Distribution Alignment Optimization through Neural Collapse for Long-tailed Classification

Jintong Gao [1] **He Zhao** [2] **Dandan Guo** [1] **Hongyuan Zha** [3]

Abstract

A well-trained deep neural network on balanced datasets usually exhibits the Neural Collapse (NC) phenomenon, which is an informative indicator of the model achieving good performance. However, NC is usually hard to be achieved for a model trained on long-tailed datasets, leading to the deteriorated performance of test data. This work aims to induce the NC phenomenon in imbalanced learning from the perspective of distribution matching. By enforcing the distribution of last-layer representations to align the ideal distribution of the ETF structure, we develop a Distribution Alignment Optimization (DisA) loss, acting as a plug-and-play method can be combined with most of the existing long-tailed methods, we further instantiate it to the cases of fixing classifier and learning classifier. The extensive experiments show the effectiveness of DisA, providing a promising solution to the imbalanced issue. Our code is available at DisA.

1. Introduction

Deep neural networks on balanced datasets have become well-established with many techniques for solving various tasks in several domains (Zeng et al., 2022; Hu et al., 2023; Allingham et al., 2023). However, new challenges arise when we venture into imbalanced/long-tailed datasets. In long-tailed classification, for example, the majority classes have a substantial number of samples while the minority classes have fewer. Traditional classification methods without considering such class imbalance tend to overly focus on the majority classes, resulting in insufficient learning for the minority classes.

Recently, studies in Papyana et al. (2020); Kothapalli et al. (2023); Rangamani et al. (2023) show a well-trained neural network with cross-entropy loss on a given balanced classification task usually exhibits the Neural Collapse (NC) phenomenon, meaning that the last-layer features collapsing into within-class means and the classifier weight vectors converge to the simplex Equiangular Tight Frame (ETF) geometric structure. However, on long-tailed datasets, a model can hardly achieve NC if trained in the same way as on balanced datasets (Fang et al., 2021; Thrampoulidis et al., 2022). The mean feature vectors no longer form an ETF structure with Cross-Entropy (CE) or Mean Squared Error (MSE) losses (Hong & Ling, 2023; Dang et al., 2023) and an unexpected phenomenon (Minority Collapse) (Fang et al., 2021) occurs that the learned representations and the classifier weights of minority classes will converge to similar directions in ETF. As NC indicates the characteristics that a good classifier should have, it is a natural idea to "force" a model on long-tailed datasets to satisfy NC so that it may achieve better performance (Xie et al., 2023; Fang et al., 2021; Hong & Ling, 2023). For example, Yang et al. (2022) indicates that neural collapse optimally can be induced even in imbalanced learning as long as the learnable classifier is fixed as a random simplex ETF. Representation-Balanced Learning Framework (Peifeng et al., 2023) also maintains the geometric structure of ETF as the classifier but learns orthogonal matrices in ETF for feature learning instead of random directions. Different from them, Liu et al. (2023b) additionally proposes two explicit regularization terms to induce the NC phenomenon in imbalanced learning.

Leveraging the above observations of the NC phenomenon and ETF structure, we aim to induce the NC phenomenon in imbalanced learning from the perspective of distribution matching. In this work, we propose the **Dis**tribution **A**lignment Optimization (DisA) method for imbalanced classification based on Optimal Transport (OT) (Peyré & Cuturi, 2019; Cuturi, 2013). Specifically, we first consider the last-layer imbalanced representations as a discrete empirical distribution P. We then assume an ideal ETF structure, which is naturally balanced for each class, presented as another discrete empirical distribution Q. According to Yang et al. (2022) indicates that the features and classifier vectors are aligned with the same simplex ETF, we can enforce the distribution P over last-layer representations to be close to

[1]School of Artificial Intelligence, Jilin University [2]CSIRO's Data61 [3]The Chinese University of Hong Kong, Shenzhen. Correspondence to: Dandan Guo <guodandan@jlu.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

the ideal and balanced distribution Q of the ETF structure. Notably, different from existing NC-based long-tailed methods (Yang et al., 2022; Liu et al., 2023b; Xie et al., 2023; Peifeng et al., 2023), our method is able to not only fix the classifier with the above-mentioned ETF structure (only the encoder is learnable) but also optimize the classifier and encoder simultaneously. Due to its flexibility, we can easily combine DisA with commonly used long-tailed and NC-based long-tailed methods as a regularization term. The contributions of this paper can be summarized as follows:

- We propose a Distribution Alignment Optimization loss function as regularization for narrowing the gap between imbalanced learning representations and balanced ETF structure from the perspective of distribution for imbalanced classification.
- We point out that our method can be readily integrated into cross-entropy loss, imbalanced loss functions, and NC-based frameworks.
- Extensive experiments show our method can effectively improve the performance with various baselines on imbalanced classification against most existing methods.

2. Related Work

2.1. Imbalanced Classification

The common imbalanced classification strategies include re-weighting, re-sampling, two-stage algorithms, and others. The re-weighting approaches allocate different weights to each instance according to the sample numbers of different classes (Hong et al., 2021; Ren et al., 2020; Guo et al., 2022a; Shu et al., 2019). For example, Label-Distribution-Aware Margin (LDAM) loss (Cao et al., 2019) minimizes a margin-based generalization bound to encourage larger margins for minority classes. Furthermore, the re-sampling methods mainly consist of under-sampling and over-sampling. Under-sampling (Buda et al., 2018; Haixiang et al., 2017) discards a large portion of the data, normally causes under-fitting and deletes valuable data of the majority class by mistake. Over-sampling expands the sampling frequency and effectively improves the generalization of minority classes (Kim et al., 2020; Liu et al., 2019; Chou et al., 2020; Zhang et al., 2021; Gao et al., 2023). Moreover, two-stage algorithms are also in the spotlight (Shu et al., 2019; Li et al., 2021; Wang et al., 2021b; Zhong et al., 2021). Bilateral Branch Network (BBN) (Zhou et al., 2020) adjusts the whole model to learn from the conventional learning branch and dynamically move to the re-balancing branch. Besides, several methods also utilize contrastive learning, which implements unsupervised learning for imbalanced classification (Kang et al., 2021; Li et al., 2022; Wang et al., 2021a). The key idea is to learn a hidden space by closing the distance between similar samples and shrinking the distance between different samples. For example, Targeted Supervised Contrastive learning (TSC) (Li et al., 2022) makes the features of different classes converge to these distinct and uniformly distributed targets on the hypersphere. Our approach as regularization can be combined effectively with these various imbalanced loss functions.

2.2. Neural Collapse

Neural Collapse (NC) was observed that a linear classification model trained on a balanced dataset experienced a phenomenon (Papyana et al., 2020) where the last-layer features collapse into within-class means and the classifier weight vectors converge to the simplex Equiangular Tight Frame (ETF) geometric structure during the final stage of training. In Section 3.2, we give a detailed description of NC. Since then, researchers have been dedicated to theoretically digging deeper into this phenomenon (Liu et al., 2023a; Xu & Liu, 2023; Yang et al., 2023a). Recently, NC methods for imbalanced data have been proposed in class-incremental learning (Yang et al., 2023c;b), semantic segmentation (Zhong et al., 2023), large-scale vision-language (Zhu et al., 2023), and specifically long-tailed classification (Fang et al., 2021; Thrampoulidis et al., 2022; Hong & Ling, 2023; Dang et al., 2023).

In long-tailed classification, Yang et al. (2022) studies the potential of training a network with the last-layer linear classifier randomly initialized as a random simplex ETF and fixed during training. Considering the fixed directions of ETF can affect the generalization of the deep model, Representation-Balanced Learning Framework (RBL) (Peifeng et al., 2023) introduces orthogonal matrices to learn directions while maintaining the geometric structure of ETF. Notably, these two approaches initialize and fix the linear classifier in imbalanced learning with simplex ETF structure. Different from them, Xie et al. (2023) analyzes the reason for the performance drop under long-tailed distributions and proposes Attraction-Repulsion-Balanced Loss (ARB) to balance the gradients among the components from different classes. To induce Neural Collapse in deep long-tailed learning, Liu et al. (2023b) proposes two explicit feature regularization terms to learn compact within-class and maximally distinct between-class features for class-imbalanced data. Conversely, our approach matches imbalanced learning representation distribution and balanced ETF structure distribution from the perspective of distribution. We need no restriction on classifiers and can be trained on either the learnable classifier or using ETF as a classifier.

2.3. Optimal Transport

Optimal Transport (OT) was originally developed to solve the problem of how to select the transport matrix that would

incur the lowest cost when transporting multiple goods (Monge, 1781). Recently, OT has drawn widespread attention in different fields (Chen et al., 2023; Guo et al., 2022c; Maretic et al., 2022; Shi et al., 2023b), such as generative model (Genevay et al., 2018; Huynh et al., 2021; Arjun et al., 2019; Lénaïc & Francis, 2018) and domain adaptation (Rakotomamonjy et al., 2022; Turrisi et al., 2022; Asadulaev et al., 2023). Simultaneously, several approaches apply OT to solve long-tailed classification task from different views, such as the automatic re-weighting method (Guo et al., 2022a) and the data augmentation method (Gao et al., 2023). Different from them, we introduce a general regularization term for imbalanced classification by matching the last-layer representation distribution and the balanced ETF distribution through the NC theory.

3. Preliminaries

3.1. Imbalanced Classification

Let $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^{N}$ be the training set for a multi-class imbalanced classification problem with K classes, where $N=\sum_{k=1}^{K}n_k$ denotes total sample size, n_k is the size of class k and we assume $n_1\geq n_2\geq\cdots\geq n_K$. We decouple the deep learning model into the feature extractor f parameterized with $\boldsymbol{\theta}$ for the learned representation $\mathbf{H}:=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\ldots,\boldsymbol{h}_N\}\in\mathbb{R}^{d\times N}$ and the classifier $\boldsymbol{W}:=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_K\}\in\mathbb{R}^{d\times K}$, where $\boldsymbol{h}_i=f(x_i;\boldsymbol{\theta})\in\mathbb{R}^d$ is the last-layer representation. Training $\{\boldsymbol{\theta},\boldsymbol{W}\}$ on $\mathcal{D}$ with standard cross-entropy ignoring such class imbalance will perform poorly on the minority classes (Fang et al., 2021).

3.2. Neural Collapse

Papyana et al. (2020) reveals the Neural Collapse (NC) phenomenon during the terminal phase of training (TPT) on balanced datasets. It indicated that the last-layer representation will converge to their within-class means, and these within-class means together with the classifier vectors will collapse to the vertices of a simplex Equiangular Tight Frame (ETF). Before describing the characteristics of Neural Collapse, we first introduce the definition of simplex ETF and statistics in deep neural networks.

Definition 1 (Simplex Equiangular Tight Frame) A general Simplex Equiangular Tight Frame matrix (ETF) $\mathbf{M}\in\mathbb{R}^{d\times K}$ is a collection of vectors specified by the columns of:

$$\mathbf{M}=\sqrt{\frac{K}{K-1}}\mathbf{U}(\mathbf{I}-\frac{1}{K-1}\mathbf{1}_K\mathbf{1}_K^\top),\tag{1}$$

where $\mathbf{I}\in\mathbb{R}^{K\times K}$is the identity matrix and $\mathbf{1}_K\in\mathbb{R}^{K\times 1}$ is an all-ones vector, and $\mathbf{U}\in\mathbb{R}^{d\times K}$ $(d\geq K)$ is a rotation orthogonal matrix ($\mathbf{U}^\top\mathbf{U}=\mathbf{I}$). $\mathbf{M}:=\{\boldsymbol{m}_1,\boldsymbol{m}_2,\boldsymbol{m}_3,\ldots,\boldsymbol{m}_K\}\in\mathbb{R}^{d\times K}$ includes K classes with the $\boldsymbol{m}_k$ weight.

Statistics For a given classification task, we can formulate the class-mean features $\boldsymbol{\mu}_k=\frac{1}{n_k}\sum_{i=1}^{n_k}\boldsymbol{h}_{i,k}$ and the global-mean feature $\boldsymbol{\mu}_G=\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\mu}_k$ on the last-layer. Then the within-class covariance matrix can be computed as $\sum_{\boldsymbol{W}}=\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{1}{n_k}(\boldsymbol{h}_{i,k}-\boldsymbol{\mu}_k)(\boldsymbol{h}_{i,k}-\boldsymbol{\mu}_k)^\top$. The between-class covariance matrix is computed as $\sum_{\boldsymbol{B}}=\frac{1}{K}\sum_{k=1}^{K}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_G)(\boldsymbol{\mu}_k-\boldsymbol{\mu}_G)^\top$. The total covariance matrix is $\sum_{\boldsymbol{T}}=\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\frac{1}{n_k}(\boldsymbol{h}_{i,k}-\boldsymbol{\mu}_G)(\boldsymbol{h}_{i,k}-\boldsymbol{\mu}_G)^\top$. We use the l_2-norm of the mean features $\tilde{\boldsymbol{\mu}}_k=\frac{\boldsymbol{\mu}_k-\boldsymbol{\mu}_G}{\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_G\|_2}$. $\widetilde{\mathbf{M}}=[\boldsymbol{\mu}_k-\boldsymbol{\mu}_G]\in\mathbb{R}^{d\times K}$ is the matrix obtained by stacking the class-means into the columns of a matrix. During the terminal phase of training, there are four interrelated characteristics of the following conditions:

(NC1) Variability Collapse. As the training progresses, the within-class variation of the last-layer will collapse to class-means: $\sum_{\boldsymbol{W}}\to 0$.

(NC2) Convergence to Simplex ETF. The vectors of the class-means converge to a simplex ETF structure with the equal l_2 norm and the same pair-wise angle:

$$\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_G\|-\|\boldsymbol{\mu}_{k'}-\boldsymbol{\mu}_G\|\to 0\quad\forall k\neq k',\tag{2}$$

$$\langle\tilde{\boldsymbol{\mu}}_k,\tilde{\boldsymbol{\mu}}_{k'}\rangle=-\frac{1}{K-1}\quad\forall k\neq k'.\tag{3}$$

(NC3) Convergence to Self-duality. The Frobenius norm of the classifier with K weights aligns with the corresponding class-means:

$$\left\|\frac{\boldsymbol{W}^\top}{\|\boldsymbol{W}\|_F}-\frac{\widetilde{\mathbf{M}}}{\|\widetilde{\mathbf{M}}\|_F}\right\|_F\to 0,\tag{4}$$

(NC4) Simplification to Nearest Class-Center (NCC). Given a feature, the network classifier converges to the nearest class-mean (NCC):

$$\arg\max_k\langle\boldsymbol{w}_k,\boldsymbol{h}(x)\rangle\to\arg\min_k\|\boldsymbol{h}(x)-\boldsymbol{\mu}_k\|_2.\tag{5}$$

3.3. Optimal Transport

Optimal Transport (OT) problem has recently been used as a powerful geometric tool to measure the minimum cost of the transport matrix between distribution, with rich applications in machine learning and related areas (Ge et al., 2021; Zhao et al., 2021; Nguyen et al., 2021; Wang et al., 2022; Guo et al., 2022b; Bui et al., 2022; Vuong et al., 2023; Zhao et al., 2023; Ye et al., 2024; Vo et al., 2024). We give a brief overview of OT and more details can be found in Peyré & Cuturi (2019). Considering two discrete probability distributions $p=\sum_{i=1}^{n}a_i\delta_{x_i}$ and $q=\sum_{j=1}^{m}b_j\delta_{y_j}$, where x_i and y_j are in the same space arbitrarily and δ is a Dirac function. Then, the optimal transport distance can be defined as: $\mathbf{OT}(p,q)=\min_{\mathbf{T}\in\Pi(p,q)}\langle\mathbf{T},\mathbf{C}\rangle$, where the cost matrix $\mathbf{C}\in\mathbb{R}_+^{n\times m}$ is constructed by $C_{ij}=\mathbf{C}(x_i,y_j)$

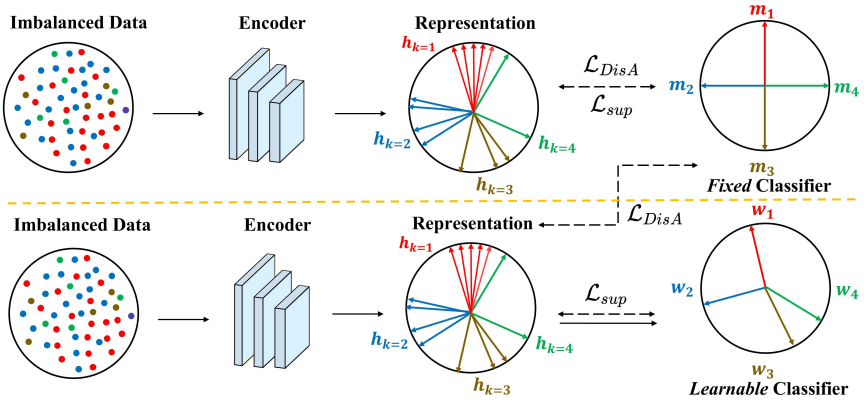


Figure 1. The training processes of our method under two scenarios. The upper part trains only with the ETF as a fixed classifier, while the bottom part trains with the learnable classifier and the ETF. (The cardinalities of the majority classes (the red and blue classes) are much larger than minority classes (the green and brown classes)).

which reflects the cost between x_i and y_j. The transport probability matrix $\mathbf{T} \in \mathbb{R}_+^{n\times m}$ satisfies $\Pi(p,q) \coloneqq \left\{\mathbf{T} \mid \sum_{i=1}^{n} T_{ij} = b_j, \sum_{j=1}^{m} T_{ij} = a_i\right\}$. The Sinkhorn algorithm (Cuturi, 2013) is usually applied by the entropy regularization constraints $H(\mathbf{T}) = -\sum_{ij} T_{ij} \log T_{ij}$ with a hyper-parameter $\epsilon > 0$ for fast optimization above problem.

4. Our Proposed Method

This work proposes a **Dis**tribution **A**lignment Optimization (DisA) method based on OT through Neural Collapse for imbalanced classification. The motivation for our work derives from NC phenomenon that the last-layer features collapse into an ETF structure in balanced learning. Nevertheless, the training model on the imbalanced dataset is difficult to activate the condition, where the learned representations of some classes are usually entangled with each other (Fang et al., 2021), resulting in poor performance for the minority classes. In the following, we introduce the details of our proposal that learn better representations with the help of the ETF structure by a distribution matching approach.

4.1. Distribution Alignment Optimization

To illustrate the above problem as a matching one between two distributions, we view the last-layer representations of N data samples within the training set as discrete N-dimensional distribution P and represent the balanced ETF structure $\mathbf{M}$ in 3.2 of K classes as another K-dimensional distribution Q. The distribution P and distribution Q can be formulated as:

$$P = \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{h}_i}, \tag{6}$$

$$Q = \frac{1}{K}\sum_{k=1}^{K} \delta_{\boldsymbol{m}_k}, \tag{7}$$

where $\boldsymbol{h}_i = f(x_i;\boldsymbol{\theta}) \in \mathbb{R}^d$ is the last-layer feature of the input x_i, and $\boldsymbol{m}_k$ is the k-th weight vector of the balanced ETF structure $\mathbf{M}$ in (1).

On the one hand, P is a discrete uniform distribution over the latent representations by construction, but the data samples in the training set are long-tailed, making P "imbalanced" in terms of the classes. On the other hand, a vector of the simplex ETF matrix $\mathbf{M}$ can be viewed as a perfect prototype of its corresponding class, and the prototypes are balanced and well separated according to the properties of NC. Therefore, Q is a balanced distribution. To learn high-quality representation by the feature extractor parameterized by $\boldsymbol{\theta}$ in an imbalanced classification task, we aim to enforce the to-be-learned distribution P over the last-layer representations to stay close to the balanced distribution Q over the ETF structure. Here, we explore the distribution alignment optimization method by minimizing the OT distance between P and Q:

$$\min_{\boldsymbol{\theta}} \mathbf{OT}(P,Q) \overset{\text{def.}}{=} \min_{\boldsymbol{\theta}} \min_{\mathbf{T}\in\Pi(P,Q)} \langle \mathbf{T}, \mathbf{C}\rangle, \tag{8}$$

where the element C_{ik} in the cost matrix $\mathbf{C} \in \mathbb{R}_+^{N\times K}$ indicates the distance between feature $\boldsymbol{h}_i$ and weight $\boldsymbol{m}_k$, which can be flexibly computed with commonly used distance measures. We define $C_{ik} = 1 - \cos(\boldsymbol{h}_i, \boldsymbol{m}_k)$ with cosine distance although other choices are possible. The transport plan matrix $\mathbf{T}$ satisfies $\Pi(P,Q) := \left\{\mathbf{T} \in \mathbb{R}_+^{N\times K} \mid \sum_{k=1}^{K} T_{ik} = \frac{1}{N}, \sum_{i=1}^{N} T_{ik} = \frac{1}{K}\right\}$.

Intuitively, minimizing this expected moving cost encourages the sample features and ETF weights to be aligned given their cost function. Since directly optimizing the transport plan in (8) requires a significant time overhead, we can adopt the entropy regularized OT loss (Cuturi, 2013) to solve the problem, where (8) can be re-written as the following optimization problem:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{DisA} = \langle \mathbf{T}^*, \mathbf{C} \rangle, \\ \text{subject to } \mathbf{T}^* = \underset{\mathbf{T} \in \Pi(P,Q)}{\arg\min} \langle \mathbf{T}, \mathbf{C} \rangle - \epsilon H(\mathbf{T}), \tag{9}$$

where $\epsilon > 0$ is a hyper-parameter of controlling the weight of the entropic regularization. We can first optimize the entropy regularized OT loss to learn the transport plan $\mathbf{T}^*$, which further allows one to learn the parameter $\boldsymbol{\theta}$ of feature extractor by minimizing the DisA loss.

Notably, minimizing the $\mathcal{L}_{DisA}$ loss defined by the representation distribution P and ETF distribution Q provides a principled and unsupervised way to encourage the feature extractor to learn more balanced representations. Therefore, our proposed the plug-and-play $\mathcal{L}_{DisA}$ loss can be easily combined with the supervised losses with a linear classifier, where $\mathcal{L}_{DisA}$ plays the role of regularizing the penultimate layer embedding. Now, the total loss for the imbalanced classification loss can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{DisA}, \tag{10}$$

where L_{sup} indicates a supervised loss function, which will be described below, and λ indicates the hyper-parameter for balancing the supervised loss and regularization loss.

4.2. Combination with Existing Methods

Thanks to its flexibility, our proposed loss can be naturally combined with existing methods. We investigate the following two ways to solve the imbalanced problem according to the concerned classifier $\boldsymbol{W}$.

Fixing Classifier $\boldsymbol{W}$ with ETF. In Yang et al. (2022) and Peifeng et al. (2023), the linear classifier is initialized as simplex ETF and fixed during training, where the former proposes a Dot-Regression (DR) loss to learn the feature extractor in imbalanced problem. When our method is incorporated with them, we define the classifier $\boldsymbol{W} \in \mathbb{R}^{d \times K}$ with an ETF structure $\mathbf{M}$ in (1), i.e., $\boldsymbol{W} = \mathbf{M}$. Fig.1 (upper) shows this setting with a fixed classifier from ETF structure, where we use the learnable feature extractor for extracting the representations and assume an ideal ETF structure as a fixed classifier. During the training process, we only need to optimize the feature extractor f parameterized by $\boldsymbol{\theta}$ without classifier $\boldsymbol{W}$, which can be expressed as follows:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{total} = \min_{\boldsymbol{\theta}} \mathcal{L}_{sup} + \lambda \cdot \min_{\boldsymbol{\theta}} \mathcal{L}_{DisA}, \tag{11}$$

where the supervised loss $\mathcal{L}_{sup}$ can be the standard cross-entropy loss or the DR loss specially designed for the fixed classifier with ETF structure by Yang et al. (2022).

Learning Classifier $\boldsymbol{W}$. We can also consider the case where the concerned classifier $\boldsymbol{W}$ is learnable, which is a more common setting for solving the long-tailed problem. Now, ours can be combined with most of the existing long-tailed methods as discussed in 2.1. Interestingly, for the NC-based long-tailed methods which learn the classifier $\boldsymbol{W}$, such as Inducing Neural Collapse (INC) (Liu et al., 2023b), ours is still applicable as an additional regularization term. Although INC introduces regularization terms to compact within-class representations and distinct between-class representations based on NC properties, our method is derived from the perspective of distribution matching optimization, which is complementary to INC. As shown in Fig.1 (bottom), the samples are fed into the trainable feature extractor followed by a learnable classifier $\boldsymbol{W}$ and an ETF structure (also termed as fixed classifier in Fig.1 (upper)), where the ETF structure is used for distribution alignment optimization loss $\mathcal{L}_{DisA}$ and the classifier $\boldsymbol{W}$ for the supervised loss $\mathcal{L}_{sup}$. Now, the model optimizes both $\boldsymbol{\theta}$ and classifier $\boldsymbol{W}$ with the following loss:

$$\min_{\boldsymbol{\theta},\boldsymbol{W}} \mathcal{L}_{total} = \min_{\boldsymbol{\theta},\boldsymbol{W}} \mathcal{L}_{sup} + \lambda \cdot \min_{\boldsymbol{\theta}} \mathcal{L}_{DisA} \tag{12}$$

Detailed formulations of L_{total} can be found in Appendix A. In summary, the main difference between equations 11 and 12 lies in the to-be-learned parameters, where the former only learns the encoder and the latter learns both the encoder and the classifier. Since the number of samples can be large in real-world datasets, it is less practical to have P over all samples in (6). We implement the proposed DisA in mini-batches by uniformly subsampling from all samples, which empirically works well in our tasks. Theoretically analyzed in Fatras et al. (2020), the mini-batch OT enjoys appealing properties of unbiased estimators, gradients, and a concentration bound around the expectation. We summarize our proposed method in Algorithm 2 of Appendix D.

4.3. Why does DisA Work for Imbalanced Datasets?

Here we intuitively analyze why DisA based on OT works for imbalanced datasets. As discussed before, P is a discrete uniform distribution over the training samples but is quite "imbalanced" in terms of the labels of the samples in a long-tailed dataset. Recalling Q is a balanced distribution over ETF vectors, DisA learns the transport plan $\mathbf{T}$ between P and Q. Specifically, T_{ik} indicates the weight between sample i and ETF vector k. With $\mathbf{L} \in \mathbb{R}^{N \times K}$ as the one-hot encoding of the ground-truth classes of N samples, we can have: $\mathbf{A} \in \mathbb{R}^{K \times K} = \mathbf{L}^\top \mathbf{T}$, where $A_{k,k'}$ indicates the weights between class k and ETF vector k' summed over all the samples belonging to class k. We show $\mathbf{A}$ computed from DisA with a fixed classifier as ETF structure on CIFAR-LT-10 in Fig.2. Note that there are much more samples from the majority classes than those from minority classes in P. If DisA assigned similar T_{ik} for every sample ignoring the imbalance, we would see $A_{k_1,k'} > A_{k_2,k'}$ where k_1 indicates one majority class and k_2 indicates a minority one.

However, Fig.2 shows that the weights of $\mathbf{A}$ focus on the diagonal (meaning that the latent representations of samples are well aligned to their corresponding ETF vectors) and the diagonal values are quite similar (i.e., $A_{k_1,k'} \approx A_{k_2,k'}$ for all k_1 and k_2). This is suggesting that DisA can balance between majority and minority classes in an adaptive way by assigning (relatively) larger weights to minority samples than majority ones.

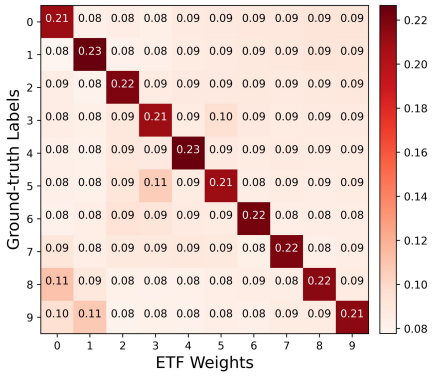


Figure 2. Our learned align matrix $\mathbf{A}$ on CIFAR-LT-10 combined with the ETF classifier of the majority classes (0-4) and minority classes (5-9).

5. Experiments

5.1. Experimental Setting

Datasets. To evaluate the effectiveness of our method, we conduct experiments on benchmark datasets for long-tailed classification, including CIFAR-LT-10 (Cui et al., 2019), CIFAR-LT-100 (Cui et al., 2019), and ImageNet-LT (Deng et al., 2009). The imbalance factor is defined as the ratio of data points between the most and least frequent classes, denoted as $\rho = \frac{n_{max}}{n_{min}}$. Specifically, CIFAR-LT-10 (CIFAR-LT-100) are derived from CIFAR-10 (CIFAR-100) (Krizhevsky et al., 2009) with imbalance factors $\rho = \{200, 100, 50, 10\}$, respectively. ImageNet-LT has 1,000 classes with the imbalance factor of $\rho = 256$, which is constructed from ImageNet (Deng et al., 2009).

Implementation Details. For all experiments, our method is implemented in PyTorch and using an SGD optimizer with a momentum of 0.9. In CIFAR-LT-10 and CIFAR-LT-100, we use ResNet-32 (He et al., 2016) as the backbone and use 200 epochs on a single Tesla A10 GPU and set the initial learning rate as 0.1, which is divided by 10 at 160th and 180th epochs. We utilize mixup (Zhang et al., 2018) as data augmentation following Yang et al. (2022) during the training with these baseline methods except Peifeng et al. (2023), which performs AutoAugment (Cubuk et al., 2019). The hyper-parameter α for sampling combination ratio in the beta distribution used for mixup is set as 1.0.

For ImageNet-LT, we conduct experiments with ResNet-50 following Yang et al. (2022); Xie et al. (2023) and ResNeXt-50 following Liu et al. (2023b); Peifeng et al. (2023). We train 200 epochs with the batchsize of 128 and weight decay of 5e-4 on four Tesla A10 GPUs. The learning rate is initialized as 0.1 and decays to zeros by cosine annealing schedule during training. According to Peifeng et al. (2023), we also report the test accuracy on three subsets: Many-shot classes with more than 100 training samples, Medium-shot classes with 20 to 100 samples, and Few-shot classes with less than 20 samples. For different baselines, we follow the same data augmentation as the combined methods.

Table 1. Top-1 test accuracy (%) of ResNet-32 on CIFAR-LT-10. * denotes our reproduced baselines with mixup augmentation. † and ‡ are reported from Yang et al. (2022) and Gao et al. (2023), respectively. The remaining methods are from the original paper.

Method	**200**	**100**	**50**	**10**
KCL	/	77.6	81.7	88.0
TSC	/	79.7	82.9	88.7
BBN	/	79.9	82.2	88.4
HCL	/	81.4	85.4	**91.1**
RIDE (3 experts)‡	/	81.6	84.0	86.3
MiSLAS	/	82.1	**85.7**	90.0
ARB	/	**83.3**	**85.7**	90.2
CE†	67.3	72.8	78.6	87.7
CE+**DisA**	69.2	74.7	79.6	88.3
CE-DRW*	75.1	80.9	81.8	88.8
CE-DRW+**DisA**	77.7	82.1	84.1	89.8
LDAM-DRW*	77.1	78.8	82.7	88.2
LDAM-DRW+**DisA**	78.0	80.4	84.8	88.3
INC-DRW	75.8*	81.9	82.7*	89.8
INC-DRW+**DisA**	78.7	82.3	84.5	90.2
INC-DRW-cRT	76.8*	82.6	83.3*	90.2
INC-DRW-cRT+**DisA**	**79.1**	**82.8**	**84.9**	**90.5**
ETF-CE	60.6	67.0	77.2	87.0
ETF-CE+**DisA**	62.4	68.4	78.9	87.8
ETF-DR	71.9	76.5	81.0	87.7
ETF-DR+**DisA**	73.7	78.5	81.4	87.8
RBL*	80.2	83.6	87.0	92.0
RBL	81.2	84.7	87.6	/
RBL+**DisA**	**82.0**	**85.1**	**87.9**	**92.6**

We report the average results of three repeated experiments with different random seeds. Besides, we set λ for regularization weight in (10) as 0.1 and ϵ for entropic regularization in (9) as 1. To straightforwardly express the setting of the implementation of baseline methods with NC phenomenon, we report the implementation details of previous methods and our method of neural collapse in Tables 8 and 9 of the Appendix E.

Baselines. We consider following baselines: (1) Cross-entropy loss (CE) and imbalanced re-weighting learning methods: DRW, LDAM-DRW (Cao et al., 2019); (2) Two-stage methods: BBN (Zhou et al., 2020), RIDE (Wang et al., 2021b), cRT (Cao et al., 2019), MiSLAS (Zhong et al., 2021); (3) Contrastive learning methods: KCL (Kang et al., 2021), TSC (Li et al., 2022), HCL (Wang et al.,

2021a); (4) Imbalanced NC-based methods: fixed classifier with ETF (ETF-CE, ETF-DR loss (Yang et al., 2022), RBL (Peifeng et al., 2023)); and learnable classifier (INC (Liu et al., 2023b), ARB (Xie et al., 2023)).

5.2. Comparison on Long-Tailed Classification

Comparison on CIFAR-LT. We first compare the performance of the proposed method with existing state-of-the-art long-tailed classification methods on the CIFAR-LT datasets. For a fair comparison, we implement CE, CE-DRW, and LDAM-DRW with mixup augmentation. We report the performance of different methods with various imbalance factors on CIFAR-LT-10 and CIFAR-LT-100.

Table 1 presents that our method shows the best overall performance, outperforming previous state-of-the-art methods significantly on CIFAR-LT-10. It confirms the validity of DisA as regularization by distribution matching optimization. Besides, combining DisA with different losses and NC-based methods can yield superior performance compared with the original baselines, no matter whether the classifier is learnable or fixed. For example, replacing the classifier with a fixed ETF classifier, ETF-DR+DisA, performs better than ETF-DR. The performance improvement is more significant when the imbalance factor is higher (the dataset is more imbalanced), where ETF-DR+DisA achieves 1.8% and 2.0% improvement with $\rho = 200, 100$, respectively. The reason behind this might be the last-layer representation has approximately converged to a balanced ETF structure in relative balance setting and vice versa in extreme imbalance, which is consistent with the phenomenon reported by Yang et al. (2022). It indicates the benefit of aligning the distribution of last-layer representation with the distribution of the ETF structure in imbalanced learning.

Moreover, DisA also outperforms previous competing methods under different imbalance factors except RBL, which demonstrates the effectiveness and flexibility of DisA in Table 2. Compared with INC-DRW, we achieve 1.1% and 2.6% improvement with $\rho = 100, 10$, respectively. CE+DisA has the 1.5% improvement over CE in both $\rho = 100, 50$. These results illustrate the proposed method can be effectively combined with most long-tailed methods. Additionally, since the discrepancy between our reproduced RBL* and the results reported by RBL in the original paper is too large, RBL+DisA is unsatisfactory. However, RBL+DisA performs significantly better than RBL* (on which RBL+DisA is implemented). We also combine ours with other classical long-tailed loss functions, such as Focal loss (Lin et al., 2017) and CB Softmax loss (Cui et al., 2019), and report the detailed results in Table 4 of Appendix B.

Comparison on ImageNet-LT. We further conduct the comparison experiments with several long-tailed classification methods on the ImageNet-LT dataset in Table 3. For a fair comparison, we use the ResNet-50 following Xie et al. (2023); Yang et al. (2022) and ResNeXt-50 following Liu et al. (2023b); Peifeng et al. (2023). We can find that our DisA surpasses related NC-based methods and various loss functions and enhances the imbalanced classification. DisA achieves higher accuracy in all classes with ResNeXt-50, which demonstrates the generalization ability to align imbalanced learning representations and balanced ETF structure. Besides, we report the performance on the balanced dataset in Table 6 and computational cost in Table 10 in Appendix.

Table 2. Top-1 test accuracy (%) of ResNet-32 on CIFAR-LT-100. * and † denote our reproduced baselines and the results from Yang et al. (2022) with mixup, respectively. The results of other methods are from their original papers.

Method	200	100	50	10
BBN	/	42.6	47.1	59.2
KCL	/	42.8	46.3	57.6
TSC	/	43.8	47.4	59.0
HCL	/	46.7	51.9	63.1
ARB	/	47.2	**52.6**	62.1
MiSLAS	/	47.0	52.3	**63.2**
RIDE (3 experts)	/	**48.6**	51.4	59.8
CE	38.7†	43.0†	48.1†	58.5*
CE+**DisA**	39.8	44.5	49.6	63.2
LDAM-DRW*	41.1	45.0	48.8	58.3
LDAM-DRW+**DisA**	43.0	47.0	52.5	64.1
CE-DRW*	41.5	45.4	51.1	61.5
CE-DRW+**DisA**	44.4	49.2	54.0	65.7
INC-DRW	42.5*	48.6	51.7*	63.1
INC-DRW+**DisA**	44.8	49.7	54.2	65.7
INC-DRW-cRT	42.8*	48.7	51.8*	63.6
INC-DRW-cRT+**DisA**	**45.2**	**49.8**	**54.4**	**65.9**
ETF-CE*	35.4	40.0	44.1	60.7
ETF-CE+**DisA**	36.5	40.7	45.2	61.2
ETF-DR	40.9	45.3	50.4	/
ETF-DR+**DisA**	41.5	45.9	51.0	63.4
RBL*	47.3	52.0	56.1	66.0
RBL	**48.9**	53.1	57.2	/
RBL+**DisA**	**48.9**	**53.2**	**57.4**	**66.6**

5.3. Analytical Experiments

To explore how our proposed DisA narrows the gap between the representation distribution and fixed ETF structure distribution, we display the visualization results and analysis.

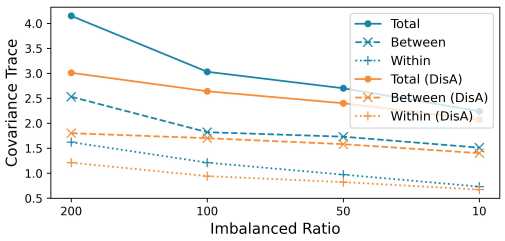


Figure 3. The covariance traces of the INC-DRW and INC-DRW+DisA with different imbalanced ratios on CIFAR-LT-10.

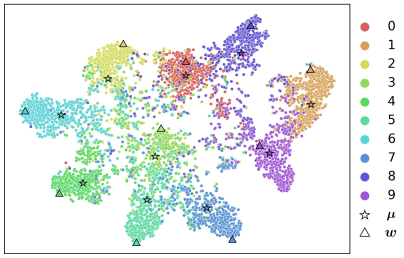
Covariance Trace. Considering the characteristics of NC, we investigate the suppression of three covariance traces of the INC-DRW and INC-DRW+DisA with different imbalanced ratios on CIFAR-LT-10 in Fig.3. We consider the normalized within-class covariance $\mathrm{Tr}(\Sigma_{\boldsymbol{W}})$ (dotted), between-class covariance $\mathrm{Tr}(\Sigma_{\boldsymbol{B}})$ (solid), and total covariance $\mathrm{Tr}(\Sigma_{\boldsymbol{T}})$ (dashed) in 3.2.

Table 3. Top-1 test accuracy (%) of ResNet-50 and ResNeXt-50 on ImageNet-LT. * denotes our reproduced baselines with mixup augmentation. The results of other methods are from their original papers.

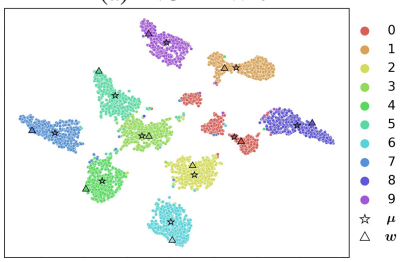
	Method	Many	Med	Few	All
ResNet-50	FCL	61.4	47.0	28.2	49.8
	KCL	61.8	49.4	30.9	51.5
	TSC	**63.5**	49.7	30.4	52.4
	MiSLAS	61.7	**51.3**	**35.8**	**52.7**
	CE*	**68.4**	37.8	4.8	44.3
	CE+**DisA**	67.7	38.6	7.3	44.8
	LDAM-DRW	57.4	47.2	23.8	47.7
	LDAM-DRW+**DisA**	57.5	48.5	25.5	48.6
	CE-DRW*	53.8	47.4	**28.8**	47.1
	CE-DRW+**DisA**	64.4	**51.4**	27.7	**53.2**
	ETF-DR*	**66.8**	37.5	10.6	44.5
	ETF-DR	/	/	/	44.7
	ETF-DR+**DisA**	65.2	**39.9**	**12.8**	**45.3**
ResNeXt-50	CE-DRW	52.6	45.7	31.5	46.4
	CE-LWS	57.1	45.2	29.3	47.7
	LADE	**65.1**	**48.9**	**33.4**	**53.0**
	RBL	**64.8**	49.6	34.2	53.3
	RBL+**DisA**	**64.8**	**49.8**	**34.7**	**53.5**
	INC-DRW*	66.4	48.5	28.7	52.7
	INC-DRW	67.1	49.7	29.0	53.6
	INC-DRW+**DisA**	**67.6**	49.2	31.6	53.9
	INC-DRW-cRT	65.6	51.2	**35.4**	54.2
	INC-DRW-cRT+**DisA**	65.0	**52.1**	33.0	**54.5**

When the model tends to achieve the NC phenomenon, the traces of within-class covariance, between-class covariance, and total covariance become lower, which means the last-layer representation will converge to their within-class means (compact within-class representation), and these within-class means will collapse to the vertices of a simplex ETF. From the normalized plots, we can observe that INC-DRW+DisA outperforms INC-DRW on all covariance traces. As the degree of imbalance ratios increases, the gaps of covariance traces between ours and baselines are larger, indicating the effectiveness of ours especially in the extremely imbalanced problems.

t-SNE Visualization. We show the t-SNE visualizations of the representations of the feature extractor learned from INC-DRW-cRT and INC-DRW-cRT+DisA on CIFAR-LT-10 in Fig.4. We can find that introducing our DisA loss can extract more compact within-class and distant between-class representations, which explains the improvement of classification performance with our method. The samples in ours are more tightly clustered around their corresponding class-mean feature (☆) and classifier weight (△). Besides, the distance between the class-mean feature and the classifier weight for each class is smaller in ours than in INC-DRW-cRT. These results illustrate that enforcing the representation distribution to be aligned with the balanced distribution from the ETF structure based on the NC theory benefits the higher-quality representation learning in imbalanced problem.



(a) INC-DRW-cRT



(b) INC-DRW-cRT+DisA

Figure 4. t-SNE visualizations of INC-DRW-cRT and INC-DRW-cRT+DisA on CIFAR-LT-10, which contain the features (dots ○), class-means (☆), and the classifier weights (△) for each class.

Transport Plan. We show the learned transport plan with the fixed ETF classifier on CIFAR-LT-10, where the rows denote the labels of 55 images from the current mini-batch and columns mark ETF weights in Fig.5. We observe that the representations can be aligned to their corresponding ETF weight vectors in the mini-batch level. DisA usually assigns larger transport plan for the minority samples than majority samples, which explains why our DisA can balance between majority and minority classes in an adaptive way. Therefore, the diagonal values of $\mathbf{A}$ in Fig. 2 are quite similar even the training set is imbalanced. To further verify whether our method ameliorates the performance of minority classes, we plot the confusion matrices and class-means angle matrices on CIFAR-LT-10 in Appendix C.

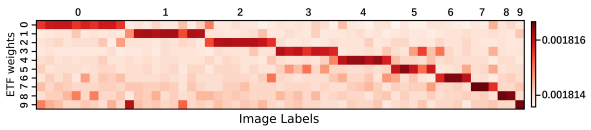


Figure 5. The learned transport plan on CIFAR-LT-10, where the rows denote the labels of 55 images from the current mini-batch and columns mark ETF weights.

6. Conclusion

In this paper, we aim to enforce the NC phenomenon for imbalanced classification and propose distribution alignment optimization based on optimal transport that encourages the alignment between the learned representations and fixed-balanced ETF structures distributionally. Our method can be used as a regularization term that is integrated into most loss functions for optimizing the classifier and encoder simultaneously. Moreover, it can be used with fixed classifier with the ideally balanced ETF structure for learning high-quality representations. Extensive experiments demonstrate that our method can enhance the generalization power of existing methods in imbalanced datasets. However, due to the intrinsic constraint of NC, when exists on high-dimension models or large-classes datasets, our method hardly satisfies the dimension condition of ETF. We believe that it is an open problem of NC, which we leave as a future work.

Acknowledgements

This work was supported by NSFC (62306125).

Impact Statement

Our research tackles a fundamental challenge in machine learning: enhancing the accuracy and reliability of image classification systems when dealing with imbalanced data. Imbalanced datasets, where some classes are underrepresented compared to others, are prevalent in real-world scenarios. This imbalance often leads to models that are biased toward the majority classes, resulting in poor performance on minority classes. Such shortcomings can undermine the effectiveness of AI systems in critical applications. Our work introduces a novel approach to imbalanced image classification by leveraging Neural Collapse. This method ensures a balanced focus on both minority and majority classes, thereby enhancing the overall classification performance and improving the representation of minority classes. For future work, we plan to extend our research to address other forms of data imbalance and explore real-time applications.

References

Allingham, J. U., Ren, J., Dusenberry, M. W., Gu, X., Cui, Y., Tran, D., Liu, J. Z., and Lakshminarayanan, B. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. In *International Conference on Machine Learning*. PMLR, 2023.

Arjun, Nitin, B., Daniel, C., and Prateek, M. Lower bounds on adversarial robustness from optimal transport. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Asadulaev, A., Shutov, V., Korotin, A., Panfilov, A., Kontsevaya, V., and Filchenkov, A. A minimalist approach for domain adaptation with optimal transport. In *Conference on Lifelong Learning Agents*, pp. 1009–1024. PMLR, 2023.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, pp. 249–259, 2018.

Bui, A. T., Le, T., Tran, Q. H., Zhao, H., and Phung, D. A unified Wasserstein distributional robustness framework for adversarial training. In *International Conference on Learning Representations*, 2022.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Chen, G., Yao, W., Song, X., Li, X., Rao, Y., and Zhang, K. Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Chou, H.-P., Chang, S.-C., Pan, J.-Y., Wei, W., and Juan, D.-C. Remix: rebalanced mixup. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 95–110, 2020.

Cubuk, E. D., Zoph, B., Mané, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Cui, Y., Jia, M., Lin, T., Song, Y., and Belongie, S. J. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Dang, H., Nguyen, T., Tran, T., Tran, H., and Ho, N. Neural collapse in deep linear network: From balanced to imbalanced data. In *International Conference on Machine Learning (ICML)*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 2009.

Fang, C., He, H., Long, Q., and Su, W. J. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118:e2103091118, 2021.

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. Learning with minibatch wasserstein: asymptotic and gradient properties. In *International Conference on Artificial Intelligence and Statistics*, pp. 2131–2141. PMLR, 2020.

Gao, J., Zhao, H., Li, Z., and Guo, D. Enhancing minority classes by mixing: An adaptative optimal transport approach for long-tailed classification. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ge, Z., Liu, S., Li, Z., Yoshie, O., and Sun, J. OTA: Optimal transport assignment for object detection. In *CVPR*, pp. 303–312, 2021.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Guo, D., Li, Z., Zheng, M., Zhao, H., Zhou, M., and Zha, H. Learning to re-weight examples with optimal transport for imbalanced classification. In *Advances in Neural Information Processing Systems*, pp. 25517–25530, 2022a.

Guo, D., Tian, L., Zhang, M., Zhou, M., and Zha, H. Learning prototype-oriented set representations for meta-learning. In *International Conference on Learning Representations*, 2022b.

Guo, D., Tian, L., Zhao, H., Zhou, M., and Zha, H. Adaptive distribution calibration for few-shot learning with hierarchical optimal transport. In *Advances in Neural Information Processing Systems*, pp. 6996–7010, 2022c.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, pp. 220–239, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

Hong, W. and Ling, S. Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data. *arXiv preprint arXiv:2309.09725*, 2023.

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

Hu, J., Guo, D., Liu, Y., Li, Z., Chen, Z., Wan, X., and Chang, T.-H. A simple yet effective subsequence-enhanced approach for cross-domain ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12890–12898, 2023.

Huynh, V., Phung, D., and Zhao, H. Optimal transport for deep generative models: state of the art and research challenges. In *International Joint Conference on Artificial Intelligence 2021*, pp. 4450–4457, 2021.

Kang, B., Li, Y., Xie, S., Yuan, Z., and Feng, J. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Kim, J., Jeong, J., and Shin, J. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13896–13905, 2020.

Kothapalli, V., Tirer, T., and Bruna, J. A neural collapse perspective on feature evolution in graph neural networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009.

Li, S., Gong, K., Liu, C. H., Wang, Y., Qiao, F., and Cheng, X. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021.

Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R., Indyk, P., and Katabi, D. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022.

Li, Z., Zhao, H., Li, Z., Liu, T., Guo, D., and Wan, X. Extracting clean and balanced subset for noisy long-tailed classification. *arXiv preprint arXiv:2404.06795*, 2024.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017.

Liu, W., Yu, L., Weller, A., and Schölkopf, B. Generalizing and decoupling neural collapse via hyperspherical uniformity gap. In *International Conference on Learning Representations (ICLR)*, 2023a.

Liu, X., Zhang, J., Hu, T., Cao, H., Yao, Y., and Pan, L. Inducing neural collapse in deep long-tailed learning. In

International Conference on Artificial Intelligence and Statistics (AISTATS), 2023b.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

Lénaïc, C. and Francis, B. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Maretic, H. P., El Gheche, M., Chierchia, G., and Frossard, P. Fgot: Graph distances based on filters and optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7710–7718, 2022.

Monge, G. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704, 1781.

Nguyen, T., Le, T., Zhao, H., Tran, Q. H., Nguyen, T., and Phung, D. Most: Multi-source domain adaptation via optimal transport for student-teacher learning. In *UAI*, pp. 225–235, 2021.

Papyana, V., Hanb, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117:24652–24663, 2020.

Peifeng, G., Xu, Q., Wen, P., Yang, Z., Shao, H., and Huang, Q. Feature directions matters: Long-tailed learning via rotated balanced representation. In *International Conference on Machine Learning (ICML)*, 2023.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning*, pp. 355–607, 2019.

Rakotomamonjy, A., Flamary, R., Gasso, G., Alaya, M. E., Berar, M., and Courty, N. Optimal transport for conditional domain matching and label shift. *Machine Learning*, pp. 1–20, 2022.

Rangamani, A., Lindegaard, M., Galanti, T., and Poggio, T. A. Feature learning in deep classifiers through intermediate neural collapse. In *International Conference on Machine Learning (ICML)*, 2023.

Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al. Balanced meta-softmax for long-tailed visual recognition. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Shi, J., Wei, T., Xiang, Y., and Li, Y. How re-sampling helps for long-tail learning? In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Shi, L., Zhen, H., Zhang, G., and Yan, J. Relative entropic optimal transport: a (prior-aware) matching perspective to (unbalanced) classification. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

Thrampoulidis, C., Kini, G. R., Vakilian, V., and Behnia, T. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35: 27225–27238, 2022.

Turrisi, R., Flamary, R., Rakotomamonjy, A., and Pontil, M. Multi-source domain adaptation via weighted joint distributions optimal transport. In *Uncertainty in Artificial Intelligence*, pp. 1970–1980. PMLR, 2022.

Vo, V., Zhao, H., Le, T., Bonilla, E. V., and Phung, D. Optimal transport for structure learning under missing data. In *International Conference on Machine Learning*, 2024.

Vuong, L. T., Le, T., Zhao, H., Zheng, C., Harandi, M., Cai, J., and Phung, D. Vector quantized wasserstein auto-encoder. In *International Conference on Machine Learning*, pp. 35223–35242. PMLR, 2023.

Wang, D., Guo, D., Zhao, H., Zheng, H., Tanwisuth, K., Chen, B., Zhou, M., et al. Representing mixtures of word embeddings with mixtures of topic embeddings. In *International Conference on Learning Representations*, 2022.

Wang, P., Han, K., Wei, X., Zhang, L., and 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 (CVPR)*, 2021a.

Wang, X., Lian, L., Miao, Z., Liu, Z., and Yu, S. X. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations (ICLR)*, 2021b.

Wei, T., Shi, J., Tu, W., and Li, Y. Robust long-tailed learning under label noise. *CoRR*, arXiv preprint arXiv:2108.11569, 2021.

Xie, L., Yang, Y., Cai, D., and He, X. Neural collapse inspired attraction-repulsion-balanced loss for imbalanced learning. *Neurocomputing*, 527:60–70, 2023.

Xu, J. and Liu, H. Quantifying the variability collapse of neural networks. In *International Conference on Machine Learning (ICML)*, 2023.

Yang, Y., Chen, S., Li, X., Xie, L., Lin, Z., and Tao, D. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yang, Y., Steinhardt, J., and Hu, W. Are neurons actually collapsed? on the fine-grained structure in neural representations. In *International Conference on Machine Learning (ICML)*, 2023a.

Yang, Y., Yuan, H., Li, X., Lin, Z., Torr, P., and Tao, D. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *International Conference on Learning Representations (ICLR)*, 2023b.

Yang, Y., Yuan, H., Li, X., Wu, J., Zhang, L., Lin, Z., Torr, P., Tao, D., and Ghanem, B. Neural collapse terminus: A unified solution for class incremental learning and its variants. *arXiv preprint arXiv:2308.01746*, 2023c.

Ye, H., Fan, W., Song, X., Zheng, S., Zhao, H., dan Guo, D., and Chang, Y. Ptarl: Prototype-based tabular representation learning via space calibration. In *International Conference on Learning Representations*, 2024.

Zeng, P., Lensen, A., and Sun, Y. Large scale image classification using gpu-based genetic programming. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 619–622, 2022.

Zhang, H., Cissé, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Zhang, M., Zhao, X., Yao, J., Yuan, C., and Huang, W. When noisy labels meet long tail dilemmas: A representation calibration method. In *International Conference on Computer Vision (ICCV)*, pp. 15844–15854, 2023.

Zhang, Y., Wei, X.-S., Zhou, B., and Wu, J. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2021.

Zhao, H., Phung, D., Huynh, V., Le, T., and Buntine, W. Neural topic model via optimal transport. In *International Conference on Learning Representations*, 2021.

Zhao, H., Sun, K., Dezfouli, A., and Bonilla, E. V. Transformed distribution matching for missing value imputation. In *International Conference on Machine Learning*, pp. 42159–42186. PMLR, 2023.

Zhong, Z., Cui, J., Liu, S., and Jia, J. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021.

Zhong, Z., Cui, J., Yang, Y., Wu, X., Qi, X., Zhang, X., and Jia, J. Understanding imbalanced semantic segmentation through neural collapse. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023.

Zhou, B., Lapedriza, À., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1452–1464, 2018.

Zhou, B., Cui, Q., Wei, X.-S., and 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 (CVPR)*, 2020.

Zhu, D., Li, Z., Zhang, M., Yuan, J., Shao, Y., Liu, J., Kuang, K., Li, Y., and Wu, C. Understanding prompt tuning for v-l models through the lens of neural collapse. *arXiv preprint arXiv:2306.15955*, 2023.

A. More Formulations of Total Classification Losses

To completely understand the combination of our approach with various imbalance loss functions, we present mathematical formulations of them in detail.

Combined the Label-Distribution-Aware Margin (LDAM) (Cao et al., 2019) Loss with Our Method:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{w}} \mathcal{L}_{Total}((x,y);f(\boldsymbol{\theta},\boldsymbol{w})) &= \mathcal{L}_{LDAM} + \lambda\mathcal{L}_{DisA} \\ &= -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\log\left(\frac{\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_k - \triangle_k)}{\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_k - \triangle_k) + \sum_{k'\neq k}^{K}\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_{k'})}\right) \\ &+ \lambda \min_{\mathbf{T}\in\Pi(P,Q)} \langle \mathbf{T}, \mathbf{C}(\mathbf{H},\mathbf{M})\rangle - \epsilon H(\mathbf{T}),\end{aligned} \tag{13}$$

where $\Delta_k = \frac{C}{n_k^{1/4}}$ for $k \in \{1,\ldots,K\}$ with constant C.

Combined the Balanced Meta-Softmax (BALMS) (Ren et al., 2020) Loss with Our Method:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{w}} \mathcal{L}_{Total}((x,y);f(\boldsymbol{\theta},\boldsymbol{w})) &= \mathcal{L}_{BALMS} + \lambda\mathcal{L}_{DisA} \\ &= -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\log\left(\frac{n_k\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_k)}{\sum_{k'=1}^{K} n_{k'}\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_{k'})}\right) \\ &+ \lambda \min_{\mathbf{T}\in\Pi(P,Q)} \langle \mathbf{T}, \mathbf{C}(\mathbf{H},\mathbf{M})\rangle - \epsilon H(\mathbf{T}).\end{aligned} \tag{14}$$

Combined the Focal Loss (Lin et al., 2017) with Our Method:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{w}} \mathcal{L}_{Total}((x,y);f(\boldsymbol{\theta},\boldsymbol{w})) &= \mathcal{L}_{Focal} + \lambda\mathcal{L}_{DisA} \\ &= -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(1 - \frac{\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_k)}{\sum_{k'=1}^{K}\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_{k'})}\right)^{\gamma}\log\left(\frac{\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_k)}{\sum_{k'=1}^{K}\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_{k'})}\right) \\ &+ \lambda \min_{\mathbf{T}\in\Pi(P,Q)} \langle \mathbf{T}, \mathbf{C}(\mathbf{H},\mathbf{M})\rangle - \epsilon H(\mathbf{T}),\end{aligned} \tag{15}$$

where γ is the modulating factor (We adopt $\gamma = 2$ in our experiments).

Combined the Class-Balanced (CB) Sigmoid Loss (Cui et al., 2019) with Our Method:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{w}} \mathcal{L}_{Total}((x,y);f(\boldsymbol{\theta},\boldsymbol{w})) &= \mathcal{L}_{CB-Sigmoid} + \lambda\mathcal{L}_{DisA} \\ &= -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\frac{1-\beta}{1-\beta^{n_k}})\log\left(\frac{1}{1+\exp(-\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_k) + \exp(\sum_{k'\neq k}^{K}\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_{k'})}\right) \\ &+ \lambda \min_{\mathbf{T}\in\Pi(P,Q)} \langle \mathbf{T}, \mathbf{C}(\mathbf{H},\mathbf{M})\rangle - \epsilon H(\mathbf{T}),\end{aligned} \tag{16}$$

where we set hyperparameter $\beta = 0.999$.

Combined the Class-Balanced (CB) Softmax Loss (Cui et al., 2019) with Our Method:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{w}} \mathcal{L}_{Total}((x,y);f(\boldsymbol{\theta},\boldsymbol{w})) &= \mathcal{L}_{CB-Softmax} + \lambda\mathcal{L}_{DisA} \\ &= -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\frac{1-\beta}{1-\beta^{n_k}})\log\left(\frac{\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_k)}{\sum_{k'=1}^{K}\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_{k'})}\right) \\ &+ \lambda \min_{\mathbf{T}\in\Pi(P,Q)} \langle \mathbf{T}, \mathbf{C}(\mathbf{H},\mathbf{M})\rangle - \epsilon H(\mathbf{T}),\end{aligned} \tag{17}$$

where we set hyperparameter $\beta = 0.999$.

Combined the Attraction-Repulsion-Balanced (ARB) Loss (Xie et al., 2023) with Our Method:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{w}} \mathcal{L}_{Total}((x,y);f(\boldsymbol{\theta},\boldsymbol{w})) &= \mathcal{L}_{ARB} + \lambda\mathcal{L}_{DisA} \\ &= -\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\log\left(\frac{\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_k)}{\sum_{k'=1}^{K}\frac{n_{k'}}{n_k}\exp(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{w}_{k'})}\right) \\ &+ \lambda\min_{\mathbf{T}\in\Pi(P,Q)}\langle\mathbf{T},\mathbf{C}(\mathbf{H},\mathbf{M})\rangle - \epsilon H(\mathbf{T}).\end{aligned} \tag{18}$$

From the formulas, it can be seen that the equations of Attraction-Repulsion-Balanced (ARB) loss and Balanced Meta-Softmax (BALMS) loss are nearly the same, so combining our method with the Balanced Meta-Softmax (BALMS) loss can be viewed as combining our method with Attraction-Repulsion-Balanced (ARB) loss.

Combined the Dot-Regression (DR) Loss (Yang et al., 2022) with Our Method:

$$\begin{aligned}\min_{\boldsymbol{\theta}} \mathcal{L}_{Total}((x,y);f(\boldsymbol{\theta})) &= \mathcal{L}_{DR} + \lambda\mathcal{L}_{DisA} \\ &= \frac{1}{2\sqrt{E_H E_M}}\left(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{m}_k - \sqrt{E_H E_M}\right)^2 \\ &+ \lambda\min_{\mathbf{T}\in\Pi(P,Q)}\langle\mathbf{T},\mathbf{C}(\mathbf{H},\mathbf{M})\rangle - \epsilon H(\mathbf{T}),\end{aligned} \tag{19}$$

where E_H and E_M are the ℓ_2 norm constraints for feature $\boldsymbol{h}_i$ and ETF weight vector $\boldsymbol{m}_k$, respectively.

Combined the Deferred Re-Weighting (DRW) (Cao et al., 2019) with Our Method: Following the proposed training method in Cao et al. (2019), we first use the standard ERM optimization schedule in (12) until the last learning rate decay with $\beta = 0$, and then apply re-weighting for optimization in the second stage with Class-Balanced (CB) Softmax loss with $\beta = 0.999$ in (17).

B. More Experiments

B.1. Comparison on Imbalanced Classification

Combined with Various Loss Functions on CIFAR-LT and Places-LT In this section, we report complete comparison studies with imbalanced long-tailed methods, including CE, CE-DRW (Cao et al., 2019), LDAM-DRW (Cao et al., 2019), BALMS (Ren et al., 2020), Focal loss (Lin et al., 2017), CB Softmax loss (Cui et al., 2019), and CB Sigmoid loss (Cui et al., 2019) with mixup augmentation. We summarize the results of the different imbalanced ratios on CIFAR-LT-10 and CIFAR-LT-100 in Table 4. We also run more experiments on Places-LT with some baselines in Table 5. Places-LT (Liu et al., 2019) is also a long-tailed dataset constructed from Places-365 (Zhou et al., 2018) with the imbalance ratio $\rho = 996$. We can find that our DisA, combined with several loss functions, improves the performance of the original loss for imbalanced classification. These results reveal the effectiveness of our proposed method when combined with other loss functions.

B.2. Comparison on Balanced Classification

To explore whether our proposed method can be used for the balanced classification task, we report on the results of CIFAR-10 (Krizhevsky et al., 2009) and CIFAR-100 (Krizhevsky et al., 2009) in Table 6. In this experiment, we compare the performance of ResNet-32 backbone. CE+DisA has grown by 7.1 % against CE in CIFAR-100, and no significant decrease in CIFAR-10. It indicates the benefit of our method of narrowing the gap between representation distribution and fixed ETF structure distribution through the neural collapse phenomenon on the balanced datasets.

B.3. Comparison on Several Neural Network Architectures

We conduct an additional experiment on neural network architectures with different imbalanced ratios on long-tailed classification in Table 7. We can see that, with the development of the layer in network structure, the performance of CE and CE+**DisA** becomes better. It thus indicates that the quality of last-layer features is related to the neural network architecture. This is reasonable since the quality of representation of DNNs depends on the depth and width of the network. Besides, whether the network structure is complex or not, introducing our DisA loss can achieve better performance than the baseline (CE), indicating the effectiveness of our method on different network architectures.

Table 4. Comparison results on CIFAR-LT-10 and CIFAR-LT-100 with imbalanced ratios $\rho = \{200, 100, 50, 10\}$. The best is marked in bold.

Method	CIFAR-LT-10				CIFAR-LT-100			
	200	100	50	10	200	100	50	10
CB Sigmoid	59.1	70.8	78.2	87.1	35.2	39.2	44.8	56.3
CB Sigmoid+**DisA**	**59.6**$^{+0.5}$	**71.3**$^{+0.5}$	**79.0**$^{+0.8}$	**88.4**$^{+1.3}$	**38.3**$^{+3.1}$	**41.2**$^{+2.0}$	**47.1**$^{+2.3}$	**57.6**$^{+1.3}$
CB Softmax	62.8	71.2	78.0	87.4	36.0	40.0	45.0	56.1
CB Softmax+**DisA**	**63.4**$^{+0.6}$	**72.1**$^{+0.9}$	**80.7**$^{+2.7}$	**88.0**$^{+0.6}$	**38.2**$^{+2.2}$	**41.9**$^{+1.9}$	**47.3**$^{+2.3}$	**57.1**$^{+1.0}$
CE	67.3	72.8	78.6	87.7	38.7	43.0	48.1	58.5
CE+**DisA**	**69.2**$^{+1.9}$	**74.7**$^{+1.9}$	**79.6**$^{+1.0}$	**88.3**$^{+0.6}$	**39.8**$^{+1.1}$	**44.5**$^{+1.5}$	**49.6**$^{+1.5}$	**63.2**$^{+4.7}$
LDAM	68.1	73.9	78.7	86.7	38.2	40.8	44.4	54.6
LDAM+**DisA**	**68.8**$^{+0.7}$	**74.5**$^{+0.6}$	**79.5**$^{+0.8}$	**87.0**$^{+0.3}$	**38.9**$^{+0.7}$	**41.3**$^{+0.5}$	**45.1**$^{+0.7}$	**54.9**$^{+0.3}$
CE-DRW	75.1	80.9	81.8	88.8	41.5	45.4	51.1	61.5
CE-DRW+**DisA**	**77.7**$^{+2.6}$	**82.1**$^{+1.2}$	**84.1**$^{+2.3}$	**89.8**$^{+1.0}$	**44.4**$^{+2.9}$	**49.2**$^{+3.8}$	**54.0**$^{+2.9}$	**65.7**$^{+4.2}$
LDAM-DRW	77.1	78.8	82.7	88.2	41.1	45.0	48.8	58.3
LDAM-DRW+**DisA**	**78.0**$^{+0.9}$	**80.4**$^{+1.6}$	**84.8**$^{+2.1}$	**88.3**$^{+0.1}$	**43.0**$^{+1.9}$	**47.0**$^{+2.0}$	**52.5**$^{+3.7}$	**64.1**$^{+5.8}$
BALMS	77.4	79.6	84.6	88.4	39.2	44.9	49.7	62.0
BALMS+**DisA**	**78.1**$^{+0.7}$	**81.8**$^{+2.2}$	**85.1**$^{+0.5}$	**89.8**$^{+1.4}$	**41.8**$^{+2.6}$	**46.5**$^{+1.6}$	**51.3**$^{+1.6}$	**62.8**$^{+0.8}$

Table 5. Comparison results of ResNet-152 on Places-LT. The best is marked in bold.

Method	Many	Medium	Few	All
CE	45.2	21.7	5.2	26.9
CE+**DisA**	**45.7**$^{+0.5}$	**23.3**$^{+1.6}$	**5.8**$^{+0.6}$	**28.0**$^{+1.1}$
CE-DRW	40.8	35.6	22.1	34.9
CE-DRW+**DisA**	**41.6**$^{+0.8}$	**36.7**$^{+1.1}$	**23.0**$^{+0.9}$	**35.8**$^{+0.9}$
LDAM-DRW	37.0	34.9	18.8	32.5
LDAM-DRW+**DisA**	**37.4**$^{+0.4}$	**35.7**$^{+0.8}$	**19.1**$^{+0.3}$	**33.0**$^{+0.5}$
BALMS	41.2	36.5	24.3	35.8
BALMS+**DisA**	**41.6**$^{+0.4}$	**37.4**$^{+0.9}$	**24.8**$^{+0.5}$	**36.5**$^{+0.7}$

Table 6. Comparison results on balanced datasets. The results of the compared methods are obtained from their respective original papers. The best is marked in bold.

Method	CIFAR-10	CIFAR-100
CE (Yang et al., 2022)	**93.6**	/
ETF-DR (Yang et al., 2022)	92.0	/
CE (Xie et al., 2023)	92.0	**68.7**
ARB (Xie et al., 2023)	**92.6**	68.6
CE (Liu et al., 2023b)	**93.4**	71.8
INC (Liu et al., 2023b)	93.3	**72.1**
CE	**93.5**	70.1
CE+**DisA**	**93.5**	**77.2**

Table 7. The performance of several neural network architectures with different imbalanced ratios on long-tailed classification.

Network	Method	200	100	50	10
ResNet-20	CE	66.7	71.7	76.9	86.7
	CE+**DisA**	**68.8**	**72.9**	**78.4**	**86.9**
ResNet-32	CE	67.3	72.8	78.6	87.7
	CE+**DisA**	**69.2**	**74.7**	**79.6**	**88.3**
ResNet-44	CE	68.2	74.0	80.4	88.8
	CE+**DisA**	**69.8**	**76.1**	**81.9**	**89.4**
ResNet-56	CE	68.8	74.3	80.7	88.9
	CE+**DisA**	**69.0**	**75.1**	**81.4**	**89.6**
ResNet-110	CE	69.6	75.2	81.1	89.6
	CE+**DisA**	**70.2**	**75.8**	**81.7**	**90.3**

C. More Analytical Results

Confusion Matrix To verify whether our method improves the performance of minority classes, we show the confusion matrices of CE-DRW, INC-DRW, and INC-DRW+DisA on CIFAR-LT-10 with $\rho = 200$ in Fig.6. We can observe that CE-DRW can almost perfectly classify the samples in majority classes and suffers a severe performance drop in the minority classes. INC-DRW contributes to increased accuracy in the minority classes but needs improvement. The incorporation of DisA goes a step further by enhancing the generalization of the minority class while maintaining the performance of the majority class. Compared with the powerful baseline, the improvement results in superior overall performance.

Class-means Angle We compare the pair-wise angles of the centered class means with the RBL and RBL+DisA learned on CIFAR-LT-10 of imbalanced ratios $\rho = \{200, 100, 50, 10\}$. The optimal pair-wise angle for 10 classes is $\arccos \frac{-1}{10-1} \approx 96.4°$. Fig.7 shows that the pair-wise angles of RBL+DisA are more closely matched to 96.4° than RBL. For example, the angle between class 9 and 1 widens from 41° to 49° with $\rho = 200$. The angle between class 4 and 1 shrinks from 123° to 113° with $\rho = 100$. It shows the consistency of our proposed DisA between neural collapse convergence and classification performance.

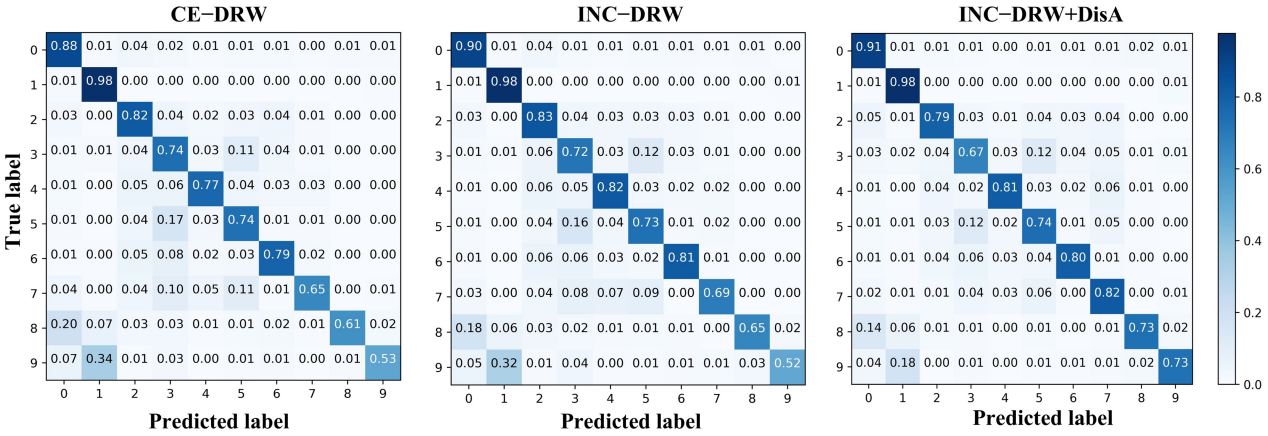


Figure 6. Confusion matrices of CE-DRW, INC-DRW, and INC-DRW+DisA on CIFAR-LT-10 with $\rho = 200$.

Hyper-parameter λ Discussion To analyze the effect of hyper-parameter λ of DisA, we conduct analytical experiments on CIFAR-LT-10 with $\rho = \{200, 100, 50, 10\}$. λ in (10) indicates the hyper-parameter for balancing the supervised loss and regularization loss. As shown in Fig.8, the larger the value of λ is, the more accurate the model. The best λ of four imbalanced ratios are $\lambda = 8$, $\lambda = 8$, $\lambda = 0.8$, and $\lambda = 6$, respectively.

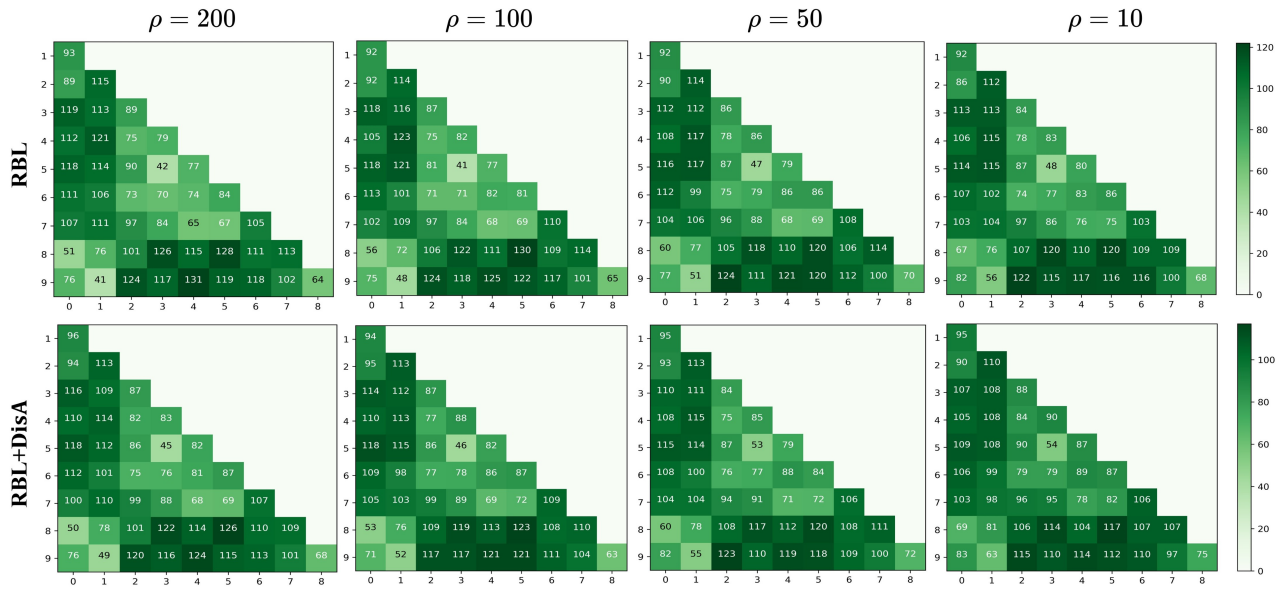
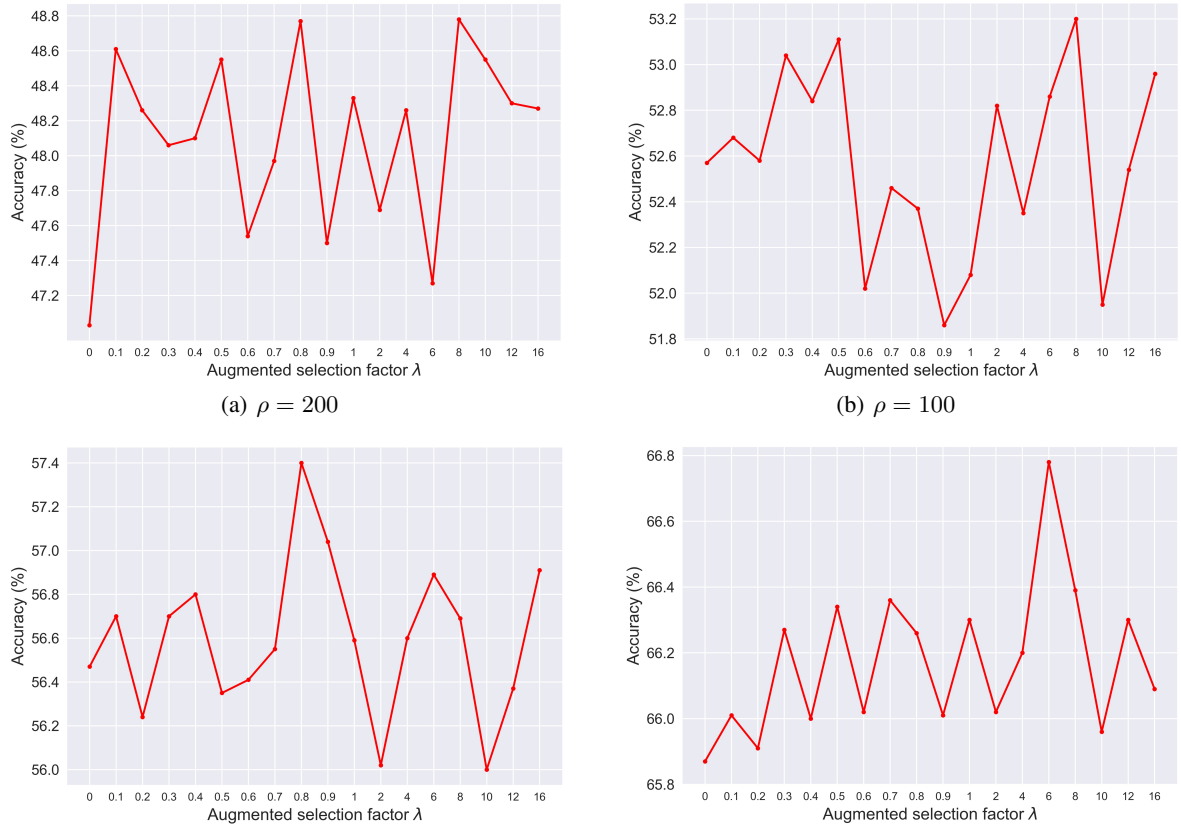


Figure 7. The pair-wise angles of the centered class means with RBL and RBL+DisA learned on CIFAR-LT-10 of different imbalanced ratios. The optimal pair-wise angle for 10 classes is $\arccos \frac{-1}{10-1} \approx 96.4°$.



(a) $\rho = 200$ (b) $\rho = 100$

(c) $\rho = 50$ (d) $\rho = 10$

Figure 8. Hyper-parameter λ discussion with different imbalanced ratios.

D. Algorithm

We summarize the complete procedure of our DisA method in Algorithm 2.

Algorithm 1 Distribution Alignment Optimization

Input: Distributions P and Q, hyper-parameter ϵ, iteration step E;
Output: $\mathcal{L}_{DisA}(P, Q)$
Compute the distance $\mathbf{C}$ between representation $\mathbf{H}$ and ETF tructure $\mathbf{M}$;
Set scaling vectors $\boldsymbol{u} \leftarrow \frac{1}{N} \cdot \mathbf{1}_N, \boldsymbol{v} \leftarrow \frac{1}{K} \cdot \mathbf{1}_K$;
for iteration $i = 1, 2, ..., E$ **do**
  $\boldsymbol{u}^{(i)} = \boldsymbol{u} / \left((\exp(-\mathbf{C}/\epsilon) \boldsymbol{v}^{(i-1)} \right)$;
  $\boldsymbol{v}^{(i)} = \boldsymbol{v} / \left((\exp(-\mathbf{C}/\epsilon)^\top \boldsymbol{u}^{(i-1)} \right)$;
end for
Compute $\mathbf{T} = \text{diag}(\boldsymbol{u}) \exp(-\mathbf{C}/\epsilon) \text{diag}(\boldsymbol{v})$;
$\mathcal{L}_{DisA}(P, Q) = \langle \mathbf{T}, \mathbf{C} \rangle$.

Algorithm 2 Overall Algorithm

Input: Training dataset $\mathcal{D}$, model f with parameter θ and classifier $\boldsymbol{W}$, ETF structure $\mathbf{M}$, a supervised learning method $\mathcal{A}$, number of epochs T, hyper-parameters $\{\beta, \lambda, \epsilon\}$;
Output: Model parameters;
Initialize θ randomly;
if fixed classifier with ETF **then**
  Define the classifier $\boldsymbol{W}$ with the ETF structure $\mathbf{M}$, $\boldsymbol{W} = \mathbf{M}$.
else
  Initialize $\boldsymbol{W}$ randomly;
end if
for epoch $t = 1, 2, ..., T$ **do**
  Build discrete distributions P with (6) and Q with (7);
  Compute $\mathcal{L}_{sup}$ according to $\mathcal{A}$ and the proposed $\mathcal{L}_{DisA}(P, Q)$ by Algorithm 1;
  if fixed classifier with ETF **then**
    Update $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta} - \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{total}$ by minimizing $\mathcal{L}_{sup} + \lambda \mathcal{L}_{DisA}$ with (11).
  else
    Update $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta} - \beta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{total}$ and $\boldsymbol{W}^* \leftarrow \boldsymbol{W} - \beta \nabla_{\boldsymbol{w}} \mathcal{L}_{total}$ by minimizing $\mathcal{L}_{sup} + \lambda \mathcal{L}_{DisA}$ with (12).
  end if
end for

E. Training Details

We report implementation details of previous methods and our method of neural collapse for long-tailed classification on the three datasets, CIFAR-LT-10, CIFAR-LT-100, and ImageNet-LT with three architectures, ResNet-32, ResNet-50, and ResNeXt-50 in Tables 8 and 9.

Table 8. Implementation details on CIFAR-LT of ResNet-32 by the SGD optimizer with a momentum of 0.9.

Method	Augment	Epochs	Batch	LR	Decay	Scheduler	Dim	GPU
ETF (Yang et al., 2022)	Mixup	200	128	0.1	2e-4	Step	64 / 128	1
INC (Liu et al., 2023b)	Mixup	200	128	0.1	5e-3	Step	/	4
+ DisA	Mixup	200	128	0.1	2e-4	Step	64 / 128	1
RBL (Peifeng et al., 2023)	AutoAug	600	256	0.1	5e-4	Cosine	256	1
+ DisA	AutoAug	600	256	0.1	5e-4	Cosine	256	1

Table 9. Implementation details on ImageNet-LT by the SGD optimizer with a momentum of 0.9 and cosine schedule.

Method	Backbone	Augment	Epochs	Batch	LR	Decay	Dim	GPU
ETF (Yang et al., 2022)	ResNet-50	Mixup	180	1024	0.1	5e-4	4096	8
+ DisA	ResNet-50	Mixup	180	128	0.1	5e-4	4096	4
INC (Liu et al., 2023b)	ResNeXt-50	Mixup + RandAug	200	256	0.05	5e-3	/	4
+ DisA	ResNeXt-50	Mixup + RandAug	200	128	0.1	5e-4	2048	4
RBL (Peifeng et al., 2023)	ResNeXt-50	CommonAug	200	64	0.25	5e-4	512	1
+ DisA	ResNeXt-50	CommonAug	200	64	0.25	5e-4	512	1

F. Computational Cost

The optimal transport (OT) problem in our method between probability distributions is computed by the Sinkhorn algorithm (Cuturi, 2013), which introduces the entropic regularization term for fast computation. We compare the computational cost of different methods on a Pentium PC with a single GTX A10 GPU. As shown in Table 10, imbalanced NC-based methods (ETF-DR, INC-DRW, and RBL) usually take more time than traditional methods (CE and CE-DRW). The reason behind this is the model computes additional information. For example, INC needs to compute the class-mean features and the global-mean feature for regularization. The process of optimizing the rotation orthogonal matrix of RBL also brings large expenses. Therefore, it is reasonable that NC-based methods have a higher computational cost. Besides, DisA spends more time than these NC-based methods since we solve an OT problem to match imbalanced and ETF distributions for representation learning. However, combining ours with others produces a better performance on long-tailed datasets with an acceptable cost. Moreover, we also conduct an experiment on the model computational overhead with various feature dimensions(d) and classes on long-tailed datasets. As shown in Table 11, we can find that when the dimensions and classes are larger, more computational cost is required by all baselines and DisA, where the model requires more memory to learn representation.

Table 10. Computational cost (s) per training epoch on long-tailed datasets.

Method	CIFAR-LT-10	CIFAR-LT-100	ImageNet-LT
CE	2.44	2.33	323.52
CE+DisA	**2.68**	**3.21**	**337.48**
CE-DRW	2.49	2.73	334.12
CE-DRW+**DisA**	**2.74**	**3.38**	**357.33**
ETF-DR	3.49	4.92	367.55
ETF-DR+**DisA**	**3.61**	**5.31**	**397.26**
INC-DRW	3.78	4.51	365.24
INC-DRW+**DisA**	**3.82**	**4.86**	**381.11**
RBL	7.40	7.91	472.13
RBL+**DisA**	**7.61**	**8.41**	**510.64**

Table 11. The computational overhead (s) per training epoch with various feature dimensionality and classes on long-tailed datasets. *: our implementation dimension in the paper.

Datasets	CIFAR-LT-10					CIFAR-LT-100				
Dimension	16	32	64	128	256	16	32	64	128	256
CE	2.41	2.42	2.44*	2.84	3.25	2.32	2.35	2.34	2.33*	2.87
CE+**DisA**	**2.67**	**2.68**	**2.68***	**3.03**	**3.47**	**2.74**	**2.81**	**2.89**	**3.21***	**3.51**
CE-DRW	2.42	2.46	2.49*	2.88	3.64	2.58	2.61	2.67	2.73*	3.79
CE-DRW+**DisA**	**2.68**	**2.71**	**2.74***	**3.06**	**3.75**	**2.79**	**2.83**	**3.11**	**3.38***	**3.94**
ETF-DR	3.21	3.33	3.49*	4.87	5.21	3.66	3.71	3.78	4.92*	5.41
ETF-DR+**DisA**	**3.34**	**3.56**	**3.61***	**5.02**	**5.35**	**3.75**	**3.91**	**4.56**	**5.31***	**5.73**
INC-DRW	3.48	3.56	3.78*	4.31	4.72	3.74	3.86	4.12	4.51*	4.97
INC-DRW+**DisA**	**3.63**	**3.79**	**3.82***	**4.63**	**4.87**	**3.93**	**4.15**	**4.56**	**4.86***	**5.22**
RBL	7.21	7.33	7.40	8.06	8.78*	7.35	7.47	7.66	7.91	8.45*
RBL+**DisA**	**7.35**	**7.58**	**7.61**	**8.36**	**8.90***	**7.78**	**7.92**	**8.24**	**8.41**	**9.03***