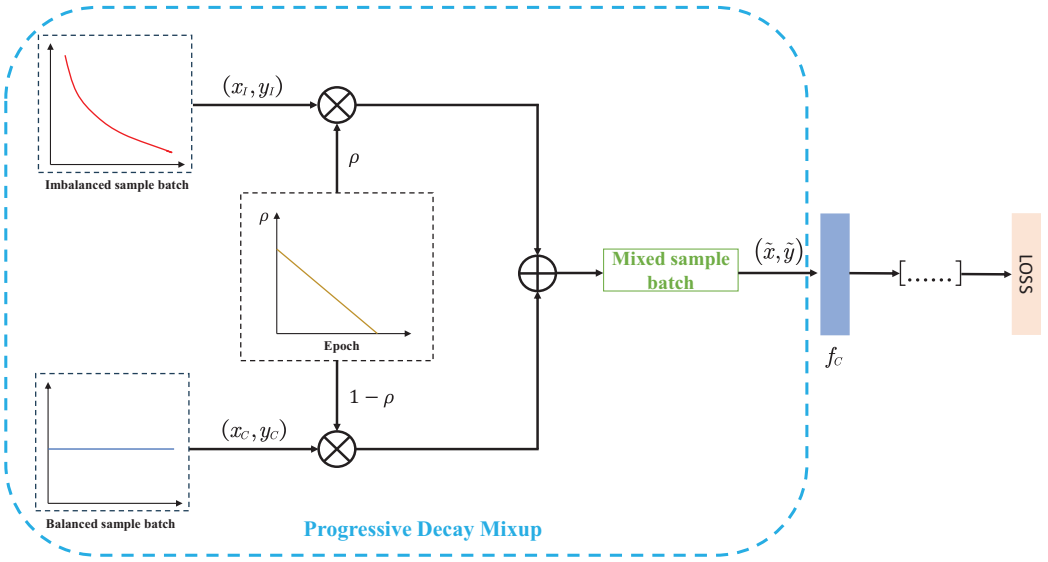


Figure 2 The specific process of the progressive decay MIxup data augmentation. Firstly, we create balanced sample batches and imbalanced sample batches by using two samplers with different sampling properties. Secondly, the two sets of sample batches are mixed to generate new batches for model training. Moreover, the proposed linear decay factor is designed as a mixing coefficient to adjust the weights of the two batches, which guides the model training.

Full-size DOI: 10.7717/peerj-cs.3261/fig-2

(*Deng et al., 2009*) and CIFAR-10. The CINIC-10 dataset is divided into three equal subsets (training, validation, and testing), each containing 90,000 images. Due to its ample number of samples, it is possible to construct the long-tailed distribution while ensuring a sufficient number of samples for each category. In our experiments, we use the training set and the validation set for training and testing, respectively.

Fashion-Mnist (*Xiao, Rasul & Vollgraf, 2017*): the Fashion-Mnist is a clothing image dataset with a training set of 60,000 samples and a testing set of 10,000 samples. The samples in the dataset are divided into 10 classes and each sample is a 28 $\times$ 28 grayscale image.

Tiny-ImageNet (*Li, Karpathy & Johnson, 2017*): the Tiny-ImageNet dataset is a common visual dataset with images derived from the ImageNet dataset. The dataset contains 200 categories, where each category is assigned 500 images for training, 50 images for validation, and 50 images for testing. In our experiments, we only use the training set and validation set.

ImageNet-LT (*Olga et al., 2015*): the ImageNet-LT dataset is a long-tailed variant of the large-scale ImageNet dataset, which was created by sampling 1,000 classes according to a Pareto distribution with $\alpha = 6$. Its sampling leads to significant class imbalance, with head classes exceeding 1,280 samples and tail classes having fewer than 5. The ImageNet-LT dataset contains approximately 115,000 images in total and is widely used as a benchmark for evaluating performance under extreme class imbalance.

Long-tailed Settings: before training, we preprocess the above dataset to construct the long-tailed version of the dataset by adjusting the sample number. The long-tailed datasets

are created based on an exponential decay function (*i.e.*, $N_i = N_{max} \times \delta^{(i/(C-1))}$). N_i denotes the sample number of class i, and N_{max} indicates the sample number of the most frequent class. C is the number of categories. The imbalance ratios δ (*i.e.*, $\delta = N_{max}/N_{min}$) we used in the experiments are 10, 50, and 100, where N_{min} is the sample number of the rarest class.

Implementation details

In the data preprocessing stage, we apply usual data augmentation strategies, including random crop and random horizontal flip. We train the model by using the mini-batch gradient descent with momentum 0.9 and weight decay 0.0002, and the batch size is set to 128. By following the work of *Cao et al. (2019)*, we set the initial learning rate to 0.1 and adopt a warm-up learning rate schedule in the first 5 epochs. The number of training epochs is 300, and we decay the learning rate at the 150^{th} epoch and 225^{th} epoch by 0.01, respectively. For CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets, we use ResNet-32 as the backbone network. For CINIC-10 and Fashion-Mnist datasets, we adopt ResNet-18 for all experiments. For ImageNet-LT dataset, we use ResNet-50 as the backbone and train the model for 200 epochs.

Experiment results

We verify the classification performance of the proposed method on the above datasets. Moreover, we compare the proposed method with the vanilla method, rebalancing strategies, and some novel Mixup-based techniques. To ensure the fairness of the experiment, all methods are set with the same random seeds. The comparison methods are as follows: (1) CE: Cross-entropy loss. (2) Focal loss (*Lin et al., 2017*): Reweighted cross-entropy loss. (3) DRW, DRS (*Cao et al., 2019*): Deferred re-weighting and deferred re-sampling. (4) LDAM-DRW (*Cao et al., 2019*): Label-distribution-aware margin loss integrate with re-weighting. (5) Mixup-based data augmentations: Mixup (*Zhang et al., 2017*) and other variants such as Remix (*Chou et al., 2020*), Balanced-Mixup (*Galdran et al., 2021*), Label-Occurrence-Balanced Mixup (*Zhang et al., 2022a*). For Balanced-Mixup, we follow the two sets of parameter settings $beta(0.2, 1)$ and $beta(0.3, 1)$ in its original article (*Galdran et al., 2021*). The experimental settings of CUDA are adopted as specified in the original article (*Ahn, Ko & Yun, 2023*).

The top-1 classification accuracies of the above methods on long-tailed CIFAR-10 and long-tailed CIFAR-100 are reported in Table 1. We discover that the proposed method achieves better classification performance than other compared methods on both long-tailed CIFAR-10 and long-tailed CIFAR-100. For example, we get 78.80% top-1 accuracy for long-tailed CIFAR-10 with $\delta = 100$, which is 2.83% higher than that of Label-Occurrence Balanced Mixup, and we get 43.90% top-1 accuracy for long-tailed CIFAR-100 with $\delta = 100$, which is 2.20% higher than that of Balanced-Mixup ($beta(0.3, 1)$).

The top-1 accuracies of the involved methods on long-tailed CINIC-10 and long-tailed Fashion-Mnist are shown in Table 2. It can be found that the proposed method still achieves better performance under most imbalanced settings. Specifically, we get 66.66%

Table 1 Experimental results (Top-1 accuracy) on long-tailed CIFAR-10 and long-tailed CIFAR-100 under different imbalance ratios. The imbalance ratio δ = 10, 50, 100 correspond to different degrees of long-tailed distribution, respectively. Bold fonts perform the best performance.

	CIFAR-10			CIFAR-100		
Methods	**10**	**50**	**100**	**10**	**50**	**100**
CE	86.94	77.36	70.99	56.95	43.83	39.01
Focal (*Lin et al., 2017*)	87.20	76.80	71.71	56.51	44.17	38.37
DRW (*Cao et al., 2019*)	87.37	77.98	74.07	57.17	44.54	39.39
DRS (*Cao et al., 2019*)	87.11	77.40	74.06	56.70	44.12	38.48
LDAM-DRW (*Cao et al., 2019*)	87.24	79.36	74.66	56.29	45.07	40.56
CUDA (*Ahn, Ko & Yun, 2023*)	87.56	77.96	74.89	59.09	46.74	38.56
Mixup (*Zhang et al., 2017*)	88.01	78.75	73.05	58.21	44.60	40.04
Remix (*Chou et al., 2020*)	88.25	79.28	72.11	58.13	45.22	40.66
Balanced-Mixup ($beta(0.2, 1)$) (*Galdran et al., 2021*)	88.42	79.39	74.38	59.15	46.25	40.11
Balanced-Mixup ($beta(0.3, 1)$) (*Galdran et al., 2021*)	88.56	79.28	74.96	59.66	46.84	41.70
Label-Occurrence Mixup (*Zhang et al., 2022a*)	88.82	81.25	75.97	59.51	47.16	40.66
Ours	**88.83**	**83.33**	**78.80**	**60.52**	**50.04**	**43.90**

Table 2 Experimental results (Top-1 accuracy) on long-tailed CINIC-10 and long-tailed Fashion-Mnist under different imbalance ratios. The imbalance ratio δ = 10, 50, 100 correspond to different degrees of long-tailed distribution, respectively. Bold fonts perform the best performance.

	CINIC-10			Fashion-Mnist		
Methods	**10**	**50**	**100**	**10**	**50**	**100**
CE	77.79	66.90	61.46	92.90	89.62	87.88
Focal (*Lin et al., 2017*)	77.45	66.21	60.79	92.81	89.07	88.44
DRW (*Cao et al., 2019*)	77.89	67.14	61.05	92.47	90.08	88.53
DRS (*Cao et al., 2019*)	77.79	66.95	61.04	92.58	89.94	88.51
LDAM-DRW (*Cao et al., 2019*)	78.21	69.42	64.92	92.76	91.90	91.48
CUDA (*Ahn, Ko & Yun, 2023*)	78.65	65.53	61.06	92.41	90.52	88.65
Mixup (*Zhang et al., 2017*)	79.70	68.37	63.35	93.96	90.49	89.11
Remix (*Chou et al., 2020*)	79.35	68.42	63.45	93.73	90.23	89.54
Balanced-Mixup ($beta(0.2, 1)$) (*Galdran et al., 2021*)	78.53	67.72	62.86	93.75	91.07	89.55
Balanced-Mixup ($beta(0.3, 1)$) (*Galdran et al., 2021*)	79.31	67.69	61.52	93.83	90.96	89.73
Label-Occurrence Mixup (*Zhang et al., 2022a*)	79.41	68.65	62.88	93.28	**92.68**	**93.28**
Ours	**81.92**	**73.06**	**66.66**	**93.99**	92.00	90.47

top-1 accuracy for long-tailed CINIC-10 with $\delta = 100$, which is 3.78% higher than that of Label-Occurrence Balanced Mixup. For long-tailed Fashion-Mnist with $\delta = 50$, we get 92.00% top-1 accuracy, which is 1.04% higher than that of Balanced-Mixup $beta(0.3, 1)$. The top-1 accuracies of the compared methods on long-tailed Tiny-ImageNet are reported in Table 3. Building upon these results, our propose method also demonstrate strong performance on the large-scale ImageNet-LT dataset, achieving a top-1 accuracy of 43.21%, which exceeded that of Mixup by 1.36%, as shown in Table 4. Although Mixup

Table 3 Experimental results (Top-1 accuracy) on long-tailed Tiny-ImageNet under different imbalance ratios. The imbalance ratio δ = 10, 50, 100 correspond to different degrees of long-tailed distribution, respectively. Bold fonts perform the best performance.

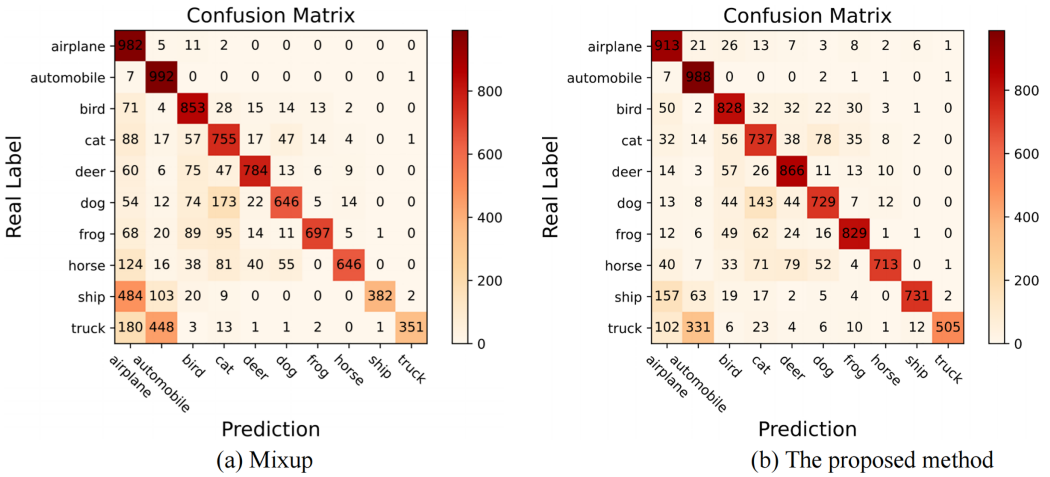
	Tiny-ImageNet		
Methods	**10**	**50**	**100**
CE	36.23	25.15	22.06
Focal (*Lin et al., 2017*)	35.73	25.26	21.79
DRW (*Cao et al., 2019*)	36.92	25.26	22.26
DRS (*Cao et al., 2019*)	36.92	25.03	22.37
LDAM-DRW (*Cao et al., 2019*)	34.14	24.71	21.49
CUDA (*Ahn, Ko & Yun, 2023*)	38.64	24.89	22.34
Mixup (*Zhang et al., 2017*)	39.61	28.19	25.00
Remix (*Chou et al., 2020*)	40.72	28.12	25.03
Balanced-Mixup ($beta(0.2, 1)$) (*Galdran et al., 2021*)	39.56	26.96	23.69
Balanced-Mixup ($beta(0.3, 1)$) (*Galdran et al., 2021*)	40.49	28.19	24.61
Label-Occurrence Mixup (*Zhang et al., 2022a*)	37.77	23.98	19.25
Ours	**42.60**	**30.88**	**25.55**

Table 4 Experimental results (Top-1 accuracy) on ImageNet-LT with pareto $\alpha = 6$. Bold fonts perform the best performance.

Methods	**ImageNet-LT**
CE	38.72
Mixup (*Zhang et al., 2017*)	41.85
Balanced Mixup (*Galdran et al., 2021*)	42.26
Ours	**43.21**

alleviates class imbalance to some extent, it still fails to sufficiently improve the performance on tail classes. In contrast, the dual-sampler strategy proposed in our approach enables more effective learning under long-tail distributions, significantly enhancing its applicability in real-world scenarios. The trend of experimental results is basically consistent with other datasets, which further confirms the wide applicability of the proposed method.

The proposed method is an improved version of Mixup in the long-tailed distribution scenario. To intuitively demonstrate that the proposed method is superior to Mixup, we present their confusion matrices on long-tailed CIFAR-10 with $\delta = 100$, and the specific visual confusion matrices are shown in Fig. 3. There into, “airplane”, “automobile” are head classes, and “ship”, “truck” are tail classes. It can be seen from the two confusion matrices that the proposed method can significantly improve the classification performance of tail classes.



(a) Mixup (b) The proposed method

Figure 3 The visualization of confusion matrices on long-tailed CIFAR-10 with the imbalance ratio of 100. (A) When the Mixup operation is carried out under long-tailed distribution, the correct classification number of tail classes is low due to insufficient representation of tail classes. (B) When the proposed method is applied, the correct classification number of tail classes is significantly increased, while the number of misclassifications is obviously reduced. It denotes that our method improves the overall classification level by improving the performance of the tail classes.

Full-size DOI: 10.7717/peerj-cs.3261/fig-3

DISCUSSIONS

Analysis of weighting factors

The proposed method adopts a linear decay factor to adjust the mixing weight of the two sample batches in the mixing process. To further illustrate the applicability of the adopted mixing strategy, this section conducts comparative experiments on different mixing factors. Referring to the work of *Zhou et al. (2020)*, different weighting coefficient strategies (including parabolic decay, cubic decay, linear decay, equal weight, and beta distribution) are used in the comparative experiments. The weighting strategies are shown in Fig. 4. The comparison experiment uses the above weighting schemes in the long-tailed CIFAR-10/100 datasets, and the experimental results are shown in Table 5.

As can be seen from Table 5, the model of linear decay strategy has the highest classification accuracy under different imbalance ratios, while that of equal weight has the lowest accuracy. Compared with the equal weight method, the top-1 accuracy of long-tailed CIFAR-10 and long-tailed CIFAR-100 datasets with $\delta = 100$ is improved by +8.97% and +5.17%, respectively. This shows that the proposed linear decay strategy has a better fit with the double sampler, and the representation learning and classifier learning are established by dynamically adjusting the label distribution of the mixed samples during the training process, so the model achieves the expected performance of classification. However, the equal weight factor loses the ability to adjust the mixing weight to some extent, and cannot be effectively combined with the double sampler. It can even damage the performance of the model, and the final performance is lower than the benchmark level.

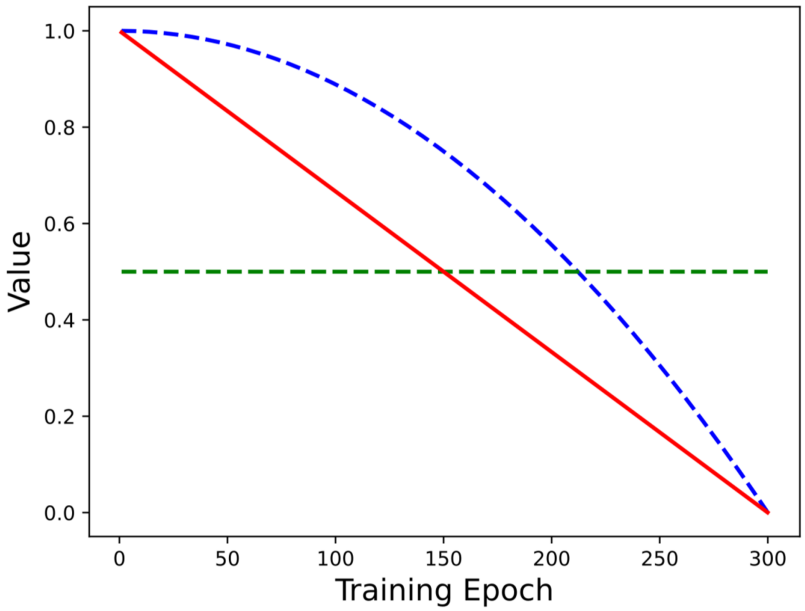


Figure 4 **The coefficients of different weighting strategies change with the number of epochs.**
Full-size DOI: 10.7717/peerj-cs.3261/fig-4

Table 5 **Performance comparison of linear decay and cubic decay in the CIFAR-10 and long-tailed CIFAR-100 datasets.** The imbalance ratio δ = 10, 50, 100 correspond to different degrees of long-tailed distribution, respectively. Bold fonts perform the best performance.

	CIFAR-10			CIFAR-100		
Methods	**10**	**50**	**100**	**10**	**50**	**100**
Equal weight ($\lambda = 0.5$)	83.34	75.16	69.83	51.63	42.77	38.73
Parabolic decay (*Zhou et al., 2020*)	87.83	80.40	74.73	58.77	45.50	39.95
Cubic decay	76.76	68.64	62.39	36.62	31.09	26.18
Radom weight ($beta(1,1)$)	**88.96**	80.82	76.68	59.88	47.09	41.30
Ours	88.83	**83.33**	**78.80**	**60.52**	**50.04**	**43.90**

According to the motivation in "Proposed decay mixing strategy" section, the purpose of decay mixing strategy is to help model learn the basic fundamental representations in the initial training phase while preventing overfitting. Then we gradually transition to classifier learning to strengthen the model's focus on tail classes. To validate our motivation, we show the accuracy of the model in different training epochs under different weighting factors. In Fig. 5, the experimental results of long-tailed CIFAR-10 and long-tailed CIFAR-100 with $\delta = 100$ are shown from left to right.

As can be seen from Fig. 5, the random weights follow the practice of *Kabra et al. (2020)* and sample from $beta(1,1)$, and it obtain a relatively good classification accuracy. The proposed linear decay trend is in good agreement with the expected training process. In the early stage of training, the imbalanced batch occupies a large mixing weight to establish the basic representation ability of the model. At this time, the classification accuracy of the

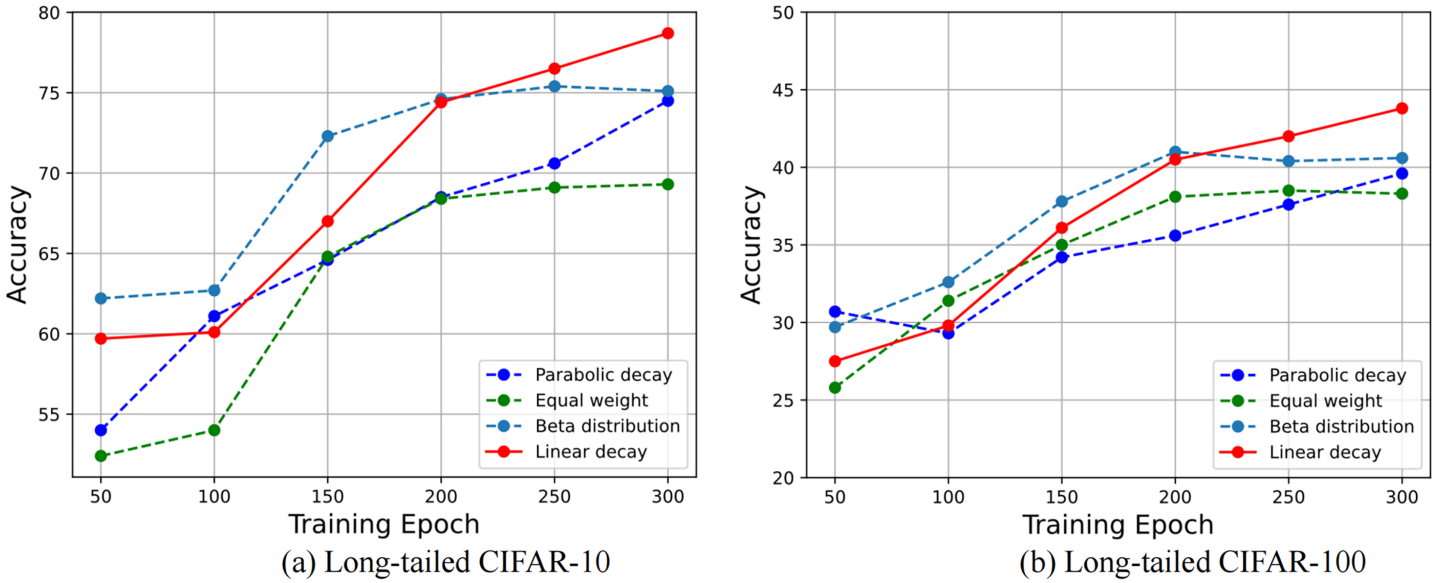


Figure 5 The graph of accuracy variation under different weight coefficient strategies. The experimental datasets based on (A) long-tailed CIFAR-10 and (B) long-tailed CIFAR-100 with the imbalance ratio of 100. Full-size DOI: 10.7717/peerj-cs.3261/fig-5

model does not improve obviously. With the increase of the training epochs, the branch of the class-balanced sampling dominates the mixing process, and the training attention of the model turns to tail classes.

In the process of gradual convergence, the classification accuracy is significantly increased, and the final classification accuracy exceeds the random weight. However, the parabolic decay strategy does not quickly transition to the rebalancing learning in the late training epochs, and the performance is lower than ours. It can be seen that a reasonable weight decay strategy is important for the mixing of double sampler samples, and an inappropriate mixing strategy will even injure the learning effect of the model.

Ablation studies

To prove the rationality of instance-based sampling combined with class-balanced sampling, this section firstly sets different control groups for the selection of sampler, namely ours (instance-based sampling) and ours (class-based sampling), which means sampling only with instance-based sampling or class-based sampling, respectively. As a Mixup-based approach, we still use the regular Mixup as one of the comparison groups. In addition, in order to eliminate the influence of decay strategies on experimental results, a new control group Label-Occurrence Balanced Mixup is set up in this section, which is the state of the art for long-tailed problem based on Mixup. It adopts two class-balanced sampling to obtain two balanced sample batches, and the $beta(1, 1)$ is used to generate the mixing factors. The above comparative experiments are trained on the long-tailed CIFAR-10/100, and the relevant results are shown in Table 6.

As can be seen from Table 6, the performance of the classifier obtained by using only instance-based sampling or only class-based sampling is inferior to that of the double sampler method, which indicates that the single sampling property cannot fit well with the

Table 6 Top-1 accuracy of different sampler combinations in long-tailed CIFAR-10 and long-tailed CIFAR-100. The imbalance ratio δ = 10, 50, 100 correspond to different degrees of long-tailed distribution, respectively. Bold fonts perform the best performance.

	CIFAR-10			CIFAR-100		
Methods	**10**	**50**	**100**	**10**	**50**	**100**
Mixup (*Zhang et al., 2017*)	88.01	78.75	73.05	58.21	44.60	40.04
Ours (instance-based sampling)	86.99	78.17	73.06	56.08	43.12	38.63
Ours (class-based sampling)	88.65	81.94	77.15	58.84	46.26	39.92
Label-Occurrence Mixup (*Zhang et al., 2022a*)	88.82	81.25	75.97	59.51	47.16	40.66
Ours	**88.83**	**83.33**	**78.80**	**60.52**	**50.04**	**43.90**

decay mixing strategy. When the two sample batches have similar class distributions, the decay mixing factor cannot adjust the label distribution to guide model training. Moreover, compared with the Label-Occurrence Balanced Mixup, the proposed method still obtains a superior effect, which certifies that the double sampler method with different sampling properties can receive better performance.

Analysis of sampler order

The proposed method has regulations on the order of the two data batches (*i.e.*, balanced sample batch and imbalanced sample batch). In the early stage of training, imbalanced sample batches from instance-based sampling occupy a higher mixing proportion to ensure the model's representation learning. With the increase of training epochs, the mixing factor decreases linearly, and the mixing weights are dominated by the balanced sample batches. In the later stage of training, the model gradually switches to classifier learning for tail classes. Regarding representation learning and classifier learning, *Zhou et al. (2020)* have made a detailed explanation. To further illustrate the effect of the sampler order, we reverse the original sampler order (*i.e.*, reversed sampler), that is, the balanced sample batches have a higher mixing weight in the early stage of training, and the imbalanced sample batch dominates the mixing weight in the late stage of training. Experimental comparison results are shown in Table 7.

As can be seen in Table 7, our method performs much better than the reverse sampler. This just confirms the mainstream view that premature application of rebalancing techniques can impair representation learning and thus reduce classification performance. In comparison, our method successfully applies the idea of decoupling learning to Mixup data augmentation and achieves good performance.

Comparison with bilateral-branch network

The proposed mixing decay strategy is motivated from the bilateral-branch network (BBN), but there are clear distinctions between the two approaches. The practice of BBN is to construct two network branches and mix the predicted scores of the network branches, while the proposed method is the mixing of the input data. As a Mixup-based data augmentation, the proposed method gets rid of the inherent network architecture of BBN, and the weights of the labels are adjusted by the linear attenuation coefficient to guide the

Table 7 Top-1 accuracy of different sampler sequences in long-tailed CIFAR-10 and long-tailed CIFAR-100. The imbalance ratio δ = 10, 50, 100 correspond to different degrees of long-tailed distribution, respectively. Bold fonts perform the best performance.

	CIFAR-10			CIFAR-100		
Methods	**10**	**50**	**100**	**10**	**50**	**100**
Ours (reversed sampler)	86.33	75.36	72.26	57.18	42.04	37.82
Ours	**88.83**	**83.33**	**78.80**	**60.52**	**50.04**	**43.90**

Table 8 Comparison of the proposed method with both BBN and DBN-Mix under long-tailed CIFAR-10/100 dataset. † denotes the experimental results cited from the original article. The imbalance ratio δ = 10, 50, 100 correspond to different degrees of long-tailed distribution, respectively. Bold fonts perform the best performance.

	CIFAR-10			CIFAR-100		
Methods	**10**	**50**	**100**	**10**	**50**	**100**
BBN† (*Zhou et al., 2020*)	88.32	82.18	79.82	59.12	47.02	42.56
Ours	88.83	83.33	78.80	60.52	50.04	43.90
DBN-Mix† (*Baik, Yoon & Choi, 2024*)	90.87	**86.82**	83.47	62.37	50.39	**45.07**
Ours + BBN	**92.55**	84.23	**83.98**	**62.45**	**50.52**	44.21

model training. To verify the advance of the proposed method, this section compares the classification performance of the proposed method with that of BBN in long-tailed CIFAR10/100, and the experimental results are shown in Table 8. Furthermore, Dual-branch Network with Bilateral Mixup (DBN-Mix) (*Baik, Yoon & Choi, 2024*) is a method that integrates the dual-branch network with Mixup. To further validate the superiority of our approach, we incorporate it with BBN and conduct a comparative analysis against DBN-Mix.

It can be seen from the experimental results in Table 8, the accuracy of our method is higher than that of the classical BBN under the long-tailed CIFAR-10 and long-tailed CIFAR-100. Although BBN has a significant improvement, this comes at the expense of additional model parameters, because the two-branch network means the doubling of model parameters, which requires more training time, resulting in low training efficiency. The proposed method effectively avoids this problem, because it does not involve the modification of the overall model architecture, and it guides the training by adjusting the label proportion of the mixed samples. Compared with the BBN, our method improves the classification performance without increasing the extra model parameters. Meanwhile, in comparison to DBN-Mix, our BBN-integrated approach still demonstrates superior performance by enabling more effective model training adjustments.

Comparison of methods using F1-score

Due to the extreme class imbalance in long-tailed classification tasks, accuracy is often insufficient to fully capture a model's effectiveness. Therefore, we adopt the F1-score, which combines both precision and recall to provide a more comprehensive

Table 9 Experimental results (F1-score) on long-tailed CIFAR-10 and long-tailed CIFAR-100 under different imbalance ratios. The imbalance ratio δ = 10, 50, 100 correspond to different degrees of long-tailed distribution, respectively. Bold fonts perform the best performance.

	CIFAR-10			CIFAR-100		
Methods	**10**	**50**	**100**	**10**	**50**	**100**
CE	84.07	68.14	59.33	52.61	31.74	27.15
Focal (*Lin et al., 2017*)	83.46	67.84	59.41	51.18	32.68	27.75
DRW (*Cao et al., 2019*)	85.42	69.24	58.24	51.61	31.54	29.15
DRS (*Cao et al., 2019*)	86.91	64.85	59.33	54.61	40.25	31.87
LDAM-DRW (*Cao et al., 2019*)	87.43	68.82	68.70	54.33	40.69	33.05
CUDA (*Ahn, Ko & Yun, 2023*)	88.96	79.62	75.48	58.36	41.67	36.21
Mixup (*Zhang et al., 2017*)	84.67	71.18	62.41	46.08	37.15	26.76
Remix (*Chou et al., 2020*)	85.05	67.64	61.94	47.36	37.31	26.46
Balanced-Mixup ($beta(0.2, 1)$) (*Galdran et al., 2021*)	90.24	80.68	73.75	58.69	41.58	36.43
Balanced-Mixup ($beta(0.3, 1)$) (*Galdran et al., 2021*)	90.67	80.77	75.14	59.34	42.74	31.80
Label-Occurrence Mixup (*Zhang et al., 2022a*)	89.48	81.02	76.42	59.71	40.01	34.55
Ours	**91.18**	**82.46**	**76.59**	**64.38**	**43.61**	**38.31**

evaluation of classification performance. As shown in Table 9, our proposed method consistently outperforms the listed Mixup-based approaches across all evaluated imbalance ratios (δ = 10, 50, 100) on both long-tailed CIFAR-10 and long-tailed CIFAR-100 datasets. In particular, on long-tailed CIFAR-100 with δ = 10, our approach achieves an F1-score of 64.38%, outperforming Balanced-Mixup $beta(0.3, 1)$ by 5.04%. These gains confirm our method's novel fusion of oversampling and progressive decay mixing, critically advancing tail-class representation learning.

CONCLUSION

This work proposes a Mixup-based data augmentation technique called progressive decay Mixup for long-tailed image classification. The method integrates Mixup augmentation with re-sampling to provide a balanced sample distribution for mixing. In addition, we design a decay mixing strategy to dynamically adjust the mixing weights during the training process. Through the above practices, we provide a more complete form of the Mixup augmentation for long-tailed distribution. Experiments illustrate that our method has achieved good results on long-tailed CIFAR-10/100, long-tailed CINIC-10, long-tailed Fashion-Mnist, long-tailed Tiny-ImageNet and ImageNet-LT.

ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude for the collaboration and support extended by Guangzhou University throughout this research endeavor. Their invaluable expertise, resources, and constructive feedback were instrumental in achieving the success of this study. We are profoundly appreciative of the opportunity to work closely with Guangzhou University; the valuable insights gained from this partnership have significantly enriched the quality of our research.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Hongyin Li conceived and designed the experiments, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Jiahao Li conceived and designed the experiments, performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Cheng Qian performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Zilin Xiang analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Rongdong Chen analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The CIFAR-10/100 dataset is available at: https://www.cs.toronto.edu/~kriz/cifar.html.

The CINIC-10 dataset is available at: https://datashare.ed.ac.uk/handle/10283/3192.

The Tiny-ImageNet dataset is available at Kaggle: https://www.kaggle.com/c/tiny-imagenet.

The Fashion-MNIST dataset is available at Kaggle: https://www.kaggle.com/datasets/zalando-research/fashionmnist.

The ImageNet-LT dataset is available at: https://www.image-net.org/challenges/LSVRC/2012.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3261#supplemental-information.

REFERENCES

Ahn S, Ko J, Yun S. 2023. CUDA: curriculum of data augmentation for long-tailed recognition. In: *International Conference on Learning Representations*. Piscataway: IEEE, 1–5.

Ando S, Huang CY. 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*. Cham: Springer.

Baik JS, Yoon I, Choi JW. 2024. DBN-Mix: training dual branch network using bilateral mixup augmentation for long-tailed visual recognition. *Pattern Recognition* **147(5)**:110107 DOI 10.1016/j.patcog.2023.110107.

Buda M, Maki A, Mazurowski MA. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* **106(7)**:249–259 DOI 10.1016/j.neunet.2018.07.011.

Cao K, Wei C, Gaidon A, Arechiga N, Ma T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In: *Advances in Neural Information Processing Systems, 32*. Cambridge.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**:321–357 DOI 10.1613/jair.953.

Chou HP, Chang S, Pan J, Wei W, Juan D. 2020. Remix: rebalanced mixup. In: *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Cham: Springer, 95–110.

Cui Y, Jia M, Lin TY, Song Y, 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.

Darlow LN, Crowley EJ, Antoniou A, Storkey AJ. 2018. CINIC-10 is not imageNet or CIFAR-10. ArXiv DOI 10.48550/arXiv.1810.03505.

Deng J, Dong W, Socher R, Li L, Li K, Li F. 2009. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Galdran A, Carneiro G, Gonzalez B, Miguel A. 2021. Balanced-mixup for highly imbalanced medical image classification. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Cham: Springer, 323–333.

Guo H, Wang S. 2021. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15089–15098.

He H, Bai Y, Garcia E, Li SA. 2008. Adaptive synthetic sampling approach for imbalanced learning. In: *IEEE International Joint Conference on Neural Networks. IEEE World Congress on Computational Intelligence.*

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang C, Li Y, Loy CC, Tang X. 2016. Learning deep representation for imbalanced classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5375–5384.

Kabra A, Chopra A, Puri N, Badjatiya P, Verma S, Gupta P, Krishnamurthy B. 2020. MixBoost: synthetic oversampling using boosted mixup for handling extreme imbalance. In: *2020 IEEE International Conference on Data Mining (ICDM)*. Piscataway: IEEE, 1082–1087.

Kang B, Xie S, Rohrbach M, Yan Z, Gordo A, Feng J, Kalantidis Y. 2019. Decoupling representation and classifier for long-tailed recognition. ArXiv DOI 10.48550/arXiv.1910.09217.

Krizhevsky A, Hinton G. 2009. Learning multiple layers of features from tiny images. Technical Report.

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems* **141**:1097–1105 DOI 10.1145/3065386.

Li S, Gong K, Liu CH, Wang Y, Qiao F, 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.

Lin T, Goyal P, Girshick R, He K, Dollar P. 2017. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Li F, Karpathy A, Johnson J. 2017. Tiny ImageNet. *Available at https://www.kaggle.com/c/tiny-imagenet*.

Morais RF, Vasconcelos GC. 2019. Boosting the performance of over-sampling algorithms through under-sampling the minority class. *Neurocomputing* **343(10)**:3–18 DOI 10.1016/j.neucom.2018.04.088.

Mullick SS, Datta S, Das S. 2019. Generative adversarial minority oversampling. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1695–1704.

Niu J, Liu Z, Lu Y, Wen Z. 2022. Evidential combination of classifiers for imbalanced data. *IEEE Transactions on Systems* **12(12)**:7642–7653 DOI 10.1109/tsmc.2022.3162258.

Olga R, Jia D, Hao S, Jonathan K, Sanjeev S, Sean M, Huang ZH, Andrej K, Aditya K, Michael B, Berg AC, Fei-Fei L. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)* **115(3)**:211–252 DOI 10.1007/s11263-015-0816-y.

Pan H, Guo Y, Yu M, Chen J. 2024. Enhanced long-tailed recognition with contrastive cutmix augmentation. *IEEE Transactions on Image Processing* **33(4)**:4215–4230 DOI 10.1109/tip.2024.3425148.

Pang S, Wang W, Zhang R, Hao W. 2023. Hierarchical block aggregation network for long-tailed visual recognition. *Neurocomputing* **549(8)**:126463 DOI 10.1016/j.neucom.2023.126463.

Park S, Hong Y, Heo B, Yun S, Choi J. 2022. The majority can help the minority: context-rich minority oversampling for long-tailed classification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6887–6896.

Park S, Lim J, Jeon Y, Choi JY. 2021. Influence-balanced loss for imbalanced visual classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 735–744.

Ren J, Yu C, Ma X, Zhao H, Yi S. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems* **33**:4175–4186.

Ren M, Zeng W, Yang B, Urtasun R. 2018. Learning to reweight examples for robust deep learning. In: *International Conference on Machine Learning*. PMLR, 4334–4343.

Santiago C, Barata C, Sasdelli M, Carneiro G, Nascimento JC. 2021. LOW: training deep neural networks by learning optimal sample weights. *Pattern Recognition* **110(7553)**:107585 DOI 10.1016/j.patcog.2020.107585.

Shu J, Xie Q, Yi L, Zhao Q, Zhou S, Xu Z, Meng D. 2019. Meta-Weight-Net: learning an explicit mapping for sample weighting. In: *Advances in Neural Information Processing Systems.* Vol. 32. Vancouver, Canada: Curran Associates, Inc.

Tan J, Wang C, Li B, Li Q, Ouyang W, Yin C, Yan J. 2020. Equalization loss for long-tailed object recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11662–11671.

Tian H, Zhang Z, Martin A, Liu Z. 2022. Reliability-based imbalanced data classification with Dempster-Shafer theory. In: *International Conference on Belief Functions*. Vol. 115, No. 7. Cham: Springer International Publishing, 77–86 DOI 10.1007/978-3-031-17801-6_8.

Wan Q, Zhou B, Wang Y. 2024. BSCGAN: structured minority class image generation under class-balanced pretraining. *The Visual Computer* **41**:3853–3865 DOI 10.1007/s00371-024-03635-5.

Wang Y, Gan W, Yang J, Wu W, Yan J. 2019. Dynamic curriculum learning for imbalanced data classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5017–5026.

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

Xiao H, Rasul K, Vollgraf R. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. ArXiv DOI 10.48550/arXiv.1708.07747.

Yang Y, Xu Z. 2020. Rethinking the value of labels for improving class-imbalanced learning. *Advances in Neural Information Processing Systems* **33**:19290–19301.

Zeng Z, Li L, Zhao Z, Liu Q. 2024. Improved fine-grained image classification in few-shot learning based on channel-spatial attention and grouped bilinear convolution. *The Visual Computer* **41**:4129–4141 DOI 10.1007/s00371-024-03650-6.

Zhang S, Chen C, Zhang X, Peng S. 2022a. Label-occurrence-balanced mixup for long-tailed recognition. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 3224–3228.

Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. 2017. mixup: beyond empirical risk minimization. ArXiv DOI 10.48550/arXiv.1710.09412.

Zhang Y, Hooi B, Hong L, Feng J. 2022b. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in Neural Information Processing Systems* **35**:34077–34090.

Zhang Y, Kang B, Hooi B, Yan S, Feng J. 2023. Deep long-tailed learning: a survey. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang Y, Wei X, Zhou B, Wu J. 2021. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence* **35(4)**:3447–3455 DOI 10.1609/aaai.v35i4.16458.

Zhang Z, Zhang Y, Tian H, Martin A, Liu Z, Ding W. 2025. A survey of evidential clustering: definitions, methods, and applications. *Information Fusion* **115**:102736 DOI 10.1016/j.inffus.2024.102736.

Zhao Y, Chen S, Chen Q, Hu Z. 2023. Combining loss reweighting and sample resampling for long-tailed instance segmentation. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 1–5.

Zhao Y, He W, Zhao H. 2025. DPA-EI: long-tailed classification by dual progressive augmentation from explicit and implicit perspective. *Knowledge-Based Systems* **311(3)**:113061 DOI 10.1016/j.knosys.2025.113061.

Zhou B, Cui Q, Wei X, Chen Z. 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 T, Liu X, Zhu E. 2022. Oversampling with reliably expanding minority class regions for imbalanced data learning. *IEEE Transactions on Knowledge and Data Engineering* **35(6)**:6167–6181 DOI 10.1109/tkde.2022.3171706.