LCReg: Long-Tailed Image Classification with Latent Categories based Recognition

Weide Liu[a,*], Zhonghua Wu[b,*], Yiming Wang[b], Henghui Ding[b], Fayao Liu[a], Jie Lin[a], Guosheng Lin[b,**]

[a]*Institute for Infocomm Research, A*STAR, Singapore 138632*
[b]*Nanyang Technological University (NTU), Singapore 639798*

Abstract

In this work, we tackle the challenging problem of long-tailed image recognition. Previous long-tailed recognition approaches mainly focus on data augmentation or re-balancing strategies for the tail classes to give them more attention during model training. However, these methods are limited by the small number of training images for the tail classes, which results in poor feature representations. To address this issue, we propose the Latent Categories based long-tail Recognition (LCReg) method. Our hypothesis is that common latent features shared by head and tail classes can be used to improve feature representation. Specifically, we learn a set of class-agnostic latent features shared by both head and tail classes, and then use semantic data augmentation on the latent features to implicitly increase the diversity of the training sample. We conduct extensive experiments on five long-tailed image recognition datasets, and the results show that our proposed method significantly improves the baselines.

1. Introduction

With the successful development of Convolution Neural Networks (CNNs), image recognition has achieved great success on the ideally collected balanced

*indicates equal contribution.
**Corresponding author. G. Lin is with School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore 639798 (e-mail: gslin@ntu.edu.sg).

datasets such as ImageNet. However, real-world applications often involve natural image data that follows a long-tail distribution, in which a small number of classes have a large number of labeled images while most classes have only a few instances or annotations. The traditional fully supervised training strategy is not effective for these unbalanced datasets, which resulting the classification performance of the tail classes dropping quickly.

Long-tailed image recognition has been proposed to address the imbalanced training data problem. The main challenges are the difficulties of handling the small-data learning problems and the extremely imbalanced classification over all the classes. Most of the long-tailed recognition methods focus on generating more data samples of tail classes via data augmentation or using the re-balancing strategy to provide higher importance weights for the tail classes. For example, widely used data augmentation techniques like cropping, flipping, and mirroring are used to generate more data samples of the tail classes during the model training. However, we argue that the diversity of the training samples for the tail classes is still inherently limited due to the limited number of training images, which leads to subtle performance improvement for the long-tailed recognition task by using these conventional data augmentation methods.

Different from the conventional data augmentation methods, semantic data augmentation [1] tries to augment the image features by adding class-aware perturbations. These perturbations are drawn from a multivariate normal distribution, with class-wise covariance matrices calculated from all available training samples. However, directly applying semantic data augmentation to the long-tailed recognition task may not be optimal, as the calculated covariance matrix for tail classes may not provide sufficient meaningful semantic directions for augmentation due to the limited number of training samples. MetaSAug [2] tries to solve the issue of imbalanced statistics by updating the class-wise covariance matrix through the minimization of the LDAM loss on the validation sets. However, the performance of this approach is still limited by the limited diversity and number of training samples for tail classes.

To overcome the limitations mentioned above, we propose to mine out and

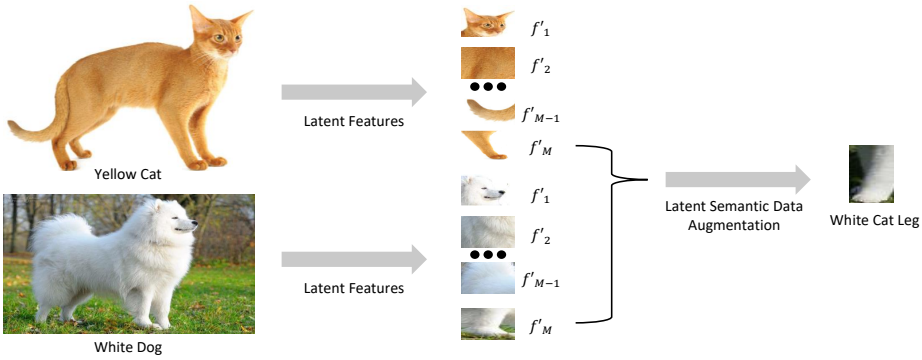

Figure 1: Our LCReg first projects the image features into the latent category features which share the commonality, such as the legs of cats and dogs. By performing the class semantic transformations along with the latent category, we aim to enrich the cat's feature by leveraging the common features, e.g., change the yellow cat leg by leveraging the dog's leg features.

augment the common features among the head and tail classes to increase the diversity of the training samples. The commonality is obtained with the assumption that objects from the same domain might share some commonalities. For instance, cats and dogs share a commonality of legs with similar shapes and appearances. Motivated by this, we argue that it is feasible to re-represent the object features with the common features belonging to the 'sub-categories,' i.e., each category contains parts of the target objects. For example, as shown in Figure 1, we can re-represent the dog and cat with a series of shared 'sub-categories' (e.g., head, leg, body, and tail) with different weights.

To address these issues, we introduce a latent feature pool that stores common features that can be learned through backpropagation during model training. As depicted in Figure 2, the latent features in the pool are class-agnostic and can be shared among all classes. To ensure that the latent features are meaningful and sufficient to represent object features, we utilize a reconstruction loss to reconstruct the original object features using the latent features, with each latent feature contributing to the reconstruction with a similarity weight. Additionally, we apply a semantic data augmentation method to the latent features in order to further increase the diversity of the training data. Our

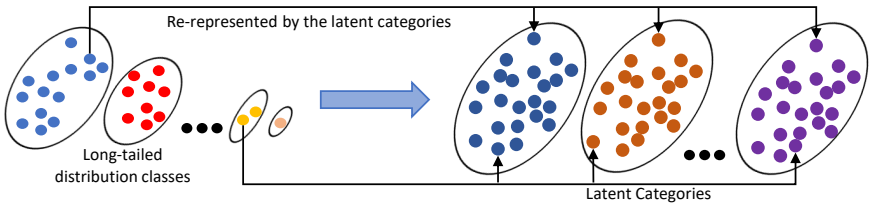

Figure 2: Our LCReg re-represent each object from the original long-tailed distribution dataset by the similarity-weighted sum of latent categories. The latent categories are shareable among the head and tailed classes and form a new balanced distributed dataset.

approach has several advantages due to the use of shareable latent features: 1) All object features are transferred to the shareable latent categories, making the latent features class-agnostic and no longer constrained by the imbalanced distribution. 2) Tail class objects can benefit from the diversity of head classes through the shareable latent features. 3) Tail classes can benefit from the increased diversity provided by data augmentation in the latent space, allowing for the development of latent semantic data augmentation.

In this work, we present a novel approach for long-tail recognition, referred to as Latent categories-based long-tail Recognition (LCReg). The main contribution of this method is the explicit learning of commonalities shared between head and tail classes, leading to improved feature representations. To further enhance the diversity of our training samples, we also propose the use of a semantic data augmentation method on our latent category features.

We conducted extensive experiments on a range of long-tailed recognition benchmark datasets, including CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, iNaturalist 2018, and Places-LT, to demonstrate the effectiveness of LCReg.

2. Related Work

2.1. Long-Tailed Recognition.

Most existing imbalanced classification works can be broadly classified into four categories: re-sampling, re-weighting, decoupling, and data augmentation.

Re-sampling and Re-weighting Data re-sampling and loss re-weighting are common approaches for long-tailed recognition tasks. The core idea of data re-sampling is to forcibly re-balance the datasets by either under-sampling head classes [3] or over-sampling tail classes [4, 5].

The over-sampling technique focuses on increasing the number of instances belonging to the less represented classes (also known as 'tail classes') in order to address the imbalance between the more represented classes (also known as 'head classes'). This has been demonstrated to be effective in previous studies such as [6]. Shen *et al.* [7] introduces a sampling approach called 'Class-Aware Sampling (CAS)', which aims to maintain an equal probability of occurrence for each class in each batch as much as possible. Dhruv *et al.* [8] developed a method for rebalancing the training data by calculating a replication factor for each image based on the distribution of labels and then repeating the images multiple times according to the replication factor. Gupta *et al.* [9] built on this idea by proposing 'Repeat Factor Sampling (RFS),' which increases the sampling frequency of images containing instances from the less represented classes (also known as 'tail instances'). 'Soft-balance Sampling with Hybrid Training' [10] combines traditional sampling techniques with 'Class-Aware Sampling (CAS),' starting by training the detector using a conventional strategy and then introducing hyperparameters to control the amount of traditional sampling.

Unlike the over-sampling method, the under-sampling approach addresses the imbalance between the more represented classes (also known as 'head classes') and the less represented classes (also known as 'tail classes') by reducing the number of samples from the head classes [11]. Random under-sampling involves randomly deleting instances from the more represented classes (also known as 'head classes') until they have the same number of instances as the less represented classes [11].

Over-sampling and under-sampling are popular techniques for addressing imbalanced data, but they come with some limitations. For instance, over-sampling the tail classes can lead to overfitting [12] and may exacerbate any errors or noise present in the tail class samples [13, 14]. Under-sampling, on the

other hand, can result in under-learning of the head classes [14, 15, 16] and may cause valuable data to be lost in the head classes. When dealing with extremely long-tailed data, under-sampling can often result in significant information loss due to the large difference in the amount of data between the head class and the tail class [17].

In addition to re-sampling techniques, loss re-weighting approaches, such as [18, 19, 20, 21], aim to balance the loss of different classes based on the number of samples they have. However, these re-balancing methods require careful calibration of weights to avoid overfitting to the tail classes or underfitting to the head classes. Both re-sampling and re-weighting approaches can suffer from drawbacks: re-sampling often results in insufficient training of the head classes or overfitting to the tail classes, while re-weighting approaches can lead to unstable optimization during training [22, 23]. The latent feature representation has been proposed to enhance the feature representation. For instance, VQ-VAE [24]) introduced a discrete latent representation to address the problem of "posterior collapse" and generate high-quality images. In contrast, our proposed method, LCReg, addresses imbalanced data by transferring the unbalanced object features to shared, balanced latent categories to learn the commonalities among both the head and tail classes.

Decoupled Training According to the decoupled training scheme [25], training the feature extractor with the entire long-tailed dataset is beneficial, but harmful to the classifier. As a result, this two-stage approach involves first training the feature extractor and classifier on the entire long-tailed dataset, and then fine-tuning the classifier using data re-sampling to balance the weight norm of each class. The bilateral-branch network proposes a similar decoupled training scheme [25] around the same time, but adds an extra classifier for fine-tuning to make the two-stage process into a single stage. Kang *et al.* [26] proposes a method for overcoming the problem of unbalanced data in machine learning tasks. They suggest that by separating the learning process into representation learning and classifier learning, it is possible to achieve strong long-tailed recognition. The representation learning phase can be conducted using either

instance-balanced sampling or class-balanced sampling, with the results indicating that instance-balanced sampling yields the best results. The BAGS [27] also explores the idea of decoupling representation learning and classifier learning. They introduce the balanced group softmax module into the classification head of a detection framework, grouping classes according to the number of instances and executing a softmax operation group by group. This allows for the separation of classes with disparate numbers of instances, effectively balancing the classifiers in the detection framework and reducing the control of the head classes over the tail classes. EDAL [28] proposes a novel AL framework tailored for Scene Graph Generation (SGG) task. The framework uses Evidential Deep Learning (EDL) coupled with a global relationship mining approach to estimate uncertainty and seeks diversity-based methods to alleviate context-level bias and image-level bias. The LPT [29] proposes to prompt the frozen pretrained model to adapt the long-tailed data. The prompts are divided into two groups: a shared prompt for the entire long-tailed dataset to learn general features, and group-specific prompts to gather group-specific features for samples with similar features. The two-phase training paradigm involves training the shared prompt in the first phase and optimizing group-specific prompts in the second phase with a dual sampling strategy and asymmetric Gaussian Clouded Logit loss. However, our method proposes the LCReg for long-tailed image recognition that aims to improve feature representation by learning common latent features shared by head and tail classes.

SimCal [30] presents a method for addressing biases in the classification head through the use of a decoupled learning scheme. The model is initially trained normally, and then a bi-level sampling scheme is used to collect class-balanced training instances through the combination of image-level and instance-level sampling. These samples are used to calibrate the classification head, improving the performance of the tail classes. To mitigate the potential negative effects of this calibration on the head classes, SimCal also introduces a Dual Head Inference architecture which selects predictions for both the tail and head classes directly from the new balanced classifier head and the original head. In addition

to the two-stage training scheme, the causal approach [31] proposes learning long-tailed datasets in an end-to-end manner by removing the negative impact of the lousy momentum effect from the causal graph. As shown later, our proposed approach can also be used in conjunction with the decoupled training scheme.

Data Augmentation Data augmentation is a common approach used to improve long-tailed recognition by creating more augmented samples to address the imbalanced distribution of data. There are several data augmentation techniques, such as generating new samples using similar samples [12, 32] or other data sources [33], image flipping, scaling, rotating, and cropping. However, these techniques may not be sufficient for tail classes, which have few samples and sparse features. Mixup techniques [22] have been shown to help tail classes by providing enriched information from head classes. In particular, label-aware smoothing [22] can be used to boost classification ability during finetuning. Another approach is semantic data augmentation [1], which has been explored in domain adaptation [34] and aims to enrich the features of tail classes and create clearer decision boundaries through feature synthesis. The ECRT [35] proposes a novel approach based on the invariance principles of causality, which allows for efficient knowledge transfer from dominant to under-represented classes using a causal data augmentation procedure. Meta-learning has also been proposed for capturing category-wise covariance for better augmentation in long-tailed recognition [2]. Our proposed method, which is built upon the work of Zhong *et al.* [22], also uses a latent semantic augmentation loss to diversify training samples in the latent category space, providing a complementary approach to data augmentation.

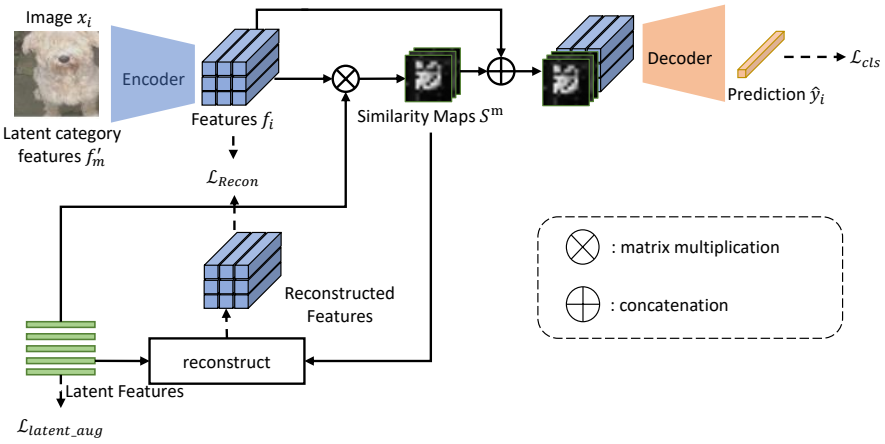

Figure 3: Our proposed LCReg pipeline is as follows: given an input image, we first encode its features with the Encoder. These encoded image features are then compared to the shareable latent category features, which are initially random but trainable embeddings, in order to generate similarity maps. To enhance the diversity of the latent features, we apply a latent implicit augmentation loss to the shareable latent category features. Additionally, to encourage latent features that contain more object information, we reconstruct the image features using the latent features and employ a reconstruction loss. Finally, by combining the similarity maps with the original image features, we utilize a decoder to obtain the final prediction.

3. Method

In this work, we aim to optimize the performance of a classifier on a long-tail distributed dataset containing N training samples with C classes. Given a training sample x_i with label y_i, our classifier uses an object feature $f_i \in \mathbb{R}^{D\times H\times W}$, generated by an encoder with parameters θ, to make a prediction $\hat{y_i}$. Our goal is to minimize the distance between the prediction and the ground truth label by optimizing both the classifier and the encoder parameters θ.

However, on long-tail distributed datasets, the majority of the object features f_i are typically generated from head classes, leading to a bias in the classification model towards these classes and poor performance on tail classes. To address this issue, we introduce a set of class-agnostic latent features f', which capture common features shared among all classes, weighted by a similarity score. In addition, we apply semantic data augmentation to these latent categories to further increase the diversity of our training samples. The overall pipeline of our proposed Latent categories-based long-tail Recognition (LCReg) method is depicted in Figure 3.

3.1. Latent category features

Firstly, we introduce a set of shareable latent features $f'_0, f'_1, ...f'_m, ...f'_M$. Each latent feature represents a latent category that captures part of the object features and is initialized as a random learnable embedding with a dimension of D. These latent features can be trained through back-propagation and have a shape of $f'_m \in \mathbb{R}^{D\times 1}$, so all the latent feature shape is $\mathbb{R}^{D\times M}$.

We further calculate the similarity maps between latent features $f' \in \mathbb{R}^{D\times M}$ and image features $f \in \mathbb{R}^{D\times HW}$ from the image encoder, which benefits the following reconstruction process.

$$S^m = \sigma(\mathcal{FC}(f'_m)^T f), \tag{1}$$

where $S^m \in \mathbb{R}^{1\times H\times W}$ indicates the m_{th} similarity map obtained by the m_{th} encoded latent feature $\mathcal{FC}(f'_m) \in \mathbb{R}^{D\times 1}$ and the image feature f. The $\mathcal{FC}$ is a

1×1 convolutional layer to encode the latent features. We normalize the map with a Sigmoid function $\sigma(\cdot)$ and then reshape the similarity map.

3.2. Reconstruction Loss

To encourage the latent features containing more object information, we use the latent features to reconstruct the image features f by employing a reconstruction loss. Specifically, with the similarity maps $S \in \mathbb{R}^{M \times H \times W}$ generated by latent features, we apply a Softmax function over all the M similarity maps to identify the most discriminative object parts for each latent category $S^m \in \mathbb{R}^{1 \times H \times W}$:

$$\hat{S}^m = \frac{\exp(S^m)}{\sum_{k=1}^{M} \exp(S^k)}. \tag{2}$$

Then we reconstruct image features f by summarizing all the latent categories with the weights from the normalized similarity maps:

$$\hat{f} = \sum_{m=1}^{M} \mathcal{FC}(f'_m)\hat{S}^m. \tag{3}$$

To compare the reconstructed features $\hat{f} \in \mathbb{R}^{D \times HW}$ and the origin features $f \in \mathbb{R}^{D \times HW}$, we calculate the correlation matrix $C_f = \hat{f}^T f$, where $C_f \in \mathbb{R}^{HW \times HW}$ and H, W are the feature size. Finally, we employ a cross-entropy loss to maximize the log-likelihood of the diagonal elements of the correlation matrix $diag(C_f)$ to encourage each latent feature to learn distinct features:

$$\mathcal{L}_{Recon} = -\sum_{j=1}^{HW} t_j log(\psi(diag(C_f))_j), \tag{4}$$

where j is the j^{th} diagonal element of the correlation matrix, and $t_j \in 1, 2, ..., HW$ is the pseudo ground truth of the diagonal element. In particular, the correlation matrix $diag(C_f)$ is a matrix of size $HW * HW$. The first diagonal element of the correlation matrix is considered as the pseudo ground truth and labeled as 1, the second diagonal element is labeled as 2, and so on. The $\psi(diag(C_f))_j$ denotes the Softmax probability for the j^{th} category.

3.3. Latent Feature Augmentation

Data augmentation is a powerful technique that has been widely used in recognition tasks to increase training samples to reduce the over-fitting problem. Traditional data augmentation, such as rotation, flipping, and color-changing, are utilized to increase the training samples by changing the image itself. In contrast to conventional data augmentation techniques, semantic data augmentation augments the semantic features by adding class-wise conditional perturbations [1]. The performance of such class-conditional semantic augmentation heavily relies on the diversity of the training samples to calculate significant, meaningful co-variance matrices for perturbation sampling. However, in the long-tail recognition task, the diversity of tail classes is low due to the limited training samples. The calculated class-conditional statistics will not include sufficient meaningful semantic direction for feature augmentation, which causes negative effects on long-tailed recognition tasks. The details are shown in Section 4.4 and Table 10.

Latent implicit semantic data augmentation. In contrast with ISDA [1], we propose to augment the latent categories to implicitly generate more training samples. To implement the semantic augmentation in the latent feature categories directly, we calculate the covariance matrices ($\mathbf{\Sigma} = \{\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, ..., \mathbf{\Sigma}_M\}$) for each latent category by updating the latent features f'_m at each iteration over total M classes. In particular, for the t^{th} training iteration, we have total $n_m^{(t)} = n_m^{(t-1)} + n_m^{'(t)}$ training samples for m_{th} latent category, where the $n_m^{'(t)}$ denotes the number of training samples at the current t^{th} iteration for m_{th} latent category. Then we estimate the average latent feature value $\mu_m^{(t)}$ of m_{th} latent category for total t iteration with:

$$\mu_m^{(t)} = \frac{n_m^{(t-1)}\mu_m^{(t-1)} + n_m^{'(t)}\mu'^{(t)}_m}{n_m^{(t)}}, \tag{5}$$

where the $\mu'^{(t)}_m = \frac{1}{n_m^{'(t)}} \sum_1^{n_m^{'(t)}} f'_m$ denotes the current average values of the latent m_{th} class features at t^{th} iteration. Then we can update the m_{th} latent category

covariance matrices for total t training iteration with:

$$\Sigma_m^{(t)} = \frac{n_m^{(t-1)}\Sigma_m^{(t-1)} + n_m^{'(t)}\Sigma_m^{'(t)}}{n_m^{(t)}} + \frac{n_m^{(t-1)} n_m^{'(t)} \Delta(\mu)\Delta(\mu)^T}{(n_m^{(t)})^2}, \tag{6}$$

where $\Delta(\mu) = (\mu_m^{(t-1)} - \mu'^{(t)}_m)$, and the $\Sigma'^{(t)}_m$ denotes the m_{th} latent category covariance matrices at current t^{th} iteration.

Then, we augment the features by sampling a semantic transformation perturbation from a Gaussian distribution $\mathcal{N}(0, \lambda\mathbf{\Sigma}_{y'_m})$, where λ indicates the hyperparameter of the augmentation strength and $y'_m \in 1, ..., M$ indicates the pseudo ground truth of the M latent categories. In particular, we set the first latent category as the first class, the second one as the second class, and the rest in the same manner. For each augmented latent feature f_m^a we have

$$f_m^a \sim \mathcal{N}(f'_m, \lambda\Sigma_{y'_m}). \tag{7}$$

Furthermore, when we sample infinite times to explore all the possible meaningful perturbations in the $\mathcal{N}(0, \lambda\mathbf{\Sigma}_{y'_m})$, there is an upper bound of the cross-entropy loss [1] on all the augmented features over N training samples:

$$\begin{aligned} \mathcal{L}_{latent\_aug} &= \sum_{i=1}^{N} L_\infty(f(\boldsymbol{x_i};\theta), y'_m; \mathbf{\Sigma}) \\ &= \frac{1}{N}\sum_{i=1}^{N} log(\sum_{j=1}^{M} e^{z_j}) \end{aligned} \tag{8}$$

$$\begin{aligned} z_j = (w_j^T - w_{y'_m}^T) f_m^a + (b_j - b_{y'_m}) + \\ \frac{\lambda}{2}(w_j^T - w_{y'_m}^T)\Sigma_{y'_m}(w_j - w_{y'_m}), \end{aligned} \tag{9}$$

where θ indicates the encoder parameters for the latent category features. w and b are the weight and biases corresponding to the a 1×1 convolution layer $\mathcal{FC}$ motioned above. Following ISDA [1], we let $\lambda = (t/T) \times \lambda_0$ to reduce the augmentation impact in the beginning of the training stage, where T indicates the total iteration.

With the augmented latent category features, we are able to increase the diversity of training samples by reconstructing the augmented latent features back to the image features f with the reconstruction loss $\mathcal{L}_{Recon}$.

3.4. Training Process

We adopt decoupled training for the long-tailed task as in [22]. Specifically, in the first stage of the training process, our training objective includes the reconstruction loss $\mathcal{L}_{Recon}$ which is applied on the latent category features, a latent augmentation loss $\mathcal{L}_{latent\_aug}$ that augments the latent features, and a cross-entropy classification loss which is applied on final prediction $\hat{y}_i$ generated with the decoder. We optimize the network parameter by combining all the losses:

$$\mathcal{L} = \alpha\mathcal{L}_{latent\_aug} + \beta\mathcal{L}_{Recon} + \gamma\mathcal{L}_{cls}, \tag{10}$$

where L_{cls} indicates the final classification loss (CE loss) between the ground truth y and the prediction $\hat{y}_i$. α, β, and γ are the trade-off parameters, which have been set to 0.1, 0.1, and 1, respectively. In the second stage of training, following [22], we finetune the network.

4. Experiments

4.1. Implementation Details

We follow the training pipeline described in previous works [22, 36] to conduct experiments on five datasets: CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, iNaturalist 2018, and Places-LT. We use the SGD optimizer and apply data augmentation techniques such as random scaling, cropping, and flipping during training. Unless otherwise stated, we use a batch size of 128 for all experiments.

4.2. Dataset

CIFAR-10-LT and CIFAR-100-LT. We conduct experiments on the long-tailed versions of the CIFAR datasets, as described in [37]. The CIFAR-10 and CIFAR-100 datasets consist of 50,000 and 10,000 training and validation images,

respectively, across 10 and 100 categories. To create a long-tailed dataset, we discard some of the training samples and rearrange the remaining ones to create an imbalance factor (IF) = N_{max}/N_{min}, where N_{max} and N_{min} are the numbers of training samples for the largest and smallest classes, respectively. Following previous works [37, 22, 36], we conduct experiments on the CIFAR-LT datasets with IF values of 10, 50, and 100.

ImageNet-LT. Liu *et al.* [38] propose the ImageNet-LT dataset, which contains 115,846 training images and 50,000 validation images, including 1000 categories, with the imbalance factor(IF) of 1280/5. This dataset is a subset of ImageNet [39]. They follow the Pareto distribution with power value = 6 to sample the images and rearrange to a new unbalanced dataset.

iNaturalist 2018. iNaturalist 2018 [40] is a large-scale dataset collected from the real world, whose distribution is extremely unbalanced. It contains 435,713 images for 8142 categories with an imbalanced factor(IF) of 1000/2.

Places-LT. Places-LT is a long-tailed distribution dataset generated from the large-scale scene classification dataset Places [41]. It consists of 184.5K images for 365 categories with an imbalanced factor(IF) of 4980/5.

4.3. Comparisons with State-of-the-art methods

Experiments on CIFAR-LT. Following previous works [22, 31, 42, 36], we conduct experiments on the CIFAR-10-LT and CIFAR-100-LT datasets with imbalance factors of 10, 50, and 100. The latent categories are set to 40 and 50 for CIFAR-10-LT and CIFAR-100-LT, respectively. As shown in Table 1, our proposed method outperforms all previous methods.

Experiments on large-scale datasets. We further evaluate the effectiveness of our method on the large-scale, imbalanced datasets ImageNet-LT, iNaturalist 2018, and Places-LT. The latent category numbers are set to 100 for ImageNet-LT and 200 for iNaturalist 2018, and 100 for the Places-LT dataset. As shown in Tables 2, 3, and 4, our proposed method improves the baseline methods by leveraging the shared commonalities between head and tail classes and employing semantic data augmentation on latent category features, achieving comparable

Method	CIFAR-10-LT			CIFAR-100-LT		
	100	50	10	100	50	10
CE (Cross Entropy)	70.4	74.8	86.4	38.4	43.9	55.8
mixup [43]	73.1	77.8	87.1	39.6	45.0	58.2
LDAM+DRW [42]	77.1	81.1	88.4	42.1	46.7	58.8
BBN(include mixup) [36]	79.9	82.2	88.4	42.6	47.1	59.2
Remix+DRW [44]	79.8	-	89.1	46.8	-	61.3
MiSLAS [22]	82.1	85.7	90.0	47.0	52.3	63.2
MetaSAug CE[2]	80.5	84.0	89.4	46.9	51.9	61.7
MetaSAug LDAM [2]	80.7	84.4	89.7	48.0	52.2	61.2
PaCo [45]	-	-	-	**52.0**	**56.0**	64.2
Ours	**83.1**	**86.5**	**91.2**	47.6	53.1	**64.2**

Table 1: The top-1 accuracy (in %) for ResNet-32 based models trained on the CIFAR-10-LT and CIFAR-100-LT datasets.

performance to previous state-of-the-art methods on all large-scale datasets.

4.4. Ablation Studies

Number of the latent categories. We conduct experiments to analyze the impact of the number of latent categories on performance for different datasets. As shown in Table 5, we experiment with both small and large-scale datasets to explore the effectiveness of the number of latent categories. For larger datasets, which have more training samples and classes, we suggest using more latent categories to better represent the original image features and achieve better performance. However, simply increasing the number of latent categories does not always lead to improved performance. For example, 40 categories yield the best performance on the CIFAR-10-LT dataset, while further increasing the number of categories leads to a rapid decline in performance. We speculate that having too many latent categories may result in the object features being split too finely, resulting in a loss of meaningful parts. Specifically, datasets similar in size and categories to CIFAR-10-LT or CIFAR-100-LT tend to exhibit better performance when the latent categories are set to approximately 20-60. Similarly, datasets comparable to ImageNet-LT demonstrate enhanced performance when the latent categories are set to around 100-300. For datasets similar to

Method	ResNet-50
CE	44.6
CE+DRW [42]	48.5
Focal+DRW [46]	47.9
LDAM+DRW [42]	48.8
NCM [25]	44.3
τ-norm [25]	46.7
cRT [25]	47.3
LWS [25]	47.7
MiSLAS [22]	52.7
MetaSAug CE [2]	47.4
PaCo* [45]	51.0
PaCo [45]	**57.0**
RIDE(2 experts) [47]	54.4
Ours	55.3

Table 2: The top-1 accuracy (in %) for the ResNet-50 based models trained on ImageNet-LT. * denotes without RandAugment method.

iNaturalist in terms of size and categories, setting the latent categories to a value larger than 200 yields improved performance.

Performance on different splits of classes. We also report the classification accuracy for classes with a large number of images (more than 100 per class), a medium number of images (20 to 100 per class), and a small number of images (less than 20 per class). In particular, we set the number of latent categories to 40 for CIFAR-10-LT, 50 for CIFAR-100-LT, 100 for ImageNet-LT, and 200 for iNaturalist 2018. As shown in Table 6, our method consistently outperforms all other methods by a large margin for all classes on all datasets.

Effect of each component. We investigate the contribution of each component of our proposed method - the latent categories, the latent augmentation loss, and the latent reconstruction loss - by conducting ablation experiments on both small and large-scale datasets. Specifically, we choose an imbalance factor (IF) of 100 and set the number of latent categories to 40 for CIFAR-10-LT and 50 for CIFAR-100-LT. For the experiments on the large challenge datasets (iNaturalist 2018), we set the number of latent categories to 100 but used a smaller training batch size of 16 due to resource constraints. As shown in

Method	ResNet-50
CB-Focal [48]	61.1
LDAM+DRW [42]	68.0
OLTR [38]	63.9
cRT [25]	65.2
τ-norm [25]	65.6
LWS [25]	65.9
BBN(include mixup) [36]	69.6
Remix+DRW [44]	70.5
MiSLAS [22]	71.6
MetaSAug CE [2]	68.8
PaCo [45]	**73.2**
RIDE(2 experts) [47]	71.4
Ours	72.6

Table 3: The top-1 accuracy (in %) for the ResNet-50 based models trained on iNaturalist 2018.

Method	ResNet-152
Range Loss [49]	35.1
FSLwF [50]	34.9
OLTR [38]	35.9
OLTR+LFME [51]	36.2
PaCo [45]	**41.2**
Ours	40.2

Table 4: The top-1 accuracy (in %) for the ResNet-152 based models trained on Places-LT.

Table 7, adding our proposed latent categories alone significantly improves the performance of the baseline method on all datasets. The performance is further improved by applying the latent augmentation loss and latent reconstruction loss.

Visualization of the latent categories. As shown in Figure 4, we visualize the latent category histogram for the ImageNet-LT dataset with 100 latent categories. We reconstruct the image features using the latent categories, with each latent category contributing a normalized similarity weight generated by equation 2. As shown in the figure, the 79th latent category (green) is highlighted for the hare' and dogs' (Images E and F), maybe due to their similar limb patterns. Additionally, the cow', human arm', and 'fisher' also share some

Dataset	Number of latent class					Dataset	Number of latent class			
	20	30	40	50	60		20	60	100	200
CIFAR-10-LT	81.9	82.4	83.1	82.5	79.6	ImageNet-LT	54.5	55.0	55.3	55.2
CIFAR-100-LT	47.1	47.2	47.4	47.6	46.1	iNaturalist 2018	-	71.6	71.6	72.6

Table 5: The results of ablation studies on the effectiveness of the number of latent categories for long-tailed image recognition tasks. We conduct experiments on both small datasets (CIFAR-10-LT and CIFAR-100-LT with imbalance factor (IF) 100) and large datasets (ImageNet-LT and iNaturalist 2018). The results show that as the size of the dataset increases (with more training samples and classes), a larger number of latent categories is generally required to achieve better performance. However, it is also noted that continuously increasing the number of latent categories beyond a certain point may not necessarily lead to further improvement and may even result in a decrease in performance.

commonalities captured by the 98th latent category (red).

As shown in Figure 5, we have included the examples from the CIFAR-10-LT dataset. The figure illustrates that objects from many classes, medium classes, and few classes share some commonalities. To further analyse the histogram of objects from different classes, we have calculated their KL divergence loss in Table 8. As the table demonstrates, the KL divergence loss between objects of different classes can still be small, such as the objects of the automobile class (from many classes) to the ship class (from few classes). This validates our assumption of shareable commonalities.

The effectiveness of hyper-parameter. As shown in Table 9, we investigate the impact of the hyperparameters α, β, and γ in Equation 10 on the performance of our method. The results in the table demonstrate that our method is relatively insensitive to these hyperparameters. This suggests that our method is robust and can achieve stable performance across a wide range of hyperparameter values.

Latent augmentation vs. ISDA As shown in Table 10, we directly apply the ISDA method [1] to the original class features and conduct experiments on CIFAR-10-LT and CIFAR-100-LT with different imbalance impacts. Our proposed latent augmentation method, which augments the features within the

Dataset	Methods	Many	Medium	Few
CIFAR10-LT IF 100	Ours*	90.9	80.8	73.7
	Ours	**92.6**	**81.5**	**75.4**
CIFAR100-LT IF 100	OLTR [38]	61.8	41.4	17.6
	LDAM + DRW [37]	61.5	41.7	20.2
	τ-norm [25]	65.7	43.6	17.3
	cRT [25]	64.0	44.8	18.1
	Ours*	63.1	48.6	25.0
	Ours	**64.2**	**49.2**	**25.4**
ImageNet-LT	cRT [25]	62.5	47.4	29.5
	LWS [25]	61.8	48.6	33.5
	Ours*	61.7	51.2	35.6
	Ours	**66.1**	**52.8**	**36.2**
iNaturalist 2018	cRT [25]	73.2	68.8	66.1
	τ-norm [25]	71.1	68.9	69.3
	LWS [25]	71.0	69.8	68.8
	Ours*	73.2	72.4	70.4
	Ours	**73.8**	**73.4**	**71.5**

Table 6: We evaluate the accuracy of our proposed methods on three different splits of classes: Many, Medium, and Few. To validate the effectiveness of our approach, we conduct experiments on a variety of datasets, including small-scale datasets such as CIFAR10-LT and CIFAR100-LT with IF 100, as well as large-scale datasets like ImageNet-LT and iNaturalist 2018. For comparison, we also report the results of our baseline approach, indicated as"Ours*," which does not incorporate latent category features or the reconstruction loss $\mathcal{L}Recon$ and latent augmentation loss $\mathcal{L}latent_aug$. This allows us to assess the impact of these additional components on the overall performance of our method.

Components			CIFAR-10-LT			CIFAR-100-LT			iNaturalist 2018
latent category	latent aug	latent recon	100	50	10	100	50	10	-
			82.1	85.7	90.0	47.0	52.3	63.2	68.9
✓			82.2	85.8	90.7	47.2	52.6	63.9	69.4
✓	✓		82.5	86.0	91.0	47.4	53.0	64.1	69.8
✓		✓	83.0	86.2	91.1	47.3	52.5	64.0	70.0
✓	✓	✓	**83.1**	**86.5**	**91.2**	**47.6**	**53.1**	**64.2**	**70.5**

Table 7: The results of ablation studies on the effectiveness of each component of our proposed method for long-tailed image recognition tasks. We conduct experiments on both small datasets (CIFAR-10-LT and CIFAR-100-LT with imbalance factor (IF) 100, 50, and 10) and a large dataset (iNaturalist 2018). The results show that each of our proposed components (utilizing latent categories, latent augmentation loss, and latent reconstruction loss) individually improves the performance of the baseline (without any of the proposed components) on all datasets. This demonstrates the effectiveness of each component in improving the performance of long-tailed image recognition tasks.

latent categories, significantly improves the performance on all long-tail recognition datasets, compared to directly using the unbalanced features. This demonstrates the effectiveness of our latent augmentation method in improving the performance of long-tailed image recognition tasks. Specifically, we set the number of latent categories to 40 for CIFAR-10-LT and 50 for CIFAR-100-LT.

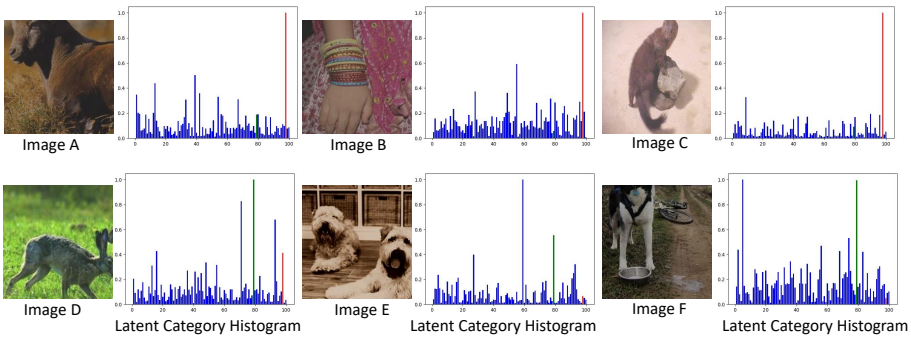

Figure 4: The weight histogram of latent categories contributing to the reconstruction of image features for a sample of images from the ImageNet-LT dataset. As depicted in the figure, the 79^{th} latent category (green) is highlighted by the hare' (Image D), dogs' (Image E and F), which may be because of containing similar shapes of limbs. Additionally, the cow' (Image A), human arm' (Image B), and 'fisher' (Image C) share some commonalities captured by the 98^{th} latent category (red). It is important to note that our proposed method aims to learn commonalities between images belonging to latent classes, which are not necessarily denoted by appearances from a human perspective. A common characteristic can be any characteristic of an object, such as color, structure, or shape.

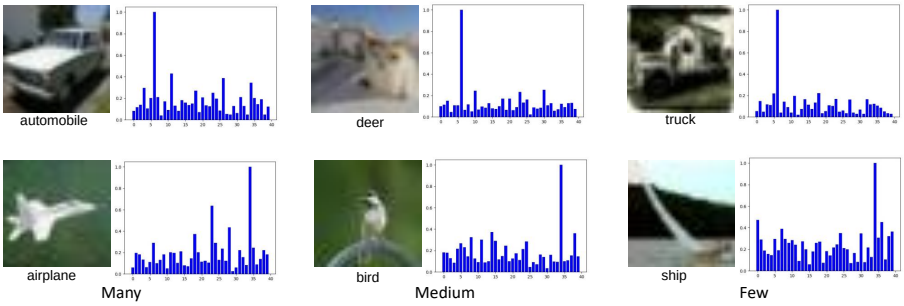

Figure 5: The weight histogram of latent categories contributing to the reconstruction of image features for a sample of images from the CIFAR-10 dataset. The figure illustrates that objects from many classes, medium classes, and few classes still share some commonalities. These results suggest that our approach is able to capture shared features across different classes.

	automobile	deer	truck	airplane	bird	ship
automobile	0.00	4.39	4.90	1.58	2.04	0.52
deer	-0.30	0.00	1.85	0.44	0.76	-0.87
truck	-0.99	0.35	0.00	-0.08	-0.03	-0.80
airplane	2.89	6.54	7.76	0.00	1.71	0.07
bird	2.63	5.94	6.60	1.14	0.00	-0.66
ship	6.37	9.95	11.92	5.01	4.71	0.00

Table 8: The KL divergence loss.

			CIFAR-10-LT			CIFAR-100-LT		
α	β	γ	100	50	10	100	50	10
1	1	1	82.0	85.6	91.0	47.0	52.8	63.8
1	0.1	1	82.2	85.8	89.8	47.3	52.6	63.7
1	1	0.1	82.4	85.8	**91.1**	47.1	52.7	63.9
0.1	0.1	1	**82.5**	**86.0**	91.0	**47.4**	**53.0**	**64.1**

Table 9: The results of ablation studies on the parameter selection for our proposed method on long-tailed image recognition tasks. We report the top-1 accuracy (%) for different values of the parameters α, β and γ on various datasets. The results show that our method is relatively insensitive to the choice of these parameters, achieving good performance across a wide range of parameter values. This demonstrates the robustness and flexibility of our proposed method in handling long-tailed image recognition tasks.

	CIFAR-10-LT			CIFAR-100-LT		
Methods	100	50	10	100	50	10
Baseline	82.1	85.7	90.0	47.0	52.3	63.2
+ ISDA	79.8	82.7	87.8	43.5	47.8	57.7
+ L_{Aug}	**82.5**	**86.0**	**91.0**	**47.4**	**53.0**	**64.1**

Table 10: Ablation studies comparing the normal feature augmentation with ISDA to the latent feature augmentation. The top-1 accuracy (%) is reported for various datasets. The results show that using the latent augmentation method on the latent category features, denoted as L_{Aug}, significantly improves the performance compared to using normal feature augmentation with ISDA on the original class features. This suggests that augmenting the latent category features, which capture commonalities among different classes, is more effective than augmenting the original class features directly.

5. Conclusion

In this work, we have proposed a novel approach for addressing the challenges of long-tailed image recognition, called latent category-based long-tail recognition (LCReg). This method aims to increase the diversity of training samples for long-tailed recognition tasks by mining and augmenting common features among head and tail classes. To achieve this, we introduce latent category features, which are class-agnostic and can be shared among all classes. These features are learned through backpropagation during model training and are used to reconstruct the original object features using a reconstruction loss. In addition, we apply a semantic data augmentation method to these latent features to further increase their diversity. Our approach has several advantages over traditional methods, including the ability to handle small-data learning problems and the use of class-agnostic features that can be shared among all classes. Our experimental results on several long-tailed recognition benchmarks demonstrate the effectiveness of our method.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2018-003), the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (RG95/20). This research is partly supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funds (Grant No. A20H6b0151).

References

[1] Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, C. Wu, Implicit semantic data augmentation for deep networks, Advances in Neural Information Processing Systems 32 (2019) 12635–12644.

[2] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, X. Cheng, Metasaug: Meta semantic augmentation for long-tailed visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5212–5221.

[3] M. Buda, A. Maki, M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Networks 106 (2018) 249–259.

[4] L. Shen, Z. Lin, Q. Huang, Relay backpropagation for effective learning of deep convolutional neural networks, in: European Conference on Computer Vision, 2016, pp. 467–482.

[5] N. Sarafianos, X. Xu, I. A. Kakadiaris, Deep imbalanced attribute classification using visual attention aggregation, in: European Conference on Computer Vision, Vol. 11215, Springer, 2018, pp. 708–725.

[6] J. Byrd, Z. Lipton, What is the effect of importance weighting in deep learning?, in: ICML, 2019, pp. 872–881.

[7] L. Shen, Z. Lin, Q. Huang, Relay backpropagation for effective learning of deep convolutional neural networks, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 467–482.

[8] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, L. Van Der Maaten, Exploring the limits of weakly supervised pretraining, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 181–196.

[9] A. Gupta, P. Dollar, R. Girshick, Lvis: A dataset for large vocabulary instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5356–5364.

[10] J. Peng, X. Bu, M. Sun, Z. Zhang, T. Tan, J. Yan, Large-scale object detection in the wild from imbalanced multi-labels, in: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9709–9718.

[11] M. Buda, A. Maki, M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Networks (2018) 249–259.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research (2002) 321–357.

[13] J. Cui, S. Liu, Z. Tian, J. Jia, Reslt: Residual learning for long-tailed recognition, arXiv preprint arXiv:2101.10633 (2021).

[14] S. Sinha, H. Ohashi, K. Nakamura, Class-wise difficulty-balanced loss for solving class-imbalance, in: Proceedings of the Asian Conference on Computer Vision, 2020.

[15] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: CVPR, 2019.

[16] M.-L. Zhang, X.-Y. Zhang, C. Wang, C.-L. Liu, Towards prior gap and representation gap for long-tailed recognition, Pattern Recognition 133 (2023) 109012.

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

[18] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent data analysis 6 (5) (2002) 429–449.

[19] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, J. Yan, Equalization loss for long-tailed object recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2020, pp. 11659–11668.

[20] W. Zhao, H. Zhao, Hierarchical long-tailed classification based on multi-granularity knowledge transfer driven by multi-scale feature fusion, Pattern Recognition (2023) 109842.

[21] X. Zhou, J. Zhai, Y. Cao, Feature fusion network for long-tailed visual recognition, Pattern Recognition 144 (2023) 109827.

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

[23] X. Zhao, J. Xiao, S. Yu, H. Li, B. Zhang, Weight-guided class complementing for long-tailed image recognition, Pattern Recognition 138 (2023) 109374.

[24] A. Van Den Oord, O. Vinyals, et al., Neural discrete representation learning, Advances in neural information processing systems 30 (2017).

[25] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, Y. Kalantidis, Decoupling representation and classifier for long-tailed recognition, in: International Conference on Learning Representations, 2020.

[26] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, Y. Kalantidis, Decoupling representation and classifier for long-tailed recognition, in: ICLR, 2020.

[27] Y. Li, T. Wang, B. Kang, S. Tang, C. Wang, J. Li, J. Feng, Overcoming classifier imbalance for long-tail object detection with balanced group softmax, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10991–11000.

[28] S. Sun, S. Zhi, J. Heikkilä, L. Liu, Evidential uncertainty and diversity guided active learning for scene graph generation, in: The Eleventh International Conference on Learning Representations.

[29] B. Dong, P. Zhou, S. Yan, W. Zuo, Lpt: Long-tailed prompt tuning for image classification, arXiv preprint arXiv:2210.01033 (2022).

[30] T. Wang, Y. Li, B. Kang, J. Li, J. Liew, S. Tang, S. Hoi, J. Feng, The devil is in classification: A simple framework for long-tail instance segmentation, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 728–744.

[31] K. Tang, J. Huang, H. Zhang, Long-tailed classification by keeping the good and removing the bad momentum causal effect, Advances in neural information processing systems 33 (2020).

[32] J. Li, Q.-F. Wang, K. Huang, X. Yang, R. Zhang, J. Y. Goulermas, Towards better long-tailed oracle character recognition with adversarial data augmentation, Pattern Recognition 140 (2023) 109534.

[33] H. He, Y. Bai, E. A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks, 2008, pp. 1322–1328.

[34] S. Li, M. Xie, K. Gong, C. H. Liu, Y. Wang, W. Li, Transferable semantic augmentation for domain adaptation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11516–11525.

[35] J. Chen, Z. Xiu, B. Goldstein, R. Henao, L. Carin, C. Tao, Supercharging imbalanced data learning with energy-based contrastive representation transfer, Advances in neural information processing systems 34 (2021) 21229–21243.

[36] B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9719–9728.

[37] K. Cao, C. Wei, A. Gaidon, N. Arechiga, T. Ma, Learning imbalanced datasets with label-distribution-aware margin loss, arXiv preprint arXiv:1906.07413 (2019).

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

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (3) (2015) 211–252.

[40] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, S. Belongie, The iNaturalist species classification and detection dataset, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8769–8778.

[41] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (6) (2017) 1452–1464.

[42] K. Cao, C. Wei, A. Gaidon, N. Arechiga, T. Ma, Learning imbalanced datasets with label-distribution-aware margin loss, in: Advances in neural information processing systems, 2019, pp. 1567–1578.

[43] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, International Conference on Learning Representations (2018).

[44] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, D.-C. Juan, Remix: Rebalanced mixup, in: European Conference on Computer Vision Workshop, 2020.

[45] J. Cui, Z. Zhong, S. Liu, B. Yu, J. Jia, Parametric contrastive learning, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 715–724.

[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: International Conference on Computer Vision, 2017, pp. 2980–2988.

[47] X. Wang, L. Lian, Z. Miao, Z. Liu, S. Yu, Long-tailed recognition by routing diverse distribution-aware experts, in: International Conference on Learning Representations, 2021.
URL https://openreview.net/forum?id=D9I3drBz4UC

[48] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268–9277.

[49] X. Zhang, Z. Fang, Y. Wen, Z. Li, Y. Qiao, Range loss for deep face recognition with long-tailed training data, in: International Conference on Computer Vision, 2017, pp. 5409–5418.

[50] S. Gidaris, N. Komodakis, Dynamic few-shot visual learning without forgetting, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4367–4375.

[51] L. Xiang, G. Ding, Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification, in: European Conference on Computer Vision, 2020.