Latent-based Diffusion Model for Long-tailed Recognition

Pengxiao Han[1], Changkun Ye[1,2], Jieming Zhou[1,2], Jing Zhang[1], Jie Hong[1,2], Xuesong Li[1,2*]
[1]Australian National University, [2]CSIRO, Australia
{pengxiao.han, changkun.ye, jieming.zhou, jing.zhang, jie.hong, xuesong.li}@anu.edu.au

Abstract

Long-tailed imbalance distribution is a common issue in practical computer vision applications. Previous works proposed methods to address this problem, which can be categorized into several classes: re-sampling, re-weighting, transfer learning, and feature augmentation. In recent years, diffusion models have shown an impressive generation ability in many sub-problems of deep computer vision. However, its powerful generation has not been explored in long-tailed problems. We propose a new approach, the Latent-based Diffusion Model for Long-tailed Recognition (LDMLR), as a feature augmentation method to tackle the issue. First, we encode the imbalanced dataset into features using the baseline model. Then, we train a Denoising Diffusion Implicit Model (DDIM) using these encoded features to generate pseudo-features. Finally, we train the classifier using the encoded and pseudo-features from the previous two steps. The model's accuracy shows an improvement on the CIFAR-LT and ImageNet-LT datasets by using the proposed method.

1. Introduction

Long-tailed recognition is a crucial task in deep computer vision because the imbalanced data distributions are close to real-world applications [4, 21, 24]. In many real-world datasets which have limited-labelled data, some classes of data have many samples, while others have few. In healthcare, many disease instances follow a long-tailed distribution [4, 11, 42, 46]. As for fraud detection [24, 36, 53], the number of samples of fraudulent transactions is much smaller than that of legitimate ones. Addressing long-tailed recognition could enhance the robustness of deep-learning visual models on real-world applications. However, long-tailed recognition is challenging since deep neural networks are more prone to overfitting the majority class while damaging the prediction accuracy for minority classes.

So far, extensive research has been conducted to address dataset imbalance issues [2, 7, 22, 27, 52]. Class-sensitive learning effectively solves the long-tailed distribution problem [2, 5, 22, 52]. Class-sensitive learning addresses this issue by re-adjusting the training loss for different classes. Logit adjustment methods adjust the prediction logits based on label frequencies [12, 17, 25, 40, 47]. However, class-sensitive learning and logit adjustment methods rely too heavily on training label frequencies. There are also representation learning methods, such as prototype learning [23, 50], metric learning [9, 49], and sequential learning [27, 50]. These methods use the feature representation of each class to make the distinction between classes more significant. Although these methods can effectively improve the accuracy of deep learning networks on long-tailed distributed datasets, they often need to be carefully designed and have limited improvements in tail classes.

Data or feature augmentation is another effective solution in long-tailed recognition [37, 44]. Some augmentation methods attempt to transfer the knowledge from head classes to tail to augment the samples of tail classes [3, 15, 44]. Others use over-sampling and under-sampling to re-balance the datasets [37]. Additionally, it is natural to come that using a high-quality generative model to augment a long-tailed distributed dataset might improve the performance. The generative models have been well developed [6, 10, 38], and variational autoencoder (VAE) and generative adversarial network (GAN) have been applied in the long-tailed problems [1, 41]. In recent years, diffusion model series have shown superior ability than other generative models [10, 26, 33, 39]. However, the application of such powerful generative models is underexplored in long-tailed problems.

Inspired by the observation above, our work attempts to leverage the diffusion model for feature augmentation to address long-tailed distribution recognition. We propose the approach named Latent-based Diffusion Model for Long-tailed Recognition (LDMLR). Specifically, we first train a baseline on the long-tailed data and obtain the encoded features. Then, we use a Denoising Diffusion Implicit Model (DDIM) model to generate pseudo-features in the latent space to augment the long-tailed training dataset.

*Corresponding author

The augmentation in latent space reduces the computational cost and speeds up the training process. Finally, we use both the encoded and pseudo-features to train the classifier of LDMLR to predict the long-tailed data. The proposed LDMLR has been validated on CIFAR-LT [19, 23] and ImageNet-LT [23, 34]. The experimental results demonstrate that our method is beneficial for long-tailed recognition. Our contribution can be summarized as below:

- Our method applies the diffusion model to enrich the feature embeddings for the long-tailed problem, offering a new solution to this challenging problem. To the best of my knowledge, we are the first to explore the capability of the diffusion model in the long-tailed recognition problem.
- When using the diffusion model, we propose to do the augmentation in the latent space instead of the image space, which reduces the computational cost and speeds up the training process.
- The experiments demonstrate that LDMLR has improved the performance of long-tailed recognition tasks on different datasets using various baselines. We achieve the essential improvements over the baselines.

2. Related Works

2.1. Long-tailed Recognition

Due to the class imbalance within datasets, long-tailed recognition is a challenging problem. A neural network trained on long-tailed datasets is prone to be biased towards head (or majority) classes with enough training data, resulting in poor performance on tail (or minority) classes. Class-sensitive learning aims to address the class imbalance problem by readjusting the traditional softmax cross-entropy loss. Traditional cross-entropy tends to provide more gradients to head classes, while tail classes receive fewer gradients. To ensure that each class has a balanced impact on the neural network during training, class-sensitive learning proposes adjusting the training loss weights for each class based on given training label frequencies, such as Focal loss [22] and Label-distribution-aware-margin (LDAM) loss [2].

Logit adjustment is another method for addressing the recognition of imbalanced datasets. This approach shares similarities with class-sensitive learning, where most of both methods require training class frequencies to rebalance the influence of head and tail classes on the model. For example, [25] modifies logit adjustment based on label frequencies, which can be implemented by posthoc or enforcement of a large relative margin between the logits of tail versus head labels. Label Shift Compensation (LSC) [40, 43, 47] is another type of logit adjustment method. Most logit adjustment methods require training label frequencies, even when not required. Some logit adjustment methods introduce an additional model, which makes the entire process slower and more complex. Representation learning involves bringing images into the feature space to learn discrimination classes. WCDAS [7] is a representation learning-based method that incorporates data-wise Gaussian-based kernels into the angular correlation between feature representation and classifier weights. In addition to class-sensitive learning and logit adjustment, there is a method that utilizes knowledge distillation to address the long-tailed recognition problem like Self-Supervision to Distillation (SSD) [20].

2.2. Generative Models for Feature Augmentation in Long-tailed Recognition

Many recent works demonstrate that feature or sample augmentation can effectively address the long-tailed recognition problem [1, 13, 41, 45]. By enriching the features or samples for the tail class, this approach is highly compatible and can easily be combined with normal baseline models. Using a powerful generative model to do augmentation can effectively diversify an imbalanced dataset [1, 41]. MFC-GAN [1] proposes a conditional GAN with fake class labels to generate a small number of minority class instances, thereby re-balancing the entire dataset. This method performs data augmentation at the image level, which is feasible to apply to large-scale image datasets. IDA-GAN [41] also uses GANs for data augmentation. First, they train a VAE encoder to model the training set distribution in the latent space. Then, they use GANs to generate tail class images to re-balance the long-tailed distributed dataset.

It is known that the diffusion models [10, 26, 33, 39] are relatively underexplored in long-tailed recognition. Since the diffusion model is a probabilistic generative model, it can generate more diverse samples than GANs, which might be more effective for solving long-tailed recognition. Moreover, it has a better ability to generate high-quality samples than other generative methods. As such, in this paper, we propose a diffusion-based augmentation method, LDMLR. We hope to exploit the excellent ability of the diffusion model for the long-tailed problem.

2.3. Diffusion Model

Due to the remarkable generative results of diffusion models, they have been a recent emerging topic in computer vision [10, 26, 29, 31, 33, 35, 39]. As a likelihood-based generative model, a diffusion model can produce exact likelihood computation, showing a powerful generation ability. DDIM [39] removes the Markov chain constraint from Denoising Diffusion Probabilistic Models (DDPM) [10], which allows fewer steps to accelerate the sampling process. DDIM significantly speeds up the sampling process without damaging the generative quality. Despite DDIM's acceleration of the sampling speed, the training speed of a

diffusion model is still considerably slow, which prevents its widespread adoption for large-scale images. The Latent Diffusion Model (LDM) [33] perfectly addresses this issue. LDM uses a VAE encoder to compress large-scale images into latent features that sequentially are used as a feature dataset for training a diffusion model in latent space.

Our goal in using data augmentation for the long-tailed recognition task is to generate diverse samples, which might enrich the low-density region of the category distribution. In order to reduce the model complexity and the training time, we choose to augment features in the latent space. It is natural to use LDM. However, LDM is specifically designed for large-scale, high-resolution images. In our method, the diffusion model is used to generate low-dimensional features. Therefore, we propose to modify DDIM for augmenting features in the latent space.

3. Approach

We propose a three-stage model called LDMLR for the long-tailed classification problem, as shown in Figure 1. We first train a neural network model on the long-tailed dataset and extract ground truth encoded features of the ground truth images. A diffusion model is then trained to generate pseudo-features of each class. Finally, the classifier is fine-tuned with the encoded and pseudo-features. The algorithm is also described in Algorithm 1.

3.1. Preliminaries

Notations: Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the image space, $\mathcal{Y} = \{1, 2, ..., K\}$ be the label space and $\mathcal{Z} \subseteq \mathbb{R}^c$ be the feature space. In the training time, a long-tailed dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ is available, with $(x_i, y_i) \sim_{i.i.d} P(x, y)$ drawn from the data distribution $P(x, y)$.

Diffusion Model: The diffusion model is a type of generative model. A typical diffusion model consists of two stages – a forward process that gradually adds noise to a clean sample and a backward process that gradually recovers a clean sample from noise. This paper mainly considers a special type of diffusion model named DDIM [39].

In the forward process, given a clean sample $x_0 \sim q(x_0)$, a series of noise samples $x_1, x_2, ..., x_T$ are generated by:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t} \cdot x_t, (1 - \alpha_t)\mathbf{I}), \quad (1)$$

for $\forall t \in \{1, 2, ..., T\}$, where T is the number of diffusion steps, $1 - \alpha_t$ is the noise variance at step t and $\mathbf{I}$ is the identity matrix with the same dimension as x_0. It is expected that $q(x_T|x_0) \approx \mathcal{N}(0, \mathbf{I})$ is close to the Gaussian noise.

In the backward process, starting from $x_T \sim \mathcal{N}(0, \mathbf{I})$, DDIM aims to recover the clean sample x_0 and $x_{\tau_1}, ..., x_{\tau_s}$, where $\{\tau_1, ..., \tau_s\} \subseteq \{1, 2, ..., T\}$ is a subset of the forward steps. The backward process satisfies:

$$p(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \mu(x_t, x_0, t), \Sigma(x_t, x_0, t)), \quad (2)$$

where:

$$\mu(x_t, x_0, t) = \sqrt{\alpha_{t-1}}x_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1 - \alpha_t}}, \quad (3)$$

$$\Sigma(x_t, x_0, t) = \sigma_t^2 \mathbf{I}. \quad (4)$$

While $\mu_\theta(x_t, x_0)$ can be modeled directly by a Neural Network, a recent study suggests that modeling the noise term $\epsilon_\theta(x_t, t)$ instead leads to better performance [10]. In this case, the backward step then satisfies:

$$\begin{aligned} x_{t-1} =& \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1 - \alpha_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) \\ &+ \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(x_t, t) + \sigma_t \epsilon_t, \end{aligned} \quad (5)$$

where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$ is an independent random noise.

After creating noise samples $x_1, x_2, ..., x_T$ in the forward process and learning $\epsilon_\theta(x_t, t)$ in the backward process, the diffusion model is able to generate data samples from Gaussian noises based on Equation 5.

3.2. Stage 1: Image Encoding

In order to augment image features with a latent diffusion model, the first step of LDMLR is to learn a neural network feature extractor $\mathcal{E} : \mathcal{X} \to \mathcal{Z}$ and obtain decent feature representations of the images in the dataset $\mathcal{D}$ (see Figure 1 (a)).

Since $\mathcal{D}$ is a labelled dataset, we propose to construct a soft classifier $f : \mathcal{X} \to \Delta^{K-1}$ based on the feature extractor $\mathcal{E}$ by adding a classification head $\mathcal{G} : \mathcal{Z} \to \Delta^{K-1}$ over $\mathcal{E}$:

$$f(x) = \mathcal{G}(\mathcal{E}(x)), \quad (6)$$

where Δ^{K-1} is the space of softmax predictions.

The feature extractor $\mathcal{E}$ can then be jointly trained with $\mathcal{G}$ over the dataset $\mathcal{D}$ with a cross entropy loss:

$$\mathcal{L}_{CE} := -\mathbb{E}_{(x,y)} \log f(x)_y, \quad (7)$$

where $(x, y) \sim p(x, y)$ follows train set distribution.

After training the classifier f, the encoder $\mathcal{E}$ can output features using ground truth images from the dataset $\mathcal{D}$. The set of the labelled encoded features is denoted as:

$$\mathcal{D}_z = \{(z_i, y_i)|z_i = \mathcal{E}(x_i), (x_i, y_i) \in \mathcal{D}\}. \quad (8)$$

With $\mathcal{D}_z$ available, we can train the latent diffusion model in the next stage to generate pseudo-features.

3.3. Stage 2: Representation Generation

As shown in Figure 1 (b), in the second stage of LDMLR, we propose to train a class-conditional latent diffusion

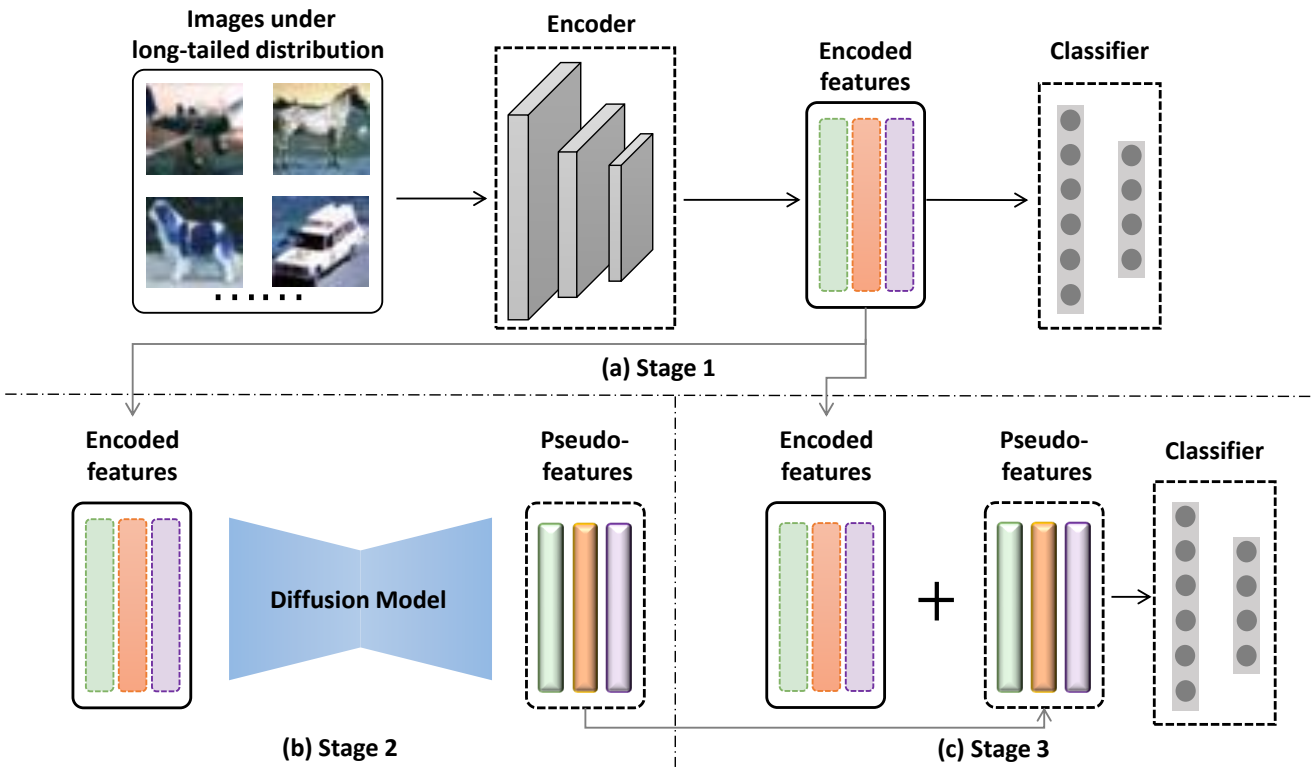

Figure 1. Overview of the proposed framework, LDMLR. The figure describes the training of the framework: (a) obtain encoded features by a pre-training convolutional neural network on the long-tailed training set, (b) Generate pseudo-features by the diffusion model using encoded features, and (c) Train the fully connected layers using encoded and pseudo-features. The encoder from (a) and the classifier from (c) are used to predict long-tailed data in the evaluation stage.

model (LDM) to generate pseudo-feature representations for different classes. The latent diffusion model has several advantages over image generation diffusion models: 1) LDM operates on one-dimensional image features instead of the original high-dimensional images. Therefore, LDM is more efficient than the image generation diffusion model in terms of training and inference time; 2) Standard diffusion model trained on a long-tailed dataset can suffer from low diversity and fidelity problems due to insufficient images in tail classes [30, 48], while LDM may suffer less from this problem because the data has lower dimensionality.

In the proposed LDMLR, we adopt the DDIM [39] diffusion model approach to train the model with the encoded $\mathcal{D}_z$ and generate pseudo-features. In the forward process, following Equation 1, an encoded feature $z_0 \in \mathcal{D}_z$ is perturbed with Gaussian noise to create $z_1, ..., z_T$ with:

$$q(z_t|z_{t-1}) := \mathcal{N}(z_t; \sqrt{\alpha_t} \cdot z_t, (1-\alpha_t)\mathbf{I}). \tag{9}$$

In the backward process, we propose to train a class-conditional neural network model $\epsilon_\theta(z_t, t, y)$ to approximate the Gaussian noise $\epsilon(z_t, t)$. The input of the neural network model includes the noisy feature z_t and the condition embedding determined by the step t and label $y \in \mathcal{Y}$.

The neural network model is trained with an MSE loss that minimizes the L2 distance between the predicted noise and the ground truth noise, which is defined as:

$$\mathcal{L}_{LDM} := \mathbb{E}_{z,y,\epsilon,\tau}\left[\|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2\right], \tag{10}$$

where $z = \mathcal{E}(x), (x, y) \sim p(x, y), \epsilon \sim \mathcal{N}(0, \mathbf{I})$ and t is drawn uniformly from $\{1, 2, ..., T\}$.

After training the LDM, we can use $\epsilon_\theta(z_t, t, y)$ to generate pseudo-feature representations from $z_T \sim \mathcal{N}(0, \mathbf{I})$ for each class $y \in \mathcal{Y}$ based on the Equation 5:

$$\begin{aligned}\hat{z}_{t-1} =& \sqrt{\alpha_{t-1}}\left(\frac{\hat{z}_t - \sqrt{1-\alpha_t}\cdot\epsilon_\theta(\hat{z}_t, t, y)}{\sqrt{\alpha_t}}\right) \\ &+ \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(\hat{z}_t, t, y) + \sigma_t\epsilon_t,\end{aligned} \tag{11}$$

These labelled pseudo-features are then collected as:

$$\mathcal{D}_{\hat{z}} = \{(\hat{z}_i, y_i)|y_i \sim p(y), \hat{z}_i \text{ generated by LDM}\} \tag{12}$$

where $p(y)$ can be different distributions on demand. For example, we can choose $p(y)$ to be none zero on tail classes so that the LDM only generates tail class features.

3.4. Stage 3: Classifier Training

In the final stage of LDMLR (see Figure 1 (c)), we fine-tune the classification head $\mathcal{G}$ with labelled encoded feature $\mathcal{D}_z$ obtained from stage 1 and labelled pseudo-feature $\mathcal{D}_{\hat{z}}$ obtained in stage 2 by the latent diffusion model.

$$\mathcal{L}_{FT} := -\mathbb{E}_{(z,y)}\left[\log \mathcal{G}(z)_y\right] - \gamma \mathbb{E}_{(\hat{z},y)}\left[\log \mathcal{G}(\hat{z})_y\right], \quad (13)$$

where $z = \mathcal{E}(x), (x, y) \sim p(x, y)$ and $(\hat{z}, y) \in \mathcal{D}_{\hat{z}}$, and γ is the hyperparameter that determines the relative contribution of the two terms in the loss.

After fine-tuning the classification head $\mathcal{G}$, we can combine the feature extractor $\mathcal{E}$ and $\mathcal{G}$ to construct the final classifier $f(x) = \mathcal{G}(\mathcal{E}(x))$ for the long-tailed classification problem. The general structure of our model has been summarised in Algorithm 1.

Algorithm 1 Proposed LDMLR

Input: Data $\mathcal{D}$, Model $\mathcal{E}(x), \mathcal{G}(z), \epsilon_\theta(z_t, t, y)$.
Stage 1: Image Encoding:
- Train $f(x) = \mathcal{G}(\mathcal{E})(x)$ with $\mathcal{D}$ and loss $\mathcal{L}_{CE}$.
- Extract labelled features $\mathcal{D}_z$.

Stage 2: Representation Generation
- Train $\epsilon_\theta(z_t, t, y)$ with $\mathcal{D}_z$ and loss $\mathcal{L}_{LDM}$.
- Generate labelled pseudo-features $\mathcal{D}_{\hat{z}}$.

Stage 3: Classifier Training
- Fine-tuning $\mathcal{G}$ with $\mathcal{D}_z + \mathcal{D}_{\hat{z}}$ and loss $\mathcal{L}_{FT}$.

Output: Classifier $f(x) = \mathcal{G}(\mathcal{E}(x))$.

4. Experiments

4.1. Setup

Dataset. We evaluate our method on CIFAR-LT [23] and ImageNet-LT [23] datasets. The CIFAR-LT experiments involve four scenarios: CIFAR-10 [19] with an imbalance factor of 100, CIFAR-10 with an imbalance factor of 10, CIFAR-100 [19] with an imbalance factor of 100, and CIFAR-100 with an imbalance factor of 10. ImageNet-LT is a subset derived from ImageNet-2012 [34], containing 1000 categories. The training set includes 115.8K images. The number of images per category ranges from 1280 to 5. The validation and test sets are balanced datasets containing 20K and 50K images, respectively.

Implementation details. For CIFAR-LT, before training the diffusion model, a ResNet-32 [8] is pre-trained with a learning rate of $1e-4$ and a dropout rate of 0.5 in the first stage of the model training. We then train the diffusion model for 200 epochs using the Adam optimizer [16] with a learning rate of $1e-3$ in the second stage. The number of diffusion steps is $1,000$, and 500 for the reverse steps. The input and output sizes of ResNet-32 are $3 \times 3 \times 32$ and 64×1, respectively. In the third stage, we reduce the learning rate to $5e-4$ when fine-tuning the fully connected layer of the classifier. For ImageNet-LT, ResNet-10 [8] and the diffusion model are trained for 200 epochs, while the final fine-tuning process uses 100 epochs. The learning rate and optimizer are the same as those used for CIFAR-LT experiments. The batch size is 128 for all training processes. γ is set as 0.05. We conduct all experiments using an NVIDIA GTX 4080 with 16 GB Graphic RAM.

4.2. Results on CIFAR-LT

Three baselines, i.e., Cross Entropy (CE), label shift [40] and WCDAS [7] are selected as baselines for the experimental comparison, and they are trained on the CIFAR-LT dataset [23]. The experimental results are presented in Table 1, from which we can find that the proposed method has improved classification accuracy over the baselines. Notably, the WCDAS+LDMLR method achieves the highest classification accuracy across both datasets and all imbalance factors, with the best performance highlighted in bold. For CIFAR-10-LT with an IF of 100, it reaches an accuracy of 86.29% (an improvement of 1.62% over the baseline WCDAS method), and for CIFAR-100-LT with an IF of 100, it achieves an accuracy of 51.92% (an improvement of 0.97% over the baseline). These results show the effectiveness of combining WCDAS with LDMLR for addressing class imbalance in image classification tasks. It is also observed that our method brings more benefits over the highly imbalanced dataset. For example, on the CIFAR-10-LT, the accuracy gain (3.80%, 1.89%, 1.62%) with IF 100 for CE, Label shift, and WCDAS are much higher than that (0.91%, 0.24%, 0.10%) with IF 10, and the similar improvement is found for the CIFAR-100-LT as well. This helps demonstrate the effectiveness of using feature-generation approaches to tackle the challenges of long-tailed recognition.

4.3. Results on ImageNet-LT

We conduct comparison experiments on the ImageNet-LT dataset [23] with the same baselines used on CIFAR-LT. The experimental results are presented in the Table 2. The baselines are first trained to learn the image feature representation and then combined with our method for augmented latent features. The WCDAS+LDMLR method showcases the highest overall accuracy of 44.8% among the augmented approaches, indicating a modest improvement of 0.2% over the non-augmented WCDAS method. Interestingly, the CE+LDMLR method shows a more pronounced overall improvement of 1.4%, suggesting that the impact of LDMLR varies with the underlying method. It can also be observed from experimental results on the ImageNet-LT dataset that our method is good at improving the classification accuracy for tail classes.

Table 1. Experimental results on CIFAR-LT [23]. The classification accuracies in % are provided. "↑" indicates the improvements over the baseline. The best numbers are in bold. The results of CE, Label Shift, and WCDAS are obtained by self-implemented networks. FASA [45] and SAFA [13] are feature-augmentation-based methods.

Method	CIFAR-10-LT		CIFAR-100-LT	
	IF=10	IF=100	IF=10	IF=100
BALMS [32]	91.3	84.9	63.0	50.8
LWS [14]	91.1	83.7	63.4	50.5
SSD [20]	-	-	62.3	46.0
t-vMF [18]	91.2	83.8	64.7	50.3
CE+DRS [51]	-	78.78	-	45.53
RIDE+CMO [28]	-	-	60.2	50.0
FASA [45]	-	-	-	45.2
SAFA [13]	88.94	80.48	59.11	46.04
CE	88.22	72.46	58.70	41.28
Label shift [40]	89.46	80.88	61.81	48.58
WCDAS [7]	92.48	84.67	65.92	50.95
CE+LDMLR	89.13 (↑0.91)	76.26 (↑3.80)	60.10 (↑1.40)	43.34 (↑2.06)
Label shift+LDMLR	89.70 (↑0.24)	82.77 (↑1.89)	62.67 (↑0.86)	49.76 (↑1.18)
WCDAS+LDMLR	**92.58** (↑0.10)	**86.29** (↑1.62)	**66.32** (↑0.40)	**51.92** (↑0.97)

Table 2. Experiemtal results on ImageNet-LT [23]. The encoder is ResNet-10 [8]. The classification accuracies in % are provided. "↑" indicates the improvements over the baseline. The best numbers are in bold.

Method	ImageNet-LT			
	Many	Medium	Few	All
cRT [14]	49.9	37.5	23.0	40.3
LWS [14]	48.0	37.5	22.8	39.6
BALMS [32]	48.0	38.3	22.9	39.9
t-vMF [18]	55.4	39.9	22.5	43.5
CE	57.7	26.6	4.4	35.8
Label shift [40]	52.0	39.3	20.3	41.7
WCDAS [7]	57.1	40.9	23.3	44.6
CE+LDMLR	57.2	29.2	7.3	37.2 (↑1.4)
Label shift+LDMLR	50.9	39.4	23.7	42.2 (↑0.5)
WCDAS+LDMLR	57.0	41.2	23.4	**44.8** (↑0.2)

4.4. Analysis

Augmentation on the image level. We study the effectiveness of image augmentation using a diffusion model and demonstrate the importance of data augmentation in latent space. To accomplish this, we train a conditional diffusion model on CIFAR-LT and use it to generate new images. These generated images are combined with the original long-tailed data to create a new image dataset, and we examine the accuracy with and without these augmented images. As presented in Table 3, feature-level augmentation (CE+LDMLR and Label shift+LDMLR) consistently outperforms image-level augmentation (CE+DM and Label shift+DM) across all settings. Specifically, in the CIFAR-10-LT dataset with IF of 100, the Label shift+LDMLR method achieves the highest classification accuracy of 82.77%, demonstrating a significant improve-

Table 3. Ablation study: augmentation on the image level. The classification accuracies in % are provided. The best numbers are in bold. The CE+DM and Lable shift+DM denote that the diffusion model is applied to generate image-level data for augmentation.

Method	CIFAR-10-LT		CIFAR-100-LT	
	IF=10	IF=100	IF=10	IF=100
CE	88.22	72.46	58.70	41.28
Label shift [40]	89.46	80.88	61.81	48.58
(Image level)				
CE+DM	88.88	73.91	59.19	42.41
Label shift+DM	89.63	82.10	61.96	48.93
(Feature level)				
CE+LDMLR (Ours)	**89.13**	**76.26**	**60.10**	**43.34**
Label shift+LDMLR (Ours)	**89.70**	**82.77**	**62.67**	**49.76**

ment over the Label shift+DM method with 82.10%. Similarly, in the CIFAR-100-LT dataset with IF of 100, the Label shift+LDMLR method also records the highest accuracy of 49.76%, surpassing the Label shift+DM method (48.93%). The limited accuracy gain of image-level augmentation could be caused by the difficulty in generating high-fidelity image samples from a limited-scale training set. The feature representation in a latent space with lower dimensions might be more easily learned than that in the image space.

Augmentation ratio. The number of generated features is important to the performance. The augmentation ratio represents the proportion between the generated and the encoded features, and we investigate the impact of this ratio on dataset CIFAR-10-LT and CIFAR-100-LT with IF of 10, as shown in Figure 2. As for CIFAR-10-LT, the classification accuracy degrades when the generation ratio is over 20%, while for CIDAR-100-LT, the accuracy goes up with the generation ratio until it passes 40%. This difference could be caused by the smaller number of tail classes in CIFAR-100-LT than in CIFAR-10-LT.

Effects of tail category: many/medium/few. Here, we investigate the effect of augmenting features from different class distributions—many, medium, few—on CIFAR-LT when IF $= 100$, as shown in Table 4. We compare the baseline method, which does not use any augmented features, with strategies that augment features for all classes and selectively for many, medium, or few classes. For both settings, augmenting features for "few" classes consistently yields the highest classification accuracy, highlighting the effectiveness of focusing augmentation efforts on underrepresented classes. Specifically, on CIFAR-10-LT, the WC-DAS+LDMLR method with augmentation on "few" classes achieves a top accuracy of 86.29%, demonstrating a significant improvement over the baseline accuracy of 84.67%. Similarly, on CIFAR-100-LT, the same method and augmentation strategy lead to the best accuracy of 51.92%, compared to the baseline's 50.95%. These results highlight the performance of targeted feature augmentation in generating "few" classes. However, this is specific to CIFAR-LT, and we augment "all" classes for ImageNet-LT.

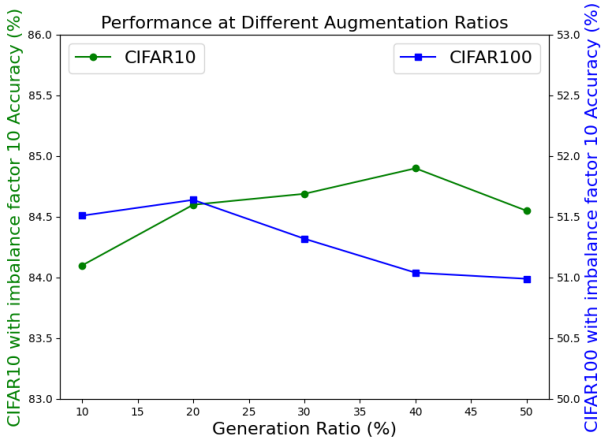

Figure 2. The impact of generation ratio on classification accuracy. The evaluation is conducted on CIFAR-10-LT and CIFAR-100-LT with IF $= 10$.

Impact of encoded and generated features. This ablation study focuses on the impact of encoded and generated features on classification accuracy, as shown in Table 5. The result encompasses experiments across two datasets: CIFAR-10-LT and CIFAR-100-LT, with IF of 10 and 100. The comparison includes approaches utilizing only encoded features, augmentations with generated features, and combining both encoded and generated features. These results demonstrate the effectiveness of the combination of encoded and generated features in addressing the long-tailed classification problems.

Table 4. Ablation study: many/medium/few. The classification accuracies in % are provided. The best numbers are in bold. "All", "Many", "Medium" and "Few" indicate augmenting features of "all", "many", "medium" and "few" classes, respectively.

Method	CIFAR-10 -LT (IF=100)					CIFAR-100-LT (IF=100)				
	Baseline	**"All"**	**"Many"**	**"Medium"**	**"Few"**	**Baseline**	**"All"**	**"Many"**	**"Medium"**	**"Few"**
CE+LDMLR	72.46	76.26	74.03	74.85	76.26	41.28	42.82	41.22	42.79	43.34
Label shift+LDMLR	80.88	81.24	78.22	78.41	82.77	48.58	48.71	44.57	48.71	49.76
WCDAS+LDMLR	84.67	84.90	83.51	83.62	86.29	50.95	51.64	49.38	51.64	51.92

Table 5. Ablation study: encoded and generated features. The classification accuracies in % are provided. The best numbers are in bold.

Method	CIFAR-10 -LT		CIFAR-100-LT	
	IF=10	**IF=100**	**IF=10**	**IF=100**
(Only encoded features)				
Label shift [40]	89.46	80.88	61.81	48.58
WCDAS [7]	92.48	84.67	65.92	50.95
(Only generated features)				
Label shift+LDMLR	89.43	82.18	54.52	32.15
WCDAS+LDMLR	91.98	83.83	64.93	50.42
(Both features)				
Label shift+LDMLR	**89.70**	**82.77**	**62.67**	**49.76**
WCDAS+LDMLR	**92.58**	**86.29**	**66.32**	**51.92**

Visualization of generated features In Figure 3, we visualize feature embeddings during the model training. The lower figure shows the encoded and generative feature distributions of the tail class for CIFAR-10 with an imbalance factor of 0.1. By comparing the distribution of encoded features and those generated by the diffusion model, we observe that the generated features can overlap with parts of the distribution of the encoded features while moderately enriching the original distribution, thereby achieving the goal of feature augmentation effectively.

Future works. The training of our LDMLR requires multiple stages. Hence, one future work could be the simplification of its training process. Another future work could be the exploration of more diffusion models. Lastly, the quality of feature augmentation depends on the diffusion model's generation quality on long-tailed distributed data. Therefore, enhancing the quality of feature augmentation depending on the generation of the diffusion model on long-tailed distributed datasets might be an important future task.

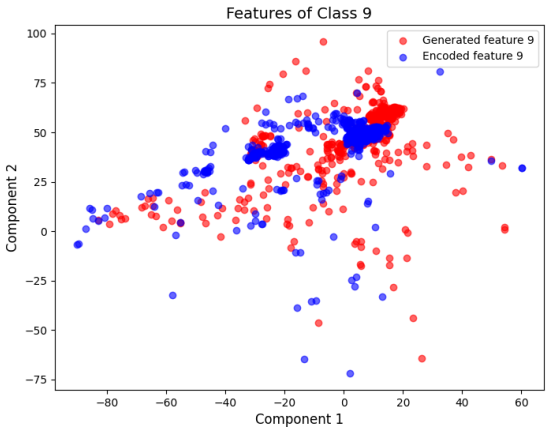

Figure 3. The encoded and generated features of tail class (class 9) in CIFAR-10-LT during the model training. From the figure, the generated features (blue points) can overlay the encoded features (red points) from the original training dataset while slightly enriching the feature space.

5. Conclusion

This work proposes a novel framework, LDMLR, to address the challenge of long-tailed recognition. The LDMLR leverages the powerful generative capabilities of diffusion models for latent-level data augmentation, aiming to balance long-tailed distributed datasets. To the best of our knowledge, we are the first to adopt the diffusion model in long-tailed problems. The experimental outcomes show our method improves in several datasets. We hope that this work could motivate more practical uses of the diffusion model.

References

[1] Adamu Ali-Gombe and Eyad Elyan. Mfc-gan: Class-imbalanced dataset classification using multiple fake class generative adversarial network. *Neurocomputing*, 361:212–221, 2019. 1, 2

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 1, 2

[3] Jiahao Chen and Bing Su. Transfer knowledge from head to tail: Uncertainty calibration under long-tailed distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19978–19987, 2023. 1

[4] Krystallenia Drosou, Stelios Georgiou, Christos Koukouvinos, and Stella Stylianou. Support vector machines classification on class imbalanced data: a case study with real medical data. *Journal of Data Science*, 12(4):727–753, 2014. 1

[5] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001. 1

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 1

[7] Boran Han. Wrapped cauchy distributed angular softmax for long-tailed visual recognition. *arXiv preprint arXiv:2305.18732*, 2023. 1, 2, 5, 6, 8

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 5, 6

[9] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1, 2, 3

[11] Gregory Holste, Song Wang, Ziyu Jiang, Thomas C Shen, George Shih, Ronald M Summers, Yifan Peng, and Zhangyang Wang. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 22–32. Springer, 2022. 1

[12] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021. 1

[13] Yan Hong, Jianfu Zhang, Zhongyi Sun, and Ke Yan. Safa: Sample-adaptive feature augmentation for long-tailed image classification. In *European Conference on Computer Vision*, pages 587–603. Springer, 2022. 2, 6

[14] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition, 2020. 6

[15] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13896–13905, 2020. 1

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[17] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34: 18970–18983, 2021. 1

[18] Takumi Kobayashi. T-vmf similarity for regularizing intra-class feature distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6616–6625, 2021. 6

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 5

[20] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition, 2021. 2, 6

[21] Xuesong Li, Ngaiming Kwok, Jose E Guivant, Karan Narula, Ruowei Li, and Hongkun Wu. Detection of imaged objects with estimated scales. In *VISIGRAPP (5: VISAPP)*, pages 39–47, 2019. 1

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 2

[23] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019. 1, 2, 5, 6

[24] Sara Makki, Zainab Assaghir, Yehia Taher, Rafiqul Haque, Mohand-Said Hacid, and Hassan Zeineddine. An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access*, 7:93010–93022, 2019. 1

[25] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 1, 2

[26] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1, 2

[27] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 864–873, 2016. 1

[28] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification, 2022. 6

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 2

[30] Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18434–18443, 2023. 4

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[32] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition, 2020. 6

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 2, 5

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[36] Haziqah Shamsudin, Umi Kalsom Yusof, Andal Jayalakshmi, and Mohd Nor Akmal Khalid. Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset. In *2020 IEEE 16th International Conference on Control & Automation (ICCA)*, pages 803–808. IEEE, 2020. 1

[37] Mayuri S Shelke, Prashant R Deshmukh, and Vijaya K Shandilya. A review on imbalanced data handling using undersampling and oversampling technique. *Int. J. Recent Trends Eng. Res*, 3(4):444–449, 2017. 1

[38] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 1

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 1, 2, 3, 4

[40] Junjiao Tian, Yen-Cheng Liu, Nathaniel Glaser, Yen-Chang Hsu, and Zsolt Kira. Posterior re-calibration for imbalanced datasets. *Advances in Neural Information Processing Systems*, 33:8101–8113, 2020. 1, 2, 5, 6, 7, 8

[41] Hao Yang and Yun Zhou. Ida-gan: A novel imbalanced data augmentation gan. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8299–8305. IEEE, 2021. 1, 2

[42] Zhixiong Yang, Junwen Pan, Yanzhan Yang, Xiaozhou Shi, Hong-Yu Zhou, Zhicheng Zhang, and Cheng Bian. Proco: Prototype-aware contrastive learning for long-tailed medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 173–182. Springer, 2022. 1

[43] Changkun Ye, Russell Tsuchida, Lars Petersson, and Nick Barnes. Label shift estimation for class-imbalance problem: A bayesian approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1073–1082, 2024. 2

[44] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019. 1

[45] Yuhang Zang, Chen Huang, and Chen Change Loy. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3457–3466, 2021. 2, 6

[46] Min Zeng, Beiji Zou, Faran Wei, Xiyao Liu, and Lei Wang. Effective prediction of three common diseases by combining smote with tomek links technique for imbalanced medical data. In *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, pages 225–228. IEEE, 2016. 1

[47] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 1, 2

[48] Tianjiao Zhang, Huangjie Zheng, Jiangchao Yao, Xiangfeng Wang, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed diffusion models with oriented calibration. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[49] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE international conference on computer vision*, pages 5409–5418, 2017. 1

[50] Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Unequal-training for deep face recognition with long-tailed noisy data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7812–7821, 2019. 1

[51] Allan Zhou, Fahim Tajwar, Alexander Robey, Tom Knowles, George J. Pappas, Hamed Hassani, and Chelsea Finn. Do deep networks transfer invariances across classes?, 2022. 6

[52] Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77, 2005. 1

[53] Honghao Zhu, Guanjun Liu, Mengchu Zhou, Yu Xie, Abdullah Abusorrah, and Qi Kang. Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection. *Neurocomputing*, 407: 50–62, 2020. 1