Probabilistic Contrastive Learning for Long-Tailed Visual Recognition

Chaoqun Du, Yulin Wang, *Graduate Student Member, IEEE*, Shiji Song, *Senior Member, IEEE*, and Gao Huang, *Member, IEEE*

***Abstract*—Long-tailed distributions frequently emerge in real-world data, where a large number of minority categories contain a limited number of samples. Such imbalance issue considerably impairs the performance of standard supervised learning algorithms, which are mainly designed for balanced training sets. Recent investigations have revealed that supervised contrastive learning exhibits promising potential in alleviating the data imbalance. However, the performance of supervised contrastive learning is plagued by an inherent challenge: it necessitates sufficiently large batches of training data to construct contrastive pairs that cover all categories, yet this requirement is difficult to meet in the context of class-imbalanced data. To overcome this obstacle, we propose a novel probabilistic contrastive (ProCo) learning algorithm that estimates the data distribution of the samples from each class in the feature space, and samples contrastive pairs accordingly. In fact, estimating the distributions of all classes using features in a small batch, particularly for imbalanced data, is not feasible. Our key idea is to introduce a reasonable and simple assumption that the normalized features in contrastive learning follow a mixture of von Mises-Fisher (vMF) distributions on unit space, which brings two-fold benefits. First, the distribution parameters can be estimated using only the first sample moment, which can be efficiently computed in an online manner across different batches. Second, based on the estimated distribution, the vMF distribution allows us to sample an infinite number of contrastive pairs and derive a closed form of the expected contrastive loss for efficient optimization. Other than long-tailed problems, ProCo can be directly applied to semi-supervised learning by generating pseudo-labels for unlabeled data, which can subsequently be utilized to estimate the distribution of the samples inversely. Theoretically, we analyze the error bound of ProCo. Empirically, extensive experimental results on supervised/semi-supervised visual recognition and object detection tasks demonstrate that ProCo consistently outperforms existing methods across various datasets.**

***Index Terms*—Contrastive learning, long-tailed visual recognition, representation learning, semi-supervised learning.**

Manuscript received 1 June 2023; revised 29 December 2023; accepted 18 February 2024. Date of publication 23 February 2024; date of current version 6 August 2024. This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0140407, in part by the National Natural Science Foundation of China under Grant 62276150 and Grant 42327901. Recommended for acceptance by J. Wang. *(Corresponding author: Gao Huang.)*

The authors are with the Department of Automation, BNRist, Tsinghua University, Beijing 100084, China (e-mail: dcq20@mails.tsinghua.edu.cn; wang-yl19@mails.tsinghua.edu.cn; shijis@tsinghua.edu.cn; gaohuang@tsinghua.edu.cn).

Our code is available at https://github.com/LeapLabTHU/ProCo.

Digital Object Identifier 10.1109/TPAMI.2024.3369102

I. Introduction

With the accelerated progress in deep learning and the emergence of large, well-organized datasets, significant advancements have been made in computer vision tasks, including image classification [1], [2], [3], object detection [4], and semantic segmentation [5]. The meticulous annotation of these datasets ensures a balance among various categories during their development [6]. Nevertheless, in practical applications, acquiring comprehensive datasets that are both balanced and encompass all conceivable scenarios remains a challenge. Data distribution in real-world contexts often adheres to a long-tail pattern, characterized by an exponential decline in the number of samples per class from head to tail [7]. This data imbalance presents a considerable challenge for training deep models, as their ability to generalize to infrequent categories may be hindered due to the limited training data available for these classes. Furthermore, class imbalance can induce a bias towards dominant classes, leading to suboptimal performance on minority classes [8], [9]. Consequently, addressing the long-tail distribution issue is crucial for the successful application of computer vision tasks in real-world scenarios.

In tackling the long-tailed data conundrum, a multitude of algorithms have been developed by adapting the traditional cross-entropy learning objective [10], [11], [12], [13]. Recent studies, however, have unveiled that supervised contrastive learning (SCL) [14] may serve as a more suitable optimization target in terms of resilience against long-tail distribution [15], [16]. More precisely, SCL deliberately integrates label information into the formulation of positive and negative pairs for the contrastive loss function. Unlike self-supervised learning, which generates positive samples through data augmentation of the anchor, SCL constructs positive samples from the same class as the anchor. Notably, initial exploration of this approach has already yielded performance surpassing most competitive algorithms designed for long-tail distribution [17], [18], [19].

Despite its merits, SCL still suffers from an inherent limitation. To guarantee performance, SCL necessitates a considerable batch size for generating sufficient contrastive pairs [14], resulting in a substantial computational and memory overhead. Notably, this issue becomes more pronounced with long-tailed data in real-world settings, where tail classes are infrequently sampled within a mini-batch or memory bank. Consequently, the loss function's gradient is predominantly influenced by head classes, leading to a lack of information from tail classes and an

0162-8828 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

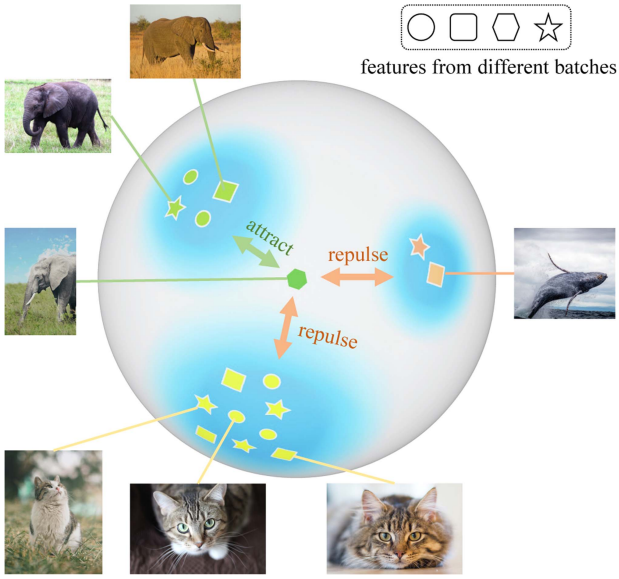


Fig. 1. Illustration of Probabilistic Contrastive Learning. ProCo estimates the distribution of samples based on the features from different batches and samples contrastive pairs from it. Moreover, a closed form of expected contrastive loss is derived by sampling an infinite number of contrastive pairs, which eliminates the inherent limitation of SCL on large batch sizes.

inherent tendency for the model to concentrate on head classes while disregarding tail classes [16], [17]. As an example, in the Imagenet-LT dataset, a typical batch size of 4096 and memory size of 8192 yield an average of fewer than one sample per mini-batch or memory bank for 212 and 89 classes, respectively.

In this study, we address the aforementioned issue with a simple but effective solution. Our primary insight involves considering the sampling of an infinite number of contrastive pairs from the actual data distribution and solving the expected loss to determine the optimization objective. By directly estimating and minimizing the expectation, the need for maintaining a large batch size is obviated. Moreover, since all classes are theoretically equivalent in the expectation, the long-tail distribution problem in real-world data is naturally alleviated.

However, implementing our idea is not straightforward due to two obstacles: 1) the methodology for modeling the actual data distribution is typically complex, e.g., training deep generative models [20], [21], [22]; and 2) calculating the expected training loss in a closed form is difficult. In this paper, we simultaneously tackle both challenges by proposing a novel probabilistic contrastive learning algorithm as illustrated in Fig. 1. Our method is inspired by the intriguing observation that deep features generally contain rich semantic information, enabling their statistics to represent the intra-class and inter-class variations of the data [23], [24], [25]. These methods model unconstrained features with a normal distribution from the perspective of data augmentation and obtain an upper bound of the expected cross-entropy loss for optimization. However, due to the normalization of features in contrastive learning, direct modeling with a normal distribution is not feasible. In addition, it is not possible to estimate the distributions of all classes in a small batch for long-tailed data. Therefore, we adopt a reasonable and simple von Mises-Fisher distribution on the unit sphere in $\mathbb{R}^n$ to model the feature distribution, which is commonly considered as an extension of the normal distribution to the hypersphere. It brings two advantages: 1) the distribution parameters can be estimated by maximum likelihood estimation using only the first sample moment, which can be efficiently computed across different batches during the training process; and 2) building upon this formulation, we theoretically demonstrate that a closed form of the expected loss can be rigorously derived as the sampling number approaches infinity rather than an upper bound, which we designate as the ProCo loss. This enables us to circumvent the necessity of explicitly sampling numerous contrastive pairs and instead minimize a surrogate loss function, which can be efficiently optimized and does not introduce any extra overhead during inference.

Furthermore, we extend the application of the proposed ProCo algorithm to more realistic imbalanced semi-supervised learning scenarios, where only a small portion of the entire training data set possesses labels [26], [27], [28], [29]. Semi-supervised algorithms typically employ a strategy that generates pseudo-labels for unlabeled data based on the model's predictions, using these labels to regularize the model's training process. Consequently, the ProCo algorithm can be directly applied to unlabeled samples by generating pseudo-labels grounded in the ProCo loss, which can subsequently be used to estimate the feature distribution inversely.

Despite its simplicity, the proposed ProCo algorithm demonstrates consistent effectiveness. We perform extensive empirical evaluations on supervised/semi-supervised image classification and object detection tasks with CIFAR-10/100-LT, ImageNet-LT, iNaturalist 2018, and LVIS v1. The results demonstrate that ProCo consistently improves the generalization performance of existing competitive long-tailed recognition methods. Furthermore, since ProCo is independent of imbalanced class distribution theoretically, experiments are also conducted on balanced datasets. The results indicate that ProCo achieves enhanced performance on balanced datasets as well.

The primary contributions of this study are outlined as follows:

- We propose a novel probabilistic contrastive learning (ProCo) algorithm for long-tailed recognition problems. By adopting a reasonable and simple von Mises-Fisher (vMF) distribution to model the feature distribution, we can estimate the parameters across different batches efficiently. ProCo eliminates the inherent limitation of SCL on large batch size by sampling contrastive pairs from the estimated distribution, particularly when dealing with imbalanced data (see Fig. 2).
- We derive a closed form of expected supervised contrastive loss based on the estimated vMF distribution when the sampling number tends to infinity and theoretically analyze the error bound of the ProCo loss (see Section III-C). This approach eliminates the requirement of explicitly sampling numerous contrastive pairs and, instead, focuses on minimizing a surrogate loss function. The surrogate loss function can be efficiently optimized without introducing any additional overhead during inference.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

- We employ the proposed ProCo algorithm to address imbalanced semi-supervised learning scenarios in a more realistic manner. This involves generating pseudo-labels based on the ProCo loss. Subsequently, we conduct comprehensive experiments on various classification datasets to showcase the efficacy of the ProCo algorithm.

II. Related Work

Long-tailed Recognition. To address the long-tailed recognition problem, early rebalancing methods can be classified into two categories: re-sampling and re-weighting. Re-sampling techniques aid in the acquisition of knowledge pertaining to tail classes by adjusting the imbalanced distribution of training data through either undersampling [30], [31] or oversampling [32]. Re-weighting methods adapt the loss function to promote greater gradient contribution for tail classes [33], [34] and even samples [10]. Nevertheless, Kang et al. [11] demonstrated that strong long-tailed recognition performance can be attained by merely modifying the classifier, without rebalancing techniques. Furthermore, post-hoc normalisation of the classifier weights [11], [35], [36] and loss margin modification [12], [13], [37], [38] have been two effective and prevalent methods. Post-hoc normalisation is motivated by the observation that the classifier weight norm tends to correlate with the class distribution, which can be corrected by normalising the weights. Loss margin modification methods incorporate prior information of class distribution into the loss function by adjusting the classifier's margin. Logit Adjustment [13] and Balanced Softmax [37] deduce that the classifier's decision boundary for each class corresponds to the $\log$ of the prior probability in the training data from the probabilistic perspective, which is demonstrated to be a straightforward and effective technique. Moreover, another common technique involves augmenting the minority classes by data augmentation techniques [2], [3], [23], [24], [39], [40], [41], [42], [43]. MetaSAug [24] employs meta-learning to estimate the variance of the feature distribution for each class and utilizes it as a semantic direction for augmenting single sample, which is inspired by implicit semantic data augmentation (ISDA) [23]. The ISDA employs a normal distribution to model unconstrained features for data augmentation and obtains an upper bound of the expected cross-entropy loss, which is related to our method. Nevertheless, features' normalization makes direct modeling infeasible in contrastive learning with a normal distribution. Hence, we adopt a mixture of von Mises-Fisher distributions on the unit sphere, allowing us to derive a closed-form of the expected contrastive loss rather than an upper bound. The vMF distribution [44] is a fundamental probability distribution on the unit hyper-sphere $\mathcal{S}^{p-1}$ in $\mathbb{R}^p$, which has been successfully used in deep metric learning [45], [46], supervised learning [47], [48], and unsupervised learning [49]. A recent study [50] introduces a classifier that utilizes the von Mises-Fisher distribution to address long-tailed recognition problems. Although this approach exhibits similarities to our method in terms of employing the vMF distribution, it specifically emphasizes the quality of representation for classifiers and features, considering the distribution overlap coefficient.

Contrastive Learning for Long-tailed Recognition. Recently, researchers have employed the contrastive learning to tackle the challenge of long-tailed recognition. Contrastive learning is a self-supervised learning approach that leverages contrastive loss function to learn a more discriminative representation of the data by maximizing the similarity between positive and negative samples [51], [52], [53], [54]. Khosla et al. [14] extended the contrastive learning to the supervised contrastive learning (SCL) paradigm by incorporating label information. However, due to the imbalance of positive and negative samples, contrastive learning also faces the problem that the model over-focusing on head categories in long-tailed recognition [9], [17], [19]. To balance the feature space, KCL [19] uses the same number positive pairs for all the classes. Recent studies [15], [16], [17], [18], [55], [56] have proposed to introduce class complement for constructing positive and negative pairs. These approaches ensure that all classes appear in every training iteration to rebalance the distribution of contrast samples. A comprehensive comparison is provided in Section III-E.

Furthermore, recent advancements in multi-modal foundation models based on contrastive learning, such as CLIP [57], have demonstrated remarkable generalization capabilities across various downstream tasks. Inspired by this, researchers have begun to incorporate multi-modal foundational models into long-tail recognition tasks. VL-LTR [58] develops a class-level visual-linguistic pre-training approach to associate images and textual descriptions at the class level and introduces a language-guided recognition head, effectively leveraging visual-linguistic representations for enhanced visual recognition.

Knowledge Distillation for Long-tailed Recognition. Knowledge distillation involves training a student model using the outputs of a well-trained teacher model [59]. This approach has been increasingly applied to long-tailed learning. For instance, LFME [60] trains multiple experts on various, less imbalanced sample subsets (e.g., head, middle, and tail sets), subsequently distilling these experts into a unified student model. In a similar vein, RIDE [61] introduces a knowledge distillation method to streamline the multi-expert model by developing a student network with fewer experts. Differing from the multi-expert paradigm, DiVE [62] demonstrates the efficacy of using a class-balanced model as the teacher for enhancing long-tailed learning. NCL [63] incorporates two main components: Nested Individual Learning and Nested Balanced Online Distillation. NIL focuses on the individual supervised learning for each expert, while NBOD facilitates knowledge transfer among multiple experts. Lastly, xERM [64] aims to develop an unbiased, test-agnostic model for long-tailed classification. Grounded in causal theory, xERM seeks to mitigate bias by minimizing cross-domain empirical risk.

Imbalanced Semi-supervised Learning (SSL). Semi-supervised learning is a subfield of machine learning that addresses scenarios where labeled training samples are limited, but an extensive amount of unlabeled data is available [26], [27], [28], [29], [65]. This scenario is directly relevant to a multitude of practical problems where it is relatively expensive to produce labeled data. The main approach in SSL is leveraging labeled data to generate pseudo-labels for unlabeled data, and then train

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

the model with both pseudo-labeled and labeled data [66]. In addition, consistency regularization or cluster assumption can be combined to further constrain the distribution of unlabeled data. For long-tailed datasets, due to class imbalance, SSL methods will be biased towards head classes when generating pseudo-labels for unlabeled data. Recently, researchers have proposed some methods to address the problem of pseudo-label generation in imbalanced SSL. DARP [67] is proposed to softly refine the pseudo-labels generated from a biased model by formulating a convex optimization problem. CReST [68] adopts an iterative approach to retrain the model by continually incorporating pseudo-labeled samples. DASO [69] focuses on the unknown distribution of the unlabeled data and blends the linear and semantic pseudo-labels in different proportions for each class to reduce the overall bias.

III. Method

A. Preliminaries

In this subsection, we start by presenting the preliminaries, laying the basis for introducing our method. Consider a standard image recognition problem. Given the training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)_{i=1}^N\}$, the model is trained to map the images from the space $\mathcal{X}$ into the classes from the space $\mathcal{Y} = \{1, 2, \ldots, K\}$. Typically, the mapping function φ is modeled as a neural network, which consists of a backbone feature extractor $F : \mathcal{X} \to \mathcal{Z}$ and a linear classifier $G : \mathcal{Z} \to \mathcal{Y}$.

Logit Adjustment [13] is a loss margin modification method. It adopts the prior probability of each class as the margin during the training and inference process. The logit adjustment method is defined as:

$$\mathcal{L}_{\text{LA}}(\boldsymbol{x}_i, y_i) = -\log \frac{\pi_{y_i} e^{\varphi_{y_i}(\boldsymbol{x}_i)}}{\sum_{y' \in \mathcal{Y}} \pi_{y'} e^{\varphi_{y'}(\boldsymbol{x}_i)}}, \tag{1}$$

where π_y is the class frequency in the training or test set, and φ_y is the logits of the class y.

Supervised Contrastive Learning (SCL) [14] is a generalization of the unsupervised contrastive learning method. SCL is designed to learn a feature extractor F that can distinguish between positive pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with the same label $y_i = y_j$ and negative pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with different labels $y_i \neq y_j$. Given any batch of sample-label pairs $B = \{(\boldsymbol{x}_i, y_i)_{i=1}^{N^B}\}$ and a temperature parameter τ, two typical ways to define the SCL loss are [14]:

$$\mathcal{L}_{\text{out}}^{\text{sup}}(\boldsymbol{x}_i, y_i) = \frac{-1}{N_{y_i}^B} \sum_{p \in A(y_i)} \log \frac{e^{\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau}}{\sum_{j=1}^K \sum_{a \in A(j)} e^{\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau}}, \tag{2}$$

$$\mathcal{L}_{\text{in}}^{\text{sup}}(\boldsymbol{x}_i, y_i) = -\log \left\{ \frac{1}{N_{y_i}^B} \sum_{p \in A(y_i)} \frac{e^{\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau}}{\sum_{j=1}^K \sum_{a \in A(j)} e^{\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau}} \right\}, \tag{3}$$

where $A(j)$ is the set of indices of the instances in the batch $B \setminus \{(x_i, y_i)\}$ with the same label j, $N_{y_i}^B = |A(y_i)|$ is its cardinality, and $\boldsymbol{z}$ denotes the normalized features of $\boldsymbol{x}$ extracted by F:

$$\boldsymbol{z}_i = \frac{F(\boldsymbol{x}_i)}{\|F(\boldsymbol{x}_i)\|}, \quad \boldsymbol{z}_p = \frac{F(\boldsymbol{x}_p)}{\|F(\boldsymbol{x}_p)\|}, \quad \boldsymbol{z}_a = \frac{F(\boldsymbol{x}_a)}{\|F(\boldsymbol{x}_a)\|}.$$

In addition, $\mathcal{L}_{\text{out}}^{\text{sup}}$ and $\mathcal{L}_{\text{in}}^{\text{sup}}$ denote the sum over the positive pairs relative to the location of the log. As demonstrated in [14], the two loss formulations are not equivalent and Jensen's inequality [70] implies that $\mathcal{L}_{\text{in}}^{\text{sup}} \leq \mathcal{L}_{\text{out}}^{\text{sup}}$. Therefore, SCL adopts the latter as the loss function, since it is an upper bound of the former.

B. Probabilistic Contrastive Learning

As aforementioned, for any example in a batch, SCL considers other examples with the same labels as positive samples, while the rest are viewed as negative samples. Consequently, it is essential for the batch to contain an adequate amount of data to ensure each example receives appropriate supervision signals. Nevertheless, this requirement is inefficient as a larger batch size often leads to significant computational and memory burdens. Furthermore, in practical machine learning scenarios, the data distribution typically exhibits a long-tail pattern, with infrequent sampling of the tail classes within the mini-batches. This particular characteristic necessitates further enlargement of the batches to effectively supervise the tail classes.

To address this issue, we propose a novel probabilistic contrastive (ProCo) learning algorithm that estimates the feature distribution and samples from it to construct contrastive pairs. Our method is inspired by [23], [24], which employ normal distribution to model unconstrained features from the perspective of data augmentation and obtain an upper bound of the expected loss for optimization. However, the features in contrastive learning are constrained to the unit hypersphere, which is not suitable for directly modeling them with a normal distribution. Moreover, due to the imbalanced distribution of training data, it is infeasible to estimate the distribution parameters of all classes in a small batch. Therefore, we introduce a simple and reasonable von Mises-Fisher distribution defined on the hypersphere, whose parameters can be efficiently estimated by maximum likelihood estimation across different batches. Furthermore, we rigorously derive a closed form of expected SupCon loss rather than an upper bound for efficient optimization and apply it to semi-supervised learning.

Distribution Assumption: As previously mentioned, the features in contrastive learning are constrained to be on the unit hypersphere. Therefore, we assume that the features follow a mixture of von Mises–Fisher (vMF) distributions [44], which is often regarded as a generalization of the normal distribution to the hypersphere. The probability density function of the vMF distribution for the random p-dimensional unit vector $\boldsymbol{z}$ is given by:

$$f_p(\boldsymbol{z}; \boldsymbol{\mu}, \kappa) = \frac{1}{C_p(\kappa)} e^{\kappa \boldsymbol{\mu}^\top \boldsymbol{z}}, \tag{4}$$

$$C_p(\kappa) = \frac{(2\pi)^{p/2} I_{(p/2-1)}(\kappa)}{\kappa^{p/2-1}}, \tag{5}$$

where $\boldsymbol{z}$ is a p-dimensional unit vector, $\kappa \geq 0$, $\|\boldsymbol{\mu}\|_2 = 1$ and $I_{(p/2-1)}$ denotes the modified Bessel function of the first kind

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

at order $p/2-1$, which is defined as:

$$I_{(p/2-1)}(z)=\sum_{k=0}^{\infty}\frac{1}{k!\Gamma(p/2-1+k+1)}\left(\frac{z}{2}\right)^{2k+p/2-1}. \quad (6)$$

The parameters $\boldsymbol{\mu}$ and κ are referred to as the mean direction and concentration parameter, respectively. A higher concentration around the mean direction $\boldsymbol{\mu}$ is observed with greater κ, and the distribution becomes uniform on the sphere when $\kappa=0$.

Parameter Estimation: Under the above assumption, we employ a mixture of vMF distributions to model the feature distribution:

$$P(\boldsymbol{z})=\sum_{y=1}^{K}P(y)P(\boldsymbol{z}|y)=\sum_{y=1}^{K}\pi_y\frac{\kappa_y^{p/2-1}e^{\kappa_y\boldsymbol{\mu}_y^{\top}\boldsymbol{z}}}{(2\pi)^{p/2}I_{p/2-1}(\kappa_y)}, \quad (7)$$

where the probability of a class y is estimated as π_y, which corresponds to the frequency of class y in the training set. The mean vector $\boldsymbol{\mu}_y$ and concentration parameter κ_y of the feature distribution are estimated by maximum likelihood estimation.

Suppose that a series of N independent unit vectors $\{(\boldsymbol{z}_i)_i^N\}$ on the unit hyper-sphere $\mathcal{S}^{p-1}$ are drawn from a vMF distribution of class y. The maximum likelihood estimates of the mean direction $\boldsymbol{\mu}_y$ and concentration parameter κ_y satisfy the following equations:

$$\boldsymbol{\mu}_y=\bar{\boldsymbol{z}}/\bar{R}, \quad (8)$$

$$A_p(\kappa_y)=\frac{I_{p/2}(\kappa_y)}{I_{p/2-1}(\kappa_y)}=\bar{R}, \quad (9)$$

where $\bar{\boldsymbol{z}}=\frac{1}{N}\sum_i^N\boldsymbol{z}_i$ is the sample mean and $\bar{R}=\|\bar{\boldsymbol{z}}\|_2$ is the length of sample mean. A simple approximation [71] to κ_y is:

$$\hat{\kappa}_y=\frac{\bar{R}(p-\bar{R}^2)}{1-\bar{R}^2}. \quad (10)$$

Furthermore, the sample mean of each class is estimated in an online manner by aggregating statistics from the previous mini-batch and the current mini-batch. Specifically, we adopt the estimated sample mean of the previous epoch for maximum likelihood estimation, while maintaining a new sample mean from zero initialization in the current epoch through the following online estimation algorithm:

$$\bar{\boldsymbol{z}}_j^{(t)}=\frac{n_j^{(t-1)}\bar{\boldsymbol{z}}_j^{(t-1)}+m_j^{(t)}\bar{\boldsymbol{z}}_j'^{(t)}}{n_j^{(t-1)}+m_j^{(t)}}, \quad (11)$$

where $\bar{\boldsymbol{z}}_j^{(t)}$ is the estimated sample mean of class j at step t and $\bar{\boldsymbol{z}}_j'^{(t)}$ is the sample mean of class j in current mini-batch. $n_j^{(t-1)}$ and $m_j^{(t)}$ represent the numbers of samples in the previous mini-batches and the current mini-batch, respectively.

Loss Derivation: Built upon the estimated parameters, a straightforward approach may be sampling contrastive pairs from the mixture of vMF distributions. However, we note that sampling sufficient data from the vMF distributions at each training iteration is inefficient. To this end, we leverage mathematical analysis to extend the number of samples to infinity and rigorously derive a closed-form expression for the expected contrastive loss function.

Proposition 1: Suppose that the parameters of the mixture of vMF distributions are π_y, $\boldsymbol{\mu}_y$, and $\kappa_y, y=1,\ldots,K$ and let the sampling number $N\to\infty$. Then we have the expected contrastive loss function, which is given by:

$$\mathcal{L}_{\text{out}}(\boldsymbol{z_i},y_i)=\frac{-\boldsymbol{z}_i\cdot A_p(\kappa_{y_i})\boldsymbol{\mu}_{y_i}}{\tau}+\log\left(\sum_{j=1}^{K}\pi_j\frac{C_p(\tilde{\kappa}_j)}{C_p(\kappa_j)}\right), \quad (12)$$

$$\mathcal{L}_{\text{in}}(\boldsymbol{z_i},y_i)=-\log\left(\pi_{y_i}\frac{C_p(\tilde{\kappa}_{y_i})}{C_p(\kappa_{y_i})}\right)+\log\left(\sum_{j=1}^{K}\pi_j\frac{C_p(\tilde{\kappa}_j)}{C_p(\kappa_j)}\right), \quad (13)$$

where $\tilde{\boldsymbol{z}}_j\sim\text{vMF}(\boldsymbol{\mu}_j,\kappa_j),\tilde{\kappa}_j=||\kappa_j\boldsymbol{\mu}_j+\boldsymbol{z}_i/\tau||_2,\tau$ is the temperature parameter.

Proof: According to the definition of supervised contrastive loss in (2), we have

$$\mathcal{L}_{\text{out}}^{\text{sup}}=\frac{-1}{N_{y_i}}\sum_{p\in A(y_i)}\boldsymbol{z}_i\cdot\boldsymbol{z}_p/\tau+\log\left(\sum_{j=1}^{K}N\frac{N_j}{N}\frac{1}{N_j}\sum_{a\in A(j)}e^{\boldsymbol{z}_i\cdot\boldsymbol{z}_a/\tau}\right), \quad (14)$$

where N_j is the sampling number of class j and satisfies $\lim_{N\to\infty}N_j/N=\pi_j$.

Let $N\to\infty$ and omit the constant term $\log N$, we have the following loss function:

$$\mathcal{L}_{\text{out}}=\frac{-\boldsymbol{z}_i\cdot\mathbb{E}\left[\tilde{\boldsymbol{z}}_{y_i}\right]}{\tau}+\log\left(\sum_{j=1}^{K}\pi_j\mathbb{E}\left[e^{\boldsymbol{z}_i\cdot\tilde{\boldsymbol{z}}_j/\tau}\right]\right) \quad (15)$$

$$=\frac{-\boldsymbol{z}_i\cdot A_p(\kappa_{y_i})\boldsymbol{\mu}_{y_i}}{\tau}+\log\left(\sum_{j=1}^{K}\pi_j\frac{C_p(\tilde{\kappa}_j)}{C_p(\kappa_j)}\right). \quad (16)$$

Eq. (16) is obtained by leveraging the expectation and moment-generating function of vMF distribution:

$$\mathbb{E}(\mathbf{z})=A_p(\kappa)\boldsymbol{\mu},A_p(\kappa)=\frac{I_{p/2}(\kappa)}{I_{p/2-1}(\kappa)}, \quad (17)$$

$$\mathbb{E}\left(e^{\mathbf{t}^{\top}\mathbf{z}}\right)=\frac{C_p(\tilde{\kappa})}{C_p(\kappa)},\tilde{\kappa}=||\kappa\boldsymbol{\mu}+\mathbf{t}||_2. \quad (18)$$

Similar to (12), we can obtain the other loss function from (3) as follows:

$$\mathcal{L}_{\text{in}}=-\log\left(\pi_{y_i}\frac{C_p(\tilde{\kappa}_{y_i})}{C_p(\kappa_{y_i})}\right)+\log\left(\sum_{j=1}^{K}\pi_j\frac{C_p(\tilde{\kappa}_j)}{C_p(\kappa_j)}\right). \quad (19)$$

□

Based on the above derivation, we obtain the expected formulations for the two SupCon loss functions. Since $\mathcal{L}_{\text{in}}$ enforces the

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

margin modification as shown in (1), we adopt it as the surrogate loss:

$$\mathcal{L}_{\text{ProCo}} = -\log\left(\pi_{y_i}\frac{C_p(\tilde{\kappa}_{y_i})}{C_p(\kappa_{y_i})}\right) + \log\left(\sum_{j=1}^{K}\pi_j\frac{C_p(\tilde{\kappa}_j)}{C_p(\kappa_j)}\right). \tag{20}$$

The empirical comparison of $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ is shown in Table IV.

Instead of costly sampling operations, we implicitly achieve infinite contrastive samples through the surrogate loss and can optimize it in a much more efficient manner. This design elegantly addresses the inherent limitations of the SCL, i.e., relying on large batch sizes (see Fig. 2). Furthermore, the assumption of feature distribution and the estimation of parameters can effectively capture the diversity of features among different classes, which enables our method to achieve stronger performance even without the sample-wise contrast as SCL (see Table III).

Numerical Computation: Due to PyTorch only providing the GPU implementation of the zeroth and first-order modified Bessel functions, one approach for efficiently computing the high-order function in ProCo is using the following recurrence relation:

$$I_{\nu+1}(\kappa) = \frac{2\nu}{\kappa}I_\nu(\kappa) - I_{\nu-1}(\kappa). \tag{21}$$

However, this method exhibits numerical instability when the value of κ is not sufficiently large.

Hence, we employ the Miller recurrence algorithm [72]. To compute $I_{p/2-1}(\kappa)$ in ProCo, we follow these steps: First, we assign the trial values 1 and 0 to $I_M(\kappa)$ and $I_{M+1}(\kappa)$, respectively. Here, M is a chosen large positive integer, and in our experiments, we set $M = p$. Then, using the inverse recurrence relation:

$$I_{\nu-1}(\kappa) = \frac{2\nu}{\kappa}I_\nu(\kappa) + I_{\nu+1}(\kappa), \tag{22}$$

we can compute $I_\nu(\kappa)$ for $\nu = M-1, M-2, \ldots, 0$. The value of $I_{p/2-1}(\kappa)$ obtained from this process is denoted as $\tilde{I}_{p/2-1}(\kappa)$, and $I_0(\kappa)$ is denoted as $\tilde{I}_0(\kappa)$. Finally, we can then compute $I_{p/2-1}(\kappa)$ as follows:

$$I_{p/2-1}(\kappa) = \frac{I_0(\kappa)}{\tilde{I}_0(\kappa)}\tilde{I}_{p/2-1}(\kappa). \tag{23}$$

Overall Objective: Following the common practice in long-tailed recognition [16], [17], [18], we adopt a two-branch design. The model consists of a classification branch based on a linear classifier $G(\cdot)$ and a representation branch based on a projection head $P(\cdot)$, which is an MLP that maps the representation to another feature space for decoupling with the classifier. Besides, a backbone network $F(\cdot)$ is shared by the two branches. For the classification branch and the representation branch, we adopt the simple and effective logit adjustment loss $\mathcal{L}_{\text{LA}}$ and our proposed loss $\mathcal{L}_{\text{ProCo}}$ respectively. Finally, the loss functions of the two branches are weighted and summed up as the overall loss function:

$$\mathcal{L} = \mathcal{L}_{\text{LA}} + \alpha\mathcal{L}_{\text{ProCo}}, \tag{24}$$

where α is the weight of the representation branch.

Algorithm 1: ProCo Algorithm.

Input: Training set $\mathcal{D}$, loss weight α
Randomly initialize the parameters $\boldsymbol{\Theta}$ of backbone F, projection head P and classifier G
for $t = 0$ **to** T **do**
  Sample a mini-batch $\{\boldsymbol{x}_i, y_i\}_{i=1}^{B}$ from $\mathcal{D}$
  Compute $\boldsymbol{z}_i = \frac{P(F(\boldsymbol{x}_i))}{\|P(F(\boldsymbol{x}_i))\|}$ and $G(F(\boldsymbol{x}_i))$
  Estimate $\boldsymbol{\mu}$ and κ according to (8) and (10)
  Compute $\mathcal{L}$ according to (24)
  Update $\boldsymbol{\Theta}$ with SGD
end for
Output: $\boldsymbol{\Theta}$

In general, by introducing an additional feature branch during training, our method can be efficiently optimized with stochastic gradient descent (SGD) algorithm along with the classification branch and does not introduce any extra overhead during inference.

Compatibility with Existing Methods: In particular, our approach is appealing in that it is a general and flexible framework. It can be easily combined with existing works applied to the classification branch, such as different loss functions, multi-expert framework, etc (see Table IX).

The pseudo-code of our algorithm is shown in Algorithm 1.

C. Theoretical Error Analysis

To further explore the theoretical foundations of our approach, we establish an upper bound on the generalization error and excess risk for the ProCo loss, as defined in (20). For simplicity, our analysis focuses on the binary classification scenario, where the labels y belong to the set $\{-1, +1\}$.

Assumption 1: $p\tau \gg 1$, with τ representing the temperature parameter and p the dimensionality of the feature space.

Proposition 2 (Generalization Error Bound): Under the Assumption 1, the following generalization bound is applicable with a probability of at least $1-\delta/2$. For every class $y \in \{-1, 1\}$ and for estimated parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\kappa}$, the bound is expressed as:

$$\mathbb{E}_{\boldsymbol{z}|y}\mathcal{L}_{\text{ProCo}}(y, \boldsymbol{z}; \hat{\boldsymbol{\mu}}, \hat{\kappa}) - \frac{1}{N_y}\sum_i \mathcal{L}_{\text{ProCo}}(y, \boldsymbol{z}_i; \hat{\boldsymbol{\mu}}, \hat{\kappa})$$
$$\leq \sqrt{\frac{2}{N_y}\boldsymbol{w}^\top\Sigma_y\boldsymbol{w}\ln\frac{2}{\delta}} + \frac{\ln(2/\delta)}{3N_y}\log\left(1 + e^{\|\boldsymbol{w}\|_2 - by}\right). \tag{25}$$

The generalization bound across all classes, with a probability of at least $1-\delta$, is thus:

$$\mathbb{E}_{(\boldsymbol{z},y)}\mathcal{L}_{\text{ProCo}}(y, \boldsymbol{z}; \hat{\boldsymbol{\mu}}, \hat{\kappa})$$
$$\leq \sum_{y\in\{-1,1\}}\frac{P(y)}{N_y}\sum_i \mathcal{L}_{\text{ProCo}}(y, \boldsymbol{z}_i; \hat{\boldsymbol{\mu}}, \hat{\kappa})$$
$$+ \sum_{y\in\{-1,1\}}\frac{P(y)\ln(2/\delta)}{3N_y}\log\left(1 + e^{\|\boldsymbol{w}\|_2 - by}\right)$$

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

$$+ \sum_{y \in \{-1,1\}} P(y) \sqrt{\frac{2}{N_y} \boldsymbol{w}^\top (\Sigma_y) \boldsymbol{w} \ln \frac{2}{\delta}}, \tag{26}$$

where N_y denotes the number of samples in class y, $\boldsymbol{w} = (\hat{\boldsymbol{\mu}}_{+1} - \hat{\boldsymbol{\mu}}_{-1})/\tau$, $b = \frac{1}{2\tau^2}(\frac{1}{\kappa_{+1}} - \frac{1}{\kappa_{-1}}) + \log \frac{\pi_{+1}}{\pi_{-1}}$, and Σ_y is the covariance matrix of $\boldsymbol{z}$ conditioned on y.

In our experimental setting, $\tau \approx 0.1$ and $p > 128$, thus Assumption 1 is reasonable in practice. Proposition 2 indicates that the generalization error gap is primarily controlled by the sample size and the data distribution variance. This finding corresponds to the insights from [13], [73], affirming that our method does not introduce extra factors in the error bound, nor does it expand the error bound. This theoretically assures the robust generalizability of our approach.

Furthermore, our approach relies on certain assumptions regarding feature distribution and parameter estimation. To assess the influence of these parameters on model performance, we derive an excess risk bound. This bound measures the deviation between the expected risk using estimated parameters and the Bayesian optimal risk, which is the expected risk with parameters of the ground-truth underlying distribution.

Assumption 2: The feature distribution of each class follows a von Mises-Fisher (vMF) distribution, characterized by parameters $\boldsymbol{\mu}^\star$ and $\kappa^\star$.

Proposition 3 (Excess Risk Bound): Given Assumptions 1 and 2, the following excess risk bound holds:

$$\begin{aligned} &\mathbb{E}_{(\boldsymbol{z},y)} \mathcal{L}_{\text{ProCo}}(y, \boldsymbol{z}; \hat{\boldsymbol{\mu}}, \hat{\kappa}) - \mathbb{E}_{(\boldsymbol{z},y)} \mathcal{L}_{\text{ProCo}}(y, \boldsymbol{z}; \boldsymbol{\mu}^\star, \kappa^\star) \\ &= \mathcal{O}\left(\Delta\boldsymbol{\mu} + \Delta\frac{1}{\kappa}\right), \end{aligned} \tag{27}$$

where $\Delta\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\star$, $\Delta\frac{1}{\kappa} = \frac{1}{\hat{\kappa}} - \frac{1}{\kappa^\star}$.

Assumption 2 is the core assumption of our method. Building upon this, Proposition 3 demonstrates that the excess risk associated with our method is primarily governed by the first-order term of the estimation error in the parameters.

D. ProCo for Semi-Supervised Learning

In order to further validate the effectiveness of our method, we also apply ProCo to semi-supervised learning. ProCo can be directly employed by generating pseudo-labels for unlabeled data, which can subsequently be utilized to estimate the distribution inversely. In our implementation, we demonstrate that simply adopting a straightforward approach like FixMatch [66] to generate pseudo-labels will result in superior performance. FixMatch's main concept lies in augmenting unlabeled data to produce two views and using the model's prediction on the weakly augmented view to generate a pseudo-label for the strongly augmented sample. Specifically, owing to the introduction of feature distribution in our method, we can compute the ProCo loss of weakly augmented view for each class to represent the posterior probability $P(y|\mathbf{z})$, thus enabling the generation of pseudo-labels.

TABLE I
COMPARISON WITH CONTRASTIVE LEARNING METHODS FOR LONG-TAILED RECOGNITION

Method	CC	CR	MM
SCL [14] & BF [74] & KCL [19]	✗	✗	✗
ELM [73]	✗	✗	✓
TSC [15]	✓	✓	✗
BCL [16] & Hybrid-PSC [18]	✓	✗	✗
PaCo [17] & GPaCo [56]	✓	✗	✓
DRO-LT [55]	✓	✓	✗
ProCo	✓	✓	✓

CC denotes class complement, CR denotes calculable representation, and MM denotes margin modification.

E. Connection With Related Work

In the following, we discuss the connections between our method and related works on contrastive learning for long-tailed recognition. Recent studies proposed to incorporate class complement in the construction of positive and negative pairs. These methods ensure that all classes appear in every iteration of training to rebalance the distribution of contrast samples. Moreover, researchers also introduced margin modification in contrastive learning due to its effectiveness. Therefore, we mainly discuss three aspects, namely class complement, calculable representation, and margin modification. Class complement represents introducing global representations of each class in contrastive learning for constructing positive and negative samples. Calculable representation corresponds to the class complements are computed from the features rather than learnable parameters. Margin modification represents adjusting the contrastive loss according to the prior frequency of different classes in the training set as shown in (1).

In our method, we introduce the class complement based on the feature distribution by estimation from the features, which is not a learnable parameter. If the class representation is a learnable parameter $\boldsymbol{w}$ and we ignore the contrast between samples, then we have the following contrastive loss:

$$\mathcal{L}(\boldsymbol{z}_i, y_i) = -\log \frac{e^{\boldsymbol{w}_{y_i}^\top \boldsymbol{z}_i/\tau}}{\sum_{j=1}^{N} e^{\boldsymbol{w}_j^\top \boldsymbol{z}_i/\tau}}, \tag{28}$$

where $\boldsymbol{w}_j$ and $\boldsymbol{z}_i$ are normalized to unit length, and τ is the temperature parameter. This is equivalent to cosine classifier (normalized linear activation) [75]—a variant of cross-entropy loss, which has been applied to long-tailed recognition algorithms [16], [76], [77]. Therefore, the sole introduction of the learnable parameter $\boldsymbol{w}$ is analogous to the role played by the weight in the classification branch, which is further validated by the empirical results in Table III.

The related works are summarized in Table I. BatchFormer (BF) [74] is proposed to equip deep neural networks with the capability to investigate the sample relationships within each mini-batch. This is achieved by constructing a Transformer Encoder Network among the images, thereby uncovering the interrelationships among the images in the mini-batch. BatchFormer facilitates the propagation of gradients from each label

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

to all images within the mini-batch, a concept that mirrors the approach used in contrastive learning.

Embedding and Logit Margins (ELM) [73] proposes to enforce both embedding and logit margins, analyzing in detail the benefit of introducing margin modification in contrastive learning. TSC [15] introduces class representation by pre-generating and online-matching uniformly distributed targets to each class during the training process. However, the targets do not have class semantics. Hybrid-PSC [18], PaCo [17], GPaCo [56] and BCL [16] all introduce learnable parameters as class representation. PaCo also enforces margin modification when constructing contrastive samples. DRO-LT [55] computes the centroid of each class and utilizes it as the class representation, which is the most relevant work to ours. The loss function and uncertainty radius in DRO-LT are devised by heuristic design from the metric learning and robust perspective. Moreover, DRO-LT considers a sample and corresponding centroid as a positive pair. But in contrast to SCL, the other samples in the batch and the centroid are treated as negative pairs, disregarding the label information of the other samples, which is somewhat pessimistic.

Compared with the above methods, ProCo is derived rigorously from contrastive loss and a simple assumption on feature distribution. Furthermore, ProCo also enforces margin modification, which is a key component for long-tailed recognition. Section IV-D demonstrates the superiority of ProCo over the above methods.

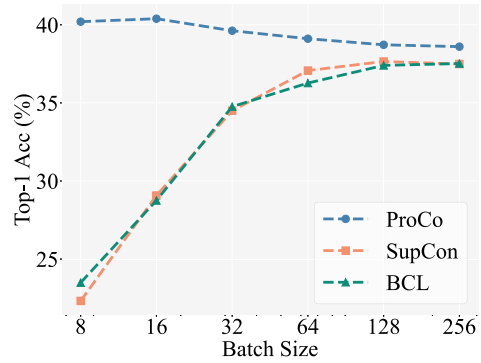


Fig. 2. Performance of the representation branch. We train the model for 200 epochs.

TABLE II
COMPARISON OF DIFFERENT CLASS COMPLEMENTS

Class Complement	Top-1 Acc.
EMA Prototype	51.6
Centroid Prototype	52.0
Normal Distribution	52.1
ProCo	**52.8**

EMA denotes exponential moving average.

IV. EXPERIMENT

In this section, we validate the effectiveness of our method on supervised/semi-supervised learning. First, we conduct a range of analytical experiments to confirm our hypothesis and analyze each component of the method, including 1) performance of the representation branch, 2) comparison of more settings, 3) comparison between two formulations of loss, 4) sensitivity analysis of hyper-parameters, 5) data augmentation strategies. Subsequently, we compare our method with existing supervised learning methods on long-tailed datasets such as CIFAR-10/100-LT, ImageNet-LT, and iNaturalist 2018. Finally, experiments on balanced datasets, semi-supervised learning, and long-tailed object detection tasks are conducted to confirm the broad applicability of our method.

A. Dataset and Evaluation Protocol

We perform long-tailed image classification experiments on four prevalent long-tailed image classification datasets: CIFAR-10/100-LT, ImageNet-LT, and iNaturalist. Following [11], [76], we partition all categories into three subsets based on the number of training samples: Many-shot (> 100 images), Medium-shot ($20 - 100$ images), and Few-shot (< 20 images). The top-1 accuracy is reported on the respective balanced validation sets. In addition, we conducted experiments on balanced image classification datasets and long-tailed object detection datasets to verify the broad applicability of our method. The effectiveness of instance segmentation was assessed using the mean Average Precision (AP_m) for mask predictions, calculated at varying Intersection over Union (IoU) thresholds ranging from 0.5 to 0.95, and aggregated across different categories. The AP values for rare, common, and frequent categories are represented as AP_r, AP_c, and AP_f, respectively, while the AP for detection boxes is denoted as AP_b.

CIFAR-10/100-LT: CIFAR-10-LT and CIFAR-100-LT are the long-tailed variants of the original CIFAR-10 and CIFAR-100 [78] datasets, which are derived by sampling the original training set. Following [10], [12], we sample the training set of CIFAR-10 and CIFAR-100 with an exponential function $N_j = N \times \lambda^j$, where $\lambda \in (0, 1)$, N is the size of the original training set, and N_j is the sampling quantity for the j-th class. The original balanced validation sets of CIFAR-10 and CIFAR-100 are used for testing. The imbalance factor $\gamma = \max(N_j)/\min(N_j)$ is defined as the ratio of the number of samples in the most and the least frequent class. We set γ at typical values 10,50,100 in our experiments.

ImageNet-LT: ImageNet-LT is proposed in [76], which is constructed by sampling a subset of ImageNet following the Pareto distribution with power value $\alpha_p = 6$. It consists of 115.8k images from 1000 categories with cardinality ranging from 1280 to 5.

iNaturalist 2018: iNaturalist 2018 [7] is a severely imbalanced large-scale dataset. It contains 8142 classes of 437.5k images, with an imbalance factor $\gamma = 500$ with cardinality ranging from 2 to 1000. In addition to long-tailed image classification, iNaturalist 2018 is also utilized in fine-grained image classification.

CUB-200-2011: The Caltech-UCSD Birds-200-2011 [79] is a prominent resource for fine-grained visual categorization tasks. Comprising 11,788 images across 200 bird subcategories, it is split into two sets: 5,994 images for training and 5,794 for testing.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

TABLE III
COMPARISON OF MORE SETTINGS

SC	CC	CR	MM	Top-1 Acc.
✗	✗	✗	✗	50.5
✗	✓	✗	✗	50.6
✓	✗	✗	✗	51.8
✓	✓	✗	✗	51.9
✓	✓	✓	✓	52.6
✗	✓	✓	✓	**52.8**

SC denotes supervised contrastive loss, CC denotes class complement, CR denotes calculable representation, and MM denotes margin modification.

TABLE IV
COMPARISON OF DIFFERENT LOSS FORMULATIONS

Imbalance Factor	100	50	10
$\mathcal{L}_{\text{out}}$	52.0	56.6	65.1
$\mathcal{L}_{\text{in}}$	**52.8**	**57.1**	**65.5**

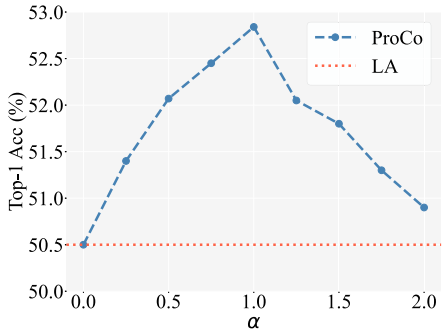


Fig. 3. Parameter analysis of the loss weight ratio α. LA denotes the logit adjustment method.

TABLE V
COMPARISON OF TRAINING THE NETWORK WITH ('W/') AND WITHOUT ('W/O') EMPLOYING AUTOAUGMENT

Method	w/o AutoAug	w/ AutoAug
Logit Adj. [13]	49.7	50.5
BCL [16]	50.1	51.9
ProCo	**50.5**	**52.8**

LVIS v1: The Large Vocabulary Instance Segmentation (LVIS) dataset [80] is notable for its extensive categorization, encompassing 1,203 categories with high-quality instance mask annotations. LVIS v1 is divided into three splits: a training set with 100,000 images, a validation (val) set with 19,800 images, and a test-dev set, also with 19,800 images. Categories within the training set are classified based on their prevalence as rare (1-10 images), common (11-100 images), or frequent (over 100 images).

B. Implementation Details

For a fair comparison of long-tailed supervised image classification, we strictly follow the training setting of [16]. All models are trained using an SGD optimizer with a momentum set to 0.9.

CIFAR-10/100-LT: We adopt ResNet-32 [2] as the backbone network. The representation branch has a projection head with an output dimension of 128 and a hidden layer dimension of 512. We set the temperature parameter τ for contrastive learning to 0.1. Following [16], [17], we apply AutoAug [85] and Cutout [86] as the data augmentation strategies for the classification branch, and SimAug [51] for the representation branch. The loss weights are assigned equally ($\alpha = 1$) to both branches. We train the network for 200 epochs with a batch size of 256 and a weight decay of 4×10^{-4}. We gradually increase the learning rate to 0.3 in the first 5 epochs and reduce it by a factor of 10 at the 160th and 180th epochs. Unless otherwise specified, our components analysis follows the above training setting. For a more comprehensive comparison, we also train the model for 400 epochs with a similar learning rate schedule, except that we warm up the learning rate in the first 10 epochs and decrease it at 360th and 380th epochs.

In order to evaluate the effectiveness of ProCo in semi-supervised learning tasks, we partition the training dataset into two subsets: a labeled set and an unlabeled set. The division is conducted through a random elimination of the labels. We predominantly adhere to the training paradigm established by DASO [69]. DASO builds its semantic classifier by computing the average feature vectors of each class for generating pseudo-labels. The semantic classifier in DASO is substituted with our representation branch for training. The remaining hyperparameters are kept consistent with those in DASO.

ImageNet-LT & iNaturalist 2018: For both ImageNet-LT and iNaturalist 2018, we employ ResNet-50 [2] as the backbone network following the majority of previous works. The representation branch consists of a projection head with an output dimension of 2048 and a hidden layer dimension of 1024. The classification branch adopts a cosine classifier. We set the temperature parameter τ for contrastive learning to 0.07. RandAug [87] is employed as a data augmentation strategy for the classification branch, and SimAug for the representation branch. We also assign equal loss weight ($\alpha = 1$) to both branches. The model is trained for 90 epochs with a batch size of 256 and a cosine learning rate schedule. For ImageNet-LT, we set the initial learning rate to 0.1 and the weight decay to 5×10^{-4}. In addition, the model is trained for 90 epochs with ResNeXt-50-32x4d [88] as the backbone network and for 180 epochs with ResNet-50. For iNaturalist 2018, we set the initial learning rate to 0.2 and the weight decay to 1×10^{-4}.

C. Components Analysis

Performance of Representation Branch: Since ProCo is based on the representation branch, we first analyze the performance of a single branch. Fig. 2 presents the performance curves of different methods without a classification branch as the batch size changes. We follow the two-stage training strategy and hyper-parameters of SCL [14]. In the first stage, the model

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
TOP-1 ACCURACY OF RESNET-32 ON CIFAR-100-LT AND CIFAR-10-LT

Dataset	CIFAR-100-LT			CIFAR-10-LT		
Imbalance Factor	100	50	10	100	50	10
CB-Focal [10]	39.6	45.2	58.0	74.6	79.3	87.5
LDAM-DRW* [12]	42.0	46.6	58.7	77.0	81.0	88.2
BBN [81]	42.6	47.0	59.1	79.8	81.2	88.3
SSP [82]	43.4	47.1	58.9	77.8	82.1	88.5
TSC [63]	43.8	47.4	59.0	79.7	82.9	88.7
Casual Model [83]	44.1	50.3	59.6	80.6	83.6	88.5
Hybrid-SC [18]	46.7	51.9	63.1	81.4	85.4	91.1
MetaSAug-LDAM [24]	48.0	52.3	61.3	80.7	84.3	89.7
ResLT [84]	48.2	52.7	62.0	82.4	85.2	89.7
Logit Adjustment† [13]	50.5	54.9	64.0	84.3	87.1	90.9
BCL† [16]	51.9	56.4	64.6	84.5	87.2	91.1
ProCo†	**52.8**	**57.1**	**65.5**	**85.9**	**88.2**	**91.9**

*denotes results borrowed from [81]. † denotes models trained in the same setting. We report the results of 200 epochs.

TABLE VII
TOP-1 ACCURACY OF RESNET-32 ON CIFAR-100-LT WITH AN IMBALANCE FACTOR OF 100

Method	Many	Med	Few	All
200 epochs				
DRO-LT [55]	64.7	50.0	23.8	47.3
RIDE [61]	68.1	49.2	23.9	48.0
Logit Adj.† [13]	67.2	51.9	29.5	50.5
BCL† [16]	67.2	**53.1**	32.9	51.9
ProCo†	**69.0**	52.7	**34.1**	**52.8**
400 epochs				
PaCo† [17]	-	-	-	52.0
GPaCo† [56]	-	-	-	52.3
Logit Adj.† [13]	68.1	53.0	32.4	52.1
BCL† [16]	69.2	53.1	34.4	53.1
ProCo†	**70.1**	**53.4**	**36.4**	**54.2**

†denotes the models trained in the same setting

is trained only through the representation branch. In the second stage, we train a linear classifier to evaluate the model's performance. From the results, we can directly observe that the performance of BCL [16] and SupCon [14] is significantly limited by the batch size. However, ProCo effectively mitigates SupCon's limitation on the batch size by introducing the feature distribution of each class. Furthermore, a comparison of different class complements is presented in Table II. Here, EMA and centroid prototype represent that the class representations are estimated by the exponential moving average and the centroid of the feature vectors, respectively. Normal distribution represents that we employ the normal distributions to model the feature distribution of each class, even though they are in the normalized feature space. Compared with merely estimating a class prototype, ProCo yields superior results by estimating the distribution of features for each class. Empirical results also confirm the theoretical analysis in Section III-B that the vMF distribution is better suited for modeling the feature distribution than the normal distribution.

Comparison of More Settings: The results are summarized in Table III, where we compare ProCo with several other settings to clarify the effectiveness of relevant components. As demonstrated in the table, the logit adjustment method fails to achieve improvement in performance when combined with a learnable class representation (50.5% versus 50.6%, 51.8% versus 51.9%). This empirical evidence suggests a certain degree of equivalence between the two approaches. Furthermore, incorporating ProCo with SupCon loss [14] yields slightly performance degradation (52.6% versus 52.8%). This result emphasizes the effectiveness of the assumption on feature distribution and parameter estimation in capturing feature diversity.

Loss Formulations: We analyze the impact of different loss formulations on the performance of ProCo. These two formulations are derived from different SupCon losses [14] as shown in Proposition 1. Table IV shows the results of different loss formulations on CIFAR-100-LT. We can observe that $\mathcal{L}_{\text{in}}$ outperforms $\mathcal{L}_{\text{out}}$.

Sensitivity Test: To investigate the effect of loss weight on model performance, we conducted experiments on CIFAR-100-LT dataset with an imbalance factor of 100. The results are presented in Fig. 3. It is evident that the model exhibits optimal performance when α is set to 1.0. This suggests that the optimal performance can be achieved by setting the loss weight of both branches equally. Therefore, we opt not to conduct an exhaustive hyper-parameter search for each dataset. Instead, we extrapolate the same loss weight to the other datasets.

Data Augmentation Strategies: Data augmentation is a pivotal factor for enhancing model performance. To investigate the impact of data augmentation, we conducted experiments on the CIFAR100-LT dataset with an imbalance factor of 100. Table V demonstrates the influence of AutoAug strategy. Under strong data augmentation, our method achieves greater performance enhancement, suggesting the effectiveness of our approach in leveraging the benefits of data augmentation.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
COMPARISONS ON IMAGENET-LT AND INATURALIST 2018 WITH DIFFERENT BACKBONE NETWORKS

Method	ImageNet-LT		iNaturalist 2018
	Res50	ResX50	Res50
90 epochs			
τ-norm [11]	46.7	49.4	65.6
MetaSAug [24]	47.4	–	68.8
SSP [82]	51.3	–	68.1
KCL [11]	51.5	–	68.6
DisAlign [89]	52.9	53.4	69.5
vMF Classifier [50]	–	53.7	–
SSD [90]	–	53.8	69.3
ICCL [91]	–	54.0	70.5
ResLT [84]	–	56.1	70.2
GCL [92]	54.9	–	72.0
RIDE [61]	54.9	56.4	72.2
Logit Adj.† [13]	55.1	56.5	–
BCL† [16]	56.0	56.7	71.8
ProCo†	**57.3**	**58.0**	**73.5**

†denotes the models trained in the same setting

TABLE IX
TOP-1 ACCURACY OF RESNET-50 ON IMAGENET-LT AND INATURALIST 2018

Method	Many	Med	Few	All
180 epochs, ImageNet-LT				
Logit Adj.† [13]	65.8	53.2	34.1	55.4
PaCo† [17]	64.4	**55.7**	33.7	56.0
BCL† [16]	67.6	54.6	36.8	57.2
NCL‡ [63]	68.2	53.9	36.3	57.0
NCL (ensemble)‡ [63]	69.1	56.4	38.9	59.2
ProCo†	**68.2**	55.1	**38.1**	**57.8**
ProCo+NCL‡	**68.4**	**54.9**	**38.6**	**57.9**
ProCo+NCL (ensemble)‡	**70.6**	**57.4**	**40.8**	**60.2**
400 epochs, iNaturalist 2018				
PaCo† [17]	70.3	73.2	73.6	73.2
BatchFormer† [74]	72.8	75.3	75.3	75.1
GPaCo† [56]	73.0	75.5	75.7	75.4
ProCo†	**74.0**	**76.0**	**76.0**	**75.8**

†and ‡ denote models trained in the same settings.

TABLE X
TOP-1 ACCURACY FOR RESNET-32 AND RESNET-50 ON BALANCED DATASETS

	CIFAR-100	CIFAR-10	CUB-200-2011
CrossEntropy	71.5	93.4	81.6
ProCo	**73.0**	**94.6**	**82.9**

The ResNet-32 model is trained on CIFAR-100/10 for 200 epochs from scratch. For CUB-200-2011, the pre-trained ResNet-50 model is fine-tuned for 30 epochs.

D. Long-Tailed Supervised Image Classification

CIFAR-10/100-LT: Table VI presents the results of ProCo and existing methods on the CIFAR-10/100-LT. ProCo demonstrates superior performance compared to other methods in handling varying levels of class imbalance, particularly in situations with higher imbalance factors. With the same training scheme, we mainly compare with Logit Adj. [13] and BCL [16]. Combining the representation branch based on contrastive learning with the classification branch can significantly improve the performance of the model. Our method further enhances the performance and effectively alleviates the problem of data imbalance. Moreover, we report longer and more detailed results on the CIFAR-100-LT dataset with an imbalance factor of 100 in Table VII. With 200 and 400 epochs, especially for tail classes, our method has a 1.2% and 2.0% improvement compared to BCL, respectively. Meanwhile, our method also maintains an improvement on head categories.

ImageNet-LT: We report the performance of ProCo on the ImageNet-LT dataset in Table VIII compared with existing methods. ProCo surpasses BCL by 1.3% on ResNet-50 and ResNeXt-50 with 90 training epochs, which demonstrates the effectiveness of our distribution-based class representation on datasets of large scale. Furthermore, Table IX lists detailed results on more training settings for ImageNet-LT dataset. ProCo has the most significant performance improvement on tail categories. In addition to combining with typical classification branches, ProCo can also be combined with other methods to further improve model performance, such as different loss functions and model ensembling methods. We also report the results of combining with NCL [63]. NCL is a multi-expert method that utilizes distillation and hard category mining. ProCo demonstrates performance improvements for NCL.

iNaturalist 2018: Table VIII presents the experimental comparison of ProCo with existing methods on iNaturalist 2018 over 90 epochs. iNaturalist 2018 is a highly imbalanced large-scale dataset, thus making it ideal for studying the impact of imbalanced datasets on model performance. Under the same training setting, ProCo outperforms BCL by 1.7%. Furthermore, to facilitate a comparison with state-of-the-art methodologies, an extended training schedule of 400 epochs is conducted. The results in Table IX indicate that ProCo is capable of effectively scaling to larger datasets and longer training schedules.

E. Balanced Supervised Image Classification

The foundational theory of our model is robust against data imbalance, meaning that the derivation of ProCo is unaffected by long-tailed distributions. In support of this, we also perform experiments on balanced datasets, as illustrated in Table X. For the CIFAR-100/10 dataset, augmentation and training parameters identical to those used for CIFAR-100/10-LT are employed. Additionally, we expand our experimentation to the fine-grained classification dataset CUB-200-2011. These results demonstrate that while our method, primarily designed for imbalanced datasets, mitigates the inherent limitations of contrastive learning in such contexts, additional experiments also highlight its effectiveness on balanced training sets. This versatility underlines the strength of our method in addressing not only imbalances in $P(y)$ but also intra-class distribution variances in $P(\boldsymbol{z}|y)$. These aspects correspond to the factors

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

TABLE XI
COMPARISON OF ACCURACY (%) ON CIFAR100-LT UNDER $\gamma_l = \gamma_u$ SETUP

Method	$\gamma = \gamma_l = \gamma_u = 10$		$\gamma = \gamma_l = \gamma_u = 20$	
	$N_1 = 50$ $M_1 = 400$	$N_1 = 150$ $M_1 = 300$	$N_1 = 50$ $M_1 = 400$	$N_1 = 150$ $M_1 = 300$
Supervised	29.6	46.9	25.1	41.2
w/ LA [13]	30.2	48.7	26.5	44.1
FixMatch [66]	45.2	56.5	40.0	50.7
w/ DARP [67]	49.4	58.1	43.4	52.2
w/ CReST+ [68]	44.5	57.4	40.1	52.1
w/ DASO† [69]	49.8	59.2	43.6	52.9
w/ ProCo†	**50.9**	**60.2**	**44.8**	**54.8**
FixMatch [66] + LA [13]	47.3	58.6	41.4	53.4
w/ DARP [67]	50.5	59.9	44.4	53.8
w/ CReST+ [68]	44.0	57.1	40.6	52.3
w/ DASO† [69]	50.7	60.6	44.1	55.1
w/ ProCo†	**52.1**	**61.3**	**46.9**	**55.9**

γ_l and γ_u are the imbalance factors for labeled and unlabeled data, respectively. N_1 and M_1 are the size of the most frequent class in the labeled data and unlabeled data, respectively. LA denotes the logit adjustment method [13]. † denotes models trained in the same setting.

TABLE XII
RESULTS ON DIFFERENT FRAMEWORKS WITH RESNET-50 BACKBONE ON LVIS v1

	ProCo	AP_b	AP_r	AP_c	AP_f	AP_m
Faster R-CNN [4]	✗	22.1	9.0	21.0	29.2	–
	✓	**24.7**	**15.5**	**24.2**	**29.3**	–
Mask R-CNN [93]	✗	22.5	9.1	21.1	30.1	21.7
	✓	**25.2**	**16.1**	**24.5**	**30.0**	**24.7**

We conduct experiments with 1x schedule.

N_y and Σ_y in Proposition 2. Overall, the results imply the broad utility of our approach across diverse datasets.

F. Long-Tailed Semi-Supervised Image Classification

We present experimental results of semi-supervised learning in Table XI. Fixmatch [66] is employed as the foundational framework to generate pseudo labels, and it is assessed effectiveness in comparison to other methods in long-tailed semi-supervised learning. We mainly follow the setting of DASO [69] except for substituting the semantic classifier based on the centroid prototype of each class with our representation branch for training. ProCo outperforms DASO across various levels of data imbalance and dataset sizes while maintaining the same training conditions. Specifically, in cases of higher data imbalance ($\gamma = 20$) and the ratio of unlabeled data ($N_1 = 50$), our proposed method exhibits a significant performance enhancement (with LA) of up to $\mathbf{2.8}\%$ when compared to DASO.

G. Long-Tailed Object Detection

In addition to image classification, we extend ProCo to object detection tasks. Specifically, we utilize Faster R-CNN [4] and Mask R-CNN [93] as foundational frameworks, integrating our proposed ProCo loss into the box classification branch. This method was implemented using mmdetection [94], adhering to the training settings of the original baselines. As depicted in Table XII, our approach yields noticeable enhancements on the LVIS v1 dataset, with both Faster R-CNN and Mask R-CNN demonstrating improved performance across various categories.

V. CONCLUSION

In this paper, we proposed a novel probabilistic contrastive (ProCo) learning algorithm for long-tailed distribution. Specifically, we employed a reasonable and straightforward von Mises-Fisher distribution to model the normalized feature space of samples in the context of contrastive learning. This choice offers two key advantages. First, it is efficient to estimate the distribution parameters across different batches by maximum likelihood estimation. Second, we derived a closed form of expected supervised contrastive loss for optimization by sampling infinity samples from the estimated distribution. This eliminates the inherent limitation of supervised contrastive learning that requires a large number of samples to achieve satisfactory performance. Furthermore, ProCo can be directly applied to semi-supervised learning by generating pseudo-labels for unlabeled data, which can subsequently be utilized to estimate the distribution inversely. We have proven the error bound of ProCo theoretically. Extensive experimental results on various classification and object detection datasets demonstrate the effectiveness of the proposed algorithm.

ACKNOWLEDGMENT

We would like to thank High-Flyer AI for the generous donation of computing resources.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1106–1114.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[6] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[7] G. Van Horn et al., "The inaturalist species classification and detection dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8769–8778.

[8] F. Graf, C. Hofer, M. Niethammer, and R. Kwitt, "Dissecting supervised constrastive learning," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 3821–3830.

[9] C. Fang, H. He, Q. Long, and W. J. Su, "Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training," *Proc. Nat. Acad. Sci.*, vol. 118, no. 43, 2021, Art. no. e2103091118.

[10] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9268–9277.

[11] B. Kang et al., "Decoupling representation and classifier for long-tailed recognition," in *Proc. Int. Conf. Learn. Representations*, 2020.

[12] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 1565–1576.

[13] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *Proc. Int. Conf. Learn. Representations*, 2021.

[14] P. Khosla et al., "Supervised contrastive learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[15] T. Li et al., "Targeted supervised contrastive learning for long-tailed recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6908–6918.

[16] J. Zhu, Z. Wang, J. Chen, Y.-P. P. Chen, and Y.-G. Jiang, "Balanced contrastive learning for long-tailed visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6898–6907.

[17] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia, "Parametric contrastive learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 695–704.

[18] P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang, "Contrastive learning based hybrid networks for long-tailed image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 943–952.

[19] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng, "Exploring balanced feature spaces for representation learning," in *Proc. Int. Conf. Learn. Representations*, 2020.

[20] I. Goodfellow et al., "Generative adversarial networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[21] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[22] J. Guo, C. Du, J. Wang, H. Huang, P. Wan, and G. Huang, "Assessing a single image in reference-guided image synthesis," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 753–761.

[23] Y. Wang, G. Huang, S. Song, X. Pan, Y. Xia, and C. Wu, "Regularizing deep networks with semantic data augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3733–3748, Jul. 2021.

[24] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "MetaSAug: Meta semantic augmentation for long-tailed visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5208–5217.

[25] Q. Cai, Y. Wang, Y. Pan, T. Yao, and T. Mei, "Joint contrastive learning with infinite possibilities," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12638–12648.

[26] X. Jia et al., "Semi-supervised multi-view deep discriminant representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2496–2509, Jul. 2021.

[27] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[28] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3581–3589.

[29] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 3546–3554.

[30] M. Kubat et al., "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. Int. Conf. Mach. Learn.*, 1997, pp. 179–186.

[31] B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos, "Class imbalance, redux," in *Proc. Int. Conf. Data Mining*, 2011, pp. 754–763.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[33] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5375–5384.

[34] A. Menon, H. Narasimhan, S. Agarwal, and S. Chawla, "On the statistical consistency of algorithms for binary classification under class imbalance," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 603–611.

[35] B. Kim and J. Kim, "Adjusting decision boundary for class imbalanced learning," *IEEE Access*, vol. 8, pp. 81674–81685, 2020.

[36] J. Zhang, L. Liu, P. Wang, and C. Shen, "To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions," 2019, *arXiv:1912.04486*.

[37] J. Ren, C. Yu, X. Ma, H. Zhao, and S. Yi, "Balanced Meta-Softmax for long-tailed visual recognition," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 4175–4186.

[38] J. Tan et al., "Equalization loss for long-tailed object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11659–11668.

[39] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 694–710.

[40] Y. Zang, C. Huang, and C. C. Loy, "FASA: Feature augmentation and sampling adaptation for long-tailed instance segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3437–3446.

[41] Y. Wang, Z. Ni, S. Song, L. Yang, and G. Huang, "Revisiting locally supervised learning: An alternative to end-to-end training," in *Proc. Int. Conf. Learn. Representations*, 2021.

[42] Y. Wang et al., "Efficienttrain: Exploring generalized curriculum learning for training visual backbones," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 5829–5841.

[43] G. Huang et al., "Glance and focus networks for dynamic visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4605–4621, Apr. 2023.

[44] K. V. Mardia, P. E. Jupp, and K. Mardia, *Directional Statistics*. Hoboken, NJ, USA: Wiley Online Library, 2000.

[45] X. Zhe, S. Chen, and H. Yan, "Directional statistics-based deep metric learning for image classification and retrieval," *Pattern Recognit.*, vol. 93, pp. 113–123, 2019.

[46] K. Roth, O. Vinyals, and Z. Akata, "Non-isotropy regularization for proxy-based deep metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7410–7420.

[47] T. R. Scott, A. C. Gallagher, and M. C. Mozer, "von Mises-Fisher loss: An exploration of embedding geometries for supervised learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10592–10602.

[48] S. Li, J. Xu, X. Xu, P. Shen, S. Li, and B. Hooi, "Spherical confidence learning for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15629–15637.

[49] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Mach. Learn. Res.*, vol. 6, pp. 1345–1382, 2005.

[50] H. Wang, S. Fu, X. He, H. Fang, Z. Liu, and H. Hu, "Towards calibrated hyper-sphere representation via distribution overlap coefficient for long-tailed learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 179–196.

[51] C. Ting, K. Simon, N. Mohammad, and H. Geoffrey, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[52] G. Jean-Bastien et al., "Bootstrap your own latent–A new approach to self-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[53] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 776–794.

[54] H. Kaiming, F. Haoqi, W. Yuxin, X. Saining, and G. Ross, "Momentum contrast for unsupervised visual representation learning," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

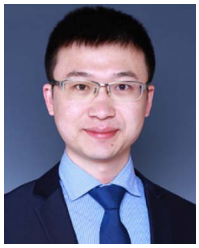
[55] D. Samuel and G. Chechik, "Distributional robustness loss for long-tail learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9475–9484.
[56] J. Cui, Z. Zhong, Z. Tian, S. Liu, B. Yu, and J. Jia, "Generalized parametric contrastive learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 22, 2023, doi: 10.1109/TPAMI.2023.3278694.
[57] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[58] C. Tian, W. Wang, X. Zhu, J. Dai, and Y. Qiao, "Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 8748–8763.
[59] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[60] L. Xiang, G. Ding, and J. Han, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 247–263.
[61] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu, "Long-tailed recognition by routing diverse distribution-aware experts," in *Proc. Int. Conf. Learn. Representations*, 2020.
[62] Y.-Y. He, J. Wu, and X.-S. Wei, "Distilling virtual examples for long-tailed recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 235–244.
[63] J. Li, Z. Tan, J. Wan, Z. Lei, and G. Guo, "Nested collaborative learning for long-tailed visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6939–6948.
[64] B. Zhu, Y. Niu, X.-S. Hua, and H. Zhang, "Cross-domain empirical risk minimization for unbiased long-tailed classification," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3589–3597.
[65] G. Huang and C. Du, "The high separation probability assumption for semi-supervised learning," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 12, pp. 7561–7573, Dec. 2022.
[66] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 596–608.
[67] J. Kim, Y. Hur, S. Park, E. Yang, S. J. Hwang, and J. Shin, "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 14567–14579.
[68] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, "CReST: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10852–10861.
[69] Y. Oh, D.-J. Kim, and I. S. Kweon, "DASO: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9776–9786.
[70] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Math.*, vol. 30, pp. 175–193, 1906.
[71] S. Sra, "A short note on parameter approximation for von Mises-Fisher distributions: And a fast implementation of," *Comput. Statist.*, vol. 27, pp. 177–190, 2012.
[72] M. Abramowitz et al., *Handbook of Mathematical Functions*, Chelmsford, MA, USA: Courier Corporation, 1964.
[73] W. Jitkrittum, A. K. Menon, A. S. Rawat, and S. Kumar, "ELM: Embedding and logit margins for long-tail learning," 2022, *arXiv:2204.13208*.
[74] Z. Hou, B. Yu, and D. Tao, "BatchFormer: Learning to explore sample relationships for robust representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7246–7256.
[75] H. Wang et al., "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5265–5274.
[76] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Open long-tailed recognition in a dynamic world," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1836–1851, Mar. 2024.
[77] J. Wang et al., "Seesaw loss for long-tailed instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9695–9704.
[78] A. Krizhevsky et al., "Learning multiple layers of features from tiny images," Univ. Toronto, Tech. Rep., 2009.
[79] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200–2011 dataset," California Inst. Technol., Tech. Rep. CNS-TR-2010-001, 2011.
[80] A. Gupta, P. Dollar, and R. Girshick, "LVIS: A dataset for large vocabulary instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5356–5364.
[81] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9716–9725.
[82] Y. Yang and Z. Xu, "Rethinking the value of labels for improving class-imbalanced learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 19290–19301.
[83] K. Tang, J. Huang, and H. Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1513–1524.
[84] J. Cui, S. Liu, Z. Tian, Z. Zhong, and J. Jia, "ResLT: Residual learning for long-tailed recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3695–3706, Mar. 2023.
[85] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 113–123.
[86] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.
[87] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 3008–3017.
[88] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995.
[89] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2361–2370.
[90] T. Li, L. Wang, and G. Wu, "Self supervision to distillation for long-tailed visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 610–619.
[91] A. M. H. Tiong, J. Li, G. Lin, B. Li, C. Xiong, and S. C. Hoi, "Improving tail-class representation with centroid contrastive learning," *Pattern Recognit. Lett.*, vol. 168, pp. 123–130, 2023.
[92] Y. L. Mengke Li and Y.-M. Cheung, "Long-tailed visual recognition via Gaussian clouded logit adjustment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6919–6928.
[93] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[94] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv: 1906.07155*.
[95] A. Maurer and M. Pontil, "Empirical bernstein bounds and sample-variance penalization," in *Proc. Annu. Conf. Comput. Learn. Theory*, 2009.

Chaoqun Du received the BS degree from the Department of Physics, Tsinghua University, Beijing, China, in 2020. He is currently working toward the PhD degree in control science and engineering with the Department of Automation, Institute of Industrial Intelligence and System in Tsinghua University. His current research interests include machine learning and computer vision.

Yulin Wang (Graduate Student Member, IEEE) received the BS degree in automation from Beihang University, Beijing, China, in 2019. He is currently working toward the PhD degree with the Department of Automation, Tsinghhua university. He was a visiting student with UC. Berkeley, Berkeley, CA, USA, in 2018. His research interests include computer vision and deep learning.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.

Shiji Song (Senior Member, IEEE) received the BS degree in mathematics from Harbin Normal University, Harbin, China, in 1986, and the MS and PhD degrees in mathematics from the Harbin Institute of Technology, Harbin, in 1989 and 1996, respectively. He is currently a professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include system modeling, control, and optimization, computational intelligence, and pattern recognition.



Gao Huang (Member, IEEE) received the BS degree from the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China, in 2009, and the PhD degree from the Department of Automation, Tsinghua University, Beijing, in 2015. He was a visiting research scholar with the Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO, USA, in 2013 and was a post-doctoral researcher with the Department of Computer Science, Cornell University, Ithaca, USA from 2015 to 2018. He is currently an associate professor with the Department of Automation, Tsinghua University. His research interests include machine learning and deep learning.

Authorized licensed use limited to: SHANGHAI UNIVERSITY. Downloaded on October 11,2024 at 03:26:59 UTC from IEEE Xplore. Restrictions apply.