Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Haoji Hu, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature: Haoji Hu Date: 01/28/2023

List any pre-prints: NA

Relevant Conference publication(s) (submitted, accepted, or published): NA

Justification for re-publication: NA

Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- We complete the diffusion-based framework to alleviate the long-tailed problem.
- DiffRC, a novel diffusion-based data augmentation algorithm is proposed.
- Two loss items are devised to promote the performance of the diffusion model.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Unlocking the Power of Diffusion Probabilistic Models for Long-tailed Recognition via Data Synthesis

Siming Fu[a], Xiaoxuan He[a], Haoji Hu[a,**]

[a]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, 310027, China

Article history:
Received 28 January 2023

Keywords:
Diffusion model
Data synthesis
Long-tailed recognition
Data augmentation

ABSTRACT

Long-tail learning seeks to address the key issue of head classes dominating the learning process under extreme class imbalance in real-world circumstances. Data augmentation, which tries to pack a set of augmentation approaches to increase the size and quality of datasets for model training, has shown to be a worthwhile research topic. The long-tail problem cannot be solved using the current data augmentation techniques. The subject of how to undertake data augmentation for long-tailed learning more effectively is yet unanswered. The diffusion-based data augmentation method, referred to as DiffuRC, which enables the diffusion model to generate varied synthetic images for tail categories by translating generated samples along with various semantically meaningful directions, was proposed to address the aforementioned problems. To our knowledge, this is the first approach to use the diffusion model to add a specific type of semantic data augmentation to the unbalanced dataset. In addition to the long-tailed problem's inherent characteristics, we enhance the diversity of generation during training and change the sampling procedure to incorporate an extra guidance signal that directs the sampling process toward tail-category neighborhoods. Our approach significantly surpasses prior studies and achieves cutting-edge performance on long-tailed recognition.

1. Introduction

Real-world data tends to be long-tailed, with a small number of head classes contributing the vast majority of the data and the majority of tail classes making up relatively little of it. An undesired phenomenon is models (Cao et al., 2019) trained with long-tailed data perform better on head classes while exhibiting extremely low accuracy on tail ones. Numerous studies have been carried out in recent years to address this issue, producing encouraging advancements in the field of deep long-tailed learning. One of the solutions to the long-tail learning problem is to create a diversified sample of tail categories. This type of solution is more applicable in real-world applications because data augmentation is a reasonably simple technique that can be used to a number of long-tailed situations.

One of the widely accepted insights (Zhong et al., 2021a; Chou et al., 2020) relies on traditional data augmentation techniques. Another common practice (Zang et al., 2021; Li et al., 2021) is the transfer of knowledge from head classes to improve model performance on tail classes. However, learning high-quality representations with the available data augmentation techniques is not acceptable, which confounds the optimization strategy for better long-tailed learning(Zhang et al., 2021b). Given the flexibility, learnability, and sampling benefits of diffusion models, there have been various attempts to expand the methodology to produce high-quality images with significant diversity (Dhariwal and Nichol, 2021a). This inspires us to conditionally generate image synthetics for long-tailed learning using the diffusion model paradigm.

In this paper, we propose a novel diffusion-based framework called DiffRC, which enables estimating the class-conditional diverse and semantically meaningful directions to generate the semantic and diversified augmented samples for tail cate-

**Corresponding author
e-mail: fusiming@zju.edu.cn (Siming Fu), Xiaoxuan_He@zju.edu.cn (Xiaoxuan He), haoji_hu@zju.edu.cn (Haoji Hu)

gories. Our approach combines the pair-wise diversity diffusion training process providing greater generation diversity for tail classes, and the modified diffusion sampling process guided by the imbalanced latent space properties. Specifically, in the training process, we extend the pre-trained diffusion model to extract the data distribution of the majority class samples which are information-rich. Building on this foundation, we propose pair-wise diversity loss to promote the target diffusion model to achieve greater generation diversity for tail classes. To further ease the limitations of tail category likelihood estimates, we develop the calibration sampling algorithm based on the properties of latent embedding via the modified prototype contrastive loss. The newly generated images by diffusion models are sampled from the calibrated distribution and help to obtain the discriminative representation space by improving the performance of severely under-represented tail classes.

We extensively validate our model on benchmarks (CIFAR-10-LT (Krizhevsky et al., 2009), CIFAR-100-LT (Krizhevsky et al., 2009), and ImageNet-LT (Liu et al., 2019a). The experimental results indicate that our method consistently beats cutting-edge methods on all benchmarks. The contributions of this paper can be summarized as follows:

- To the best of our acknowledge, we are the first in long-tailed learning to complete the diffusion-based framework to generate the semantic and diversified image samples for tail categories.
- On top of the long-tailed learning's properties, we improve the diversity of generation in the training stage by pair-wise diversity loss and modify the sampling process via the guidance of modified prototype contrastive loss to steer it towards tail-category neighborhoods.
- Our method outperforms previous works by a large margin and achieves state-of-the-art performance on the long-tailed image classification task.

2. Related Work

Data Augmentation for Long-tailed Learning. Non-transfer augmentation and transfer-based augmentation techniques have both been extensively studied in long-tailed learning. In order to overcome long-tailed issues, non-transfer augmentation aims to enhance or create traditional data augmentation approaches. For long-tailed learning, Remix (Chou et al., 2020) and MiSLAS (Zhong et al., 2021a) proposed to use data mixup and presented a re-balanced mixup strategy to improve tail classes in particular. Transfer-based augmentation aims to improve model performance on tail classes by transferring knowledge from head classes. TailCalibX (Vigneswaran et al., 2021) and GLAG (Zhang and Xiang, 2022) explored a direction that attempted to generate meaningful features by estimating the tail category's distribution. RSG (Wang et al., 2021a) dynamically evaluated a set of feature centers and added features based on the latent embedding distance. However, the existing data enhancement methods are still unsatisfactory. It is worth exploring the high fidelity and diversity data synthesis method for solving the long-tailed problem.

Diffusion Model. The diffusion models are an emerging subject in computer vision that have produced amazing outcomes in generative modeling. The probabilistic diffusion model's original intent was to simulate a given distribution using random noise. Diffusion models have beaten GANs in the image creation work and obtained both greater sample quality and better distribution coverage due to their advantageous characteristics including steady training and simple scalability (Ho et al., 2020a). The groundbreaking work of diffusion models is Denoising Diffusion Implicit Models (DDPM) (Ho et al., 2020a). However, the sampling process needs to follow the Markov chain step by step to produce one sample, which is extremely slow (Ho et al., 2020a; Nichol and Dhariwal, 2021)). Denoising Diffusion Implicit Models (DDIM) (Song et al., 2020) accelerates the sampling process via an iterative non-Markovian way while keeping the same training process unchanged. Considering the high fidelity and diversity of the diffusion model, we adopt it to generate samples, which relieves the data imbalance.

3. DiffRC

3.1. Revisiting Diffusion Models

Denoising diffusion probabilistic models (DDPM) (Ho et al., 2020b) model the data distribution by reversing a gradual noising process. The forward diffusion process is defined as a Markov chain, which adds a small amount of noise to the sample in T steps, producing a sequence of noisy samples $x_1, ..., x_T$ (x_0 is an uncorrupted image). The noised versions are obtained according to the following Markov process:

$$p\left(x_t \mid x_{t-1}\right)=\mathcal{N}\left(x_t ; \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t \cdot I\right), \tag{1}$$

where $\{\beta_t\}_{i=1}^{T}$ is a predefined incremental variance schedule and I is the identity matrix having the same dimensions as the input image x_0. We can sample from $q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right)$ in a closed form:

$$\boldsymbol{x}_t=\sqrt{\bar{\alpha}_t} \boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t} \boldsymbol{z}, \tag{2}$$

where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $\alpha_t=1-\beta_t$ and $\overline{\alpha_t}:=\prod_{s=1}^{t} \alpha_s$.

During the reverse or sample process, we start from a sample $x_T \sim \mathcal{N}(0, \mathbf{I})$ and follow the reverse steps $p\left(x_{t-1} \mid x_t\right)=\mathcal{N}\left(x_{t-1} ; \mu\left(x_t, t\right), \Sigma\left(x_t, t\right)\right)$. We can model the process with a deep neural network which learns the Gaussian transition $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right)$. DDPM fixes the covariance to a constant value and rewrites the mean as a function of noise, predicting the noise component of a noisy sample x_t using the function $z_\theta\left(x_t, t\right)$. So a simple loss for training the diffusion model is given:

$$\mathcal{L}_{\text {base }}=E_{t, x_0, z}\left\|z-z_\theta\left(x_t, t\right)\right\|^2, \tag{3}$$

where θ is the diffusion model parameters. Furthermore, Dhariwal *et al*. (Dhariwal and Nichol, 2021b) propose a method for the conditional diffusion model, whose sample process is referred to as:

$$x_{t-1} \leftarrow \mathcal{N}\left(\mu+s \Sigma \nabla_{x_t} \log p_\phi\left(y \mid x_t\right), \Sigma\right), \tag{4}$$

where $\leftarrow$ is denoted as the sampling process. The diffusion model during sampling is guided by using gradients from a classifier $p_\phi\left(y \mid x_t\right)$ (x_t is the noisy sample and y is the label).

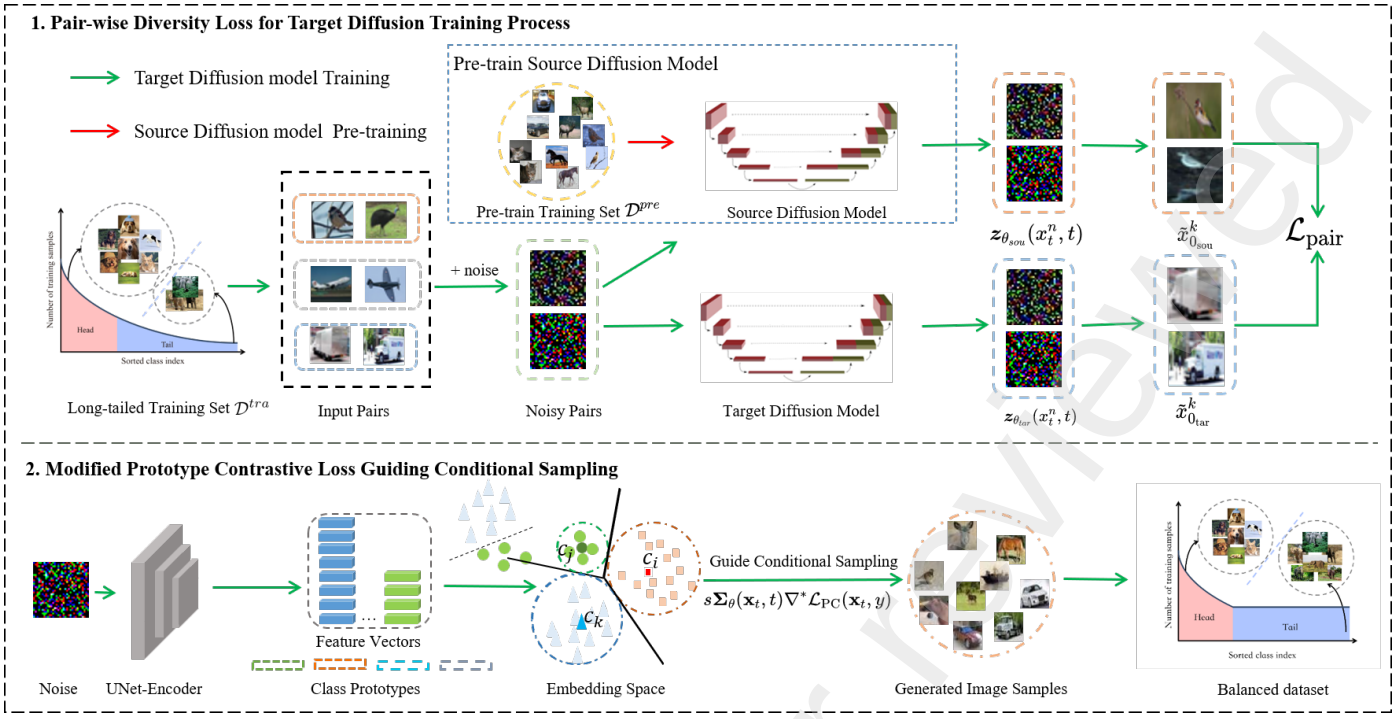


Fig. 1. Overview of the diffusion training and sampling phase of our proposed DiffRC. Upper box demonstrates the unconditional diffusion model, which consists of the following steps: training an unconditional source diffusion model with $\mathcal{L}_{\text{base}}$ on head classes, and training the target diffusion model with $\mathcal{L}_{\text{pair}}$ for enhancing the diversity of medium and tail classes. Bottom box introduces the modified prototype contrastive loss $\mathcal{L}_{\text{PC}}$ which guides the diffusion model to generate samples for the insufficient data. The generated samples ease the imbalance of the long-tailed training dataset.

3.2. Pair-wise Diversity Loss for Target Diffusion Training Process

Diffusion models tend to overfit seriously and fail to produce high-quality images with considerable diversity when training data of tail categories is limited. They generate reasonable images but still lack diversity and can only replicate the training samples for tail categories. For long-tailed learning, considering $\mathcal{D}^{tra} = \{\boldsymbol{x}^i, y^i\}$, $i \in \{1, \cdots, K\}$ be the training set, where $\boldsymbol{x}^i$ denotes an image sample and y^i indicates its class label. Let K be the total numbers of classes and N_i be the number of samples in class i, where $\sum_{i=1}^{K} N_i = N$. A long-tail setup can be defined by ordering the number of samples per category, i.e. $N_1 \geq N_2 \geq \ldots \geq N_K$ and $N_1 \gg N_K$ after sorting of N_i. To pre-train the source diffusion model, we conduct the pre-train dataset $\mathcal{D}^{pre} = \{\boldsymbol{x}^i, y^i\}$, $i \in \{1, \cdots, H\}$, where samples are from the head classes of $\mathcal{D}^{tra}$, and H is the numbers of head classes. As shown in Fig. 1, we pre-train the source diffusion model to extract the rich real distribution of head class samples as:

$$\mathcal{L}_{\text{base}} = E_{t,x_0^i,z}\left\|z - z_{\theta_{sou}}\left(x_t^i, t\right)\right\|^2, \quad \forall x_0^i \in \mathcal{D}^{pre}. \tag{5}$$

On top of the pre-trained diffusion model θ_{sou}, the target model θ_{tar} is initialized to it consistently. Then, we sample a batch of noised images by randomly adding Gaussian noises to enable the target diffusion model to maintain consistent diversity for the medium and tail classes. Furthermore, the source and target models are applied to predict the fully denoised images $x_0 \sim q(x_0)$. The prediction $\tilde{x}_0$ is defined as:

$$\tilde{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} x_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} z_\theta(x_t, t), \tag{6}$$

To measure the relative distances between the predicted samples $\tilde{x}_0^{k,i}$ and $\tilde{x}_0^{k,j} (0 \leq i, j \leq N_k)$ from the same category k, we propose the method to measure the probability distribution in the source and target models as follows:

$$p_i^{\text{sou}} = \text{Softmax}\left(\left\{s\left(\tilde{x}_{0_{\text{sou}}}^{k,i}, \tilde{x}_{0_{\text{sou}}}^{k,j}\right)\right\}_{\forall i \neq j}\right) \tag{7}$$

$$p_i^{\text{tar}} = \text{Softmax}\left(\frac{1}{1+\gamma^{N_k}}\left\{s\left(\tilde{x}_{0_{\text{tar}}}^{k,i}, \tilde{x}_{0_{\text{tar}}}^{k,j}\right)\right\}_{\forall i \neq j}\right) \tag{8}$$

where $s(\cdot)$ is the cosine similarity function, and N_k is the training sample number of class k. γ is the loss-related parameter, which is commonly set as 0.1 (Cui et al., 2019). Compared with pre-trained diffusion model θ_{sou}, the diversity degradation of target diffusion model θ_{tar} mainly comes from the shortened relative distances between generated samples. Formally, we design the pair-wise diversity loss as:

$$\begin{aligned}\mathcal{L}_{\text{pair}}\left(z_{\text{sou}}, z_{tar}\right) = &\mathbb{E}_{t,x_0,z}\sum_k \sum_i D_{\text{KL}}\left(p_i^{\text{tar}} \| p_i^{\text{sou}}\right), \\ &k \notin \{1, \cdots, H\}\end{aligned} \tag{9}$$

where D_{KL} represents the KL-divergence function. The $\mathcal{L}_{\text{pair}}$ enables the target diffusion model θ_{tar} to preserve the relative distances between generated samples during source and target diffusion models while achieving greater generation diversity for medium/tail classes. The overall training loss function:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda \mathcal{L}_{\text{pair}} \tag{10}$$

Table 1. Top 1 accuracy of CIFAR-10-LT / CIFAR-100-LT with various imbalance factors (100, 50, 10). RL, DT, and DA indicate representation learning, decouple training, and data augmentation, respectively.

Type	Method	CIFAR-10-LT			CIFAR-100-LT		
		100	50	10	100	50	10
Baseline	Softmax	70.5	74.8	86.3	38.3	43.9	55.7
RL	KCL (Kang et al., 2021)	77.6	81.7	88.0	42.8	46.3	57.6
	DRO-LT (Samuel and Chechik, 2021)	82.6	-	-	47.3	57.6	63.4
	TSC (Li et al., 2022)	79.7	82.9	88.7	43.8	47.4	59.0
	Hybrid-SC (Wang et al., 2021b)	78.8	83.8	90.0	46.7	51.9	63.1
DT	Decoupling (Kang et al., 2019)	79.8	82.2	88.3	43.3	47.4	57.9
	De-confound (Tang et al., 2020)	80.6	83.6	88.5	44.1	50.3	59.6
	MiSLAS (Zhong et al., 2021b)	82.1	85.7	90.0	47.0	52.3	63.2
	Bag of tricks (Zhang et al., 2021c)	71.9	78.4	-	47.8	51.7	-
DA	MetaSAug (Li et al., 2021)	79.5	84.0	89.5	48.1	52.3	61.3
	RSG (Wang et al., 2021a)	79.6	82.8	-	44.6	48.5	-
	GLAG (Zhang and Xiang, 2022)	-	-	-	51.7	55.3	64.5
Ours	Baseline + DiffRC	82.8	86.1	90.3	52.3	56.0	65.2

The proposed pair-wise diversity loss $\mathcal{L}_{\text{pair}}$ is added to relieve overfitting and preserve generation diversity during adaptation when training data is limited.

3.3. Modified Prototype Contrastive Loss Guiding the Tail-category Sampling

Most of the previous work on long-tail recognition was on improving the feature space because the latent embedding indicates high-level semantic information which is vital to reconstruct the image samples. We shift our attention to leverage the relationship of latent embedding to incorporate the guiding signals in the sampling process. Supervised contrastive learning has been proven to be an effective method for relieving the data imbalance in long-tailed recognition, which introduces cluster-based prototypes and encourages embeddings to gather around their corresponding prototypes. Motivate by this, we construct prototypes for learning a better feature space. Our original feature prototypes follow the MoPro (Li et al., 2020), adopting the exponential-moving-average (EMA) algorithm during training:

$$\boldsymbol{c}_k \leftarrow m\boldsymbol{c}_k + (1-m)f(x_t^{k,i}), \quad \forall i \in \{0, \cdots, N_K\}, \tag{11}$$

where $\boldsymbol{c}_k$ is the prototype for class k and m is the momentum coefficient, usually set as 0.999. Leveraging the properties of the tail category features, we devise the modified prototype contrastive loss to focus on marking them:

$$\mathcal{L}_{\text{PC}} = -\log\left[\frac{\pi_k \exp\left(f(x_t^{k,i}) \cdot \boldsymbol{c}^i/\tau\right)}{\sum_{j=1}^{K} \pi_j \exp\left(f(x_t^{k,i}) \cdot \boldsymbol{c}^j/\tau\right)}\right], \tag{12}$$

where π denotes the vector of sample frequencies ($\pi_k = N_k/N$), and τ is the temperature, setting as 0.07 in our work. We cooperate the modified prototype contrastive loss function with our sampling strategy. We modify the sampling process as follows:

$$\begin{aligned}\mathbf{x}_{t-1} = \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) + \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\,\mathbf{z} \\ + s\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\,\nabla^* \mathcal{L}_{\text{PC}}(\mathbf{x}_t, y)\end{aligned} \tag{13}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, ∇^* refer to normalized gradients, and s is a scaling hyper-parameter. The guidance term $s\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\,\nabla^* \mathcal{L}_{\text{PC}}(\mathbf{x}_t, y)$ dominates the Gaussian transition term from the diffusion model and steers the sampling process off data-manifold. With the modification, the sampling process is guided to generate samples belonging to the medium and tail classes, which are used to remedy the imbalance.

4. Experiments

4.1. Experimental Settings

4.1.1. Datasets and Setup

We conduct experiments on long-tailed image classification datasets, including the CIFAR-10-LT / CIFAR-100-LT (Krizhevsky et al., 2009), and ImageNet-LT (Liu et al., 2019a).

- **CIFAR-10-LT / CIFAR-100-LT** is based on the original CIFAR-10 / CIFAR-100 dataset, whose training samples per class are constructed by imbalance ratio (the imbalance ratios we adopt in our experiment are 10, 50 and 100).
- **ImageNet-LT** is a long-tailed version of the dataset by sampling a portion of the ImageNet dataset using the Pareto distribution and a power value of 6. With class cardinality ranging from 5 to 1,280, it has 115.8K images spanning 1,000 categories.

4.1.2. Implementation Details

Following (Ren et al., 2020), we employ ResNet-32 (He et al., 2016) as the feature extractor for all techniques on CIFAR-10-LT / CIFAR-100-LT, respectively. We train the SGD optimizer with a batch size of 256 and a momentum of 0.9 for three imbalance ratios (10, 50 and 100). We employ ResNetXt-50 (Xie et al., 2016) as the feature extractor for each method on ImageNet-LT. Based on batch size 512 and momentum 0.9,

Table 2. Results for ImageNet-LT under 90 and 200 training epochs in terms of accuracy (Acc). Class re-balancing, decoupled training, and representation learning are denoted in this table by the letters CR, DT, and RL, respectively.

Type	Method	90 epochs				200 epochs			
		Many	Med.	Few	All	Many	Med.	Few	All
Baseline	Softmax	66.5	39.0	8.6	45.5	66.9	40.4	12.6	46.8
CR	BALMS (Ren et al., 2020)	61.7	48.0	29.9	50.8	62.4	47.7	32.1	51.2
	LDAM (Cao et al., 2019)	62.3	47.4	32.5	51.1	60.0	49.2	31.9	51.1
	LADE (Hong et al., 2021)	62.2	48.6	31.8	51.5	63.1	47.7	32.7	51.6
	DisAlign (Zhang et al., 2021a)	62.7	52.1	31.4	53.4	-	-	-	-
DT	Decoupling (Kang et al., 2019)	62.4	39.3	14.9	44.9	60.9	36.9	13.5	43.0
	MiSLAS (Zhong et al., 2021b)	62.1	48.9	31.6	51.4	65.3	50.6	33.0	53.4
	De-confound (Tang et al., 2020)	63.0	48.5	31.4	51.8	64.9	46.9	28.1	51.3
	xERM_{TDE} (Zhu et al., 2022)	-	-	-	-	68.6	50.0	27.5	54.1
RL	OLTR (Liu et al., 2019b)	58.2	45.5	19.5	46.7	62.9	44.6	18.8	48.0
	DRO-LT (Samuel and Chechik, 2021)	-	-	-	-	64.0	49.8	33.1	53.5
	PaCo (Cui et al., 2021)	59.7	51.7	36.6	52.7	63.2	51.6	39.2	54.4
DA	RSG (Wang et al., 2021a)	68.7	43.7	16.2	49.6	65.0	49.4	31.1	52.9
	SSP (Yang and Xu, 2020)	65.6	49.6	30.3	53.1	67.3	49.1	28.3	53.3
Ours	Baseline + DiffRC	64.3	52.1	32.5	54.1	66.4	53.4	35.9	55.4

we do training with the SGD optimizer. A cosine scheduler (Loshchilov and Hutter, 2016) decays the learning rate in both training epochs (90 and 200 training epochs) from 0.2 to 0.0.

4.1.3. Training and Sampling Settings of Diffusion Model

For the diffusion model (Dhariwal and Nichol, 2021b), we employ a U-Net-based architecture with adaptive group normalization. We take into account the U-Net encoder when designing the classifier architecture. The diffusion process timestep is a criterion for the classifier and diffusion model. For the diffusion process, $T = 1000$ is taken into account. We sample with 250 timesteps because it expedites the procedure while barely degrading the image quality.

4.2. Comparison to State-of-the-art Methods

4.2.1. Experimental Results on CIFAR-10-LT / CIFAR-100-LT

As shown in Tab. 1, to prove the versatility of our method, we employ our method on the CIFAR-10-LT / CIFAR-100-LT dataset with three imbalance ratios. We compare against the most relevant methods and choose methods that are recently published and representative of different types, such as class re-balancing, decouple training and data augmentation. Our method surpasses the DRO-LT (Samuel and Chechik, 2021) under various imbalance factors, especially on the largest imbalance factor (82.8% vs 82.6%) and (52.3% vs 47.3%) on CIFAR-10-LT / CIFAR-100-LT, respectively. Furthermore, compared with the data augmentation methods GLAG (Zhang and Xiang, 2022), our model achieves competitive performance (52.3% vs 51.7% with 100 imbalance factor) on CIFAR-100-LT.

4.2.2. Experimental Results on ImageNet-LT

The long-tailed results on ImageNet-LT are displayed in Tab. 2. To establish a fair comparison, we use the performance data from the deep long-tailed survey (Zhang et al., 2021b) for several approaches at 90 and 200 training epochs, respectively. Our strategy surpasses state-of-the-art methods by a wide margin, achieving overall accuracy of 54.1 and 55.4 percent. Our approach outperforms representation learning technique SSP by 1.0% (54.1% vs 53.1%) at 90 training epochs and by 2.1% (55.4% vs 53.3%) at 200 training epochs. In addition, during 90 and 200 training epochs, respectively, our technique outperforms PaCo by 1.4% (54.1% vs. 52.7%) and 1.0% (55.4% vs. 54.4%).

4.3. Discussion

4.3.1. Visualization of the Generated Images

We randomly choose the generated images for each original image. It is worth noting that in contrast to the homogenized image samples generated by the original diffusion model, our method is able to generate context-rich minority samples that have diverse contexts. For example, while the original 'black swan' contains two swans on the lake, our diffusion model generates kinds of swans that swim in the water, demonstrating that they are able to preserve the information from the condition image. As shown in Fig. 2, our method DiffRC can generate results that are semantically consistent with the original input samples in high perceptual quality. Moreover, our DiffRC can transform the shape, size or pose automatically to fit the original simples. For example, our model generates diverse images of 'cheetah' for tail classes, which contain various background and foreground information. Overall, our method can generate results that are semantically consistent with the original input images in high perceptual quality.

Table 3. Ablation study of the proposed DiffRC when optimizing different terms with respect to performance on ImageNet-LT. The components contain pair-wise diversity loss (PDL) and modified prototype contrastive loss (MPCL). The "baseline" means ResNetXt-50 trained with CrossEnntropy loss by 90 epochs. To assess imbalance, the accuracy of the three groups' standard deviations (Std) (Jiang et al., 2021) is used. ↓ means the metric is the lower the better.

Baseline	PDL	MPCL	Many	Med.	Few	Std (↓) (imbalancedness)	All
✓	✗	✗	66.5	39.0	8.6	23.6	45.5
✓	✓	✗	65.2	49.3	26.9	15.7	52.7
✓	✗	✓	63.7	51.3	31.1	13.4	53.3
✓	✓	✓	64.3	52.1	32.5	13.0	54.1

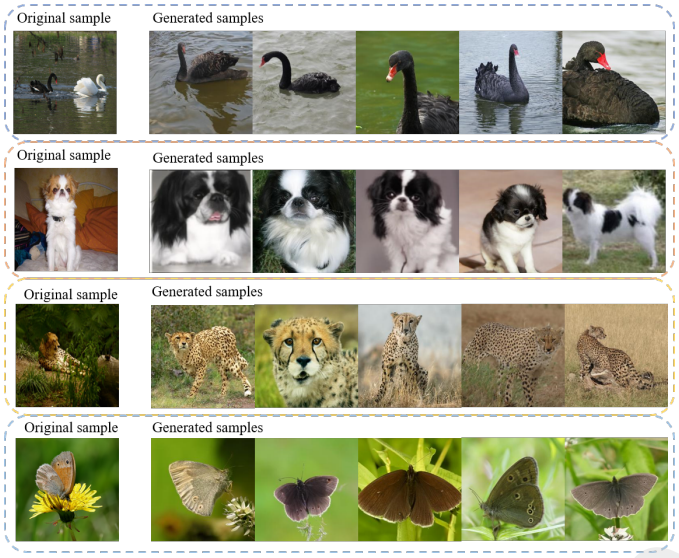
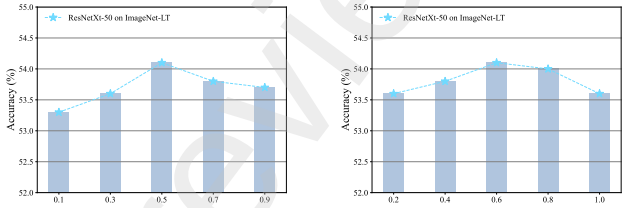


Fig. 2. Visualization of the original image and generated images from the same tail classes in ImageNet-LT (black swan, Japanese spaniel, cheetah, ringlet). The plot on the left is the original sample, and the plot on the right is the generated sample.



(a) λ in Eq. 10 (b) α in Eq. 13

Fig. 3. Ablation study on λ in Eq. 10 and s in Eq. 13.

4.3.2. Ablation Study

λ in overall training loss. The λ in Eq. 10, which controls the amount of adjustment in overall training loss, is a significant hyper-parameter in our approach. We investigate how sensitive the accuracy is to λ values. We set the hyper-parameter $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The impact of the trade-off parameter λ on validation accuracy is quantified in Fig. 3(a). It demonstrates that the best outcomes are obtained by combining the $\mathcal{L}_{pair}$ and $\mathcal{L}_{base}$ with an ideal λ of 0.5.

s in image sampling generation. In Eq. 13, we introduce a a scaling hyper-parameter s which controls the degree of distribution calibration. We initialize s to 0.2 for each tail class and it changes adaptively during sampling. In order to do ablation experiments, we placed the hyper-parameters in the range of 0.2 to 1 with a stride of 0.2 as shown in Fig. 3(b). Overall, the larger s means more confidence to modify the original distribution of sampling. The optimal α for ImageNet-LT is 0.6.

Effectiveness of PDL and MPCL. Tab. 3 confirms the crucial functions of our adaptive modules for modified prototype contrastive loss (PDL) and pair-wise diversity loss (MPCL). The baseline only performs decoupled training pipelines without using any components of our methods. Firstly, when we adopt PDL on the baseline, it significantly surpasses the performance over the baseline (52.7% vs 45.5%). Moreover, in the sampling process, our MPCL module further boosts the performance, especially in the tail classes (53.3% vs 45.5%). In addition, with our proposed PDL and MPCL, the Std of the three groups is considerably lower, which indicates that the balance between categories is considerably enhanced. The outcomes point to the efficiency of both the PDL and MPCL components in enhancing long-tailed learning performance.

5. Conclusion

In this paper, we have proposed a novel diffusion-based framework to tackle the long-tail challenge, denoted as DiffRC. Our framework includes pair-wise diversity loss for the target diffusion training process and modified prototype contrastive loss guiding the tail-category sampling. The first module of our method completes the pair-wise diversity loss to guide the target diffusion model to preserve the relative distances between samples in the medium/tail classes and keep similar distributions with the head class. The second module leverages the property of the latent prototype to incorporate the guiding signals in the sampling process. The generated samples are not simplified because of the small amount of samples in medium and tail classes, but have a certain diversity. Then, the sampled images ease the dominance of the head classes in classification decisions. The experimental outcomes demonstrate that our approach achieves cutting-edge performance in a variety of long-tailed learning settings.

References

Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T., 2019. Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems 32.

Chou, H.P., Chang, S.C., Pan, J.Y., Wei, W., Juan, D.C., 2020. Remix: rebalanced mixup, in: European Conference on Computer Vision, Springer. pp. 95–110.

Cui, J., Zhong, Z., Liu, S., Yu, B., Jia, J., 2021. Parametric contrastive learning. arXiv:2107.12028.

Cui, Y., Jia, M., Lin, T.Y., 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, pp. 9268--9277.

Dhariwal, P., Nichol, A., 2021a. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems 34, 8780--8794.

Dhariwal, P., Nichol, A., 2021b. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems 34, 8780--8794.

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, pp. 770--778.

Ho, J., Jain, A., Abbeel, P., 2020a. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems 33, 6840--6851.

Ho, J., Jain, A., Abbeel, P., 2020b. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems 33, 6840--6851.

Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., Chang, B., 2021. Disentangling label distribution for long-tailed visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6626--6636.

Jiang, Z., Chen, T., Mortazavi, B.J., Wang, Z., 2021. Self-damaging contrastive learning, in: International Conference on Machine Learning, PMLR. pp. 4927--4939.

Kang, B., Li, Y., Xie, S., Yuan, Z., Feng, J., 2021. 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., Kalantidis, Y., 2019. Decoupling representation and classifier for long-tailed recognition .

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

Li, J., Xiong, C., Hoi, S.C., 2020. Mopro: Webly supervised learning with momentum prototypes. arXiv preprint arXiv:2009.07995 .

Li, S., Gong, K., Liu, C.H., 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, pp. 5212--5221.

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

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X., 2019a. Large-scale long-tailed recognition in an open world, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X., 2019b. Large-scale long-tailed recognition in an open world, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2537--2546.

Loshchilov, I., Hutter, F., 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 .

Nichol, A.Q., Dhariwal, P., 2021. Improved denoising diffusion probabilistic models, in: International Conference on Machine Learning, PMLR. pp. 8162--8171.

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

Samuel, D., Chechik, G., 2021. Distributional robustness loss for long-tail learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Song, J., Meng, C., Ermon, S., 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 .

Tang, K., Huang, J., Zhang, H., 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. Advances in Neural Information Processing Systems 33, 1513--1524.

Vigneswaran, R., Law, M.T., Balasubramanian, V.N., Tapaswi, M., 2021. Feature generation for long-tail classification, in: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 1--9.

Wang, J., Lukasiewicz, T., Hu, X., Cai, J., Xu, Z., 2021a. Rsg: A simple but effective module for learning imbalanced datasets, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3784--3793.

Wang, P., Han, K., Wei, X.S., Zhang, L., Wang, L., 2021b. Contrastive learning based hybrid networks for long-tailed image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 943--952.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2016. Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431 .

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

Zang, Y., Huang, C., Loy, C.C., 2021. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3457--3466.

Zhang, S., Li, Z., Yan, S., He, X., Sun, J., 2021a. Distribution alignment: A unified framework for long-tail visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2361--2370.

Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J., 2021b. Deep long-tailed learning: A survey. arXiv preprint arXiv:2110.04596 .

Zhang, Y., Wei, X.S., Zhou, B., Wu, J., 2021c. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks, in: Proceedings of the AAAI conference on artificial intelligence, pp. 3447--3455.

Zhang, Z., Xiang, X., 2022. Long-tailed classification with gradual balanced loss and adaptive feature generation.

Zhong, Z., Cui, J., Liu, S., Jia, J., 2021a. Improving calibration for long-tailed recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16489--16498.

Zhong, Z., Cui, J., Liu, S., Jia, J., 2021b. Improving calibration for long-tailed recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16489--16498.

Zhu, B., Niu, Y., Hua, X.S., Zhang, H., 2022. Cross-domain empirical risk minimization for unbiased long-tailed classification, in: AAAI Conference on Artificial Intelligence.