Towards Calibrated Hyper-Sphere Representation via Distribution Overlap Coefficient for Long-Tailed Learning

Hualiang Wang[1,3], Siming Fu[1], Xiaoxuan He[1], Hangxiang Fang[1], Zuozhu Liu[1,2], and Haoji Hu[1(✉)]

[1] College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
{hualiang_wang,fusiming,Xiaoxiao_He,fhx,chuhp,haoji_hu}@zju.edu.cn
[2] ZJU-UIUC Institute, Zhejiang University, Hangzhou, China
zuozhuliu@intl.zju.edu.cn
[3] Angelalign Inc., Shanghai, China

Abstract. Long-tailed learning aims to tackle the crucial challenge that head classes dominate the training procedure under severe class imbalance in real-world scenarios. However, little attention has been given to how to quantify the dominance severity of head classes in the representation space. Motivated by this, we generalize the cosine-based classifiers to a von Mises-Fisher (vMF) mixture model, denoted as vMF classifier, which enables to quantitatively measure representation quality upon the hyper-sphere space via calculating distribution overlap coefficient. To our knowledge, this is the first work to measure representation quality of classifiers and features from the perspective of distribution overlap coefficient. On top of it, we formulate the inter-class discrepancy and class-feature consistency loss terms to alleviate the interference among the classifier weights and align features with classifier weights. Furthermore, a novel post-training calibration algorithm is devised to zero-costly boost the performance via inter-class overlap coefficients. Our method outperforms previous work with a large margin and achieves state-of-the-art performance on long-tailed image classification, semantic segmentation, and instance segmentation tasks (e.g., we achieve 55.0% overall accuracy with ResNetXt-50 in ImageNet-LT). Our code is available at https://github.com/VipaiLab/vMF_OP.

Keywords: von Mises-Fisher Distribution · Distribution overlap coefficient · Long-tailed learning · Representation learning

H. Wang and S. Fu—These authors contributed equally.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-20053-3_11.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13684, pp. 179–196, 2022.
https://doi.org/10.1007/978-3-031-20053-3_11

1 Introduction

Most real-world data comes with a long-tailed nature: a few head classes contribute the majority of data, while most tail classes comprise relatively few data.An undesired phenomenon is models [2,36,42] trained with long-tailed data perform better on head classes while exhibiting extremely low accuracy on tail ones.

To remedy it, one of the mainstream insights works on devising balanced classifiers [16,44,45] against imbalanced data. The cosine-based classifier discards the norms that have been proven to be larger on head classes [53]. The τ-norm classifier [16] manually shrinks the discrepancy among the norms of classifier weights through a τ-normalization function. In addition, some works [2,13,23,32] attach extra margin or scale terms on output scores to prompt classifiers to focus on data-scarce classes. Another prevailing method devotes to learning discriminative features using imbalanced data [5,31,34,43,51]. Range loss [51] is proposed to enlarge the inter-class feature distance and reduce the intra-class feature variation within the mini-batch data. Unsupervised discovery (UD) [43] uses self-supervised learning to help the model highlight tail classes from the feature level. In addition, LDA [31] transfers the learned feature distribution from the training domain to an ideal balanced domain.

While achieving promising performance, there lack of measures to quantitatively evaluate to what extent these classifiers or features can achieve the presumed "balanced" classifiers or "discriminative" features. Hence, one cannot measure how severely head classes dominate the features and classifiers in the high-dimensional representation space, resulting in confusions to guide further optimization for improved long-tailed learning.

To this end, we first extend cosine-based classifier as a von Mises-Fisher (vMF) distribution mixture model on hyper-sphere, denoted as the vMF classifier. Second, based on the representation space constructed by the vMF classifier, we mathematically define a novel measure between two probability density fuctions, denoted as distribution overlap coefficient o_Λ, to quantify to what extent the classifiers are "balanced" or features are "discriminative". A high o_Λ means that the two distributions (classes) are severely intertwined together. We suppose that o_Λ among classes in a "balance" classifier should be low enough, i.e., one class is not overwhelmingly dominated by other ones. "Discriminative" features means o_Λ between features and the corresponding classifier weights is high enough, i.e., features are well matched with correct classes.

On top of o_Λ, we provide an explicit optimization objective to boost the representation quality on hyper-sphere, i.e., to allow classifier weights to be distributed separately while aligning the weights of classifiers with features. Specifically, we propose two loss terms: the inter-class discrepancy and class-feature consistency loss. The first one minimizes the overlap among classifier weights, and the second one maximizes the overlap between features and the corresponding classifier weights. To further ease dominance of the head classes in classification decisions during inference, we develop a post-training calibration algorithm for classifier at zero cost based on the learned class-wise overlap coefficients.

We extensively validate our model on three typical visual recognition tasks, including image classification on benchmarks (ImageNet-LT [25] and iNaturalist2018 [39]), semantic segmentation on ADE20K dataset [55], and instance segmentation on LVIS-v1.0 dataset [9]. The experimental results and ablative study demonstrate our method consistently outperforms the state-of-the-art approaches on all the benchmarks.

Summary of Contributions:

- To the best of our acknowledge, we are the first in long-tailed learning to define the distribution overlap coefficient to evaluate representation quality for features and the proposed vMF classifiers.
- We formulate overlap-based inter-class discrepancy and class-feature consistency loss terms to alleviate the interference among the classifier weights and align features with classifier weights.
- We develop a post-training calibration algorithm for classifier at zero cost based on the learned class-wise overlap coefficients to ease dominance of the head classes in classification decisions during inference.
- Our models outperform previous work with a large margin and achieve state-of-the-art performance on long-tailed image classification, semantic segmentation and instance segmentation tasks.

2 Related Works

Classifier Design for Deep Long-Tailed Learning. In generic visual problems [11,54], the common practice of deep learning is to use linear classifier. However, long-tailed class imbalance often results in larger classifier weight norms for head classes than tail classes, which makes the linear classifier easily biased to dominant classes. To address long-tailed class imbalance, researchers design different types of classifiers. Scale-invariant cosine classifier [44] is proposed, where both the classifier weights and sample features are normalized. The τ-normalized classifier [16] rectifies the imbalance of decision boundaries by introducing the τ temperature factor for normalization [48]. Realistic taxonomic classifier (RTC) [45] addresses the issue with hierarchical classification where different samples are classified adaptively at different hierarchical levels. GistNet classifier [24] leverages the over-fitting to the popular classes to transfer class geometry from popular to few-shot classes. Causal classifier [37] records the bias by computing the exponential moving average features during training, and then removes the bad causal effect by subtracting the bias from prediction logits during inference.

Representation Learning for Long-Tailed Learning. Existing representation learning methods for long-tailed learning mainly focus on metric learning, prototype learning. Metric learning based methods [17,34,41] explore distance-based losses to learn a more discriminative feature space. LMLE [14] introduces a quintuple loss to learn representations that maintain both inter-cluster and

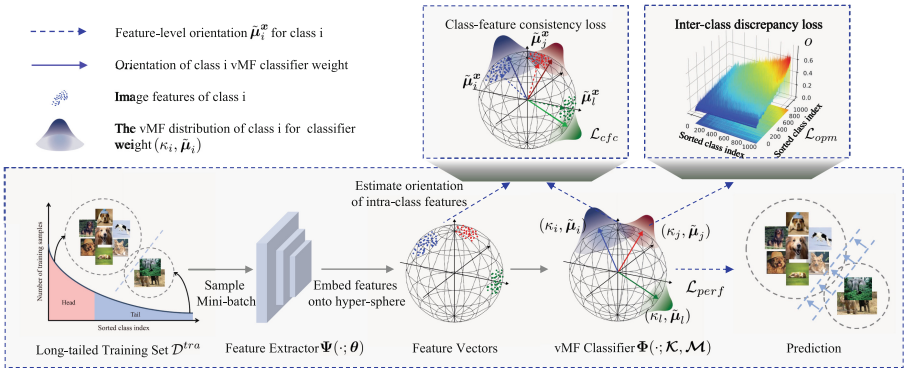


Fig. 1. Overview of our proposed method during the training period. **Bottom box** consists of the following steps in sequence: sampling a mini-batch images $\mathcal{B}$ from training set $\mathcal{D}^{tra}$, learning features by the feature extractor $\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta})$, embedding features onto hyper-sphere, predicting output via our proposed vMF classifier $\boldsymbol{\Phi}(\cdot;\boldsymbol{\mathcal{K}},\boldsymbol{\mathcal{M}})$ and calculating the performance loss value. **Upper boxes** introduce our proposed the class-feature consistency loss term $\mathcal{L}_{cfc}$ and inter-class discrepancy loss term $\mathcal{L}_{icd}$.

inter-class margins. Prototype learning based methods [26,58] seek to learn class-specific feature prototypes to enhance long-tailed learning performance. Open long-tailed recognition (OLTR) [26] innovatively explores the idea of feature prototypes to handle long-tailed recognition in an open world. Self-supervised pre-training (SSP) [47] uses self-supervised learning for model pre-training, followed by standard training on long-tailed data.

von Mises-Fisher Distribution. In directional statistics, the von Mises-Fisher distribution [15] is a probability distribution on the hyper-sphere. There are a lot of methods built on von Mises-Fisher distribution in machine learning and deep learning. The vMF Mixture Model (vMFMM) [10] proposes SFR model which assumes that the facial features are unit vectors and distributed according to a mixture of vMFs. The vMF k-means algorithm [28] is proposed based on the mixture vMF distribution to unsupervisedly evaluate the compactness and orientation of clusters. More recently, the t-vMF similarity [19] rebuilds the classifier by the proposed similarity based on vMF distribution to regularize features within deteriorated data. Sphere Confidence Face [20] minimizes KL divergence between spherical Dirac delta and r-radius vMF to achieve superior performance on face uncertainty learning.

Different from all them, to our best acknowledge, we are the first to quantify the distribution overlap coefficient between vMF distributions. Benefiting from it, we conduct a series of comprehensive and in-depth analyses to explore how to achieve high-quality representation space built upon vMF distribution.

3 Methodology

First, we briefly review the canonical pipeline of long-tailed learning, exemplified by long-tailed image classification, and elaborate on our proposed vMF classifier.

Afterward, we mathematically define the distribution overlap coefficient. On top of it, we further present the proposed the inter-class discrepancy loss and class-feature consistency loss terms. Finally, a post-training calibration algorithm is devised to zero-costly boost performance.

3.1 Build vMF Classifier on Hyper-Sphere

Let $\mathcal{D}^{tra} = \{\boldsymbol{I}^l, y^l\}$, $l \in \{1, \cdots, N\}$ be the training set, where $\boldsymbol{I}^l$ denotes an image sample and $y^l = i$ indicates it belongs to class i. Let C be the total numbers of classes, n_i be the number of samples in class i, where $\sum_{i=1}^{C} n_i = N$. The class prior distribution on training set can be defined as $p_{\mathcal{D}}^{tra}(i) = n_i/N$.

As shown in Fig. 1, given a pair $(\boldsymbol{I}^l, y^l)$ sampled from a mini-batch $\mathcal{B} \subset \mathcal{D}^{tra}$, feature vector $\boldsymbol{x}^l = \boldsymbol{\Psi}(\boldsymbol{I}^l; \boldsymbol{\theta}) \in \mathbb{R}^{1\times d}$ is extracted by the feature extractor $\boldsymbol{\Psi}(\cdot; \boldsymbol{\theta})$, of which learnable parameter $\boldsymbol{\theta}$ is instantiated by a neural network (e.g., ResNet). Then $\boldsymbol{x}^l$ is projected onto the unit hyper-sphere $\mathbb{S}^{d-1}$ via $\tilde{\boldsymbol{x}}^l = \boldsymbol{x}^l/\|\boldsymbol{x}^l\|_2$ and subsequently fed into the vMF classifier.

We depict the classifier with C classes as a mixture model with C von Mises-Fisher distributions on $\mathbb{S}^{d-1}$, each class containing two variables: the compactness $\kappa_i \in \mathbb{R}^+$ and the unit orientation vector $\tilde{\boldsymbol{\mu}}_i \in \mathbb{R}^{1\times d}$. Consequently, vMF classifier is well-defined as $\boldsymbol{\Phi}(\cdot; \boldsymbol{\mathcal{K}}, \boldsymbol{\mathcal{M}})$, where $\boldsymbol{\mathcal{K}} = \{\kappa_1, ..., \kappa_C\}$ and $\boldsymbol{\mathcal{M}} = \{\tilde{\boldsymbol{\mu}}_1, ..., \tilde{\boldsymbol{\mu}}_C\}$ are learnable compactness and orientation vectors for C classes, respectively. The probability density function (PDF) $p(\tilde{\boldsymbol{x}}|\kappa_i, \tilde{\boldsymbol{\mu}}_i)$ of i-th class is mathematically defined as:

$$p(\tilde{\boldsymbol{x}}|\kappa_i, \tilde{\boldsymbol{\mu}}_i) = C_d(\kappa_i) e^{\kappa_i \cdot \tilde{\boldsymbol{x}} \tilde{\boldsymbol{\mu}}_i^\top} = \frac{{\kappa_i}^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} \cdot I_{\frac{d}{2}-1}(\kappa_i)} e^{\kappa_i \cdot \tilde{\boldsymbol{x}} \tilde{\boldsymbol{\mu}}_i^\top}, \tag{1}$$

where $I_v(\kappa)$ is the modified Bessel function [18] of the first kind of real order v and $C_d(\kappa)$ is a normalization constant.

From the view of Bayes Theorem [29], given the class prior distribution $p_{\mathcal{D}}^{tra}(i)$ and $p(\tilde{\boldsymbol{x}}^l|\kappa_i, \tilde{\boldsymbol{\mu}}_i)$, the probability p_i^l for $\boldsymbol{I}^l$ belonging to class i can be formulated by the posterior probability $p(y^l = i|\tilde{\boldsymbol{x}}^l)$ as:

$$p_i^l = p(y^l = i|\tilde{\boldsymbol{x}}^l) = \frac{p_{\mathcal{D}}^{tra}(i) \cdot p(\tilde{\boldsymbol{x}}^l|\kappa_i, \tilde{\boldsymbol{\mu}}_i)}{\sum_{j=1}^{C} p_{\mathcal{D}}^{tra}(j) \cdot p(\tilde{\boldsymbol{x}}^l|\kappa_j, \tilde{\boldsymbol{\mu}}_j)}. \tag{2}$$

Equation 2 is the formulation of our vMF classifier. Our vMF classifer degrades to a balanced cosine classifier [32] with a temperature σ, when $\kappa_i = const\ \sigma, \forall i \in [1, C]$.

The performance loss $\mathcal{L}_{perf}$ of the mini-batch $\mathcal{B}$ is calculated by the cross-entropy function as follows:

$$\mathcal{L}_{perf} = -\frac{1}{N'} \sum_{l=1}^{N'} \sum_{i=1}^{C} \mathbb{1}[y^l = i] \cdot \log p_i^l, \tag{3}$$

Table 1. Derivatives for compactness and orientation of vMF classifier.

	∂o_Λ	$\partial \log p_i^l$
$\partial \kappa_i$	$o_\Lambda^2 \cdot \frac{\partial A_d(\kappa_i)}{\partial \kappa_i} \cdot (\kappa_j \cdot \tilde{\boldsymbol{\mu}}_i \tilde{\boldsymbol{\mu}}_j^\top - \kappa_i)$	$(1 - p_i^l) \cdot (\tilde{\boldsymbol{x}}^l \tilde{\boldsymbol{\mu}}_i^\top - A_d(\kappa_i))$
$\partial \kappa_j$	$o_\Lambda^2 \cdot (A_d(\kappa_i) \cdot \tilde{\boldsymbol{\mu}}_i \tilde{\boldsymbol{\mu}}_j^\top - A_d(\kappa_j))$	$-p_j^l \cdot (\tilde{\boldsymbol{x}}^l \tilde{\boldsymbol{\mu}}_j^\top - A_d(\kappa_j))$
$\partial \tilde{\boldsymbol{\mu}}_i$	$o_\Lambda^2 \cdot \kappa_j \cdot A_d(\kappa_i) \cdot \tilde{\boldsymbol{\mu}}_j$	$(1 - p_i^l) \cdot \kappa_i \cdot \tilde{\boldsymbol{x}}$
$\partial \tilde{\boldsymbol{\mu}}_j$	$o_\Lambda^2 \cdot \kappa_j \cdot A_d(\kappa_i) \cdot \tilde{\boldsymbol{\mu}}_i$	$-p_j^l \cdot \kappa_j \cdot \tilde{\boldsymbol{x}}$

where $\mathbb{1}[y = i]$ is the binary indicator that denotes whether the corresponding image comes from the i-th class and N' is the number of samples in a mini-batch.

The total loss $\mathcal{L}$ for mini-batch $\mathcal{B}$ in one iteration is calculated as:

$$\mathcal{L} = \mathcal{L}_{perf} + \lambda \cdot (\mathcal{L}_{icd} + \mathcal{L}_{cfc}), \tag{4}$$

where $\mathcal{L}_{icd}$ and $\mathcal{L}_{cfc}$ are proposed additional loss terms to regularize feature and classifier, which will be introduced in the subsequent subsection. λ is a hyper-parameter to adjust the weight of additional loss terms.

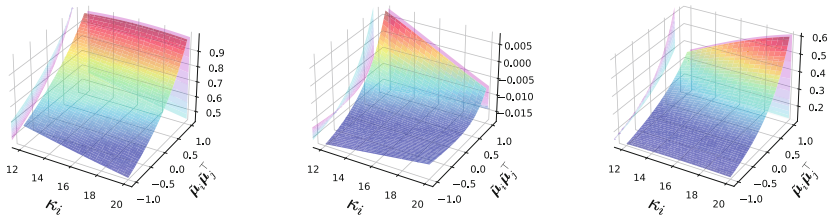
3.2 Quantify Distribution Overlap Coefficient on Hyper-Sphere

As aforementioned, we geometrically depict the classifier as a set of vMF distributions on $\mathbb{S}^{d-1}$. The distribution overlap coefficient [7] is mathematically explained as the area of intersection between two probability density functions. Based on it, we mathematically quantify distribution overlap coefficient to measure the intersection degree of two classes (vMF distribution) in the $\mathcal{S}^{d-1}$. In this paper, we provide the analytic expression o_Λ based on Kullback-Leibler divergence [30] for the vMF distribution [8]. Specifically, o_Λ is defined as:

$$o_\Lambda(\kappa_i, \kappa_j, \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\mu}}_j) = \frac{1}{1 + KL\{p(\tilde{\boldsymbol{x}}|\kappa_i, \tilde{\boldsymbol{\mu}}_i), p(\tilde{\boldsymbol{x}}|\kappa_j, \tilde{\boldsymbol{\mu}}_j)\}}, \tag{5}$$

where $KL\{p(\tilde{\boldsymbol{x}}|\kappa_i, \tilde{\boldsymbol{\mu}}_i), p(\tilde{\boldsymbol{x}}|\kappa_j, \tilde{\boldsymbol{\mu}}_j)\}$ is the Kullback-Leibler divergence between two vMF distributions, abbreviated as KL_{ij}:

$$\begin{aligned} KL_{ij} &= -\int_{\tilde{\boldsymbol{x}}} p(\tilde{\boldsymbol{x}}|\kappa_i, \tilde{\boldsymbol{\mu}}_i) \cdot \ln \frac{p(\tilde{\boldsymbol{x}}|\kappa_j, \tilde{\boldsymbol{\mu}}_j)}{p(\tilde{\boldsymbol{x}}|\kappa_i, \tilde{\boldsymbol{\mu}}_i)} \, d\tilde{\boldsymbol{x}} \\ &= \ln \frac{C_d(\kappa_i)}{C_d(\kappa_j)} \cdot \underbrace{\int_{\tilde{\boldsymbol{x}}} C_d(\kappa_i) \cdot e^{\kappa_i \cdot \tilde{\boldsymbol{x}} \tilde{\boldsymbol{\mu}}_i^\top} \, d\tilde{\boldsymbol{x}}}_{=1} \\ &+ \underbrace{(\int_{\tilde{\boldsymbol{x}}} \tilde{\boldsymbol{x}} \cdot C_d(\kappa_i) \cdot e^{\kappa_i \cdot \tilde{\boldsymbol{x}} \tilde{\boldsymbol{\mu}}_i^\top} \, d\tilde{\boldsymbol{x}})}_{=\mathbb{E}[\tilde{\boldsymbol{x}}] = A_d(\kappa_i) \cdot \tilde{\boldsymbol{\mu}}_i} (\kappa_i \cdot \tilde{\boldsymbol{\mu}}_i^\top - \kappa_j \cdot \tilde{\boldsymbol{\mu}}_j^\top) \\ &= \ln \frac{C_d(\kappa_i)}{C_d(\kappa_j)} + A_d(\kappa_i) \cdot (\kappa_i - \kappa_j \tilde{\boldsymbol{\mu}}_i \tilde{\boldsymbol{\mu}}_j^\top), \end{aligned} \tag{6}$$



(a) $o_\Lambda(\kappa_i, \kappa_j, \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\mu}}_j)$ (b) $\frac{\partial o_\Lambda(\kappa_i,\kappa_j,\tilde{\boldsymbol{\mu}}_i,\tilde{\boldsymbol{\mu}}_j)}{\partial \kappa_i}$ (c) $\frac{\partial o_\Lambda(\kappa_i,\kappa_j,\tilde{\boldsymbol{\mu}}_i,\tilde{\boldsymbol{\mu}}_j)}{\partial \tilde{\boldsymbol{\mu}}_i \tilde{\boldsymbol{\mu}}_j^\top}$

Fig. 2. Visualization of overlap coefficient $o_\Lambda(\kappa_i, \kappa_j, \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\mu}}_j)$ and partial derivatives for κ_i and $\tilde{\boldsymbol{\mu}}_i \tilde{\boldsymbol{\mu}}_j^\top$. To exhibit them in 3D coordination, κ_j is fixed to a certain value, instantiated as 16. κ_i and $\tilde{\boldsymbol{\mu}}_i \tilde{\boldsymbol{\mu}}_j^\top$ ($\tilde{\boldsymbol{\mu}}_i \in \mathbb{R}^{1\times 512}$) are uniformly sampled 100 values from range [12, 20] and range $[-1, 1]$, respectively.

where $A_d(\kappa_i) = I_{d/2}(\kappa_i)/I_{d/2-1}(\kappa_i)$ is non-decreasing and $0<A_d(\kappa_i)<1$. $\mathbb{E}[\tilde{\boldsymbol{x}}]$ is the expectation vector for $\tilde{\boldsymbol{x}} \sim p(\tilde{\boldsymbol{x}}|\kappa_i, \tilde{\boldsymbol{\mu}}_i)$ [33]. Generally $0<o_\Lambda \leq 1$. $o_\Lambda = 1$ (i.e., $\kappa_i = \kappa_j$ and $\tilde{\boldsymbol{\mu}}_i \tilde{\boldsymbol{\mu}}_j^\top = 1$) means they are completely congruent. $o_\Lambda \to 0$ indicates there is nearly no intersection between two distributions.

The derivatives of κ_i, κ_j, $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\boldsymbol{\mu}}_j$ for o_Λ are listed as the Col 1 of Table 1. And visualization for them is demonstrated in Fig. 2. Specifically, the partial derivative with respect to $\tilde{\boldsymbol{\mu}}_i \tilde{\boldsymbol{\mu}}_j^\top$ is non-negative.

The partial derivatives with respect to κ_i or κ_j are non-monotonous. An empirical conclusion is that κ_i and κ_j need to be kept at the same order of magnitude to achieve guaranteed performance, when using o_Λ as the optimization objective.

3.3 Improve Representation of Feature and Classifier via o_Λ

Inter-class Discrepancy Loss. To achieve the discriminative representation space in long-tailed learning, we seek to optimize our vMF classifier via shrinking the overlap among classes as much as possible to alleviate the overwhelm of the head classes on the tail ones. We denote the above optimization objective as the inter-class discrepancy loss term $\mathcal{L}_{icd}$, which acts function on the weights $\boldsymbol{\mathcal{K}}$ and $\boldsymbol{\mathcal{M}}$ of the vMF classifier.

First, we measure the average overlap coefficient o_i among class i and all the other classes, formulated by:

$$o_i = \frac{1}{C-1} \sum_{j=1, j\neq i}^{C} o_\Lambda(\kappa_i, \kappa_j, \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\mu}}_j). \tag{7}$$

Furthermore, we define the inter-class discrepancy loss term $\mathcal{L}_{icd}$ as:

$$\mathcal{L}_{icd} = \frac{1}{C} \sum_{i=1}^{C} o_i, \tag{8}$$

The proposed $\mathcal{L}_{icd}$ minimizes the average distribution overlap coefficient to regularize distributions, contributing to a more distinction-prone classifier on $\mathbb{S}^{d-1}$.

Class-Feature Consistency Loss. In addition, the poorly matching between the feature vectors and the corresponding classifier weights derives unsatisfied performance, especially for the sample-starved classes. Class-feature consistency loss term $\mathcal{L}_{cfc}$ is proposed to alleviate the above issue by aligning features with the corresponding classifier weights as far as possible.

Specifically, we first fit the class-wise feature distribution $(\kappa^{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}^{\boldsymbol{x}})$ within the mini-batch $\mathcal{B}$. The class set involved in $\mathcal{B}$ is denote as $\mathcal{C}'$. For a certain class $i \in \mathcal{C}'$, the feature-level orientation vector $\tilde{\boldsymbol{\mu}}_i^{\boldsymbol{x}}$ is defined as:

$$\tilde{\boldsymbol{\mu}}_i^{\boldsymbol{x}} = \frac{\sum_{l=1,y^l=i}^{N'} \boldsymbol{x}^l}{\| \sum_{l=1,y^l=i}^{N'} \boldsymbol{x}^l \|_2}. \tag{9}$$

Considering that the compactness κ is over-sensitive to sample number and intractable to be estimated [10], κ is shared between the feature and the corresponding classifier weight, i.e., feature-level compactness $\kappa_i^{\boldsymbol{x}}$ for class i is equal to κ_i. Then, $\mathcal{L}_{cfc}$ is formulated as following:

$$\mathcal{L}_{cfc} = \mathbb{E}_{i \in \mathcal{C}'}[1 - o_{\Lambda}(\kappa_i, \kappa_i^{\boldsymbol{x}}, \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\mu}}_i^{\boldsymbol{x}})], \tag{10}$$

where $\mathbb{E}$ indicates the average function. $\mathcal{L}_{cfc}$ is, in effect, equivalent to maximizing the distribution overlap coefficient between features and the corresponding classifier weights.

3.4 Calibrate Classifier Weight Beyond Training via o_{Λ}

Despite exerting additional loss terms to regularize features and classifiers, the overwhelm of the head classes on the tail ones is, in effect, tough to eradicate under a highly imbalanced dataset. We visualize the compactness of the classifier and the average overlap coefficients from a well-trained model, as demonstrated in Col 1 of Fig. 3. The head classes share larger compactness and smaller overlap coefficients, however, the case for tail ones is reversed.

A general summary of the calibration strategy is that increase the compactness for classes that are severely overlapped with other classes. Specifically, given a well-trained vMF classifier $\boldsymbol{\Phi}(\cdot; \boldsymbol{\mathcal{K}}, \boldsymbol{\mathcal{M}})$, we first apply Eq. 7 to obtain the average overlap coefficient for each class, denoted as $\boldsymbol{\mathcal{O}} = \{o_1, ..., o_C\}$. Then we use a maximum-minimum normalization strategy to reconcile $\boldsymbol{\mathcal{O}}$ to the same value range as $\boldsymbol{\mathcal{K}}$, to make sure that both are on the same order of magnitude by:

$$\hat{o}_i = \frac{o_i - o^{min}}{o^{max} - o^{min}} \cdot (\kappa^{max} - \kappa^{min}) + \kappa^{min}, \tag{11}$$

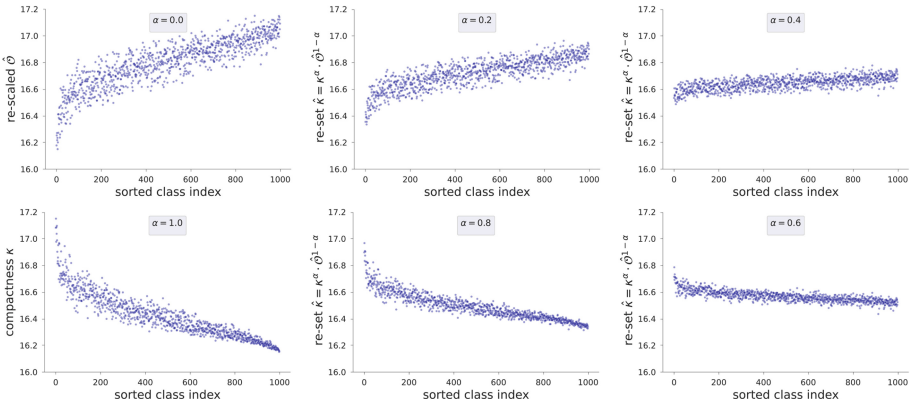


Fig. 3. The calibrated compactness of vMF classifier (trained on ImageNet-LT with ResNetXt-50 feature extractor). Under different α settings, we adjust κ via Eq. 11 and Eq. 12. Each picture represents the value of re-scaled $\hat{\boldsymbol{\mathcal{K}}}$ when α equals to the corresponding value. When $\alpha = 0$, it indicates $\hat{\kappa}_i = \hat{o}_i$, while $\alpha = 1$, $\hat{\kappa}_i = \kappa_i$.

where o^{max} and o^{min} are maximum and minimum values of set $\boldsymbol{\mathcal{O}}$, respectively, as well as κ^{max} and κ^{min}. We reset compactness vector as $\hat{\boldsymbol{\mathcal{K}}} = \{\hat{\kappa}_1, ..., \hat{\kappa}_C\}$, formulated as following:

$$\hat{\kappa}_i = \kappa_i^{\alpha} \cdot \hat{o}_i^{1-\alpha}, \tag{12}$$

$\alpha \in [0, 1]$ is a hyper-parameter to balance the importance contribution to the re-scaled $\hat{\boldsymbol{\mathcal{K}}}$ as shown in Fig. 3. In the inference period, we comply with a canonical assumption that the classes on the test set follow the uniform distribution, i.e., $p_{\mathcal{D}}^{test}(i) = 1/C$. Consequently, we replace $p_{\mathcal{D}}^{tra}(i)$ by $p_{\mathcal{D}}^{test}(i)$ in Eq. 2, and the vMF classifier is calibrated as $\boldsymbol{\Phi}(\cdot; \hat{\boldsymbol{\mathcal{K}}}, \boldsymbol{\mathcal{M}})$.

Moreover, our post-training calibration algorithm is capable of extending to several wide-used classifiers for cost-free performance boosting. Next, we instantiate how to apply the algorithm above to calibrate the weights of τ-norm [16], causal classifiers [38] and linear classifiers. Given the weight vector $\boldsymbol{w}_i^{\tau}$ of class i from a well-trained τ-norm classifier $\boldsymbol{W}^{\tau}$, we equivalently convert $\boldsymbol{w}_i^{\tau}$ into compactness $\kappa_i = \|\boldsymbol{w}_i^{\tau}\|_2^{1-\tau}$ and orientation vector $\tilde{\boldsymbol{\mu}}_i = \boldsymbol{w}_i^{\tau} / \|\boldsymbol{w}_i^{\tau}\|_2$. After calibration via Eq. 11 and Eq. 12, $\boldsymbol{w}_i^{\tau}$ is rebuilt by producting orientation vector and re-balanced compactness together. Along the same lines, the weight vector $\boldsymbol{w}_i^{cau}$ for a well-trained causal classifier $\boldsymbol{W}^{cau}$ is converted to $\kappa_i = \|\boldsymbol{w}_i^{cau}\|_2 / (\|\boldsymbol{w}_i^{cau}\|_2 + \gamma)$ and $\tilde{\boldsymbol{\mu}}_i = \boldsymbol{w}_i^{cau} / \|\boldsymbol{w}_i^{cau}\|_2$. The weight vector $\boldsymbol{w}_i^{lin}$ for a well-trained linear classifier $\boldsymbol{W}^{lin}$ is converted to $\kappa_i = \|\boldsymbol{w}_i^{lin}\|_2$ and $\tilde{\boldsymbol{\mu}}_i = \boldsymbol{w}_i^{lin} / \|\boldsymbol{w}_i^{lin}\|_2$. γ and τ are both the hyper-parameters for classifiers above. (Detail proofs in Appendix **A.2**)

4 Experiments

In this section, we conduct a series of experiments to validate the effectiveness of our method. Below we present our experimental analysis and ablation study on the image classification task in Sect. 4.1, followed by our results on semantic segmentation task and instance segmentation task in Sect. 4.2.

Table 2. Results on ImageNet-LT in terms of accuracy (Acc) under 90 and 200 training epochs. In this table, CR, DT, RL and CD indicate class re-balancing, decouple training, representation learning and classifier design, respectively. † indicates only vMF classifier is applied w/o additional loss terms and post-training calibration algorithm.

Type	Method	90 epochs				200 epochs			
		Many	Med	Few	All	Many	Med	Few	All
Baseline	Softmax	66.5	39.0	8.6	45.5	66.9	40.4	12.6	46.8
CR	Focal Loss [23]	66.9	39.2	9.2	45.8	67.0	41.0	13.1	47.2
	BALMS [32]	61.7	48.0	29.9	50.8	62.4	47.7	32.1	51.2
	LDAM [2]	62.3	47.4	32.5	51.1	60.0	49.2	31.9	51.1
	LADE [13]	62.2	48.6	31.8	51.5	63.1	47.7	32.7	51.6
	DisAlign [50]	62.7	52.1	31.4	53.4	–	–	–	–
DT	IB-CRT [16]	62.6	46.2	26.7	49.9	64.2	46.1	26.0	50.3
	CB-CRT [16]	62.4	39.3	14.9	44.9	60.9	36.9	13.5	43.0
	MiSLAS [54]	62.1	48.9	31.6	51.4	65.3	50.6	33.0	53.4
	xERM_{TDE} [57]	–	–	–	–	68.6	50.0	27.5	54.1
RL	OLTR [26]	58.2	45.5	19.5	46.7	62.9	44.6	18.8	48.0
	SSP [47]	65.6	49.6	30.3	53.1	67.3	49.1	28.3	53.3
	DRO-LT [34]	–	–	–	–	64.0	49.8	33.1	53.5
	PaCo [5]	59.7	51.7	36.6	52.7	63.2	51.6	39.2	54.4
CD	τ-norm [16]	61.8	46.2	27.4	49.6	–	–	–	–
	TDE [38]	63.0	48.5	31.4	51.8	64.9	46.9	28.1	51.3
	Ours†	64.2	49.8	26.9	52.2	65.9	50.5	28.1	53.4
	Ours	**64.2**	**51.4**	**31.8**	**53.7**	**65.1**	**52.8**	**34.2**	**55.0**

4.1 Long-Tailed Image Classification Task

Datasets and Setup. We perform experiments on long-tailed image classification datasets, including the ImageNet-LT [25] and iNaturalist2018 [39].

- ImageNet-LT is a long-tailed version of the ImageNet dataset by sampling a subset following the Pareto distribution with power value 6. It contains 115.8K images from 1,000 categories, with class cardinality ranging from 5 to 1,280.
- iNaturalist2018 is the largest dataset for long-tailed visual recognition. It contains 437.5K images from 8,142 categories. It is extremely imbalanced with an imbalance factor of 512.

Experimental Details. For image classification on ImageNet-LT, we implement all experiments in PyTorch. Following [5,13,38], we use ResNetXt-50 [46] as the feature extractor for all methods. We conduct model training with the SGD optimizer based on batch size 512, momentum 0.9. In both training epochs

Table 3. Benchmarking on iNaturalists2018 in accuracy (%). DT, CD and RL indicate decouple training, classifier design and representation learning, respectively.

Type	Method	iNaturalist2018			
		Many	Med	Few	All
Baseline	CE	72.2	63.0	57.2	61.7
DT	Decoupling [16]	65.6	65.3	65.5	65.6
	BBN [56]	49.4	70.8	65.3	66.3
CD	TDE [38]	–	–	–	68.7
	τ-norm [16]	65.6	65.3	65.5	65.6
RL	TSC [21]	72.6	70.6	67.8	69.7
	DisAlign [50]	69.0	71.1	70.2	70.6
	Ours	**72.8**	**71.7**	**70.0**	**71.0**

(90 and 200 training epochs), the learning rate is decayed by a cosine scheduler [27]. On iNaturalist2018 [39] dataset, we use ResNet-50 [46] as the feature extractor for all methods with 200 training epochs, with the same experimental parameters set for the other. By default, learnable κ for all categories are initialized as 16 and λ is 0.2. Moreover, we use the same basic data augmentation (i.e., random resize and crop to 224, random horizontal flip, color jitter, and normalization) for all methods.

Comparison with State of the Arts. In our paper, the comparison methods use single models. Note that there are also ensemble models for long-tailed classification, e.g., RIDE [42] and TADE [52]. For fair comparisons, following xERM [57], we will not include their results in the experiments. Table 2 shows the long-tailed results on ImageNet-LT. We adopt the performance data from the deep long-tailed survey [53] for various methods at 90 and 200 training epochs to make a fair comparison. Our approach achieves 53.7% and 55.0% in overall accuracy, which outperforms the state-of-the-art methods by a significant margin at 90 and 200 training epochs, respectively. Compared with representation learning methods, our method surpasses SSP by 0.6% (53.7% vs 53.1%) at 90 training epochs and outperforms SSP by 1.7% (55.0% vs 53.3%) at 200 training epochs. In addition, our method obtains higher performance by 1.0% (53.7% vs 52.7%) and 0.6% (55.0% vs 54.4%) comparing to PaCo at 90 and 200 training epochs, respectively. We observe that our vMF classifier (w/o proposed additional loss terms and post-training calibration algorithm) still achieves better performance than previous classifier design strategies, i.e., our vMF classifier surpasses τ-norm and TDE which by 2.6% (52.2% vs 49.6%) and 0.4% (52.2% vs 51.8%) at 90 epochs. Moreover, our vMF classifier performs better when training 200 epochs than 90 epochs (53.4% vs 52.2%), in contrast to TDE (51.3% vs 51.8%). This shows that our vMF classifier has more potential to fit data better and learn better representations.

Table 4. Performance of semantic segmentation on ADE20K and instance segmentation on LVIS-v1.0. R-50 and R-101 denote ResNet-50 and ResNet-101, respectively. 'Cascade-R101' is for Cascade Mask R-CNN [1].

Model	Method	ADE20K		Model	Method	LVIS-v1.0	
		mIoU	mAcc			AP	AP_b
OCRNet (HRNet-W18)	Baseline	40.8	50.9	Cascade (R101)	Cross-Entropy	22.6	25.2
	Ours	**41.5**	**52.9**		De-confound [38]	23.5	25.8
DeepLabV3+ (R-50)	Baseline	44.9	55.0		TDE [38]	27.1	30.0
	DisAlign [50]	45.7	**57.3**		EQL v2 [35]	28.8	32.3
	Ours	**45.9**	57.0		DisAlign [50]	28.9	32.7
DeepLabV3+ (R-101)	Baseline	46.4	56.7		BAGS [22]	27.9	31.5
	DisAlign [50]	47.1	59.5		Seesaw Loss [40]	29.6	32.5
	Ours	**47.2**	**59.8**		**Ours**	**29.8**	**32.9**

Furthermore, Table 3 presents the experimental results on the naturally-skewed dataset iNaturalist2018. Compared with the improvement brought by representation learning and classifier design approaches, our method achieves competitive result (71.0%) consistently.

4.2 Long-Tailed Semantic and Instance Segmentation Task

To further validate our method, weconduct comprehensive experiments on the semantic and instance segmentation datasets, i.e., ADE20K [55] and LVIS-v1.0 [9].

Dataset and Setup

- ADE20K is a scene parsing dataset covering 150 fine-grained semantic concepts and it is one of the most challenging semantic segmentation datasets. The training set contains 20,210 images with 150 semantic classes. The validation and test set contain 2,000 and 3,352 images respectively.
- LVIS-v1.0 contains 1230 categories with both bounding box and instance mask annotations. LVIS-v1.0 divides all categories into 3 groups based on the number of images that contain those categories: frequent (>100 images), common (11–100 images) and rare (<10 images). We train the models with 57K train images and report the accuracy on 5K val images.

Experimental Details. We evaluate our method using two wide-adopted segmentation models (OCRNet [49] and DeepLabV3+ [4]) based on different backbone networks. We initialize the backbones using the models pre-trained on ImageNet [6] and the framework randomly. All models are trained with an image

Table 5. Ablation on our proposed two loss terms and the loss weight λ. 'None' indicates only the performance loss term is applied to train model. '0.1' means λ is set as 0.1.

Additional Loss	All	Many	Med.	Few
Baseline	51.8	62.6	48.9	31.3
None	52.2	64.2	49.8	26.9
0.2, $\mathcal{L}_{icd}$	52.9	64.2	49.8	31.7
0.2, $\mathcal{L}_{cfc}$	53.1	65.3	50.4	27.9
0.2, $\mathcal{L}_{icd}$, $\mathcal{L}_{cfc}$	**53.5**	**65.4**	**50.8**	**29.1**
0.1, $\mathcal{L}_{icd}$, $\mathcal{L}_{cfc}$	53.2	65.0	50.7	27.9
0.4, $\mathcal{L}_{icd}$, $\mathcal{L}_{cfc}$	52.6	64.9	50.1	26.8
0.3, $\mathcal{L}_{icd}$, $\mathcal{L}_{cfc}$	53.2	65.1	50.8	28.0

Table 6. Ablation on the hyper-parameter α of post-training calibration algorithm with different classifiers. ‡ indicates the corresponding classifier is calibrated under the optimal α.

$\mathcal{K}$	All	Many	Med.	Few
Linear	43.2	66.2	35.4	6.0
Linear ‡	48.3	60.9	46.4	19.9
τ-norm [16]	48.6	69.9	42.5	10.1
τ-norm [16] ‡	53.0	66.5	50.2	24.1
Causal [38]	49.0	69.6	43.0	12.2
Causal [38] ‡	50.9	69.0	45.8	17.5
Ours	53.5	65.4	50.8	29.1
Ours ‡	**53.7**	**63.9**	**51.5**	**32.4**

size of 512×512 and 160K iterations in total. We train the models using Adam optimizer with the initial learning rate 0.01, weight decay 0.0005 and momentum 0.9. Furthermore, We implement our method on LVIS-v1.0 with mmdetection [3] and train Mask R-CNN [12] with random sampler by 2x training schedule. The model is trained with batch size of 16 for 24 epochs. The optimizer is SGD with momentum 0.9 and weight decay 0.0001. The initial learning rate is 0.02 with 500 iterations' warm up. For above two tasks, we set the optimal configuration in our experiments that is all learnable $\mathcal{K}$ are initialized to 16.

Comparison with State of the Arts. For the semantic segmentation task, The numerical results and comparison with other peer methods are reported in left part of Table 4. Our method achieves 0.7% (41.5% vs 40.8%) improvement in mIoU using OCRNet with HRNet-W18. Moreover, our method outperforms the baseline with large at 1.0% (45.9% vs 44.9%) in mIoU using DeeplabV3+ with ResNet-50 when the iteration is 160K. Even with a stronger backbone: ResNet-101, our method also achieves 0.8% (47.2% vs 46.4%) mIoU improvement than

baseline. For the instance segmentation task, we report quantitative results and compare our method with recent work in the right part of Table 3. Our method can achieve 29.8% in AP and 32.9% in AP_b when applied to the Cascade-R101. Apart from the CE loss baseline, we further compare our method with recent designs for long-tailed instance segmentation. Our method surpasses Seesaw Loss by 0.2% (29.8% vs 29.6%) AP, and surpasses DisAlign by 0.9% (29.8% vs 28.9%) AP, which reveals the effectiveness of our method (Table 6).

4.3 Ablation Study

We conduct ablation study on ImageNet-LT dataset to further understand the hyper-parameters of our methods and the effect of each proposed component.

Ablation Study on Two Additional Loss Terms and the Loss Weight λ. Firstly, we evaluate the effectiveness of the proposed $\mathcal{L}_{icd}$ and $\mathcal{L}_{cfc}$. Setting $\lambda = 0.2$ and initializing $\kappa = 16$, we train vMF classifier w/o additional loss terms, w/ $\mathcal{L}_{icd}$, w/ $\mathcal{L}_{cfc}$ and w/ both of them, respectively. Experimental results are reported in Row 1–4 of Table 5. Our baseline is the balanced cosine classifier [32]. Conclusions are **(1).** Giving additional surveillance via $\mathcal{L}_{icd}$ is beneficial to the performance on tail classes. It can be seen from the second and third rows in the Table 5. The performance of the tail of the loss term has been greatly improved (26.9% vs 31.7%). **(2).** $\mathcal{L}_{cfc}$ gains the non-trival performance improvements on all classes. **(3).** Simultaneously adopting the above two loss terms further improves the accuracy by 1.3%, further widening the performance gap up to 1.7% compared with the baseline. Secondly, we conduct four experiments on different λ. Row 5–8 of Table 5 show $\lambda = 0.2$ is the optimal setting.

Ablation Study on Post-calibration Algorithm with Different Classifier To verify the versatility of our post-training calibration algorithm, we perform it on our vMF, linear, τ-norm ($\tau = 0.7$, optimal setting in [16]) and causal [38] classifiers, following Sect. 3.4. All of them have trained on ImageNet-LT with ResNetXt-50. We set the hyper-parameter α in the interval 0 to 1 with a stride of 0.1 and take the eleven sets of values to conduct ablation experiments on above classifiers. For linear classifier, the optimal $\alpha = 0.7$, where our algorithm improves allover accuracy performance by 5.1%. For τ-norm classifier and causal classifier, under the optimal $\alpha = 0.1$, the allover accuracy is improved by 4.4% and 1.9%. When $\alpha = 0.2$, our vMF classifier achieves highest accuracy 53.7%. The reason for slight improvement on ours may be because it has already learned with proposed loss terms ($\mathcal{L}_{cfc}$ and $\mathcal{L}_{icd}$) that are also based on distribution overlap coefficient.

5 Conclusions

In this paper, we extend cosine-based classifiers as a vMF distribution mixture model on hyper-sphere, denoted as the vMF classifier. Benefiting from the

representation space constructed by the vMF classifier, we define the distribution overlap coefficient to measure the representation quality for features and classifiers. Based on distribution overlap coefficient, we formulate the inter-class discrepancy and class-feature consistency loss terms to alleviate the interference among the classifier weights and align features with classifier weights. Furthermore, we develop a novel post-training calibration algorithm to zero-costly boost the performance. Our method outperforms previous work with a large margin and achieves state-of-the-art performance on long-tailed image classification, semantic segmentation, and instance segmentation tasks.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (U21B2004), the Zhejiang Provincial key RD Program of China (2021C01119) , and the Zhejiang University-Angelalign Inc. R & D Center for Intelligent Healthcare.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: high quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **43**(5), 1483–1498 (2019)
2. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Adv. Neural Inf. Process. Syst. **32** (2019)
3. Chen, K., et al.: Mmdetection: open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
5. Cui, J., Zhong, Z., Liu, S., Yu, B., Jia, J.: Parametric contrastive learning (2021)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
7. Dhaker, H., Ngom, P., Mbodj, M.: Overlap coefficients based on kullback-leibler divergence: exponential populations case. Int. J. Appl. Math. Res. **6**(4) (2017)
8. Diethe, T.: A note on the kullback-leibler divergence for the von mises-fisher distribution. arXiv preprint arXiv:1502.07104 (2015)
9. Gupta, A., Dollar, P., Girshick, R.: Lvis: a dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5356–5364 (2019)
10. Hasnat, M., Bohné, J., Milgram, J., Gentric, S., Chen, L., et al.: von mises-fisher mixture model-based deep learning: application to face verification. arXiv preprint arXiv:1706.04264 (2017)
11. He, K., Fan, H., Wu, Y., Xie, S., 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)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
13. Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., 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)

14. Huang, C., Li, Y., Loy, C.C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5375–5384 (2016)
15. Jupp, P.E., Mardia, K.V.: Maximum likelihood estimators for the matrix von mises-fisher and bingham distributions. Ann. Stat. **7**(3), 599–606 (1979)
16. Kang, B., et al.: Decoupling representation and classifier for long-tailed recognition (2019)
17. Kang, B., Li, Y., Xie, S., Yuan, Z., Feng, J.: Exploring balanced feature spaces for representation learning. In: International Conference on Learning Representations (2021)
18. Kent, J.: Some probabilistic properties of bessel functions. Ann. Probabil., 760–770 (1978)
19. Kobayashi, T.: t-vmf similarity for regularizing intra-class feature distribution. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6612–6621 (2021)
20. Li, S., Xu, J., Xu, X., Shen, P., Li, S., Hooi, B.: Spherical confidence learning for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15629–15637 (2021)
21. Li, T., et al.: Targeted supervised contrastive learning for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6918–6928 (2022)
22. Li, Y., et al.: Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10991–11000 (2020)
23. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
24. Liu, B., Li, H., Kang, H., Hua, G., Vasconcelos, N.: Gistnet: a geometric structure transfer network for long-tailed recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8209–8218 (2021)
25. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
26. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., 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)
27. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
28. Mash'al, M., Hosseini, R.: K-means++ for mixtures of von mises-fisher distributions. In: 2015 7th Conference on Information and Knowledge Technology (IKT), pp. 1–6. IEEE (2015)
29. Nicholls, E., Stark, A.: Bayes' theorem. Med. J. Aust. **2**(26), 1335–1339 (1971)
30. Papadopoulos, C.I.: On the Kullback-Leibler information measure and statistical inference. Wayne State University (1971)
31. Peng, Z., Huang, W., Guo, Z., Zhang, X., Jiao, J., Ye, Q.: Long-tailed distribution adaptation. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3275–3282 (2021)
32. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Adv. Neural Inf. Process. Syst. **33**, 4175–4186 (2020)

33. Romanazzi, M.: Discriminant analysis with high dimensional von mises-fisher distributions. In: 8th Annual International Conference on Statistics, pp. 1–16. Athens Institute for Education and Research (2014)
34. Samuel, D., Chechik, G.: Distributional robustness loss for long-tail learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
35. Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: Equalization loss v2: a new gradient balance approach for long-tailed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1685–1694 (2021)
36. Tan, J., et al.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11662–11671 (2020)
37. Tang, K., Huang, J., Zhang, H.: Long-tailed classification by keeping the good and removing the bad momentum causal effect. In: NeurIPS (2020)
38. Tang, K., Huang, J., Zhang, H.: Long-tailed classification by keeping the good and removing the bad momentum causal effect. Adv. Neural Inf. Process. Syst. **33**, 1513–1524 (2020)
39. Van Horn, G., et al.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8769–8778 (2018)
40. Wang, J., et al.: Seesaw loss for long-tailed instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
41. Wang, P., Han, K., Wei, X.S., Zhang, L., Wang, L.: Contrastive learning based hybrid networks for long-tailed image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 943–952 (2021)
42. Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S.: Long-tailed recognition by routing diverse distribution-aware experts. In: International Conference on Learning Representations (2021)
43. Weng, Z., Ogut, M.G., Limonchik, S., Yeung, S.: Unsupervised discovery of the long-tail in instance segmentation using hierarchical self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2603–2612 (2021)
44. Wu, T., Liu, Z., Huang, Q., Wang, Y., Lin, D.: Adversarial robustness under long-tailed distribution (2021)
45. Wu, T.Y., Morgado, P., Wang, P., Ho, C.H., Vasconcelos, N.: Solving long-tailed recognition with deep realistic taxonomic classifier
46. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431 (2016)
47. Yang, Y., Xu, Z.: Rethinking the value of labels for improving class-imbalanced learning. Adv. Neural Inf. Process. Syst. **33**, 19290–19301 (2020)
48. Ye, H.J., Chen, H.Y., Zhan, D.C., Chao, W.L.: Identifying and compensating for feature deviation in imbalanced deep learning (2020)
49. Yuan, Y., Wang, J.: Ocnet: object context network for scene parsing (2018)
50. Zhang, S., Li, Z., Yan, S., He, X., Sun, J.: Distribution alignment: a unified framework for long-tail visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2361–2370 (2021)
51. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y.: Range loss for deep face recognition with long-tailed training data. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5409–5418 (2017)

52. Zhang, Y., Hooi, B., Hong, L., Feng, J.: Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. arXiv preprint arXiv:2107.09249 (2021)
53. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: a survey. arXiv preprint arXiv:2110.04596 (2021)
54. Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16489–16498 (2021)
55. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 633–641 (2017)
56. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9719–9728 (2020)
57. Zhu, B., Niu, Y., Hua, X.S., Zhang, H.: Cross-domain empirical risk minimization for unbiased long-tailed classification. In: AAAI Conference on Artificial Intelligence (2022)
58. Zhu, L., Yang, Y.: Inflated episodic memory with region self-attention for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4344–4353 (2020)