DECOUPLED AND PATCH-BASED CONTRASTIVE LEARNING FOR LONG-TAILED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

The imbalance of the dataset leads to the trained model being biased towards head classes and under-represent the tail classes, making the long-tailed recognition challenging. To address those issues, this paper proposes the decoupled and patch-based contrastive learning. Given an anchor image, the supervised contrastive learning pulls two kinds of positives together in the embedding space: the same image with different data augmentation and other images from the same classes. The weights of two kinds of positives can be influenced by the cardinality of different classes, leading to biased feature space. The decoupled supervised contrastive loss decouples the two kinds of positives, removing the influence of the imbalanced dataset. To improve the discriminative of the learned model on the tail classes, patch-based self distillation crops the small patches from the global view of an image. These small patches can encode the shared visual patterns between different images, and thus can be used to transfer similarity relationship knowledge. Experiments on several long-tailed classification benchmarks demonstrate the superiority of our method. For instance, it achieves 57.7% top-1 accuracy on the ImageNet-LT dataset. Combined with the ensemble-based method, the performance can be further boosted to 59.7%. Our code will be released.

1 INTRODUCTION

Due to the powerful deep learning methods, the computer vision field has made impressive progress in the last decades, boosting various real-scenario tasks, *e.g.*, image classification Russakovsky et al. (2015), object detection Lin et al. (2017), and semantic segmentation Long et al. (2015). Prior researches mostly focus on learning from the manually balanced dataset, *e.g.*, ImageNet1K Russakovsky et al. (2015), and COCO Lin et al. (2014). However, in real-world applications, training samples always exhibit a long-tailed class distribution, where a few head classes contribute to most of the observations, while most tail classes are associated with only a few samples Gupta et al. (2019); Van Horn et al. (2018). The heavy imbalance of the training dataset leads to poor performance of the learned model because of two main challenges: (a) the loss function designed for the balanced dataset can be easily biased towards the head classes; (b) the lack of samples for tail classes makes it difficult to learn the discriminative model for these classes, *i.e.*, the under-representation of the tail classes.

Recently, some works Kang et al. (2020); Li et al. (2021b) extend the powerful contrastive learning Hadsell et al. (2006) to deal with the representation learning of long-tailed recognition, showing its great potential. These works deal with the model biased problem of the contrastive loss on the long-tailed dataset by designing the selection of positive samples Kang et al. (2020) or making the features of different classes converge to the pre-defined uniformly distributed targets Li et al. (2021b). These works ignore that the construction of positives in the supervised contrastive loss can lead to bias and still do not well-solved the under-representation problem of tail classes.

In this work, we reformulate contrastive learning as a distribution alignment task. The target distribution of supervised contrastive loss tries to learn a uniform one-hot distribution, where different views of the same image and different images from the same classes have the same probability while the probabilities of negative samples are zero. This target distribution can lead to biased model and ignore the useful relationship information between different images, resulting unsatisfactory performance on long-tailed recognition. We thus propose decoupled supervised contrastive loss (DSCL) and patch-based self distillation (PBSD) to address those problems.

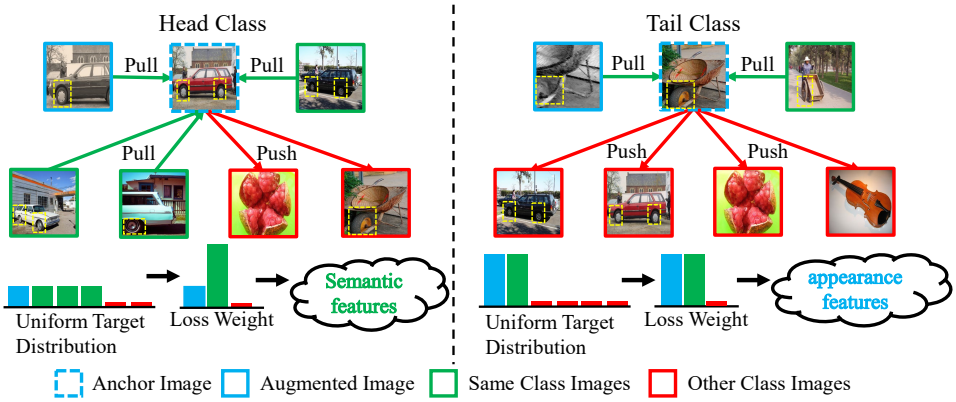

Figure 1: Given an anchor image, supervised contrastive loss pulls the anchor image with its augmented one and other images from the same class together and pushes images from different classes away. The imbalanced dataset leads to different loss weights on these two kinds of positives, making the model learn different kinds of features for head and tail classes. This simple definition of positive and negative samples also ignores the useful knowledge encoded in the shared patterns between classes (marked in the yellow rectangle).

Supervised contrastive loss pulls two kinds of positive samples together: the different views of the same image generated by the data augmentation and different images from the same classes. As discussed in Wei et al. (2020), pulling anchor with these two kinds of positives together makes the model learn different kinds of features, task-agnostic appearance features and task-specific semantic features, respectively. As shown in Fig. 1, the uniform one-hot distribution treats these two kinds of positives equally, leading that head classes and tail classes have different weights for these two kinds of positives. Therefore, the model will learn different kinds of features for head classes and tail classes. The proposed decoupled supervised contrastive loss decouples these two kinds of positive samples and re-designs the target distribution, removing the influence of the imbalanced training set on the weights of these two kinds of positive samples. This simple modification can strongly improve the performance of the model on the long-tailed dataset.

The motivation of PBSD is that many visual patterns can be shared between different classes, *e.g.*, 'wheels' as shown in Fig. 1. Therefore, the similarity relationships between visual patterns can help the learning of the tail classes. To extract visual patterns, besides the global view of an image, we also crop multi small patches from the global view. To maintain the similarity relationship knowledge, ROI pooling He et al. (2017) is used to extract features from the feature maps of the global view according to the positions of cropped small patches. Then we leverage the distillation loss to force the similarity distribution produced by small patches aligning the similarity distribution produced by the pooled feature.

We evaluate our method on several long-tailed datasets including ImageNet-LT Liu et al. (2019), and iNaturaList 2018 Van Horn et al. (2018). Experimental results show that our proposed method improves the standard supervised contrastive learning baseline by 6.5% and achieves superior performance compared with recent works, *e.g.*, outperforming recent proposed supervised contrastive learning based method TSC Li et al. (2021b) by 5.3% on long-tailed ImageNet.

To the best of our knowledge, this is an original contribution to first decouple the two kinds of positives in supervised contrastive loss and use patch-based self distillation to introduce the similarity relationship knowledge into the contrastive learning in the long-tailed recognition. Our method is easy to implement, making it has the potential to benefit the application of the computer vision model under long-tailed situations.

2 Related Work

This work is related to long-tailed recognition and contrastive learning. This section briefly reviews related works in these categories.

Long-tailed recognition aims to address the problem of the model training in the situation where a small portion of classes have massive samples but the others are associated with only a few samples Zhang et al. (2021b). Most related works can be summarized into three categories, *e.g.*, re-balancing methods, decoupling methods, transfer learning methods and ensemble-based methods.

Re-balancing methods use re-sampling or re-weighting to deal with long-tailed recognition Byrd & Lipton (2019); Shen et al. (2016); Buda et al. (2018); Japkowicz & Stephen (2002); Ren et al. (2020); Lin et al. (2017). Re-sampling methods typically include over-sampling for the tail classes Byrd & Lipton (2019); Shen et al. (2016) or under-sampling for the head classes Buda et al. (2018); Japkowicz & Stephen (2002). Besides re-sampling, re-weighting the loss function is also an effective solution, *e.g.*, LDAM Cao et al. (2019) leverages the class-dependent margins based on label frequencies and encourages tail classes to have larger margins. Balanced-Softmax Ren et al. (2020) presents the unbiased extension of Softmax based on the Bayesian estimation.

Decoupling methods find that the re-balancing methods are harmful to the discriminative of the learned backbone Kang et al. (2019). Therefore, they propose the two-stage training to decouple the representation learning and classifier training. Based on this conclusion, DisAlign Zhang et al. (2021a) develops an adaptive calibration function to adjust the classification score.

Transfer learning methods enhance the performance of the model by transferring knowledge from head classes to tail classes. RSG Wang et al. (2021) improves the performance by generating some new samples for tail classes with the help of head classes knowledge. BatchFormer Hou et al. (2022) introduces a one-layer Transformer Vaswani et al. (2017) to transfer knowledge by learning the sample relationships from mini-batch.

Ensemble-based methods leverage multi experts to solve long-tailed visual learning. RIDE Wang et al. (2020) proposes a multi-branch network to learn diverse classifiers in parallel. NCL Li et al. (2022) trains each expert individually and performs knowledge transferring between different experts. Although ensemble-based method achieves superior performance, the introduction of multi experts usually increases the number of parameters and computational complexity.

Contrastive learning has received much attention because of its superior performance on representation learning Li et al. (2020); Caron et al. (2020); Zhu et al. (2021); Feichtenhofer et al. (2021); He et al. (2020); Chen et al. (2020a); Chen & He (2021). Contrastive learning aims to find a feature space that can encode the semantic similarities by pulling the positive pair together while pushing the negative pair apart. Based on the powerful contrastive learning, some methods leverage it into the field of long-tailed recognition, *e.g.*, KCL Kang et al. (2020) finds that the self-supervised learning based on contrastive learning can learn a balanced feature space. To leverage the useful label information, they extend the supervised contrastive loss Khosla et al. (2020) by introducing a k-positive sampling method. TSC Li et al. (2021b) improves the uniformity of the feature distribution by making the features of different classes converge to the pre-defined uniformly distributed targets.

Different from existing contrastive learning based long-tail recognition methods, we firstly point out the reason of that the supervised contrastive loss can be influenced by the imbalanced dataset. Based on this analysis, we design a decoupled target distribution that is robust to the imbalanced training set. In addition, we further extend the contrastive learning by introducing patch-based self distillation to transfer knowledge between classes, mitigating the under-representation of the tailed classes.

3 METHODOLOGY

3.1 OVERVIEW

Given a training dataset $\mathcal{D} = \{\mathrm{x}_i, \mathrm{y}_i\}_{i=1}^{n}$, where x_i denotes an image and $\mathrm{y}_i \in \{1, \ldots, K\}$ is its class label. Assuming that n_k denotes the cardinality of class k in $\mathcal{D}$, and the classes are sorted by cardinality in decreasing order, *i.e.*, if $i_1 < i_2$, then $n_{i_1} \geq n_{i_2}$. In long-tailed image classification, the training dataset is extremely imbalanced, *i.e.*, $n_1 \gg n_K$, and the imbalance ratio is defined as n_1/n_K. Typically, the testing dataset $\mathcal{T}$ is balanced. The training objective of long-tailed image classification method denotes as,

$$\theta^*, \omega^* = \arg\min_{\theta,\omega} \mathcal{L}(D; \theta; \omega), \tag{1}$$

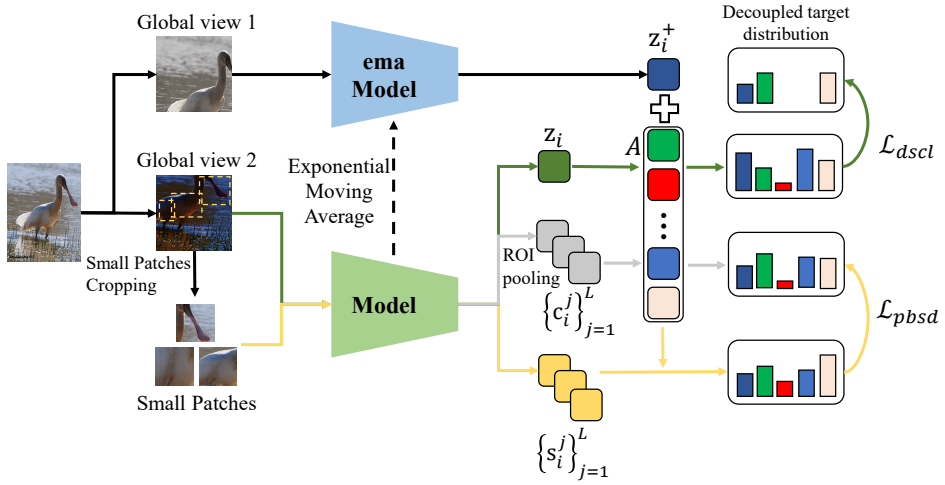

Figure 2: Illustration of the proposed method. Data augmentation is performed to get two global views of a training image. Then multi small patches are cropped from the global view. The backbone and ema backbone are used to extract the normalized features. These features are used to calculate the similarity distribution with memory queue A. $\mathcal{L}_{dscl}$ is used to minimize the divergence between the decoupled target distribution and the similarity distribution produced by the global view, while $\mathcal{L}_{pbsd}$ is used to transfer knowledge through forcing the similarity distribution produced by the small patches mimic the similarity distribution produced by the features that are pooled from the feature map of global view with ROI pooling at the same positions of the small patches.

where $\mathcal{L}$ is the loss function, θ denote the parameters of feature extraction backbone $\mathrm{f}_\theta(\mathrm{x}_i)$, which can map an image into a d-dim feature vector v_i, and ω denote the parameters of classifier $\mathrm{h}_\omega(\mathrm{v}_i)$, which classifies a feature vector to K classification scores. If $\mathcal{L}$ is the loss function designed for the balanced dataset, the large numbers of samples from head classes make the learned model biased towards the head classes and the lack of samples from tail classes leads to under-representation of these classes.

For an anchor image x_i and its normalized feature embedding $\mathrm{z}_i = \mathrm{g}_\gamma(\mathrm{f}_\theta(\mathrm{x}_i))$ extracted with the backbone and an extra projection head g_γ He et al. (2020), the supervised contrastive loss Khosla et al. (2020) optimizes the embedding space by pulling the anchor image with its another data augmentation and other images with the same class label together and pushing the images with different class labels apart,

$$\mathcal{L}_{scl} = \sum_{i=1}^{|\mathcal{D}|} \frac{1}{|P(i)|+1} \sum_{\mathrm{t}\in\{\mathrm{z}_i^+ \cup P(i)\}} -\log \frac{\exp\left(\mathrm{t}\cdot\mathrm{z}_i/\tau\right)}{\exp\left(\mathrm{z}_i^+\cdot\mathrm{z}_i/\tau\right) + \sum_{\mathrm{m}\in A}\exp\left(\mathrm{m}\cdot\mathrm{z}_i/\tau\right)}, \tag{2}$$

where τ is a predefined temperature parameter, z_i^+ is the embedding feature of another data augmentation, A is the set of samples that can be formulated by the training batch Chen et al. (2020a) or the memory queue He et al. (2020), $P(i) = \{\mathrm{m} \in A : \mathrm{y}_m = \mathrm{y}_i\}$ is the set of positive samples' embedding features of z_i, and $|P(i)|$ is the cardinality of $P(i)$.

We define the probability of the conditional distribution as,

$$p(\mathrm{t}|\mathrm{z}_i) = \frac{\exp(\mathrm{t}\cdot\mathrm{z}_i/\tau)}{\exp(\mathrm{z}_i^+\cdot\mathrm{z}_i/\tau) + \sum_{\mathrm{m}\in A}\exp\left(\mathrm{m}\cdot\mathrm{z}_i/\tau\right)}, \tag{3}$$

where $\mathrm{t} \in \{\mathrm{z}_i^+ \cup A\}$. With the definition of the conditional distribution, Eq. (2) can be formulated as a distribution alignment task,

$$\mathcal{L}_{align} = \sum_{i=1}^{|\mathcal{D}|} \sum_{\mathrm{t}\in\{\mathrm{z}_i^+\cup A\}} -\hat{p}(\mathrm{t}|\mathrm{z}_i)\log p(\mathrm{t}|\mathrm{z}_i), \tag{4}$$

where $\hat{p}(\mathrm{t}|\mathrm{z}_i)$ is the probability of the target distribution. In the self-supervised learning $\hat{p}(\mathrm{t}|\mathrm{z}_i) = 1(\mathrm{t} = \mathrm{z}_i^+)$ and $\hat{p}(\mathrm{t}|\mathrm{z}_i) = 0(\mathrm{t} \in A)$ while in supervised learning $\hat{p}(\mathrm{t}|\mathrm{z}_i) = 1/(|P(i)| + 1)(\mathrm{t} \in \{\mathrm{z}_i^+ \cup P(i)\})$ and $\hat{p}(\mathrm{t}|\mathrm{z}_i) = 0$ for other samples in A. Such target distribution is not suitable to the long-tailed recognition task, as it can lead to the biased model and ignores the valuable similarity relationship knowledge between classes.

$\hat{p}(\mathrm{z}_i^+|\mathrm{z}_i) = 1/(|P(i)| + 1)$ and $|P(i)| \propto n_{\mathrm{y}_i}$. Therefore, the imbalance of training dataset can influence the target distribution, *i.e.*, the head classes have a much smaller $\hat{p}(\mathrm{z}_i^+|\mathrm{z}_i)$ than the tail classes. To relief the influence of the imbalanced long-tailed training dataset, we decouple the target distribution in supervised contrastive learning into z_i^+ part and $\mathrm{t} \in P(i)$ part and design a new target distribution. The decoupled supervised contrastive loss (DSCL) can be written as,

$$\mathcal{L}_{dscl} = \sum_{i=1}^{|\mathcal{D}|} \sum_{\mathrm{t} \in \{\mathrm{z}_i^+ \cup A\}} -\hat{p}_{dscl}(\mathrm{t}|\mathrm{z}_i) \log p(\mathrm{t}|\mathrm{z}_i), \tag{5}$$

where $\hat{p}_{dscl}(\mathrm{t}|\mathrm{z}_i)$ denotes the probability of the decoupled target distribution.

Some visual patterns are shared between classes, *e.g.*, the pattern 'fur' in dog class is also useful to the recognition of the cat class. The target distribution of $\mathcal{L}_{scl}$ only considers pulling the positive samples together and pushing the negative samples away, ignoring the knowledge encoded in the shared patterns between classes. We introduce Patch-based Self Distillation (PBSD) to mine similarity relationship knowledge between classes. To extract visual patterns, besides the global view of an image, we also crop some small patches from its global view. The feature embeddings of these small patches denote as $\{s_i^j\}_{j=1}^L$, where L denotes the number of small patches. The conceptual loss of PBSD can be written as,

$$\mathcal{L}_{pbsd} = \sum_{i=1}^{|\mathcal{D}|} \frac{1}{L} \sum_{j=1}^{L} \sum_{\mathrm{t} \in \{\mathrm{z}_i^+ \cup A\}} -q_{pbsd}(\mathrm{t}, \mathrm{s}_i^j) \log p(\mathrm{t}|\mathrm{s}_i^j), \tag{6}$$

where $q_{pbsd}(\mathrm{t}, \mathrm{s}_i^j)$ is the probability of the PBSD target distribution that is related to s_i^j and t. We will introduce it in the following sections. Minimizing Eq. (6) can force the distribution produced by small patches maintaining the similarity relationships between classes, thus encoding the shared patterns knowledge into the model.

The overall loss function of our method can be written as,

$$\mathcal{L}_{overall} = \mathcal{L}_{dscl} + \lambda_1 \mathcal{L}_{pbsd}, \tag{7}$$

where λ_1 is the loss weight.

Following Li et al. (2021b); Kang et al. (2020), after the training of the backbone, we discard the learned projection head $\mathrm{g}_\gamma(\cdot)$ and train a linear classifier on top of the learned backbone using the standard cross-entropy loss with the class-balanced sampling strategy Kang et al. (2019).

We illustrate our method in Fig. 2. The following sections proceed to present details of decoupled supervised contrastive loss and patch-based self distillation.

3.2 Decoupled Supervised Contrastive Loss

As discussed in Section 3.1, the supervised contrastive loss minimizes the divergence between the produced conditional distribution and a designed target distribution, where positive pairs have $\hat{p}(\mathrm{t}|\mathrm{z}_i) = 1/(|P(i)| + 1)$ and negative pairs satisfy $\hat{p}(\mathrm{t}|\mathrm{z}_i) = 0$. There are two kinds of positives in the supervised contrastive loss: (a) z_i^+ the embedding feature of another data augmentation of x_i; (b) $\mathrm{t} \in P(i)$ the embedding feature of samples which have the same label with x_i in A.

As shown in Eq. (2), supervised contrastive loss treats these two kinds of positives equally. However, pulling the anchor with its another data augmentation or with the samples from the same semantic class together is trying to capture different kinds of features for visual representation Wei et al. (2020), *i.e.*, task-agnostic appearance features and task-specific semantic features, respectively. $|P(i)| \approx \frac{n_{\mathrm{y}_i}}{n}|A|$, in a balanced dataset, $n_1 \approx n_2 \approx \cdots \approx n_K$. Therefore, different classes have almost the same $|P(i)|$ and the target distribution. In an imbalanced dataset, $|P(i)|$ of different classes can be totally

different, *e.g.*, the head classes have a large $|P(i)|$ while $|P(i)|$ in the tail classes is small. In other word, $\hat{p}(\mathrm{z}_i^+|\mathrm{z}_i)$ in the head classes is much smaller than $\hat{p}(\mathrm{z}_i^+|\mathrm{z}_i)$ in the tail classes. Consequently, different classes can learn different kinds of features, *i.e.*, the head classes focus on the task-specific semantic features while the tail classes focus on the task-agnostic appearance features as shown in Fig. 4, leading to inconsistency between the learned features of the head classes and the tail classes.

To address aforementioned problem, we re-design the target distribution, where for any class, $\hat{p}_{dscl}(\mathrm{z}_i^+|\mathrm{z}_i)$ is the same. Given a pre-defined $\hat{p}_{dscl}(\mathrm{z}_i^+|\mathrm{z}_i) = \alpha, \alpha \in [0, 1]$, the decoupled target distribution and decoupled contrastive loss can be written as,

$$\hat{p}_{dscl}(\mathrm{t}|\mathrm{z}_i) = \begin{cases} \alpha, & \mathrm{t} = \mathrm{z}_i^+ \\ \dfrac{1-\alpha}{|P(i)|}, & \mathrm{t} \in P(i) \\ 0, & \mathrm{t} \in A \setminus P(i) \end{cases} \tag{8}$$

$$\mathcal{L}_{dscl} = \sum_{i=1}^{|\mathcal{D}|}(-\alpha \log p(\mathrm{z}_i^+|\mathrm{z}_i) + \frac{1-\alpha}{|P(i)|}\sum_{\mathrm{t}\in P(i)} - \log p(\mathrm{t}|\mathrm{z}_i))$$

In the decoupled target distribution, decoupling the two kinds of positives ensures that the weights of two kinds positives are not influenced by the imbalanced datasets, preventing the model from capturing different kinds of features for the head classes and tail classes. Experiments in Section 4.2 show that this simple modification can strongly improve the performance of the model on long-tailed classification.

3.3 Patch-based Self Distillation

To mine knowledge encoded in the visual patterns, inspired by the part-based methods in fine-grained image recognition Zhang et al. (2014); Quan et al. (2019); Sun et al. (2018), we introduce the image patch cropping, *i.e.*, besides the global view of an image, we crop multi small patches from the global view and extract their feature embeddings. The feature embeddings of these small parts denote as $\{\mathrm{s}_i^j\}_{j=1}^L$, where L denotes the number of small parts, and the coordinates of these patches in the global view denote as $\{\mathrm{B}_i^j\}_{j=1}^L$.

To calculate the target distribution of $\mathcal{L}_{pbsd}$, we use ROI pooling He et al. (2017) to get feature from the feature map of the global view according to $\{\mathrm{B}_i^j\}_{j=1}^L$. The pooled features after the projection head denote as $\{\mathrm{c}_i^j\}_{j=1}^L$. The probability of the target distribution can be calculated by,

$$q_{pbsd}(\mathrm{t}, \mathrm{s}_i^j) = p(\mathrm{t}|\mathrm{c}_i^j) = \frac{\exp(\mathrm{t}\cdot \mathrm{c}_i^j/\tau)}{\exp(\mathrm{z}_i^+ \cdot \mathrm{c}_i^j/\tau) + \sum_{\mathrm{m}\in A}\exp\left(\mathrm{m}\cdot \mathrm{c}_i^j/\tau\right)}. \tag{9}$$

With the target distribution, the PBSD loss can be calculated by,

$$\mathcal{L}_{pbsd} = \sum_{i=1}^{|\mathcal{D}|}\frac{1}{L}\sum_{j=1}^{L}\sum_{\mathrm{t}\in\{\mathrm{z}_i^+\cup A\}} -p(\mathrm{t}|\mathrm{c}_i^j)\log p(\mathrm{t}|\mathrm{s}_i^j), \tag{10}$$

note that $p(\mathrm{t}|\mathrm{c}_i^j)$ is detached from the computation graph to block the gradient.

A small part of an object contains visual pattern that can be shared by other classes, *e.g.*, wheels for the class 'truck' and class 'bus'. PBSD crops multi small patches from the global view of an image to extract semantic-related visual patterns. $q(\mathrm{t}, \mathrm{s}_i^j)$ is calculated to mine relationship of the shared patterns between different images. Minimizing Eq. (10) maintains the shared patterns between different images, thus transferring knowledge and mitigating the under-representation of the tailed classes. Experiments in Section 4.2 validate that the introduction of the image patch cropping is vital to the performance. Only using global view of the image with different color data augmentations to perform Eq. (10) has a poor performance compared with our implementation.

Table 1: Effectiveness of each component in our method. Supervised contrastive loss is used as baseline. * denotes using global views instead of the small patches to calculate Eq. (10).

Settings	Baseline	+ DSCL	+ PBSD	+ Both*	+ Both
Many	61.6	63.4	67.2	67.2	**68.5**
Medium	48.6	50.0	53.7	53.9	**55.2**
Few	30.3	31.4	34.6	33.7	**35.4**
Overall	51.2	52.6	56.3	56.2	**57.7**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We use two popular datasets for long-tailed recognition as follows: 1) **ImageNet-LT** Liu et al. (2019) contains 115,846 training images of 1,000 classes sampled from the ImageNet1K Russakovsky et al. (2015), with class cardinality ranging from 5 to 1,280. 2) **iNaturaList 2018** Van Horn et al. (2018) is a real-world long-tailed dataset with 437,513 training images of 8,142 classes, with class cardinality ranging from 2 to 1,000.

Evaluation Metrics. We follow the standard evaluation metrics that evaluating our models on the testing set and reporting the overall top-1 accuracy across all classes. To give a detailed analysis, we follow Liu et al. (2019) that groups the classes into splits according to their number of images: Many ($>$ 100 images), Medium (20 - 100 images), and Few ($<$ 20 images).

Implementation Details. For fair comparison, we follow the implementations of TSC Li et al. (2021b) and KCL Kang et al. (2020) that train the representation backbone at the first stage and train the linear classifier with the freezed learned backbone at the second stage. All experiments are implemented on PyTorch Paszke et al. (2019). We adopt ResNet-50 He et al. (2016) as backbone for all experiments. The α in Eq. (8) is set as 0.1 and the loss weight λ_1 in Eq. (7) is 1.5.

At the first stage, the basic framework is the same as MoCoV2 Chen et al. (2020b), the momentum value for the updating of ema model is 0.999, the temperature τ is 0.07, the size of memory queue A is 65536, and the output dimension of projection head is 128. The data augmentation is the same as MoCoV2 Chen et al. (2020b). We crop the small patches from the global view randomly with the crop scale (0.05, 0.6) and resize them to 64. The number of small parts L is 5. SGD optimizer is used with a learning rate decays by cosine scheduler from 0.1 to 0 with batch size 256 on 2 Nvidia RTX 3090 in 200 epochs. At the second stage, the parameters are the same as Li et al. (2021b). The linear classifier is trained for 40 epochs with CE loss and class-balanced sampling Kang et al. (2019) with batch size 2048 using SGD optimizer. The learning rate is initialized as 10, 30 for ImageNet-LT, and iNaturaList 2018, respectively, and multiplied by 0.1 at epoch 20 and 30. For testing, the image is first resized to 256 and *center cropping* is performed to get image of size 224.

4.2 ABLATION STUDY

Components analysis. We analyze the effectiveness of each proposed component on ImageNet-LT in Table 1. Supervised contrastive loss Khosla et al. (2020) is used as baseline. Compared with the supervised contrastive loss baseline, changing the target distribution to the decoupled form as in Eq. (8) improves the top-1 accuracy by 1.4%, which is already higher than the recent proposed contrastive learning based method TSC Li et al. (2021b). Different from many methods for long-tailed classification that improve the performance of tail classes but sacrifice the head classes performance, our proposed PBSD can transfer knowledge between classes, improving the performance of both head and tail classes. Both DSCL and PBSD are important to the final performance and the combination of them achieves the best performance. The introduction of small patches is important to the PBSD. We conduct experiment by performing different color augmentations on the same global view and use them to calculate Eq. (10). It decreases the overall accuracy by about 1.5%. It can be concluded that each component in our method is effective in boosting performance.

The impact of α in Eq. (8) is investigated in Fig. 3. α determines the weight of pulling the anchor with its data augmented one. Setting $\alpha = 0$ means only pulling the anchor with other images from the same class. Without pulling the anchor with its data augmented one together decreases the accuracy from 57.7% to 56.8%, showing the importance of involving two kinds of positives. In addition, this

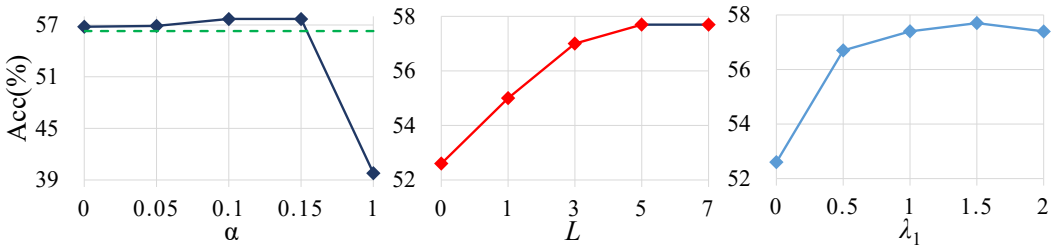

Figure 3: Evaluation of α in Eq. (8), L the number of small patches, and λ_1 the loss weight.

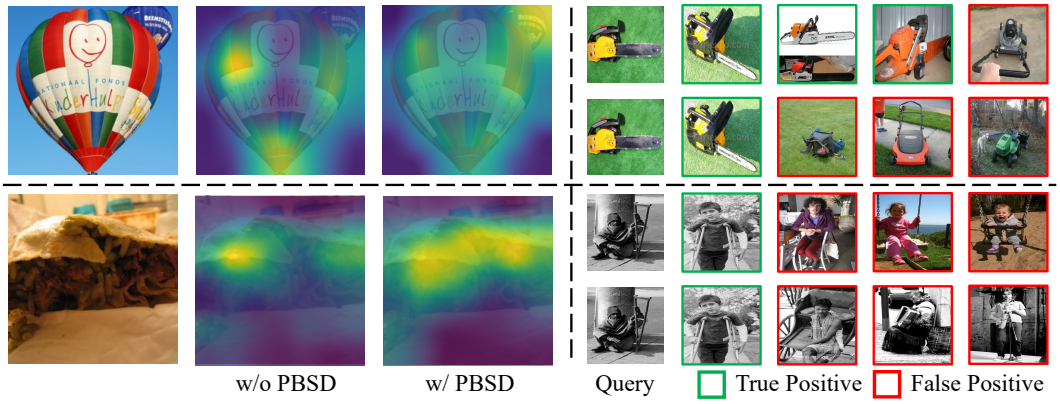

Figure 4: **Left:** CAM Zhou et al. (2016) visualizations of our method with/without PBSD. **Right:** examples of retrieval results with the features generated by model trained with DSCL and supervised contrastive learning. In each part, the first row is the results of model trained with DSCL and the second row is the results of the compared baseline.

setting still outperforms supervised contrastive loss (green dotted line in the figure), indicating that preventing the head classes and tail classes from learning different kinds of features is important. When $\alpha = 1$, the loss degenerates into the self-supervised loss. The accuracy is only 39.8% because of the lack of label information. We set α as 0.1, which gets the best performance.

The impact of the number of small patches is shown in Fig. 3. The model benefits from cropping more small patches from the global view. The top-1 accuracy improves from 55.0% to 57.7% when increasing L from 1 to 5. We set L as 5 for a good trade-off between training consuming and accuracy.

The impact of the loss weight λ_1 is shown in Fig. 3. It can be concluded that PBSD is important to the final performance and this parameter is easy to tune. We set it as 1.5 for its good performance.

Visualization analysis is conducted in Fig. 4. For the CAM Zhou et al. (2016) visualization, the CAM of the model without PBSD focuses on the small parts while PBSD makes the model capture more useful information, validating that with PBSD model learns more shared patterns between classes. For the visualization of retrieval results, the query images are both from the tail classes. In the first example, the class of query is 'chainsaw', but the retrieval results of baseline are related to the appearance features like green color. In the second example, the query image is a gray image with class 'crutch', the retrieval results of the baseline are all gray images with a person inside, not the semantic features related to 'crutch'. Although our method also has some false positives, these false positives are more related to semantic class 'crutch', *e.g.*, sticks held in hands. It is clear that supervised contrastive loss makes the features of tail classes mainly focus on the visual appearance feature. The visualization results further validate the effectiveness of our method.

4.3 Comparison with Recent Works

We compare our method with recent works on ImageNet-LT Liu et al. (2019), and iNaturaList 2018 Van Horn et al. (2018). The compared methods include the re-balancing methods Liu et al.

Table 2: Comparison with recent methods on ImageNet-LT, and iNaturaList 2018. CE denotes training the model with the cross entropy loss. † denotes the model is trained with RandAugment Cubuk et al. (2020) and 400 epochs, which has potential unfair comparison with our method.

Methods	Reference	ImageNet-LT				iNaturaList 2018			
		Many	Medium	Few	Overall	Many	Medium	Few	Overall
CE	-	64.0	33.8	5.8	41.6	72.2	63.0	57.2	61.7
OLTR Liu et al. (2019)	CVPR2019	35.8	32.3	21.5	32.2	59.0	64.1	64.9	63.9
Balanced Ren et al. (2020)	NeurIPS2020	61.1	47.5	27.6	50.1	-	-	-	-
τ-norm Kang et al. (2019)	ICLR2020	56.6	44.2	27.4	46.7	65.6	65.3	65.9	65.6
cRT Kang et al. (2019)	ICLR2020	58.8	44.4	26.1	47.3	69.0	66.0	63.2	65.2
LWS Kang et al. (2019)	ICLR2020	57.1	45.2	29.3	47.7	65.0	66.3	65.5	65.9
Logit Adjust Menon et al. (2020)	ICLR2020	-	-	-	50.4	-	-	-	68.4
DisAlign Zhang et al. (2021a)	CVPR2021	59.9	49.9	31.8	51.3	-	-	-	67.8
MetaSAug Li et al. (2021a)	CVPR2021	-	-	-	47.4	-	-	-	68.8
RSG Wang et al. (2021)	CVPR2021	-	-	-	-	-	-	-	67.9
BatchFormer Hou et al. (2022)	CVPR2022	61.4	47.8	33.6	51.1	-	-	-	-
KCL Kang et al. (2020)	ICLR2020	61.8	49.4	30.9	51.5	-	-	-	68.6
PaCo† Cui et al. (2021)	ICCV2021	65.0	55.7	38.2	57.0	73.2	70.3	73.2	73.6
TSC Li et al. (2021b)	CVPR2022	63.5	49.7	30.4	52.4	72.6	70.6	67.8	69.7
Our	This paper	68.5	55.2	35.4	57.7	74.2	72.9	70.3	72.0
CBD Iscen et al. (2021)	BMVC2021	68.5	52.7	29.2	55.6	-	-	-	73.6
RIDE Wang et al. (2020)	ICLR2021	-	-	-	55.4	70.9	72.4	73.1	72.6
ACE Cai et al. (2021)	ICCV2021	-	-	-	54.7	-	-	-	72.9
NCL† Li et al. (2022)	CVPR2022	-	-	-	59.5	72.7	75.6	74.5	74.9
Our + RIDE	This paper	70.1	57.5	37.7	59.7	74.8	74.9	73.1	74.2

(2019); Ren et al. (2020), the decoupling methods Kang et al. (2019); Menon et al. (2020); Zhang et al. (2021a), the transfer learning based methods Li et al. (2021a); Zhong et al. (2021); Hou et al. (2022), the methods that extend the supervised contrastive learning Kang et al. (2020); Li et al. (2021b); Cui et al. (2021), and the ensemble-based methods (Iscen et al., 2021; Cai et al., 2021; Li et al., 2022).

As shown in Table 2, directly using standard cross entropy loss suffers a low performance on the tail classes. Most long-tailed recognition methods can improve the performance of the methods on the testing set. However, the accuracy of Many split is sacrificed. Compared with the re-balancing methods, decoupling methods adjust the classifier after the training of the backbone and achieve a better performance, showing the effectiveness of the two-stage training. Transfer learning based-methods improve the performance of the model with less sacrifice of head classes performance, *e.g.*, BatchFormer has higher accuracy on Many split compared with DisAlign which has the same overall accuracy with it. The supervised contrastive learning based methods achieve better performance with less sacrifice of head classes accuracy and higher overall accuracy. Our method achieves the best 57.7% overall accuracy on ImageNet-LT, even outperforming PaCo Cui et al. (2021) that uses stronger data augmentation and twice training epochs. Our method is the only one that can improve the performance of all the head, medium and few classes without introducing stronger data augmentation or multi-model ensemble.

Our method is easy to use and can also be combined with ensemble-based method to further boost its performance. Combined with RIDE Wang et al. (2020), our method achieves 59.7% overall accuracy on ImageNet-LT, outperforming all the compared ensemble-based method. In summarize, our method shows superior performance among compared methods.

5 Discussion and Conclusion

This paper addresses the long-tailed recognition. Our proposed method has shown impressive improvements in long-tailed recognition. This is achieved by the decoupled supervised contrastive loss that mitigates the influence of the imbalanced dataset by designing a more reasonable target distribution and patch-based self distillation that transfers knowledge between classes by introducing useful similarity relationship knowledge into contrastive learning. Experiments on several long-tailed recognition benchmarks demonstrate the effectiveness of the proposed method.

REFERENCES

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *ICML*, pp. 872–881. PMLR, 2019.

Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV*, pp. 112–121, 2021.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 32, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607. PMLR, 2020a.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pp. 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, pp. 702–703, 2020.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, pp. 715–724, 2021.

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.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *CVPR*, pp. 3299–3309, 2021.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pp. 5356–5364, 2019.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pp. 1735–1742. IEEE, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pp. 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.

Zhi Hou, Baosheng Yu, and Dacheng Tao. Batchformer: Learning to explore sample relationships for robust representation learning. *arXiv preprint arXiv:2203.01522*, 2022.

Ahmet Iscen, André Araujo, Boqing Gong, and Cordelia Schmid. Class-balanced distillation for long-tailed visual recognition. *arXiv preprint arXiv:2104.05279*, 2021.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *ICLR*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *CVPR*, pp. 6949–6958, 2022.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 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 *CVPR*, pp. 5212–5221, 2021a.

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. *arXiv preprint arXiv:2111.13998*, 2021b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

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 *CVPR*, pp. 2537–2546, 2019.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pp. 3431–3440, 2015.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037, 2019.

Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *ICCV*, pp. 3750–3759, 2019.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 33:4175–4186, 2020.

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.

Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*, pp. 467–482. Springer, 2016.

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pp. 480–496, 2018.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

Jianfeng Wang, Thomas Lukasiewicz, Xiaolin Hu, Jianfei Cai, and Zhenghua Xu. Rsg: A simple but effective module for learning imbalanced datasets. In *CVPR*, pp. 3784–3793, 2021.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

Longhui Wei, Lingxi Xie, Jianzhong He, Jianlong Chang, Xiaopeng Zhang, Wengang Zhou, Houqiang Li, and Qi Tian. Can semantic labels assist self-supervised visual representation learning? *arXiv preprint arXiv:2011.08621*, 2020.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pp. 1492–1500, 2017.

Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*, pp. 834–849. Springer, 2014.

Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, pp. 2361–2370, 2021a.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021b.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, pp. 16489–16498, 2021.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pp. 2921–2929, 2016.

Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *ICCV*, pp. 10306–10315, 2021.

A APPENDIX

This Appendix provides more details of experiments and analysis on each proposed component.

Ablation Study on Different Imbalanced Ratios of Datasets. We conduct more experiments to further validate the effectiveness of the proposed method across different imbalanced ratios. We generate 2 datasets with different imbalanced ratios from the ImageNet1K following the Pareto distribution. The detailed statistics of the generated datasets and the experimental results are shown in Table 3 and Table 4, respectively. As shown in the Table 4, our proposed DSCL and PBSD generalize well on different imbalanced ratios. Both of them can bring performance improvement.

Table 3: Dataset statistics on training instance numbers, including the maximal and minimal instance number per class, the number of total training instances, and the imbalanced ratio.

	Max	Min	Total	Imbalanced Ratio
ImageNetLT	1280	5	115846	256
Dataset A	857	10	115852	85.7
Dataset B	343	37	115801	9.27

Ablation Study on Different Backbones. It is also important for the method to generalize well on different backbones, we thus conduct experiments that use ResNeXt50 Xie et al. (2017) as backbone. ResNeXt50 is another commonly used backbone in Long-tailed recognition. The results are shown in Table 5. It is clear that our proposed components are also effective on ResNeXt50.

Ablation Study on DSCL. We conduct more experiments to validate that DSCL is a reasonable choice to remove the bias of the supervised contrastive loss. As discussed in Section.3.1, the bias of the supervised contrastive loss is caused by the imbalanced A. Therefore, one possible solution is to maintain a balanced memory queue A. The result is shown in the second row of Table 6. The balanced memory queue does not bring performance improvement and is even harmful to the performance of the tail classes. The reason is that the balanced memory queue leads to more negative gradients to the

Table 4: Ablation study of each component in our method on datasets with different imbalanced ratios. Supervised contrastive loss Khosla et al. (2020) is used as baseline.

	ImageNetLT	Dataset A	Dataset B
Baseline	51.2	53.2	56.1
+ DSCL	52.6	54.7	57.3
+ PBSD	56.3	59.2	61.8
+ Both	57.7	59.6	62.7

Table 5: Ablation study of each component in our method on different backbones. Supervised contrastive loss Khosla et al. (2020) is used as baseline.

	ResNet50	ResNeXt50
Baseline	51.2	51.8
+ DSCL	52.6	53.2
+ PBSD	56.3	57.7
+ Both	57.7	58.7

Table 6: Ablation study on DSCL. Top-1 accuracy is used as metric. Balanced Queue denotes A is a balanced memory queue. Re-weighting denotes the loss weight is added as in Cui et al. (2019).

Settings	Many	Medium	Few	Overall
Baseline	61.6	48.6	30.3	51.2
Balanced Queue	62.3	48.9	29.2	51.4
Re-weighting	59.7	45.9	30.4	49.1
DSCL	63.4	50.0	31.4	52.6

tail classes, *i.e.*, the tail classes receive too many gradients to push them away from other samples in the feature space. While in the original imbalanced memory queue, the number of tail class samples is small, preventing too many negative gradients on the tail classes. Re-weighting is a commonly used method to remove the bias. We add the loss weight as in Cui et al. (2019),

$$\begin{aligned}\mathcal{L}_{scl} &= \sum_{i=1}^{|\mathcal{D}|} \frac{\omega_{\mathrm{y}_i}}{|P(i)|+1} \sum_{\mathrm{t}\in\{\mathrm{z}_i^+\cup P(i)\}} -\log \frac{\exp\left(\mathrm{t}\cdot \mathrm{z}_i/\tau\right)}{\exp\left(\mathrm{z}_i^+\cdot \mathrm{z}_i/\tau\right)+\sum_{\mathrm{m}\in A}\exp\left(\mathrm{m}\cdot \mathrm{z}_i/\tau\right)}, \\ \omega_{\mathrm{y}_i} &= \frac{1-\beta}{1-\beta^{n_{\mathrm{y}_i}}}\end{aligned} \tag{11}$$

where $\beta \in [0, 1)$ is a hyper-parameter, and n_{y_i} is the number of samples in the class y_i. The result shows that re-weighting method decreases the performance. The same phenomenon is also validated in Kang et al. (2019) that the re-weighting leads to a poor discriminative feature space. It can be concluded that DSCL is a reasonable choice to remove the bias of the supervised contrastive loss.