Long-Tailed Recognition by Mutual Information Maximization between Latent Features and Ground-Truth Labels

Min-Kook Suh [1] **Seung-Woo Seo** [1]

Abstract

Although contrastive learning methods have shown prevailing performance on a variety of representation learning tasks, they encounter difficulty when the training dataset is long-tailed. Many researchers have combined contrastive learning and a logit adjustment technique to address this problem, but the combinations are done ad-hoc and a theoretical background has not yet been provided. The goal of this paper is to provide the background and further improve the performance. First, we show that the fundamental reason contrastive learning methods struggle with long-tailed tasks is that they try to maximize the mutual information between latent features and input data. As ground-truth labels are not considered in the maximization, they are not able to address imbalances between classes. Rather, we interpret the long-tailed recognition task as a mutual information maximization between latent features and ground-truth labels. This approach integrates contrastive learning and logit adjustment seamlessly to derive a loss function that shows state-of-the-art performance on long-tailed recognition benchmarks. It also demonstrates its efficacy in image segmentation tasks, verifying its versatility beyond image classification. Code is available at https://github.com/bluecdm/Long-tailed-recognition.

1. Introduction

A supervised classification task has been an active research topic for decades. However, its performance is still unsatisfactory when the dataset shows a long-tailed distribution, where a few classes, i.e., head classes, contain a major number of samples while the remaining classes, i.e., tail classes, have only a small number of samples. The most straightforward remedy to this problem is to rebalance the training dataset through weighted sampling. However, it is a suboptimal strategy that may be detrimental to the accuracy of the head classes (Wang et al., 2017).

Recently, contrastive learning (Oord et al., 2018; Chen et al., 2020a; Khosla et al., 2020) is widely used in representation learning and showing state-of-the-art performance on various tasks. The contrastive learning framework learns representations by pushing latent features from the same sample closer and separating them from different samples. However, its performance degrades on long-tailed datasets as the samples are imbalanced (Cui et al., 2021). Therefore, several works have been conducted to adapt it to the long-tailed recognition tasks (Cui et al., 2021; Zhu et al., 2022) by combining it with logit adjustment techniques (Ren et al., 2020; Menon et al., 2021), which is another method to solve the long-tailed recognition task that modulates the prediction score of networks based on the appearance frequency of classes. Although previous methods empirically show that combining contrastive learning and logit adjustment boosts performance, they do not provide the theoretical meaning of the combination.

In this paper, we describe the theoretical meaning of the combination and provide an improved method for combining contrastive learning and logit adjustment. We find that the performance of previous contrastive learning methods degrade on long-tailed datasets because they aim to maximize the mutual information (MI) between the latent features and input data. These approaches do not consider the imbalance of label distribution, as the ground-truth label is not involved in the maximization. Instead, we propose maximizing the MI between latent features and ground-truth labels, allowing the consideration of label distribution.

By replacing the input data term with the ground-truth label term, we derive a general loss function that encompasses various other methods. The derived loss function includes a likelihood of latent feature term and a prior of classes term, and different ways of modeling these terms lead to different previous methods. For example, the softmax cross-entropy loss for a balanced dataset and the logit adjustment

[1]Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea. Correspondence to: Min-Kook Suh <bluecdm@snu.ac.kr>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

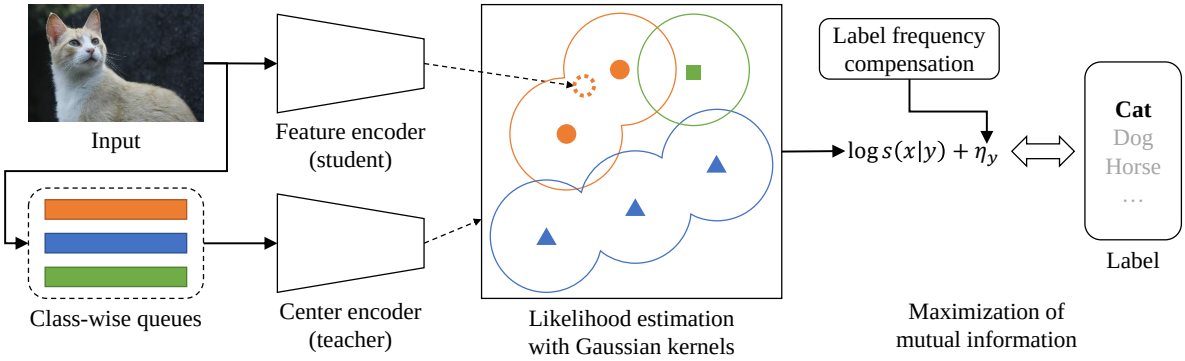

Figure 1. We address the long-tailed recognition by maximizing the mutual information of latent features, x, and ground-truth labels, y. We solve the maximization problem by dividing it into two terms: an unnormalized likelihood term, $s(x|y)$, and a logit-adjustment term to compensate for label frequency, η_y. η_y is a value that depends on the frequency of the label, and $s(x|y)$ is estimated using a neural network. The proposed loss is achieved by estimating $s(x|y)$ with Gaussian kernels using latent features of other samples.

term for an imbalanced dataset (Ren et al., 2020) can be derived by modeling the likelihood as an isotropic Gaussian. Supervised contrastive loss (Khosla et al., 2020) can be derived under the assumption of a balanced training dataset by estimating the likelihood using a sampling-based kernel density estimation with Gaussian kernels.

By removing the assumption of a balanced dataset, we derive the proposed loss function that seamlessly integrates the contrastive learning and logit adjustment. Since the kernel density estimation results in a Gaussian mixture likelihood, we refer to the loss as Gaussian Mixture Likelihood (GML) loss. We also provide an efficient method for modeling the Gaussian mixture likelihood, as shown in Fig. 1. We use contrast samples to estimate the likelihood. Similar to Momentum Contrast (MoCo) (He et al., 2020), we use queues to store contrast samples. However, because a long-tailed dataset is being handled, tail classes do not have sufficient samples to create the Gaussian mixture. To resolve this problem, we use multiple class-wise queues rather than a single queue of MoCo. However, in the case of tail classes, the update frequencies of contrast samples are significantly longer. As a result, old samples of tail classes' queues are generated by highly outdated encoders. Therefore, we use class-wise queues along with a teacher–student strategy. Unlike a momentum encoder used in MoCo, we use a pre-trained teacher to generate contrast samples.

We evaluate the proposed method on various long-tailed recognition datasets: ImageNet-LT (Liu et al., 2019), iNaturalist 2018 (Van Horn et al., 2018), and CIFAR-LT (Cui et al., 2019); the proposed method surpasses all previous long-tailed recognition methods. In addition, as the proposed method is related to supervised contrastive learning and knowledge distillation, we compare our method with them on both balanced and imbalanced datasets. Unsurprisingly, our method shows superior performance to them on imbalanced datasets, exhibiting comparable performance on balanced datasets. Finally, we evaluate our method on a commonly used semantic segmentation dataset ADE20K (Zhou et al., 2017), which also contains imbalanced pixel labels. Simply replacing the cross-entropy loss with the proposed loss also boosts the performance, indicating the proposed framework can be extended beyond a classification task.

Our main contributions are summarized as follows.

- We show that the fundamental limitation of contrastive learning methods on long-tailed tasks comes from directly maximizing the MI between latent features and input data. Instead, we propose to tackle the long-tailed recognition by MI maximization between latent features and ground-truth labels.
- While previous methods have combined contrastive learning and the logit adjustment without investigating a theoretical background, we find that contrastive learning implies a Gaussian mixture likelihood and the logit adjustment is derived from the prior of classes.
- We propose an efficient way to model the Gaussian mixture likelihood using a teacher–student framework that demonstrates its superiority in various long-tailed recognition tasks.

2. Related Work

2.1. Long-Tailed Recognition

Rebalancing Datasets. Since the most straightforward approach to the long-tailed recognition problem is rebalancing the dataset during training, several early works (Chawla et al., 2002; Japkowicz & Stephen, 2002; Drummond et al., 2003; Han et al., 2005; He & Garcia, 2009; Mikolov et al., 2013) have focused on rebalancing approaches. However, Byrd & Lipton (2019) found that the effect of rebalancing diminishes on overparameterized deep neural networks given sufficiently long training epochs.

Normalized Classifier. Networks trained on long-tailed datasets tend to have biased classifier weights (Alshammari et al., 2022); this problem can be alleviated by normalizing the weights of the classification layer. Kang et al. (2020) proposed to normalize the weights by decoupling the representation learning and classification. They first trained the network jointly using an instance-based sampling strategy. Then, they retrained the classifier only using a class-balanced sampling strategy. Gidaris & Komodakis (2018) proposed the use of a cosine similarity classifier instead of a dot-product classifier. It bypasses the biased weights problem by only considering the relative angle. Accordingly, we adopt the cosine similarity as the classifier of our network.

Logit Adjustment. Another approach to the long-tailed recognition problem is to modulate the logit values. Interestingly, Ren et al. (2020) and Menon et al. (2021) derived similar results using different approaches. Ren et al. (2020) showed that the softmax loss is a biased estimator and proposed a Balanced Softmax loss; Menon et al. (2021) proposed a post-hoc logit adjustment and a logit-adjusted softmax cross-entropy loss. Both works show that adding a logit-adjustment term proportional to the logarithm of label frequency is essential to the long-tailed recognition. In accordance with their results, many studies (Cui et al., 2021; Feng et al., 2021; Hong et al., 2021; Zhu et al., 2022) have included logit adjustment as a part of their methods.

2.2. Contrastive Learning

Cui et al. (2021) found that the performance of supervised contrastive loss (Khosla et al., 2020) significantly degrades when it is applied to a long-tailed dataset. Therefore, they proposed Parametric Contrastive learning (PaCo), which rebalances the contrast samples by adding parametric class-wise learnable centers in the samples. To integrate the logit-adjustment technique into their method, the authors added the adjustment term to the center learning. Further, Zhu et al. (2022) proposed Balanced Contrastive Learning (BCL), which utilizes the number of contrasts of each class in a mini-batch to balance the gradient contribution of classes. To integrate the logit-adjustment technique, they used a weighted sum of the logit-adjustment loss and their loss.

3. Proposed Method

In this section, we describe our approach to handle long-tailed recognition based on MI. Because the MI is intractable, we use InfoNCE loss (Oord et al., 2018) to maximize its lower bound. We adopt the notations of Poole et al. (2019) to express InfoNCE loss.

$$I(X;Y) \geq \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{\exp f(x_i,y_i)}{\frac{1}{K}\sum_{j=1}^{K}\exp f(x_i,y_j)}\right] \quad (1)$$

where the expectation is taken over subsets of a dataset and K denotes the size of the subset. Equality holds when $f(x,y) = \log p(x|y) + c(x)$ and $K \to \infty$ where $c(x)$ is a function that only depends on x.

3.1. Contrastive Learning as MI Maximization between Latent Features and Input Data

From Eq. 1, we can recover the loss functions of contrastive learning methods (Chen et al., 2020b; He et al., 2020) by substituting x_i with the query feature, $f_q(t_q(x_i), w)$, and y_j with the input image, x_j; it shows that contrastive learning methods maximize the MI between latent features and input data. Detailed proof is shown in Appendix A.1.

Since ground-truth label term is not included in the maximization, they are prone to the imbalance of label frequency. Supervised contrastive learning (Khosla et al., 2020) tries to modify the loss function to include the ground-truth label term, but it is still based on the MI maximization between latent features and input data.

3.2. Long-Tailed Recognition by MI Maximization between Latent Features and Ground-Truth Labels

To enable the label frequency considered in the MI maximization process, we formulate the long-tailed recognition problem as a maximum MI problem of latent features and ground-truth labels. However, replacing the input data term with the ground-truth label term results in a loss function that is not contrastive learning. In this section, we describe procedures to recover the contrastive loss.

Logit Adjustment. First, we substitute X in Eq. 1 with the set of latent features and Y with the set of ground-truth labels to represent the MI maximization between latent features and ground-truth labels.

$$\begin{aligned}\mathop{\mathbb{E}}_{i}\log\frac{\exp f(x_i,y_i)}{\mathbb{E}_j\exp f(x_i,y_j)} &= \mathop{\mathbb{E}}_{i}\log\frac{\exp f(x_i,y_i)}{\sum_{c\in C}\exp f(x_i,c)p(c)}\\ &= \mathop{\mathbb{E}}_{i}\log\frac{\exp(f(x_i,y_i)+\eta_{y_i})}{\sum_{c\in C}\exp(f(x_i,c)+\eta_c)} - \eta_{y_i} \quad (2)\end{aligned}$$

We realize the above term by increasing K to the size of the entire training dataset. Here, x_i and y_i denote the latent feature and ground-truth label of the i-th sample, respectively. C denotes the set of all classes and $\eta_c = \log p(c)$ denotes the logit-adjustment term, which is the logarithm of the appearance frequency of a class.

Eq. 2 is a general template and various previous methods can be derived by different modeling of the likelihood and prior. For simplicity, we define $s(x|y) = \exp f(x,y)$ and use s to denote the unnormalized likelihood of latent features. As an example, we derive a softmax cross-entropy loss with the logit-adjustment term by defining $s(x|y) = \exp(w_y \cdot x + b_y)$, a dot-product classifier.

Gaussian Mixture Likelihood Loss. The inequality on Eq. 1 becomes tighter as f approaches $\log p(x|y)$, hence, we choose to estimate s using sampling-based kernel density estimation with Gaussian kernels. This estimation leads to a Gaussian mixture likelihood, where the centers of the Gaussian mixtures are the contrast samples.

$$s(x|y) = \exp f(x,y) = \frac{1}{\|Z_y\|} \sum_{z_y \in Z_y} \exp(z_x \cdot z_y / \tau_g) \quad (3)$$

where z_x denotes the L2 normalized query feature of x, Z_y denotes a set of L2 normalized contrast features of class y, and τ_g denotes a temperature parameter for GML loss, which is quadratically proportional to the variance of the Gaussian mixture. The subscript i of x_i and y_i is omitted for simplicity. Gaussian mixture is represented using dot product instead of a quadratic term to maintain consistency with previous contrastive losses. It does not modify the meaning of Gaussian mixture, as L2 normalization is applied to both the query and contrast features. Specifically,

$$-\|z_x - z_y\|_2^2/(2\tau_g) = z_x \cdot z_y/\tau_g - 1/\tau_g \quad (4)$$

and the last constant term cancels out in the following equation.

Note that we do not estimate the centers of Gaussian mixtures, but simply use the latent features of the contrast encoder as centers. Therefore, the training burden is not increased by this procedure and remains almost the same as that of previous contrastive learning methods.

By modeling s using the Gaussian mixture, we derive the GML loss which seamlessly integrates the contrastive learning and the logit compensation. Substituting f of Eq. 2 using Eq. 3 derives the proposed loss L_{GML}.

$$-L_{GML} = \log \frac{\exp\left[\log \frac{1}{\|Z_y\|} \sum_{z_y \in Z_y} \exp(z_x \cdot z_y/\tau_g) + \eta_y\right]}{\sum_{c \in C} \exp\left[\log \frac{1}{\|Z_c\|} \sum_{z_c \in Z_c} \exp(z_x \cdot z_c/\tau_g) + \eta_c\right]} \quad (5)$$

3.3. Training with GML Loss

Class-wise Queues for Contrast Samples. The denominator of L_{GML} requires at least one contrast sample for each class. However, the strategy of MoCo (He et al., 2020) does not guarantee any minimum number of samples, because it uses a queue of randomly sampled contrast samples. Therefore, we use multiple queues with different lengths. Each class has one assigned queue, and its length is proportional to the frequency of the class plus a predefined constant.

$$\|Z_c\| = k_m + (k - k_m \times \|C\|) \times p(c) \quad (6)$$

where k_m denotes the minimum length of the queue and k denotes the total number of contrast samples.

Teacher–Student Strategy. Maintaining class-wise queues causes another problem. MoCo (He et al., 2020) uses a momentum encoder to generate contrast samples, and they are stored in a queue. Therefore, old samples generated using an outdated encoder are replaced with new samples. On the other hand, we use multiple class-wise queues and their update frequency is proportional to the ratio of the classes in the training dataset. Therefore, queues of tail classes have excessively long update frequencies and old samples of their queues are generated by highly outdated encoders. To overcome this problem, we adopt a teacher–student strategy and use a pre-trained teacher encoder to generate contrast samples.

Training Procedure. Usually, a contrastive loss is used to train a backbone network, and a classifier layer is trained separately after training the backbone network. However, in the supervised setting, the classifier layer can be trained simultaneously with the backbone network. Therefore, we attach a cosine similarity classifier to the network and train them simultaneously. The classifier is trained without contrastive loss using the following loss function.

$$-L_{cls} = \log \frac{\exp(m_x \cdot m_y/\tau_s + \eta_y)}{\sum_{c \in C} \exp(m_x \cdot m_c/\tau_s + \eta_c)} \quad (7)$$

where m_x denotes L2 normalized x, m_c denotes L2 normalized weight at class c of the classifier, and τ_s denotes a temperature hyperparameter for softmax cross-entropy loss. In contrast to the Gaussian mixture setting, m_c is a parameter that needs to be trained.

In addition, we use MLP encoders followed by L2 normalization for contrast samples z_x and z_y, similar to previous contrastive losses (Chen et al., 2020a;c), while a simple L2 normalization is used for the cosine similarity classifier.

$$z_x = \frac{\text{MLP}(x)}{\|\text{MLP}(x)\|}, \; m_x = \frac{x}{\|x\|} \quad (8)$$

In summary, our training procedure is as follows. First, we estimate the ratio of classes in the training dataset to calculate η_c. Then, we train a teacher network with a cosine similarity classifier with L_{cls} without contrastive loss. Finally, we train a student network and a cosine similarity classifier simultaneously with L_{GML} and L_{cls}.

Trainable Temperature. Unlike previous contrastive learning methods (Chen et al., 2020a; He et al., 2020; Khosla et al., 2020; Tian et al., 2020; Cui et al., 2021; Zhu et al., 2022), we train τ_g along with the network parameters to reduce the burden of hyperparameter tuning. However, L_{GML} is not suitable for training τ_g. If we estimate the variance of the Gaussian mixture when the contrast centers include the query feature, the optimal variance is too small, and the Gaussian mixture becomes spiky. Therefore, we exclude contrast features extracted from the same sample to

the query when training τ_g. For simplicity, we use a fixed τ_g throughout most of the experiments in this work. However, we show in Sec. 4.9 that τ_g can be trained, and it boosts the network performance.

By contrast, we find training τ_s results in a very low optimal value, which causes two negative effects. First, it makes L_{cls} dominate the training procedure, and second, it significantly degrades the accuracy of tail classes. Therefore, we leave τ_s as a hyperparameter.

3.4. Relation with Other Methods

As mentioned in Sec. 3.2, we can derive previous methods by adopting different likelihood and prior models.

Balanced Softmax. By modeling $s(x|y)$ as an isotropic Gaussian instead of a Gaussian mixture, we can derive the balanced softmax with a cosine similarity classifier.

$$s(x|y) = \exp f(x,y) = \exp\left(\frac{-(m_x - m_y)^2}{2\sigma^2}\right) \tag{9}$$

where m_y and σ denote the center and variance of the Gaussian model, respectively.

Because we apply L2 normalization to x and y, substituting f in Eq. 2 results in L_{cls}, which is the cosine similarity classifier with logit adjustment.

Supervised Contrastive Loss for Balanced Dataset. Supervised contrastive loss (Khosla et al., 2020) can be derived by assuming the training dataset is balanced. A balanced training dataset gives all η_cs and $\|Z_c\|$s equal. Therefore, the proposed GML loss for balanced datasets becomes as follows:

$$\begin{aligned} -L_{GML}^{(balanced)} &= \log \frac{\sum_{z_y \in Z_y} \exp(z_x \cdot z_y/\tau_g)}{\sum_{c \in C} \sum_{z_c \in Z_c} \exp(z_x \cdot z_c/\tau_g)} \\ &\geq \sum_{z_y \in Z_y} \log \frac{\exp(z_x \cdot z_y/\tau_g)}{\sum_{z \in Z} \exp(z_x \cdot z/\tau_g)} \end{aligned} \tag{10}$$

where $Z = \bigcup_c Z_c$ denotes the set of all contrast samples. By applying Jensen's inequality, we achieve supervised contrastive loss.

Visual-Linguistic Representation Learning. Visual-Linguistic Long-Tailed Recognition (VL-LTR) (Tian et al., 2021) utilizes a text sentences dataset and a pre-trained visual-linguistic model to address the problem of insufficient training samples of tail classes. Similarly, our method can be used to connect visual and text representations. In particular, we use text features as centers of the Gaussian mixture instead of contrast features and the pre-trained visual-linguistic model to extract the text features.

4. Experiments

4.1. Datasets

ImageNet-LT. Liu et al. (2019) constructed ImageNet-LT by sampling ImageNet-2012 (Russakovsky et al., 2015) following a Pareto distribution with $\alpha = 6$. The training set of ImageNet-LT contains $115.8k$ images of 1000 classes, ranging from a maximum of 1280 images to a minimum of 5 images per class. Meanwhile, the test set is balanced such that head classes and tail classes have the same impact on the accuracy evaluation. The test set of ImageNet-LT contains $50k$ images, with 50 images per class.

iNaturalist 2018. iNaturalist 2018 (Van Horn et al., 2018) is a large-scale image classification dataset containing 8142 classes. The goal of iNaturalist 2018 is to push state-of-the-art image recognition for "in the wild" data of highly imbalanced and fine-grained categories. The training set of iNaturalist 2018 contains $437.5k$ images of 8142 classes, ranging from a maximum of 1000 images to a minimum of 2 images per class. The test set of iNaturalist 2018 is also balanced, similar to ImageNet-LT.

CIFAR-10-LT and CIFAR-100-LT. Cui et al. (2019) constructed long-tailed versions of CIFAR (Krizhevsky, 2009) by reducing the number of training samples according to an exponential function. CIFAR-LT is categorized by its imbalance factor, which is the ratio of training samples for the largest class to that for the smallest.

ADE20K. ADE20K (Zhou et al., 2017) is a widely used image semantic segmentation dataset. The evaluation is conducted on 150 classes, where the most common class comprises more than 15% of total training pixels, while the rarest one comprises only 0.02%.

4.2. Implementation Details

We adopt training hyperparameter settings from previous long-tailed recognition papers (Cui et al., 2021; Tian et al., 2021; Zhu et al., 2022) with some modifications. For ImageNet-LT, we train the proposed method using an SGD optimizer whose learning rate is initially set to 0.05 and decays by a cosine scheduler. Input images are resized to 224×224, and a batch size of 128 is used. The weight decay and momentum are set to 5×10^{-4} and 0.9, respectively. The MLP of the contrast encoder has one hidden layer of 2048 channels and its output layer has 1024 channels. The total number of contrast samples (k) is 16384, and the minimum number of contrast samples per class (k_m) is 2. Randaugment (Cubuk et al., 2020) is applied for the classifier training and Simaugment (Chen et al., 2020a) for the contrastive learning. Finally, $\tau_s = 1/30$ is used throughout all experiments.

For iNaturalist 2018, we use a batch size of 128, input

Table 1. Performance comparison on the ImageNet-LT dataset.

Method	Epochs	Class frequency			
		All	Many	Med.	Few
Focal Loss	90	43.7	64.3	37.1	8.2
τ-norm	90	49.4	59.1	46.9	30.7
LWS	90	49.9	60.2	47.2	30.3
BALMS	90	51.4	62.2	48.8	29.8
LADE	90	51.9	62.3	49.3	31.2
DisAlign	90	53.4	62.7	52.1	31.4
BCL	90	56.7	67.2	53.9	36.5
Proposed	**90**	**58.3**	**68.7**	**55.7**	**38.6**
PaCo	400	58.2	68.0	56.4	37.2
Proposed	**400**	**58.8**	**68.2**	**56.7**	**39.5**

Table 2. Performance comparison on the iNaturalist 2018 dataset.

Method	Epochs	Accuracy
τ-norm	100	65.6
Hybrid-SC	100	66.7
SSP	100	68.1
KCL	100	68.6
DisAlign	100	69.5
RIDE (2 experts)	100	71.4
BCL	100	71.8
Proposed	**100**	**73.1**
RIDE (2 experts)	400	69.5
τ-norm	400	71.5
Balanced Softmax	400	71.8
PaCo	400	73.2
Proposed	**400**	**74.5**

image size of 224×224, and weight decay of 2×10^{-4}. The learning rate is set to 0.02 with the cosine scheduler. k and k_m are set to 65536 and 2, respectively. For CIFAR-LT, we use the training epochs of 200 and 400, a batch size of 64, and a weight decay of 5×10^{-4}. k and k_m are set to 4096 and 2, respectively. The learning rate is set to 0.05 and decays by a factor of 10 at 160 and 180 epochs (320 and 360 epochs if the training epochs is 400). We describe more about implementation details in Appendix A.2.

4.3. Long-Tailed Recognition

Comparison on ImageNet-LT. We first compare the performance of the proposed method with existing state-of-the-art long-tailed recognition methods on the ImageNet-LT dataset. We compare performances of the same backbone ResNeXt-50 and same number of training epochs for a fair comparison. Following the previous categorization of classes (Liu et al., 2019), we also evaluate the accuracy on subsets: many-shot (over 100 training samples), medium-shot (20-100 training samples), and few-shot (under 20 training samples).

Table 1 presents the experimental results on the ImageNet-LT dataset. The proposed method shows the best overall performance, outperforming previous state-of-the-art methods significantly. The gain is maximized on tail classes proving the efficacy of the proposed method on the long-tailed recognition task.

Table 3. Performance comparison on the CIFAR-100-LT dataset with different imbalance factors.

Method	Epochs	Imbalance factor		
		100	50	10
Focal loss	200	38.4	44.3	55.8
CB-Focal	200	39.6	45.2	58.0
LDAM-DRW	200	42.0	46.6	58.7
BBN	200	42.6	47.0	59.1
SSP	200	43.4	47.1	58.9
Casual model	200	44.1	50.3	59.6
Hybrid-SC	200	46.7	51.8	63.1
MetaSAug-LDAM	200	48.0	52.3	61.3
ResLT	200	48.2	52.7	62.0
BCL	200	51.9	56.6	64.9
Proposed	**200**	**53.0**	**57.6**	**65.7**
MiSLAS	400	47.0	52.3	63.2
Balanced Softmax	400	50.8	54.2	63.0
PaCo	400	52.0	56.0	64.2
Proposed	**400**	**54.0**	**58.1**	**67.0**

Comparison on iNaturalist 2018 dataset. We also evaluate the performance of the proposed method on the iNaturalist 2018, which is a large-scale highly imbalanced image classification dataset. For a fair comparison with previous methods, we use ResNet-50 as backbone. Table 2 shows the corresponding experimental result. The proposed method show significant performance improvements over all previous methods, including contrastive learning-based methods.

Comparison on CIFAR-100-LT dataset. Subsequently, we conduct extensive experiments on the CIFAR-100-LT dataset with different imbalance factors. We adopt imbalance factors of 100, 50, and 10, which are commonly used imbalance factors for evaluating the performance on CIFAR-LT dataset. A large imbalance factor implies a highly imbalanced dataset. In this experiment, we compare the performancess of ResNet-32 backbones.

Table 3 presents the comparison results on the CIFAR-100-LT dataset. The proposed method is robust to imbalance factors and consistently outperform previous long-tailed recognition methods on various imbalance factors by significant margins. Indeed, the robustness is also verified in experiments that compare our method with the supervised contrastive learning and knowledge distillation methods on balanced datasets. The comparisons are shown in Secs. 4.6 and 4.7.

Comparison with Visual-Linguistic Models. Visual-

Table 4. Performance comparison with visual-linguistic models. In this experiment, the backbone networks are initialized with CLIP (Radford et al., 2021) pre-trained weights.

ImageNet-LT dataset:

Method	Backbone	Accuracy
NCM	ResNet-50	49.2
cRT	ResNet-50	50.8
τ-norm	ResNet-50	51.2
LWS	ResNet-50	51.5
Zero-Shot CLIP	ResNet-50	59.8
VL-LTR	ResNet-50	70.1
Proposed	**ResNet-50**	**70.9**
VL-LTR	ViT-B	77.2
Proposed	**ViT-B**	**78.0**

iNaturalist 2018 dataset:

Method	Backbone	Accuracy
VL-LTR	ViT-B	81.0
Proposed	**ViT-B**	**82.1**

Table 5. Ablation study on the effect of each component.

Loss type	Teacher-student	Accuracy
Cross-entropy	✗	55.4
BCL	✗	56.7
BCL	✓	57.1 (+0.4)
Proposed	✗	56.0
Proposed	✓	**58.3 (+2.3)**

linguistic models utilize training samples from text modality to enhance the performance of long-tailed image classification tasks. We compare our method with visual-linguistic models by extending it to learn the visual-linguistic representation as described in Section 3.4.

Table 4 presents the comparison results with visual-linguistic models. We follow the training settings of VL-LTR (Tian et al., 2021) and use a larger input size for the iNaturalist 2018 dataset. An input image size of 384×384 is used only for the iNaturalist 2018 dataset in this experiment. The proposed method successfully connects two different modalities, even when the dataset is imbalanced. Thus, it exhibits the best performance regardless of network architecture or dataset.

4.4. Ablation Study

An ablation study is conducted on the ImageNet-LT dataset to investigate the effect of the components of the proposed method. Table 5 reveals that there is a performance gain when the proposed loss is used. However, the gain is insufficient because the contrast samples of tail classes are generated by highly outdated encoders as described in Sec. 3.3. Adopting a teacher–student framework solves the aforementioned problem, resulting in a significant gain in accuracy. To separate the effect of teacher–student framework from that of the proposed loss function, we measure the effect of teacher–student on BCL (Zhu et al., 2022), which is another contrastive learning-based method for long-tailed recognition. Since BCL does not employ queues to store contrast samples, it does not suffer from the outdated encoder problem. The teacher–student framework does not significantly improve the performance on BCL, proving that the impact of knowledge distillation is not significant. Meanwhile, it resolves the outdated encoder problem of the proposed method and leads to significant performance improvement.

Table 6. Effect of Teacher's Performance on Student.

Teacher		Student	
Backbone	Accuracy	Backbone	Accuracy
ResNet-34	50.3	ResNeXt-50	58.0
ResNet-50	55.2	ResNeXt-50	58.1
ResNeXt-50	56.4	ResNeXt-50	58.1
ResNeXt-101	57.9	ResNeXt-50	58.3

4.5. Relation with Accuracy of Teacher

To measure the effect of the accuracy of teacher and find the best one, we modulate the performance of teachers by adopting different sizes of backbone architecture. The chosen backbone architectures are ResNet-34, ResNet-50, ResNeXt-50, and ResNeXt-101, and they are trained on ImageNet-LT dataset for 90 epochs. All students use the same backbone architecture ResNeXt-50 and are also trained for 90 epochs. Table 6 shows the effect of the performance of the teacher on the student. It is observed that the accuracy teacher dramatically decreases as the backbone is changed to ResNet-34, but the accuracy of the student remains stable. Moreover, the accuracy of the student surpasses that of the teacher. We find that a teacher with better performance leads to a better student, but the impact is marginal; a sufficient result can be achieved by using the same backbone architecture for both teacher and student. This result coincides with the finding of the ablation study, which indicates that the impact of knowledge distillation is not significant. Since a low-accuracy teacher can still successfully resolve the outdated encoder problem, the proposed method shows outstanding performance regardless teacher's accuracy.

4.6. Comparison with Supervised Contrastive Learning

Table 7 presents the performance comparison of the proposed method with supervised contrastive learning (Khosla et al., 2020). Experiments are conducted using networks with ResNet-50 backbone on CIFAR-10-LT dataset with different imbalance factors. The proposed method shows

Table 7. Performance comparison with supervised contrastive learning on the CIFAR-10-LT dataset with different imbalance factors.

Imb. factor	Cross-Entropy	SupCon	**Proposed**
1	94.8	96.0	**95.9**
10	88.4	94.0	**94.5**
50	69.1	88.1	**90.6**
100	64.1	82.7	**86.7**

Table 8. Performance comparison with knowledge distillation methods.

Balanced dataset:

Method	CIFAR-100	ImageNet
None	69.1	69.8
KD	70.7	70.7
CRD	71.2	71.2
Proposed	**71.4**	**71.1**

Imbalanced dataset:

Method	CIFAR-100-LT	ImageNet-LT
None	48.6	56.3
KD	49.9	56.5
CRD	50.6	57.2
Proposed	**51.2**	**58.3**

the best accuracy with different imbalance factors, and the gap between previous methods increases as the imbalance factor increased. In addition, the proposed method shows a performance comparable with that of supervised contrastive learning when the dataset is balanced.

4.7. Comparison with Knowledge Distillation

Because we adopt a teacher–student framework to train the proposed method, comparing it with previous knowledge distillation methods is relevant. We select two knowledge distillation methods for the comparison: KD (Hinton et al.), which does not utilize contrastive learning, and CRD (Tian et al., 2020), which utilizes contrastive learning. For the CIFAR-100 and CIFAR-100-LT datasets, we train a student network with ResNet-20 backbone using a teacher network with ResNet-56 backbone for 240 epochs. We use a ResNet-18 student and a ResNet-34 teacher for ImageNet experiments, and a ResNeXt-50 student and a ResNeXt-101 teacher for ImageNet-LT experiments. For both ImageNet experiments, the networks are trained for 90 epochs.

Table 8 presents the performance comparison with knowledge distillation methods. The proposed method achieves the best performance on both balanced and imbalanced datasets. Knowledge distillation methods designed for balanced datasets show better accuracy than vanilla training on imbalanced datasets as well. This is because they provide measures to transfer the knowledge learned from head classes to tail classes, mitigating the lack of training samples. However, their gains are not the best because they do not consider the frequency of classes.

Table 9. Performance comparison on the ADE20K semantic segmentation dataset.

Method	mIoU	mAcc
Cross-Entropy	36.1	45.4
Proposed ($\alpha = 0.2$)	**38.1**	51.4
Proposed ($\alpha = 1.0$)	31.7	**59.8**

4.8. Semantic Segmentation Task

To prove the robustness of the proposed method, we replace the cross-entropy loss of semantic segmentation with the proposed method and measure the performance change. In this experiment, we measure the effect of the loss functions, not that of networks. Therefore, we perform the comparison using a widely used network FCN (Long et al., 2015) with ResNet-50 backbone. The evaluation is conducted on ADE20K (Zhou et al., 2017) using $160k$ training steps.

Table 9 presents the performance comparison of losses. mIoU refers to the mean intersection-over-union (IoU) and mAcc refers to the mean accuracy (Acc), where the mean is taken over classes. The proposed method achieves the best mIoU or mAcc depending on the hyperparameter α, which modulates the level of logit adjustment as given by Eq. 11.

$$-L_{GML}^{(\alpha)} = \log \frac{\exp\left[\log \frac{1}{\|Z_y\|}\sum_{z_y \in Z_y} \exp(z_x \cdot z_y/\tau_g) + \alpha\eta_y\right]}{\sum_{c\in C}\exp\left[\log \frac{1}{\|Z_c\|}\sum_{z_c \in Z_c} \exp(z_x \cdot z_c/\tau_g) + \alpha\eta_c\right]} \tag{11}$$

$$\text{IoU} = \text{TP}/(\text{TP} + \text{FN} + \text{FP}) \tag{12}$$

$$\text{Acc} = \text{TP}/(\text{TP} + \text{FN}) \tag{13}$$

Eqs. 12-13 give the definitions of IoU and Acc, where TP, FN, and FP denote true positive, false negative, and false positive, respectively. mAcc is the same metric as the balanced evaluation setting used in classification tasks, and $\alpha = 1.0$ gives the best mAcc. However, as FPs arise from other classes, mIoU is less sensitive to the accuracy of tail classes than mAcc. Therefore, the best mIoU is achieved by boosting the accuracy of head classes at the expense of tail classes, which is achieved by decreasing α. The effect of α on mIoU and mAcc is shown in Fig. 2.

4.9. Trainable Temperature Analysis

We examine the difference between a trainable τ_g and a fixed one. Fig. 3 shows the change in τ_g during training. τ_g

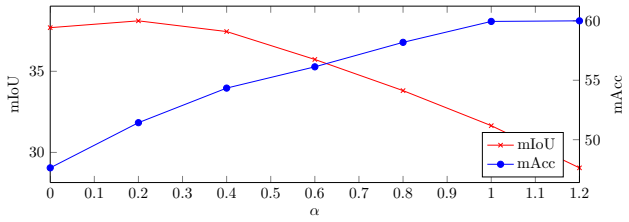

Figure 2. Effect of α on semantic segmentation performance.

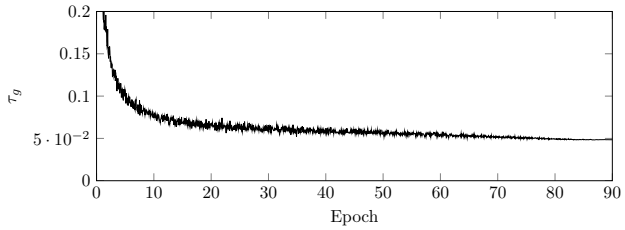

Figure 3. Change in τ_g during training.

becomes smaller as the encoder network converges. The final value is approximately 0.05, which is similar to the hyperparameter choice of other methods (He et al., 2020; Tian et al., 2020; Zhu et al., 2022), i.e., 0.07. Furthermore, training τ_g results in slightly better performance, boosting the accuracy from 58.2 to 58.3.

5. Conclusion

In this paper, we show that the fundamental problem of contrastive learning methods on long-tailed recognition comes from maximizing the mutual information between latent features and input data. To overcome this limitation, we interpret the long-tailed recognition as the mutual information maximization between latent features and ground-truth labels. This approach seamlessly integrates contrastive learning and the logit adjustment technique. It also verifies that contrastive learning implies the use of a Gaussian mixture likelihood and the logit adjustment is derived from the prior, while previous methods have combined them without understanding the theoretical background. Further, we propose an efficient way of modeling the Gaussian mixture likelihood using a teacher–student framework.

Extensive experiments on both long-tailed datasets and balanced datasets verify the superiority of the proposed method, which marks state-of-the-art performance on various benchmarks. Finally, as real-world data often show a long-tailed distribution, the proposed method can be applied to other tasks as well. As an example, we conduct experiments on a semantic segmentation dataset. The proposed method also showed a large performance gain on semantic segmentation, demonstrating its versatility.

Acknowledgements

This research was supported by the Challengeable Future Defense Technology Research and Development Program through the Agency For Defense Development(ADD) funded by the Defense Acquisition Program Administration(DAPA) in 2023(No.915027201), the Institute of New Media and Communications, the Institute of Engineering Research, and the Automation and Systems Research Institute at Seoul National University.

References

Alshammari, S., Wang, Y.-X., Ramanan, D., and Kong, S. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6897–6907, 2022.

Byrd, J. and Lipton, Z. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pp. 872–881. PMLR, 2019.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Contributors, M. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 715–724, 2021.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Drummond, C., Holte, R. C., et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pp. 1–8. Citeseer, 2003.

Feng, C., Zhong, Y., and Huang, W. Exploring classification equilibrium in long-tailed object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3417–3426, 2021.

Gidaris, S. and Komodakis, N. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4367–4375, 2018.

Han, H., Wang, W.-Y., and Mao, B.-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pp. 878–887. Springer, 2005.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hinton, G., Vinyals, O., Dean, J., et al. Distilling the knowledge in a neural network.

Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., and Chang, B. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6626–6636, 2021.

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

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

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. 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.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33: 4175–4186, 2020.

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

Tian, C., Wang, W., Zhu, X., Wang, X., Dai, J., and Qiao, Y. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. *arXiv preprint arXiv:2111.13579*, 2021.

Tian, Y., Krishnan, D., and Isola, P. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Wang, Y.-X., Ramanan, D., and Hebert, M. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Zhu, J., Wang, Z., Chen, J., Chen, Y.-P. P., and Jiang, Y.-G. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6908–6917, 2022.

A. Appendix

A.1. Contrastive Learning as MI Maximization between Latent Features and Input Data

The loss function of MoCo (He et al., 2020) is written as follows.

$$L_{MoCo} = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)} \tag{A.1}$$

where q denotes an encoded query, k_i denotes i-th key, k_+ is a key in the dictionary that matches q, and τ is a temperature hyperparameter.

As MoCo uses separated data augmentations and encoders to extract query and keys, q and k can be rewrite as follows.

$$q = f_q(t_q(x_j)) \tag{A.2}$$
$$k_+ = f_k(t_k(x_j)) \tag{A.3}$$
$$k_i = f_k(t_k(x_i)) \tag{A.4}$$

where f_q and f_k denotes the query encoder and the key encoder, and t_q and t_k denotes data augmentations for query and key, respectively.

Finally, the back-propagation is blocked at the key encoder and only the query encoder is updated by gradient. To represent this, we update Eq. A.2 to Eq. A.5.

$$q = f_q(t_q(x_j); w) \tag{A.5}$$

Summarizing above, Eq. A.1 becomes as the follow.

$$L_{MoCo} = -\log \frac{\exp(f_q(t_q(x_j); w) \cdot f_k(t_k(x_j))/\tau)}{\sum_{i=0}^{K} \exp(f_q(t_q(x_j); w) \cdot f_k(t_k(x_i))/\tau)} \tag{A.6}$$

By substituting x, y, and f in Eq. 1 using the followings, we show that MoCo maximizes the mutual information between latent features and input data.

$$x_i \leftarrow f_q(t_q(x_i); w) \tag{A.7}$$
$$y_j \leftarrow x_j \tag{A.8}$$
$$f(x, y) \leftarrow x \cdot f_k(t_k(y))/\tau \tag{A.9}$$

In other words, MoCo loss is identical to using a stochastic gradient descent to find the optimal parameter w^* that maximizes a lower bound of mutual information between latent features, $f_q(t_q(x); w)$ and input data x.

A.2. Implementation Details

Table A.1 shows the teacher architectures, training settings, augmentation strategies, and hyperparameter choices used in the experiments. γ, β, and α denote the weights used for L_{cls}, L_{GML}, and L_{KD}, respectively. We follow the settings of previous papers (Cui et al., 2021; Zhu et al., 2022) with some exceptions.

Training a network for 400 epochs on ImageNet-LT leads to overfitting when $\gamma = 1$ is used; reducing γ while increasing β alleviates the problem to improve the performance. Further, L_{KD} considerably enhances the accuracy of CIFAR-LT experiments, but its effect becomes marginal when applied to ImageNet-LT or iNatualist experiments.

For experiments in Tables 4, 6, and 7, we follow the training settings and hyperparameter choices of baseline methods (Khosla et al., 2020; Tian et al., 2020; 2021). In the experiments in Table 6, the encoder network and classifier are trained separately for a fair comparison with SupCon (Khosla et al., 2020), whereas they are trained simultaneously in other experiments. The encoder network is trained for 1000 epochs using L_{GML}. Subsequently, the classifier is trained for 100 epochs using L_{cls} with the encoder parameters fixed.

Semantic segmentation experiments in Table 8 are implemented based on open-source codebase mmsegmentation (Contributors, 2020) and follow the training hyperparameters and data augmentation settings provided in the codebase. The auxiliary

Table A.1. Hyperparameter choice of ImageNet-LT, iNaturalist, and CIFAR-100-LT experiments.

	Arch_s	Arch_t	Epochs	MLP	k	k_m	Aug_cls	Aug_GML	τ_g	τ_s	γ	β	α
Table 1. (a)	X50	X101	90	(2048, 2048, 1024)	16384	2	Randaug	Simaug	0.07	1/30	1	1	0
Table 1. (b)	X50	X101	400	(2048, 2048, 1024)	16384	2	Randaug	Simaug	0.07	1/30	0.5	2	0
Table 2. (a)	R50	R152	100	(2048, 2048, 1024)	65536	2	Simaug	Simaug	0.1	1/30	1	1	0
Table 2. (b)	R50	R152	400	(2048, 2048, 1024)	65536	2	Simaug	Simaug	0.1	1/30	1	1	0
Table 3. (a)	R32	R56	200	(64, 64, 32)	4096	2	Autoaug	Autoaug	0.1	1/30	1	1	1
Table 3. (b)	R32	R56	400	(64, 64, 32)	4096	2	Autoaug	Autoaug	0.1	1/30	1	1	1

loss of FCN (Long et al., 2015) is replaced with the cosine similarity classifier and trained using L_{cls}. The segmentation head also is replaced with the cosine similarity classifier and trained using $L_{cls} + L_{GML}$. The hyperparameter choice is $(k, k_m, \tau_s, \tau_g) = (8192, 27, 0.05, 0.07)$.