LONG-TAILED RECOGNITION BY LEARNING FROM LATENT CATEGORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we address the challenging task of long-tailed recognition. Previous long-tailed recognition methods commonly focus on data augmentation of tailed classes or re-balancing strategy to give more attention to tailed classes during training. However, due to the limited training images for tailed classes, the diversity of augmented images are still restricted, which results in poor feature representations. In this work, we argue that there are common latent features between the head and tailed classes that can be used to give better feature representation. We propose to learn a set of semantic and class-agnostic latent features shared by the head and tailed classes. Then, we implicitly enrich the training sample diversity via leveraging semantic data augmentation for the commonality features. We evaluate our methods on several popular long-tailed datasets and achieve new state-of-the-art performance consistently.

1 INTRODUCTION

With the successful development of Convolution Neural Networks (CNNs), image recognition has achieved great success in both speed and accuracy on the ideal collected balanced datasets, e.g., ImageNet (Russakovsky et al., 2015) and Oxford Flowers-102 (Nilsback & Zisserman, 2008). However, in most real-world applications, the natural data often follows a long-tail distribution, where a few classes have abundant labeled images while most classes have only a few instances or a few annotations. The classification performance of the tailed classes on such an unbalanced dataset would drop quickly with the conventional fully supervised training strategy.

The long-tailed recognition task 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 extreme imbalanced classification over all the classes. Most of the long-tailed recognition methods focus on generating more data samples of tailed classes via data augmentation or using the re-balancing strategy to provide more important weight for the tailed classes. For example, widely used data augmentation techniques like cropping, flipping, and mirroring are used to increase the training samples. However, the diversity of the training samples for the tailed classes is still inherently limited due to the limited number of training objects, which leads to subtle performance improvement by those conventional data augmentation methods.

In contrast to conventional data augmentation, semantic data augmentation (Wang et al., 2019) tries to augment the image features by adding class-wise conditional perturbations. The perturbations are sampled from the multivariate normal distribution, where the class-wise covariance matrices are calculated from all the training samples. However, directly applying semantic data augmentation to the long-tailed recognition task may not be suitable since the calculated covariance matrix of the tailed classes may not constitute satisfactory meaningful semantic directions for semantic augmentation due to the limited training samples. MetaSAug (Li et al., 2021a) tries to solve the imbalanced statistics problem by updating the class-wise covariance matrix through minimizing the LDAM loss on the validation sets. However, the performance is still constrained to the limited diversity and training samples for the tailed classes.

To overcome the limitations motioned above, we propose to mine out the commonality features among the head and tailed classes to increase the diversity of the training samples. The commonality is obtained with an assumption that objects from the same domain might share some commonalities.

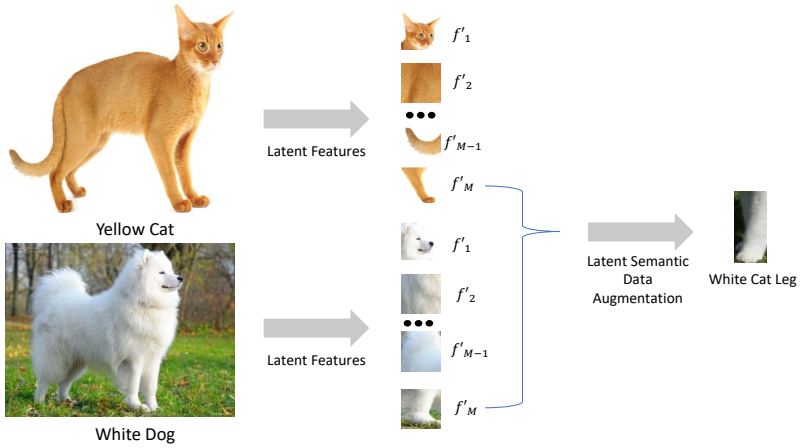

Figure 1: Our LCReg first projects the image features into the latent category features which share the commonality, such as the legs of cats and dogs. By performing the class semantic transformations along with the latent category, we aim to enrich the cat's feature by leveraging the commonality features,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

For example, the cat and dog share a commonality on legs, where they have similar shapes and appearances. It is feasible to re-represent the object features with the commonality features belonging to the 'sub-categories': 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.

In particular, we introduce a latent feature pool to store the commonality features, which can be learned through the back-propagation during the model training. As shown in Figure 2, the latent features from the pool are class-agnostic and shareable among all the classes. To ensure the latent features are meaningful and sufficient to represent object features, we apply a reconstruction loss to reconstruct the original object features with latent features. Each latent feature contributes to reconstructing the object with the similarity weights. Our method has several advantages with the shareable latent features: 1) We transfer all the object features to the shareable latent categories, making the latent features class-agnostic, which allows our approach no longer constrained to the imbalance distribution. This leads to 2) the tailed class objects can benefit from the thriving diversity of the head with the shareable latent features. 3) The tailed classes can benefit from the data augmentation technique with the increased diversity, which allows us to develop a latent semantic data augmentation in the latent space.

The main contributions of this work are concluded from three aspects:

- We propose a Latent Categories based long-tail Recognition (LCReg) method to address the long-tail issue. The proposed LCReg explicitly learns the commonalities shared among the head and tailed classes for better feature representations.
- We adopt a semantic data augmentation method on our proposed latent category features to implicitly enrich the diversity of the training samples.
- Experiments on multiple long-tailed recognition benchmark datasets (CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, iNaturalist 2018, and Places-LT) validate the effectiveness of our method and show that our method achieves state-of-the-art performance.

2 Related Work

2.1 Re-sampling and Re-weighting

Data re-sampling and loss re-weighting are common approaches to long-tailed recognition. The core idea of data re-sampling is to forcibly re-balance the datasets by either under-sampling head classes (Buda et al., 2017; More, 2016; Drummond & Holte, 2003) or over-sampling tail classes

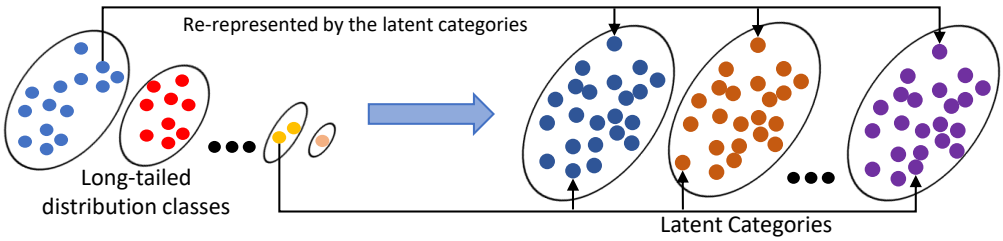

Figure 2: Our LCReg re-represent each object from the original long-tailed distribution dataset by the similarity-weighted sum of latent categories. The latent categories are shareable among the head and tailed classes and form a new balanced distributed dataset.

(Buda et al., 2017; Shen et al., 2016; Sarafianos et al., 2018). Likewise, loss re-weighting (Wang et al., 2017; Huang et al., 2016; 2018; Mikolov et al., 2013; Japkowicz & Stephen, 2002; Tan et al., 2020) approaches try to balance the loss of semantic classes according to their respective number of samples. However, these re-balancing approaches need careful calibration of weighting to prevent the training from overfitting to tail classes or underfitting to head classes. In particular, the data re-sampling approaches often result in insufficient training of head classes or overfitting to the tail classes; the loss re-weighting approaches suffer from unstable optimization during training (Zhong et al., 2021).

2.2 Decoupled Training

The decoupled training scheme (Kang et al., 2020) analyzes and finds that training with the entire long-tailed dataset is beneficial to the feature extractor but harmful to the classifier. Therefore, this two-stage approach proposes first to train the feature extractor and the classifier with the whole long-tailed datasets, and then to finetune the classifier with the data re-sampling to balance the weight norm of each semantic class in the classifier. The bilateral-branch network equivalently proposes the decoupled training scheme in the same period as (Kang et al., 2020) by adding an extra classifier for the finetuning such that the two-stage training becomes one. Besides the two-stage training scheme, a causal approach (Tang et al., 2020) proposes to learn the long-tail datasets in an end-to-end manner by removing the lousy momentum effect from the causal graph. As shown later, our proposed approach is also complementary to the decoupled training.

2.3 Data Augmentation

Data augmentation is another line of approaches to facilitate long-tailed recognition, as more augmented samples can alleviate the severely imbalanced distribution of datasets. Recent studies (Zhou et al., 2020; Zhong et al., 2021; Zhang et al., 2021) demonstrate that the mixup helps the tail classes with enriched information from the head classes. Specifically, (Zhong et al., 2021) additionally proposes label-aware smoothing for finetuning to boost the classification ability. Here, we take a step further to explore how the augmentation in the latent category space benefits the long-tailed classification. We follow another type of augmentation called semantic data augmentation (Wang et al., 2019) which has been explored recently in domain adaptation (Li et al., 2021b). In long-tailed visual recognition, (Li et al., 2021a) proposes meta-learning to capture category-wise covariance for better augmentation. Unlike the existing augmentation approaches, we augment the latent category features through a latent semantic augmentation loss to diversify the training samples. We build our proposed method upon (Zhong et al., 2021) to show that our method is also complementary to the data augmentation approaches.

3 Method

Given a long-tail distributed dataset contains N training samples with C classes, we sample the i^{th} training sample x_i and its corresponding label y_i from the dataset. The final prediction for i^{th} sample $\hat{y}_i$ is classified from the object feature $f_i \in \mathbb{R}^{D\times H\times W}$, which is generated by the encoder with

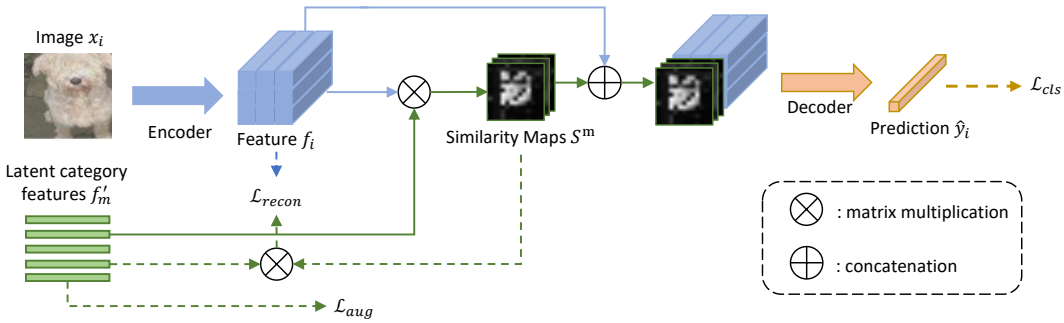

Figure 3: The pipeline of our proposed LCReg.

parameters θ. Our training objective is to optimize the parameters θ and the classifier to minimize the distance between the prediction $\hat{y}_i$ and the ground truth y_i. However, for long-tail distributed datasets, due to the imbalance distribution among each class, most of the features f_i are obtained from the head classes, which makes the classification model biased to the head classes, resulting in unsatisfactory performance on the tailed classes. To alleviate the bias problem, we introduce a set of class-agnostic latent features f'_m, which store the commonality features among all the classes. In particular, each latent feature contributes part of the object features weighted by a similarity score. Moreover, we apply semantic data augmentation on the latent categories to further enrich the diversity of the training samples. The pipeline of our proposed LCReg is shown 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 depicts a latent category representing part of the object features, which is initialized by a random learnable embedding with a dimension of D and able to be trained through back-propagation. The shape of the latent features is $D \times M$.

To enforce each latent feature to learn different object parts and distinguish with each other, we apply a 1×1 convolutional layer $\mathcal{FC}$ to encode the latent features to M classes. 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.

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

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

where $S^m(a,b)$ indicates the m^{th} similarity map at the spatial location (a,b) obtained by the m^{th} latent feature $\mathcal{FC}(f'_m) \in \mathbb{R}^{D \times 1}$ and the image feature f. We normalize the map with a Sigmoid function $\sigma(\cdot)$.

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:

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

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

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

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 ground truth of the diagonal element, we define the first element to be the first class, the second one as the second class, and the rest in the same manner.. 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 (Wang et al., 2019). 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 tailed classes is low due to the limited training samples. The calculated class-conditional statistics will not include sufficient meaningful semantic direction for feature augmentation.

Latent implicit semantic data augmentation In contrast with ISDA (Wang et al., 2019), 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 co-variance matrices for each latent class by updating the latent features f'_m at each iteration over total M classes: $\mathbf{\Sigma}' = \{\mathbf{\Sigma}'_1, \mathbf{\Sigma}'_2, ..., \mathbf{\Sigma}'_M\}$. 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 indicate the hyperparameter of the augmentation strength and $y'_m \in 1, ..., M$ indicates the groundtruth of the M latent categories. For each augmented latent feature f^a_m we have

$$f^a_m \sim \mathcal{N}(f'_m, \lambda\Sigma_{y'_m}). \tag{5}$$

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 (Wang et al., 2019) on all the augmented features over N training samples:

$$\begin{aligned}\mathcal{L}_{latent\_aug} &= \sum_{i=1}^{N} L_\infty(f(\boldsymbol{x_i};\theta), y'_m; \mathbf{\Sigma}') \\ &= \frac{1}{N}\sum_{i=1}^{N} log(\sum_{j=1}^{M} e^{(\boldsymbol{w}_j^T - \boldsymbol{w}_{y'_m}^T)f^a_m + (b_j - b_{y'_m}) + \frac{\lambda}{2}(\boldsymbol{w}_m^{T'} - \boldsymbol{w}_{y'_m}^T)\Sigma_{y'_m}(\boldsymbol{w}_j - \boldsymbol{w}_{y'_m})})\end{aligned} \tag{6}$$

where θ indicates the encoder parameters for the latent category features. $\boldsymbol{w}$ and $\boldsymbol{b}$ are the weight and biases corresponding to the a 1×1 convolution layer $\mathcal{FC}$ motioned above. Following ISDA (Wang

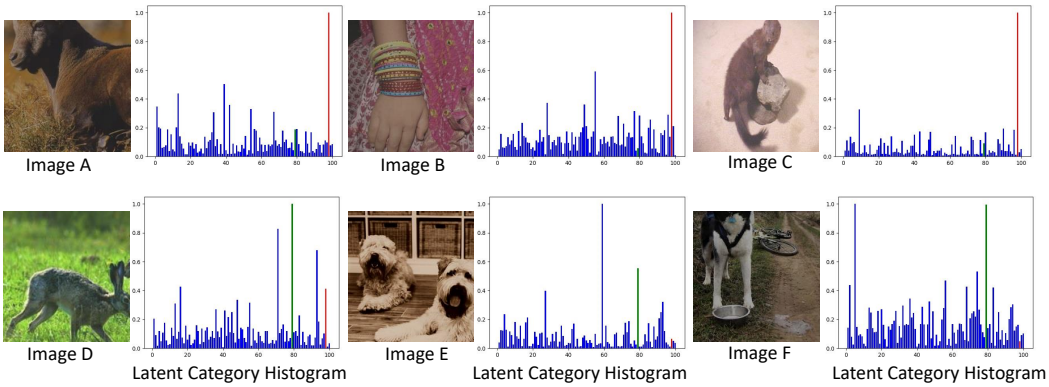

Figure 4: We visualize the weight histogram of latent categories contributing to the reconstruction of the image features. As shown in the figures, the 79^{th} latent category (green) is highlighted by the 'hare'(Image D), and 'dogs'(Image E and F), while all of them contain the similar shape of the limbs. Furthermore, the 'cow'(Image A), 'human arm'(Image B), and 'fisher'(Image C) share some commonalities captured by the 98^{th} latent category(red).

et al., 2019), 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 (Zhong et al., 2021). 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. In the second stage of training, we further add the label-aware smoothing to finetune. We optimize the network parameter by combining all the losses:

$$\mathcal{L} = \alpha\mathcal{L}_{latent\_aug} + \beta\mathcal{L}_{Recon} + \gamma\mathcal{L}_{cls}, \tag{7}$$

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.

4 Experiments

4.1 Dataset

We follow the training pipeline as in (Zhong et al., 2021; Zhou et al., 2020) and conduct experiments on five datasets, including CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, iNaturalist 2018, and Places-LT.

CIFAR-10-LT and CIFAR-100-LT. Following (Cao et al., 2019a), we use long-tail version CIFAR datasets to conduct experiments. CIFAR-10 and CIFAR-100 contain 50000 images and 10000 for training and validation, including 10 and 100 categories, respectively. In particular, we discard the training samples to reorganize a unbalanced dataset with imbalance factor(IF) $\beta = N_{max}/N_{min}$. The N_{max} and N_{min} are the numbers of training samples for the largest and the smallest classes. Following (Cao et al., 2019a; Zhong et al., 2021; Zhou et al., 2020), we conduct the experiments on the CIFAR-LT with imbalance factor(IF) $\beta = 10, 50$ and 100.

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 (Zhang et al., 2018)	73.1	77.8	87.1	39.6	45.0	58.2
LDAM+DRW (Cao et al., 2019b)	77.1	81.1	88.4	42.1	46.7	58.8
BBN(include mixup) (Zhou et al., 2020)	79.9	82.2	88.4	42.6	47.1	59.2
Remix+DRW (Chou et al., 2020)	79.8	-	89.1	46.8	-	61.3
MiSLAS (Zhong et al., 2021)	82.1	85.7	90.0	47.0	52.3	63.2
MetaSAug CE(Li et al., 2021a)	80.5	84.0	89.4	46.9	51.9	61.7
Ours	**83.1**	**86.5**	**91.2**	**47.6**	**52.5**	**63.8**

Table 1: Top-1 accuracy (%) for ResNet-32 based models trained on CIFAR-10-LT and CIFAR-100-LT.

Method	ResNet-50
CE (Cross Entropy (CE))	44.6
CE+DRW (Cao et al., 2019b)	48.5
Focal+DRW (Lin et al., 2017)	47.9
LDAM+DRW (Cao et al., 2019b)	48.8
NCM (Kang et al., 2020)	44.3
τ-norm (Kang et al., 2020)	46.7
cRT (Kang et al., 2020)	47.3
LWS (Kang et al., 2020)	47.7
MiSLAS (Zhong et al., 2021)	52.7
RIDE† (Wang et al., 2021)	54.4
MetaSAug CE (Li et al., 2021a)	47.4
Ours	**55.3**

(a) ImageNet-LT

Method	ResNet-50
CB-Focal (Cui et al., 2019)	61.1
LDAM+DRW (Cao et al., 2019b)	68.0
OLTR (Liu et al., 2019)	63.9
cRT (Kang et al., 2020)	65.2
τ-norm (Kang et al., 2020)	65.6
LWS (Kang et al., 2020)	65.9
BBN(include mixup) (Zhou et al., 2020)	69.6
Remix+DRW (Chou et al., 2020)	70.5
MiSLAS (Zhong et al., 2021)	71.6
RIDE† (Wang et al., 2021)	71.4
MetaSAug CE (Li et al., 2021a)	68.8
Ours	**72.6**

(b) iNaturalist 2018

Method	ResNet-152
Range Loss (Zhang et al., 2017)	35.1
FSLwF (Gidaris & Komodakis, 2018)	34.9
OLTR (Liu et al., 2019)	35.9
OLTR+LFME (Xiang & Ding, 2020)	36.2
Ours	**40.2**

(c) Places-LT

Table 2: Top-1 accuracy (%) on ImageNet-LT, iNaturalist 2018 and Places-LT. † indicate the results with 2 experts.

ImageNet-LT. Liu *et al.* (Liu et al., 2019) 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 (Russakovsky et al., 2015). They follow the Pareto distribution with power value $\alpha = 6$ to sample the images and rearrange to a new unbalanced dataset.

iNaturalist 2018. iNaturalist 2018 (Van Horn et al., 2018) is a large-scale dataset collected from the real world, whose distribution is extremely unbalanced. It contains 435,713 images for 8142 categories with 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 (Zhou et al., 2017). It consists of 184.5K images for 365 categories with an imbalanced factor(IF) of 4980/5.

Dataset	Number of latent class					Dataset	Number of latent class			
	20	30	40	50	60		20	60	100	200
Cifar10	81.9	82.4	83.1	82.5	79.6	ImageNet-LT	54.5	55.0	55.3	55.2
Cifar100	47.1	47.2	47.4	47.6	46.1	Naturalist 2018	-	71.6	71.6	72.6

Table 3: Ablation studies of the effectiveness of the number of latent categories. We conduct the experiments on the small dataset (CIFAR-10-LT and CIFAR-100-LT with IF 100) and large dataset (ImageNet-LT and Naturalist 2018). The larger the dataset (more training samples and classes), the more latent categories are needed to represent better performances.

Components			Dataset		
latent category	latent aug	latent recon	CIFAR-10	CIFAR-100	Naturalist 2018
			82.1	47.0	68.9
✓			82.2	47.0	69.4
✓	✓		82.5	47.4	69.8
✓		✓	83.0	47.3	70.0
✓	✓	✓	83.1	47.6	70.5

Table 4: Ablation studies of each component, including whether utilizing our proposed latent category, latent augmentation loss(latent aug) and latent reconstruction loss (latent recon). We conduct the experiments on the small dataset (CIFAR-10-LT and CIFAR-100-LT with IF 100) and large dataset (Naturalist 2018). The results show that each of our proposed components improves the baseline (without any component).

4.2 Comparisons with State-of-the-arts

Experiments on CIFAR-LT. Following (Zhong et al., 2021; Tang et al., 2020; Cao et al., 2019b; Zhou et al., 2020), we conduct the experiments on CIFAR-10-LT and CIFAR-100-LT with different IF 10, 50, and 100. As shown in Table 1. Our proposed method outperforms all previous methods.

Experiments on large-scale datasets. We further validate the effectiveness of our method on the large-scale imbalanced datasets, *i.e.*, ImageNet-LT, iNaturalist 2018, and Places-LT. Table 2 lists the experimental results. Our proposed method outperforms all the other methods and achieves the new state-of-the-art performance on all the large-scale datasets.

4.3 Ablation Studies

The number of the latent categories. We conduct the experiments to analyze how the latent categories affect the performance on different datasets. As shown in Table 3, we experiment on both small and large scale datasets to explore the effectiveness with the number of latent categories. For the larger datasets, which contain more training samples and classes, we suggest using more latent categories to represent the original image features to achieve better performances. However, enlarging the number of latent categories could not continuously increase the performances. For example, 40 categories yield the best performance on the CIFAR-10-LT dataset. Continually increasing the number of categories would drop the performances very quickly. We speculate that if there are too many latent categories, each object feature might be split too finely by the latent features, failing to obtain the meaningful parts.

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. We conduct the ablation experiments on both the small and large scale datasets to validate our method. Specifically, we choose $IF = 100$ and set the latent categories as 40 and 50 for CIFAR-10-LT and CIFAR-100-LT datasets, respectively. For the experiment on the large challenge datasets, we set the number of latent categories to 100 with a small training batch size(16) due to the resource limitation. As shown in Table 4, only adding our proposed latent categories could have a significant improve-

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 (Liu et al., 2019)	61.8	41.4	17.6
	LDAM + DRW (Cao et al., 2019a)	61.5	41.7	20.2
	τ-norm (Kang et al., 2020)	65.7	43.6	17.3
	cRT (Kang et al., 2020)	64.0	44.8	18.1
	Ours*	63.1	48.4	25.3
	Ours	**64.2**	**49.2**	**25.3**
ImageNet-LT	cRT (Kang et al., 2020)	62.5	47.4	29.5
	LWS (Kang et al., 2020)	61.8	48.6	33.5
	Ours*	61.7	51.3	35.8
	Ours	**66.2**	**52.9**	**35.8**
iNaturalist 2018	cRT (Kang et al., 2020)	73.2	68.8	66.1
	τ-norm (Kang et al., 2020)	71.1	68.9	69.3
	LWS (Kang et al., 2020)	71.0	69.8	68.8
	Ours*	73.2	72.4	70.4
	Ours	**73.8**	**73.4**	**71.5**

Table 5: We report accuracy on three splits of classes: Many, Medium, and Few. We validate our methods on multiple datasets, including small-scale datasets (CIFAR10-LT, CIFAR100-LT with IF 100) and large-scale datasets (ImageNet-LT, and iNaturalist 2018). Ours* indicates ours baseline (without the latent category features, reconstruction loss l_{recon} and latent augmentation loss l_{aug}).

ment over the baseline method (MiSLAS (Zhong et al., 2021)) for all the datasets. The performances are further improved by applying the latent augmentation loss and the latent reconstruction loss.

Visualization of the latent categories As shown in Figure 4, we visualize the latent category histogram on the ImageNet-LT dataset with 100 latent categories. We reconstruct the image features with the latent categories, and each latent category contributes with a normalized similarity weight generated by equation 2. As shown in the figure, the 79^{th} latent category (green) is highlighted by the 'hare' and 'dogs' (Image E and F), while both of them contain similar limb patterns. Furthermore, the 'cow', 'human arm', and 'fisher' also share some commonalities captured by the 98^{th} latent category(red).

4.4 Performance on different splits of classes

We further report the classification accuracy for the many (more than 100 images per class), medium (20 to 100 images per class), and the few (less than 20 images per class) classes, respectively. As shown in Table 5, our method achieves the best performance on the many, medium, and few classes by a large margin for all the datasets. Specifically, on the ImageNet-LT 'many' dataset, our LCReg achieves 4.4% accuracy gain over the previous SOTA methods while keeping the performances of medium and few classes not dropped.

5 Conclusion

In this paper, we have proposed a latent category recognition(LCReg) method to increase the diversity of the training samples for long-tailed recognition tasks by mining out the commonality features among the head and tailed classes. We apply the semantic data augmentation method on our proposed latent category features to implicitly enrich the diversity of the training samples. Experiments on several long-tailed recognition benchmarks validate the effectiveness of our method and show our method achieves state-of-the-art performance.

REFERENCES

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *CoRR*, abs/1710.05381, 2017. URL http://arxiv.org/abs/1710.05381.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*, 2019a.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, pp. 1567–1578, 2019b.

Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. In *ECCVW*, 2020.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pp. 9268–9277, 2019.

Chris Drummond and Robert Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. *Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets*, 01 2003.

Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, pp. 4367–4375, 2018.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 5375–5384. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.580. URL https://doi.org/10.1109/CVPR.2016.580.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *CoRR*, abs/1806.00194, 2018. URL http://arxiv.org/abs/1806.00194.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, 2002. URL http://content.iospress.com/articles/intelligent-data-analysis/ida00103.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5212–5221, 2021a.

Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *CVPR*, 2021b.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pp. 2980–2988, 2017.

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*, pp. 2537–2546, 2019.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 3111–3119, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Ajinkya More. Survey of resampling techniques for improving classification performance in unbalanced datasets. *CoRR*, abs/1608.06048, 2016. URL http://arxiv.org/abs/1608.06048.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pp. 708–725. Springer, 2018. doi: 10.1007/978-3-030-01252-6_42. URL https://doi.org/10.1007/978-3-030-01252-6_42.

Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, volume 9911 of *Lecture Notes in Computer Science*, pp. 467–482. Springer, 2016. doi: 10.1007/978-3-319-46478-7_29. URL https://doi.org/10.1007/978-3-319-46478-7_29.

Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 11659–11668. IEEE, 2020. doi: 10.1109/CVPR42600.2020.01168. URL https://doi.org/10.1109/CVPR42600.2020.01168.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *NeurIPS*, 33, 2020.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *CVPR*, pp. 8769–8778, 2018.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=D9I3drBz4UC.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 7029–7039, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/147ebe637038ca50a1265abac8dea181-Abstract.html.

Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32: 12635–12644, 2019.

Liuyu Xiang and Guiguang Ding. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *ECCV*, 2020.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ICLR*, 2018.

Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *ICCV*, pp. 5409–5418, 2017.

Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *AAAI*, pp. 3447–3455, 2021.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16489–16498, 2021.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 40(6):1452–1464, 2017.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pp. 9719–9728, 2020.