Inducing Neural Collapse in Deep Long-tailed Learning

Xuantong Liu[1,*] **Jianfeng Zhang**[2] **Tianyang Hu**[2] **He Cao**[1] **Lujia Pan**[2] **Yuan Yao**[1,✉]
[1] The Hong Kong University of Science and Technology, [2]Huawei Noah's Ark Lab

Abstract

Although deep neural networks achieve tremendous success on various classification tasks, the generalization ability drops sheer when training datasets exhibit long-tailed distributions. One of the reasons is that the learned representations (i.e. features) from the imbalanced datasets are less effective than those from balanced datasets. Specifically, the learned representation under class-balanced distribution will present the *Neural Collapse* ($\mathcal{NC}$) phenomena. $\mathcal{NC}$ indicates the features from the same category are close to each other and from different categories are maximally distant, showing an optimal linear separable state of classification. However, the pattern differs on imbalanced datasets and is partially responsible for the reduced performance of the model. In this work, we propose two explicit feature regularization terms to learn high-quality representation for class-imbalanced data. With the proposed regularization, $\mathcal{NC}$ phenomena will appear under the class-imbalanced distribution, and the generalization ability can be significantly improved. Our method is easily implemented, highly effective, and can be plugged into most existing methods. The extensive experimental results on widely-used benchmarks show the effectiveness of our method.

1 INTRODUCTION

Modern deep neural networks have shown the ability to outperform humans on many tasks, such as computer vision, natural language processing, playing games, etc., and keep refreshing state-of-the-art performance for complex classification tasks. However, when the training dataset is class-imbalanced, such as a long-tailed distribution, where a few majority classes occupy most of the training samples while a large number of minority classes own very limited samples, the performance of the model drops off a cliff (Van Horn & Perona, 2017; Buda et al., 2018). These imbalanced distributions are ubiquitous in real-world applications, e.g., fault diagnosis, face recognition, autonomous driving, etc. Therefore, how to improve the discriminative ability of the model trained on the imbalanced dataset has always been a topic of considerable concern.

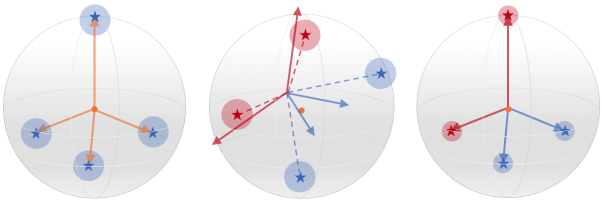


Figure 1: Illustration of geometry configuration of the zero-centered class means and classifier weights under (a) balanced dataset, (b) imbalanced dataset, and (c) imbalanced dataset with our method. The arrows represent classifier weights and the stars are the class centers. The size of the circle around the star reflects the variance of the feature from the same class. In (b) and (c), red and blue represent the majority and minority classes, respectively. Note that under imbalanced label distribution, both the centered class means and classifier weights form an asymmetric structure and are no longer parallel.

Some recent studies focus on learning more effective representation to improve the long-tailed recognition ability. Supervised contrastive loss (Khosla et al., 2020) is utilized to learn compact within-class and maximally distant between-class representation by introducing uniformly distributed class centers, which leads to improvement in long-tailed performance (Li et al., 2022; Cui et al., 2021; Zhu et al., 2022). These characteristics of the feature representation are consistent with those learned from balanced datasets (Graf et al., 2021), where the classification models can spontaneously learn tight and discriminative features. However, contrastive

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

* This work is done when Xuantong Liu was doing her internship at Huawei Noah's Ark. ✉ Corresponding author email: yuany@ust.hk.

learning is more computationally expensive and requires more iterations to converge than the standard Cross-Entropy (CE) loss.

Meanwhile, the learning behavior of deep classification models in a balanced setting has been investigated both empirically and theoretically (Papyan et al., 2020; Galanti et al., 2021; Han et al., 2022). The *Neural Collapse* ($\mathcal{NC}$) phenomenon was uncovered by Papyan et al. (2020) when investigating the last-layer embedding, i.e. the feature representation, and the corresponding classifier weights in deep classification models during training. $\mathcal{NC}$ shows that the learned features (or embedded vectors) of the same class will collapse to their class centers. Meanwhile, these class centers, after globally centered, as well as the classifier weights, will form a simplex equiangular tight frame (ETF) during the terminal phase of training (TPT), i.e. when the model achieves zero training error. The ETF structure maximizes the between-class variability so as the *Fisher discriminant ratio* (Fisher, 1936), resulting in an optimal linear separable state for classification. Subsequent studies have found more characteristics of this phenomenon, including the global optimal property (Zhu et al., 2021) and generalization ability (Galanti et al., 2021).

However, on imbalanced datasets, the deep neural networks will exhibit different geometric structures, and some $\mathcal{NC}$ phenomena will no longer occur (Fang et al., 2021; Thrampoulidis et al., 2022). The last-layer features of the same class still converge to their class means, but the class means, as well as the classifier weights, are not in the form of ETFs any more. Specifically, compared to majority classes, the learned features of minority classes will have a larger norm, and correspondingly the norm of classifier weights will be smaller (Kang et al., 2019; Fang et al., 2021). Furthermore, as the imbalance level increases, the phenomenon of *Minority Collapse* may arise, in which both the learned representations and the classifier weights on minority classes will become indistinguishable (Fang et al., 2021). The absence of some $\mathcal{NC}$ property partially explains the performance gap between the balanced and imbalanced datasets.

In this paper, we first elaborate that the appearance of $\mathcal{NC}$ can help to minimize the generalization error in the imbalanced problem. According to this property, we propose two simple yet effective regularization terms to explicitly induce all the $\mathcal{NC}$ phenomena in neural networks trained on imbalanced datasets. The regularization terms can be added to CE loss directly. Compared with supervised contrastive learning, these terms have lower computational cost. Our proposed method not only helps the $\mathcal{NC}$ to occur faster for models trained on the balanced datasets, but also drives the $\mathcal{NC}$ phenomenon to occur on datasets with imbalanced categories. The resulting model can also obtain better generalization ability and robustness without over-training as in Papyan et al. (2020). Furthermore, our proposed method is orthogonal to most existing methods dealing with long-tailed problems. It thus can be easily plugged into the objective function to obtain further improvements.

In summary, our contributions can be listed below:

- We observe that when training data is imbalanced, the class centers of minority classes move closer to those of the majority classes, making their instances difficult to distinguish.
- We demonstrate that, although some $\mathcal{NC}$ phenomena do not naturally exist in an imbalanced case, we can achieve lower generalization error when all $\mathcal{NC}$ proprieties hold. Thus we propose two simple yet effective regularization terms to manually induce the $\mathcal{NC}$ during imbalanced training.
- We experimentally show that our method can significantly improve the performance in various long-tailed tasks and boost most existing methods.

2 PROBLEM SETUP

2.1 Preliminaries

Let $f_\phi \circ g_\theta(\cdot)$ denote a neural network classifier, where $g_\theta(\cdot)$ is a feature extractor and $f_\phi(\cdot)$ is a linear classifier. We define $\boldsymbol{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_n]^T \in \mathbb{R}^{n\times P}$ to be the output of $g_\theta(\cdot)$. Here P is the dimension of the latent feature, and n is the training sample size. The weights of f_ϕ are denoted by $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K] \in \mathbb{R}^{P\times K}$, and the corresponding bias vector is $\boldsymbol{b} = [b_1, ..., b_K]$, where K is the number of classes. $\boldsymbol{h}_i \in \mathbb{R}^P$ and $y_i \in \{k\}_{k=1}^K$ denote the feature and label of the i-th sample. The label matrix is denoted by $\boldsymbol{Y} \in \mathbb{R}^{n\times K}$. In the training data, we have $n = \sum_{k=1}^K n_k$, where n_k is the sample size of class k. We use $||\cdot||_F$ and $||\cdot||$ to denote the Frobenius norm of a matrix and the l_2-norm of a vector.

Definition 1 (Simplex ETF)**.** A simplex ETF is a collection of equal-length and maximally-equiangular vectors. We call a $P \times K$ matrix $\boldsymbol{M}$ an ETF if it satisfies

$$\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M} = \alpha\left(\frac{K}{K-1}\boldsymbol{I} - \frac{1}{K-1}\mathbb{1}_K\mathbb{1}_K^{\mathrm{T}}\right) \tag{1}$$

for some non-zero scalar α. Where $\boldsymbol{I}$ is the identity matrix, and $\mathbb{1}_K$ is an all-ones vector.

Let $\boldsymbol{\mu}_k = \frac{1}{n_k}\sum_{y_i=k}\boldsymbol{h}_i$ be the center of class k and $\boldsymbol{\mu}_C = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\mu}_k$ be the arithmetic mean of the class centers. In the balanced case, where we have $n_k = \frac{n}{K}$ for each class, $\mathcal{NC}$ will appear during TPT. The phenomena can be formally described by four properties:

- ($\mathcal{NC}_1$) **Variability collapse.** Intra-class variances collapse to zero during the terminal phase of training, i.e., for any sample i from class k, we have

$$||\boldsymbol{h}_i - \boldsymbol{\mu}_k|| = 0 \tag{2}$$

- ($\mathcal{NC}_2$) **Convergence to simplex ETF.** The class centers (after zero-center normalization) converge to the vertices of an ETF, i.e.

$$\cos(\boldsymbol{\mu}_k - \boldsymbol{\mu}_C, \boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_C) = -\frac{1}{K-1}, \quad (3)$$

$$||\boldsymbol{\mu}_k - \boldsymbol{\mu}_C|| = ||\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_C||. \quad (4)$$

- ($\mathcal{NC}_3$) **Convergence to self-duality.** The weights of linear classifiers are parallel to the corresponded zero-centered class centers, i.e.

$$\boldsymbol{w}_k = \alpha(\boldsymbol{\mu}_k - \boldsymbol{\mu}_C). \quad (5)$$

- ($\mathcal{NC}_4$) **Simple decision rule.** Given a feature, the last-layer classifier's behavior is equivalent to the nearest class center (NCC) decision rule, i.e.

$$\arg\max_k \langle \boldsymbol{w}_k, \boldsymbol{h} \rangle = \arg\min_k ||\boldsymbol{h} - \boldsymbol{\mu}_k|| \quad (6)$$

2.2 Neural Collapse and Imbalanced Data

In this section, we first illustrate why $\mathcal{NC}$ disappears on imbalanced datasets using mean squared error (MSE) loss. Then we demonstrate that the $\mathcal{NC}$ properties will lead to a lower generalization error bound thus we can benefit from it under imbalanced distribution.

2.2.1 The Optimal Classifier Under Imbalanced Distribution

Some recent studies have shown that the test performance of neural networks trained with MSE loss is comparable to those trained with CE loss in classification tasks (Demirkaya et al., 2020; Fang et al., 2021; Hu et al., 2021; Han et al., 2022). Thanks to its tractability, we can use MSE loss to illustrate the absence of the $\mathcal{NC}$ phenomenon on imbalanced datasets. For linear classifiers, the MSE loss is

$$\mathcal{L}(\boldsymbol{H}, \boldsymbol{W}) = \frac{1}{2n}||\boldsymbol{Y} - (\boldsymbol{H}\boldsymbol{W} + \mathbb{1}_n \boldsymbol{b}^{\mathrm{T}})||_F^2. \quad (7)$$

Let $\bar{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}_i$ be the global feature mean, $\boldsymbol{\Sigma}_T = (\boldsymbol{H} - \mathbb{1}_n \bar{\boldsymbol{h}})^{\mathrm{T}}(\boldsymbol{H} - \mathbb{1}_n \bar{\boldsymbol{h}})$ be the total covariance matrix of $\boldsymbol{H}$ and $\dot{\boldsymbol{M}} = [\boldsymbol{\mu}_1 - \bar{\boldsymbol{h}}, ..., \boldsymbol{\mu}_K - \bar{\boldsymbol{h}}] \in \mathbb{R}^{P \times K}$. We can have the closed form of the optimal $\boldsymbol{W}$ and $\boldsymbol{b}$ under the MSE loss as follows.

Proposition 1. (Webb & Lowe, 1990). In general, for fixed features $\boldsymbol{H}$, the optimal weight matrix and the bias vector that minimize $\mathcal{L}(\boldsymbol{H}, \boldsymbol{W})$ are

$$\boldsymbol{W}_{LS} = \boldsymbol{\Sigma}_T^{\dagger} \dot{\boldsymbol{M}} \boldsymbol{\Lambda}, \quad (8)$$

$$\boldsymbol{b}_{LS} = \frac{1}{n}\mathbb{1}_n^{\mathrm{T}} \boldsymbol{Y} - \boldsymbol{\mu}_G \boldsymbol{W}_{LS}, \quad (9)$$

where † denotes the Moore-Penrose pseudoinverse, and $\boldsymbol{\Lambda} = diag(n_1, \cdots, n_K)$ is a diagonal matrix.

From Eq.(9), we can observe that the optimal weight matrix depends on the features and is strongly affected by $\boldsymbol{\Lambda}$, i.e. the proportions of classes. Specifically, the classifier weights of the majority classes will have larger norms. The $\mathcal{NC}$ phenomena reflect the intimate connection between the last layer features and the classifier weights. Thus, skewed classifiers imply that the features are also biased, and many studies have empirically investigated that the uneven label distribution can lead to an imbalanced feature space (Kang et al., 2020; Fang et al., 2021; Li et al., 2022).

Particularly, Fang et al. (2021) show the *Minority Collapse* phenomenon that reveals the skewed classifier weights encountering an imbalanced label distribution where majority classes own much more samples than the minority ones. In addition, they theoretically prove that unbiased classifiers can be obtained through over-sampling. However, empirical results show a limited performance improvement or even decline due to the over-fitting of the minority classes (Drummond et al., 2003; Weiss et al., 2007). On the other hand, the classifiers are always better tuned than the learned features (Thrampoulidis et al., 2022). Therefore, in this work, we mainly focus on regularizing the embeddings during training to get non-skewed and representative features. Then we tune a balanced classifier based on our well-learned features.

2.2.2 Importance of $\mathcal{NC}$ on Imbalanced Datasets

As we already know that when the training set is class-imbalanced, the geometric structure of the classifiers and centered class means are not symmetric, which may introduce some bias in the model and affect the performance of the test set (Kang et al., 2019, 2020; Fang et al., 2021). Recent work indicates that compact within-class representations along with evenly distributed class centers can help learn high-quality representations, and substantial practice confirms this (Li et al., 2022; Zhu et al., 2022; Cui et al., 2022). These intuitions lead to similar situations with $\mathcal{NC}$. In this section, we explain why the $\mathcal{NC}$ can be considered favorable representations and can provide reduced generalization errors under long-tailed distributions from the perspective of *domain adaptation*.

As a standard evaluation approach in long-tailed learning, models are usually tested on balanced datasets. Since the training set is imbalanced, we can regard this scenario as a label shift domain adaptation problem, where the source domain is imbalanced, and the target domain is balanced.

First, the following proposition shows that properties of $\mathcal{NC}_1$ and $\mathcal{NC}_2$ can be approximately preserved in the target domain.

Proposition 2. (Galanti et al., 2021) Let μ_k^S (resp. μ_k^T) and σ_k^S (resp. σ_k^T) be the mean and variance of the representations of class k on the source domain (resp. target domain). For any two different classes, k and k', with probability at

least $1-\delta$ over $\mathcal{D}_S$, we have

$$\frac{\sigma_k^T+\sigma_{k'}^T}{2\|\mu_k^T-\mu_{k'}^T\|^2} \leq (1+A^2)\left(\frac{\sigma_k^S+\sigma_{k'}^S}{2\|\mu_k^S-\mu_{k'}^S\|^2}+B\right), \quad (10)$$

where $A=\frac{\mathcal{O}(\sqrt{\log(1/\delta)/n_k})}{||\mu_k^T-\mu_{k'}^T||}$, $B=\frac{\mathcal{O}(\sqrt{\log(1/\delta)/n_k})}{||\mu_k^S-\mu_{k'}^S||^2}$.

The ETF geometry of $\{\mu_k\}_{k=1}^K$ indicates that the distance $\|\mu_k^S-\mu_{k'}^S\|$ achieve maximum value for all $k\neq k'$. On the other hand, $\|\mu_k^T-\mu_{k'}^T\|$ is also lower bounded by $\|\mu_k^S-\mu_{k'}^S\|$ (Galanti et al., 2021). Hence, A and B are upper bounded and diminish to zero as n_k gets larger. Therefore, we can roughly speak $\mathcal{NC}_1$ and $\mathcal{NC}_2$ can generalize to the target domain.

We then illustrate how the existence of $\mathcal{NC}_1$ and $\mathcal{NC}_2$ help to reduce the generalization error. According to Ben-David et al. (2006), for any classifier h, the error on target domain $\epsilon_T(h)$ will be bound by the empirical error on the source domain and the divergence between source and target feature domains plus a constant:

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) + \text{const}, \quad (11)$$

where $d_{\mathcal{H}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$[1] measures some 'distance' between source and target domains over the feature space $\mathcal{Z}$. Although not exactly the same, substituting $d_{\mathcal{H}}$ with the Jensen–Shannon distance d_{JS} (Endres & Schindelin, 2003) will not significantly change the result. Theoretically, minimizing d_{JS} between source and target distributions will reduce the right-hand side of Eq.(11) as well. Let $\mathcal{D}^Z$ and $\mathcal{D}^Y$ be the distributions defined over the latent feature space and label space, respectively. As $\mathcal{D}^Y$ can be induced from $\mathcal{D}^Z$ from a generative perspective, according to Zhao et al. (2019), we have

$$d_{JS}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) \geq d_{JS}(\mathcal{D}_S^Y, \mathcal{D}_T^Y), \quad (12)$$

i.e., $d_{JS}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$ is the lower bounded by $d_{JS}(\mathcal{D}_S^Y, \mathcal{D}_T^Y)$, which is a constant determined by source and target label distributions.

With $\mathcal{NC}_1$ and $\mathcal{NC}_2$, the distribution over $\mathcal{Z}$ collapses to a K-component mixture Dirac distribution. More precisely, we have $\Pr(Z=\boldsymbol{h})=\Pr(Y=\boldsymbol{y})$. In this case, $d_{JS}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$ attains its lower bound $d_{JS}(\mathcal{D}_S^Y, \mathcal{D}_T^Y)$, which is the objective of some classical domain adaptation algorithms (Long et al., 2015; Ganin et al., 2017),

3 LEARNING REPRESENTATION VIA INDUCING NEURAL COLLAPSE

The previous analysis inspires us to induce $\mathcal{NC}$ phenomena to imbalanced training. We mainly focus on the core properties, $\mathcal{NC}_1$ and $\mathcal{NC}_2$, and come up with two corresponding regularization terms.

[1] $d_{\mathcal{H}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$ denotes the $\mathcal{H}$-divergence between $\mathcal{D}_S^Z$ and $\mathcal{D}_T^Z$, a precise definition is provided in Ben-David et al. (2010).

3.1 Feature Regularization

Compact within-class features. $\mathcal{NC}_1$ underlines that the model is seeking to learn compact within-class features by pushing the last-layer embedding to be close to their class centers, which seems natural but actually hard to achieve in practice. Han et al. (2022) decomposed the MSE loss and discovered that the loss in the late training stages is dominated by the ℓ_2-distance between the feature and the corresponding class center. This indicates that although $\mathcal{NC}_1$ is the inevitable trend, it is quite difficult to realize. Therefore, we add explicit regularization to make $\mathcal{NC}_1$ more inclined to appear. Especially, for the class-imbalanced dataset, we consider the inverse ratio of class sizes as weights to avoid excessive force on the majority classes. This indicates the difference between our $\mathcal{NC}_1$ regularization and the *center loss* (Wen et al., 2016) that pushes all features equally to their class center. Formally, we define the $\mathcal{NC}_1$ regularization as the within-class feature distance, $\mathcal{L}_W$, with the formula of

$$\mathcal{L}_W = \sum_{k=1}^{K}\sum_{y_i=k}\frac{1}{n_k}||\boldsymbol{h}_i-\boldsymbol{\mu}_k||_2^2. \quad (13)$$

Distinct between-class features. $\mathcal{NC}_2$ shows that with balanced class distribution, all pairs of centered class means tend to form equal-sized angles, implying the maximally separated between-class features. However, under the imbalanced distribution, the class centers of the minority classes are close to the majority ones, leading to indistinguishable features. Therefore, we propose $\mathcal{NC}_2$ regularization to minimize the maximal pairwise cosine similarity between all the centered class means, equivalent to maximizing the minimal pairwise angle. Consider the angular version, the objective of $\mathcal{NC}_2$ regularization is:

$$\max\min_{k\neq k'}\arccos\frac{\langle\dot{\boldsymbol{\mu}}_k,\dot{\boldsymbol{\mu}}_{k'}\rangle}{||\dot{\boldsymbol{\mu}}_k||\cdot||\dot{\boldsymbol{\mu}}_{k'}||}, \quad (14)$$

where $\dot{\boldsymbol{\mu}}_k=\boldsymbol{\mu}_k-\boldsymbol{\mu}_C$. As noted in Wang et al. (2020), updating the average of each vector's maximum cosine is more efficient than just optimizing the global maximum cosine. Therefore, we define the formula for the $\mathcal{NC}_2$ regularization as

$$\mathcal{L}_B = -\frac{1}{K}\sum_{k=1}^{K}\min_{k',k'\neq k}\arccos\frac{\langle\dot{\boldsymbol{\mu}}_k,\dot{\boldsymbol{\mu}}_{k'}\rangle}{||\dot{\boldsymbol{\mu}}_k||\cdot||\dot{\boldsymbol{\mu}}_{k'}||}. \quad (15)$$

In summary, our proposed feature regularization includes two terms, $\mathcal{L}_W$ and $\mathcal{L}_B$, corresponding to minimize the within-class distance and maximize the between-class discrepancy, respectively. They can be easily coupled with supervised losses with a linear classifier to regularize the penultimate layer embedding. Finally, we have the following loss for training:

$$\mathcal{L}=\mathcal{L}_{sup}+\lambda_1\mathcal{L}_W+\lambda_2\mathcal{L}_B. \quad (16)$$

where $\mathcal{L}_{sup}$ denotes the supervised loss, *e.g.* CE loss and MSE loss. λ_1 and λ_2 are hyperparameters that control the impact of $\mathcal{L}_W$ and $\mathcal{L}_B$.

3.2 Occurrence of Neural Collapse

First, we illustrate that $\mathcal{L}_B$ will lead all pairs of the K class means to have the same cosine equals to $-\frac{1}{K-1}$, with the following proposition.

Proposition 3. (Wang et al., 2020) The minimum of the maximal pair-wise cosine similarity between n vectors is $\frac{-1}{n-1}$, which can be reached when the vectors have an equal-sized pair-wise angle and zero mean.

Therefore, denote $\hat{\boldsymbol{M}} = [\frac{\dot{\boldsymbol{\mu}}_1}{||\dot{\boldsymbol{\mu}}_1||}, ..., \frac{\dot{\boldsymbol{\mu}}_K}{||\dot{\boldsymbol{\mu}}_K||}]$. Recall that the objective of $\mathcal{L}_B$ is to minimize the maximal pair-wise cosine similarity of the centered class means, thus with $\mathcal{L}_B$, we have

$$\hat{\boldsymbol{M}}^{\mathrm{T}}\hat{\boldsymbol{M}} = \frac{K}{K-1}\boldsymbol{I} - \frac{1}{K-1}\mathbf{1}_K\mathbf{1}_K^{\mathrm{T}}. \tag{17}$$

According to **Definition** 1, $\hat{\boldsymbol{M}}$ form a simplex ETF. Furthermore, although $\mathcal{L}_W$ and $\mathcal{L}_B$ do not explicitly enforce the centered class means to have an equal norm, we empirically observe this desired result (see the experimental result in **Section** 4.2.1). Let $||\dot{\boldsymbol{\mu}}_1|| = \cdots = ||\dot{\boldsymbol{\mu}}_K|| = \alpha$ and $\bar{\boldsymbol{M}} = [\dot{\boldsymbol{\mu}}_1, \cdots, \dot{\boldsymbol{\mu}}_K]$, then we have

$$\bar{\boldsymbol{M}}^{\mathrm{T}}\bar{\boldsymbol{M}} = \alpha\left(\frac{K}{K-1}\boldsymbol{I} - \frac{1}{K-1}\mathbf{1}_K\mathbf{1}_K^{\mathrm{T}}\right), \tag{18}$$

indicating the centered class means indeed from an ETF. Therefore, with the proposed feature regularization terms $\mathcal{L}_W$ and $\mathcal{L}_B$, $\mathcal{NC}_1$ and $\mathcal{NC}_2$ can happen even when the training set is imbalanced.

In addition, we can prove that with the existence of $\mathcal{NC}_1$ and $\mathcal{NC}_2$, retrain the classifier with class-balance sampling, the classifier can become parallel with the centered feature mean, indicating the self-duality ($\mathcal{NC}_3$). Ultimately, the symmetric structure of the regularized class means brings about an unbiased linear classifier.

Proposition 4. **Proposition** 1+$\mathcal{NC}_1$+$\mathcal{NC}_2$+class-balanced sampling can lead to $\mathcal{NC}_3$.

Proof. With class-balanced sampling, the training label distribution can be regarded as balanced, and $\dot{\boldsymbol{M}} = \bar{\boldsymbol{M}}$. Then the optimal re-trained classifier $\boldsymbol{W}_r$ is

$$\boldsymbol{W}_r = \frac{n}{K}\boldsymbol{\Sigma}_T^{\dagger}\dot{\boldsymbol{M}}, \tag{19}$$

with the existence of $\mathcal{NC}_1$, we have $\boldsymbol{\Sigma}_T = \dot{\boldsymbol{M}}\dot{\boldsymbol{M}}^{\mathrm{T}}$. Thus,

$$\begin{aligned}\boldsymbol{W}_r &= \frac{n}{K}(\dot{\boldsymbol{M}}\dot{\boldsymbol{M}}^{\mathrm{T}})^{\dagger}\dot{\boldsymbol{M}} \\ &= \frac{n}{K}(\dot{\boldsymbol{M}}\dot{\boldsymbol{M}}^{\mathrm{T}})^{\dagger}\dot{\boldsymbol{M}}\dot{\boldsymbol{M}}^{\mathrm{T}}(\dot{\boldsymbol{M}}^{\mathrm{T}})^{\dagger} \\ &= \frac{n}{K}(\dot{\boldsymbol{M}}^{\mathrm{T}})^{\dagger},\end{aligned}$$

with $\mathcal{NC}_2$ which implies that $\dot{\boldsymbol{M}}$ form a simplex-ETF, thus, $(\dot{\boldsymbol{M}}^{\mathrm{T}})^{\dagger} = c\dot{\boldsymbol{M}}$ for some constant c (Papyan et al., 2020), then we can obtain $\boldsymbol{W}_r = \alpha\dot{\boldsymbol{M}}$, demonstrating the asserted self-duality ($\mathcal{NC}_3$). □

In conclusion, with the proposed $\mathcal{L}_W$ and $\mathcal{L}_B$, we can obtain *compact within-class* and *distinct between-class* representations under imbalanced-class distribution. In line with linear discriminant analysis (LDA) (Fisher, 1936), this provides an optimal solution for the linear classifier.

4 EXPERIMENTS

4.1 Classification and long-tailed recognition

In this section, we conduct various experiments on image classification tasks on both balanced and long-tailed datasets to validate the effectiveness of our method. We denote our approach as NC, indicating the occurrence of the $\mathcal{NC}$ phenomena. By default, CE is adopted as $\mathcal{L}_{sup}$.

4.1.1 Experiment Setup

Datasets. Two balanced datasets (CIFAR10 and CIFAR100) and three long-tailed datasets (CIFAR10-LT, CIFAR100-LT, and ImageNet-LT) are used in our experiments. Following Cao et al. (2019), CIFAR10/100-LT are created by downsampling each class's samples to obey an exponential decay with an imbalance ratio $r = 100$ and 10. Here $r = \max\{n_k\}/\min\{n_k\}$. ImageNet-LT (Liu et al., 2019), including 115,846 samples and 1,000 categories with size ranging from 5 to 1,280, is generated from the ImageNet-2012 (Deng et al., 2009) dataset using a Pareto distribution with the power value $\alpha = 6$.

Baselines. In addition to the typical approaches for addressing imbalanced data, such as re-sampling (RS) and re-weighting (RW) in inverse proportion to the class size, the investigation of more conducive methods that decouple representation learning and classifier training, as well as relevant methods inspired by $\mathcal{NC}$, are also carried out. To be specific, we compare traditional supervised learning methods with DRW (Cao et al., 2019), LWS (Kang et al., 2019), and cRT (Kang et al., 2019), and two recent works, namely BBN (Zhou et al., 2020) and MiSLAS (Zhong et al., 2021). Our comparison also includes supervised contrastive learning approaches, namely FCL (Kang et al., 2020), KCL (Kang et al., 2020), and TSC (Li et al., 2022). In addition, the comparison involves $\mathcal{NC}$-inspired methods such as ETF classifier+DR (Yang et al., 2022) and ARB-Loss (Xie et al., 2023).

Implementation details. We mainly follow the common training protocol. In all experiments, we adopt SGD optimizer with the momentum of 0.9, weight decay of 0.005,

and train the model for 200 epochs following Alshammari et al. (2022). We utilize mix-up (Zhang et al., 2018) during the representation learning stage for all datasets. For CIFAR10/100(-LT), we use ResNet-32 (He et al., 2016) as the backbone and a multi-step schedule that decays the learning rate as its 0.1 at the 160-th and 180-th epochs with initialization of 0.1. We use 4 GeForce GTX 2080Ti GPUs with a batch size of 128. For ImageNet-LT, we use ResNeXt-50 (Xie et al., 2017) as the backbone and cosine schedule that gradually decays the learning rate from 0.05 to 0. We use 4 Tesla V100 GPUs to train the models with a batch size of 256. We also adopt Randaugment (Cubuk et al., 2020) for ImageNet-LT. We report the average results of three independent trials with different random seeds. Our code is available at https://github.com/Pepper-lll/NCfeature.

The hyperparameters λ_1 and λ_2 need to be adjusted according to the complexity of the datasets. In general, simple datasets with few categories require a small magnitude of feature regularization, while for complex datasets with plenty of categories, we need larger λ_1 and λ_2. Besides, similar to Li et al. (2022), we also find that it is better to regularize the feature learning from half of the training process for large-scale datasets, i.e., CIFAR100 and ImageNet-LT. Our hyperparameter settings and the epoch number to start feature regularization are summarized in Table 1.

The class centers $\{\boldsymbol{\mu}_k\}_{k=1}^K$ are updated in each mini-batch, instead of in the entire training set, which has been proved not efficient in large-scale datasets (Wen et al., 2016). Besides, our regularization terms are better to combine with re-balancing strategies to ensure the matching between classifier weights and class centers. The combination can lead to a remarkable improvement. In our experiments, we choose DRW and cRT as the re-balancing strategies.

Table 1: Hyperparameter setting.

Dateset	λ_1	λ_2	start epoch
CIFAR10(-LT)	0.01	0.1	0
CIFAR100(-LT)	0.01	0.5	100
ImageNet-LT	0.05	1.0	100

4.1.2 Results

Balanced data. As we mentioned before, our method is applicable to both balanced and imbalanced datasets. First, we conduct experiments to validate our model on balanced CIFAR10 and CIFAR100 datasets. Table 2 shows that our method can reduce the generalization error with both CE and MSE loss.

Imbalanced data. Table 3 and 4 present our results on CIFAR10-LT, CIFAR100-LT, and ImageNet-LT. We can find that our method surpasses existing methods on all three datasets. For ImageNet-LT, we further test the accuracy on three groups of classes according to the sample size, including Many-shot (>100 samples), Medium-shot (20∼100 samples), and Few-shot (<20 samples). The results show that our method can substantially improve the accuracy of the Medium- and Few-shot categories with almost no impact on the accuracy of the Many-shot categories compared to the plain training with CE.

Table 2: Top-1 test accuracy (%) on the balanced datasets.

Method	CIFAR10	CIFAR100
CE	**93.4**	71.8
+NC	93.3	**72.1**
MSE	91.1	70.7
+NC	**91.7**	**71.9**

Table 3: Top-1 test accuracy (%) on CIFAR10-LT and CIFAR100-LT. The results of the compared methods are obtained from their respective original papers. The best and second-best results are marked in bold and underlined.

Method	CIFAR10-LT		CIFAR100-LT	
imbalance ratio	100	10	100	10
CE	70.4	86.4	38.4	55.7
CE-RS	72.8	87.8	36.7	57.7
CE-RW	74.4	87.9	32.5	58.2
CE-DRW	75.1	86.4	42.5	56.2
LDAM-DRW	77.0	88.2	43.5	58.7
BBNm	79.9	88.4	42.6	59.2
MiSLAS	82.1	90.0	47.0	<u>63.2</u>
KCL	77.6	88.0	42.8	57.6
TSC	79.7	88.7	43.8	59.0
ETF classifier+ DR	76.5	87.7	45.3	-
ARB-Loss	**83.3**	**90.2**	47.2	62.1
NC-DRW	81.9	<u>89.8</u>	<u>48.6</u>	63.1
NC-DRW-cRT	<u>82.6</u>	**90.2**	**48.7**	**63.6**

Combine with existing approaches. Our regularization terms can be easily plugged into most of the existing algorithms. To validate the effectiveness, in Table 5, we add the proposed regularization terms to three different types of algorithms. We follow their original experiment settings to compare the performance differences before and after adding regularization terms. The results show that our regularization terms can increase the accuracy in all three algorithms.

4.2 Discussions

In this section, to verify the correctness and further explore the properties of our method, we show the learned representations, performance robustness, and ablation study on various combinations of loss and regularizations.

4.2.1 Representation Analysis

We extensively analyze the representations learned with our method to explain the advantages relative to the baseline.

Table 4: Top-1 test accuracy (%) on ImageNet-LT.

Methods	Many	Medium	Few	All
CE	**68.2**	38.1	5.82	45.3
CE-RS	64.6	42.6	17.8	47.8
CE-RW	52.0	41.4	19.8	42.5
CE-DRW	52.6	45.7	31.5	46.4
CE-cRT	58.8	33.0	26.1	47.3
CE-LWS	57.1	45.2	29.3	47.7
MiSLAS	61.7	**51.3**	**35.8**	52.7
FCL	61.4	47.0	28.2	49.8
KCL	62.4	49.0	29.5	51.5
TSC	63.5	49.7	30.4	52.4
ETF classifier+ DR	-	-	-	44.7
ARB-Loss	60.2	51.8	38.3	52.8
NC-DRW	67.1	49.7	29.0	53.6
NC-DRW-cRT	65.6	51.2	35.4	**54.2**

Table 5: Top-1 test accuracy (%) on real-world long-tail datasets of our methods combined with others. Note that we replicated experiments of RIDE with data distributed parallel training and got results with slight differences from Wang et al. (2021). c10, c100 and iNet are short for CIFAR10, CIFAR100 and ImageNet respectively.

Method	c10-LT	c100-LT	iNet-LT
LDAM-DRW	77.0	42.0	48.8
+NC	77.1(0.1$^{\uparrow}$)	43.2(1.2$^{\uparrow}$)	49.5(0.7$^{\uparrow}$)
Logit Adjust	77.4	43.9	51.1
+NC	78.8(1.4$^{\uparrow}$)	44.6(0.7$^{\uparrow}$)	53.2(2.1$^{\uparrow}$)
RIDE (2 experts)	-	46.5	51.9
RIDE (3 experts)	-	47.5	54.2
RIDE (4 experts)	-	48.8	55.2
+NC (2 experts)	-	46.8(0.3$^{\uparrow}$)	52.2(0.3$^{\uparrow}$)
+NC (3 experts)	-	48.1(0.6$^{\uparrow}$)	54.8(0.6$^{\uparrow}$)
+NC (4 experts)	-	49.1(0.3$^{\uparrow}$)	56.0(0.8$^{\uparrow}$)

As for the corresponding analysis of classifiers, we obtained consistent findings with previous studies (Kang et al., 2019) and therefore do not repeat them here.

Maximally separated class centers. We compare the pair-wise angles of the centered class means learned on CIFAR10-LT with vanilla training, re-sampling (RS), re-weighting (RW), and the proposed regularization terms in Figure 2. We arrange the class indexes in descending order based on their sizes. Under a long-tailed distribution, the minority class centers move closer to the majority with plain model. In Figure 2(a), the angles between class 8 and 0, class 9 and 1, and class 5 and 3 are around $50°$ which is far lower than the optimal angle of $96°$. RS and RW can assist in the acquisition of more distinguishable features, as demonstrated Figure 2(b) and 2(c)). However, with our regularization terms (Figure 2(d)), we can observe that the pair-wise angles between all the class centers remain consistently close to the optimum value. In addition, the significant improvement on the experimental results indicates that the features learned by our method are more generalizable.

Zero-centered class means with the equal norm. Although neither $\mathcal{L}_{\mathcal{W}}$ nor $\mathcal{L}_B$ forces the class center to be of equal norm, we can observe it in our experiments, as shown in Figure 3. This result strongly indicates that we can successfully induce $\mathcal{NC}$ in imbalanced data.

4.2.2 Robustness

We test the robustness of our method against random noise with different neural networks on CIFAR10/100 and their long-tailed version where the imbalance ratio $r = 100$. Here Resnet-32 and ResNet-18 are employed. ResNet-18 is a wider network with the last-layer feature dimension of 512, while ResNet-32 is 64. The models are all trained with DRW. The results are reported in Table 6. We can observe that our regularization terms can improve the robustness for different model capacities.

4.2.3 Ablation Study

We conduct experiments to examine the effectiveness of two regularization terms separately over CE loss and the comparison with *center loss*. The results, presented in Table 7, demonstrate that each term can significantly improve accuracy individually, and that their combination produces the best results. Meanwhile, $\mathcal{L}_W$ consistently produces better results than *center loss*, suggesting that modifying the coefficient is crucial. We can also find that $\mathcal{NC}_2$ property is more useful, implying the importance of sufficiently distant class centers for the long-tail recognition task.

5 RELATED WORK

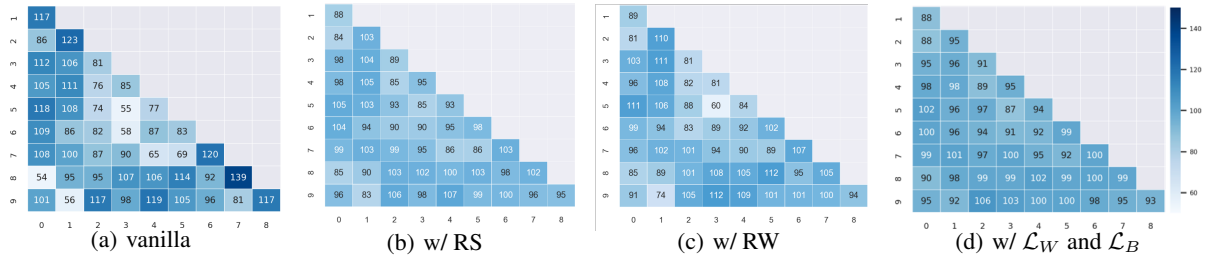
5.1 Long-tailed recognition

Long-tailed distribution is ubiquitous in the real world, which brings big challenges for most deep learning models. Classical methods dealing with this problem include data re-sampling and loss re-weighting. The former refers to re-sampling the instances to achieve relatively balanced training data, basically including over-sampling (Ando & Huang, 2017; Shelke et al., 2017), under-sampling (Shelke et al., 2017), and class-balanced sampling (Cui et al., 2019). Instead of changing the original data distribution, loss re-weighting uses cost-sensitive re-weighting strategies and assigns different weights to instances from different classes according to the sample sizes (Lin et al., 2017; Cui et al., 2019). However, although the re-sampling and re-weighting approaches can improve the performance of minority classes, they may lead to overfitting (Li et al., 2022) and hurt the representation learning (Kang et al., 2019).

Recent works also focus on representation learning under long-tailed data distribution. This stream of study mainly

Table 6: Random Noise Robustness Results. CE‡ denotes Cross-Entropy loss with the feature regularization $\mathcal{L}_W + \mathcal{L}_B$.

Gaussian noise std		0.00	0.10	0.20	0.30	0.40	0.00	0.10	0.20	0.30	0.40
Dataset	Loss	ResNet-32					ResNet-18				
CIFAR10	CE	**93.3**	89.4	76.6	54.7	35.7	94.9	91.5	**79.2**	56.9	35.9
	CE‡	93.1	**89.6**	**76.6**	**57.5**	**39.0**	**95.1**	**91.8**	78.9	**57.1**	**37.4**
CIFAR10-LT	CE	77.0	75.0	62.3	45.3	32.5	79.2	75.7	63.5	47.2	33.8
	CE‡	**79.2**	**75.5**	**63.2**	**47.8**	**35.5**	**81.0**	**77.5**	**66.5**	**51.4**	**37.9**
CIFAR100	CE	71.8	**60.2**	**40.0**	23.9	14.1	78.2	65.9	44.2	25.0	14.3
	CE‡	**72.3**	59.0	39.7	23.9	**15.1**	**78.6**	**67.8**	**46.9**	**28.2**	**16.6**
CIFAR100-LT	CE	42.5	37.2	25.4	16.3	10.2	46.8	41.1	31.6	**23.6**	**17.6**
	CE‡	**45.7**	**39.0**	**27.2**	**16.8**	**11.1**	**47.2**	**41.6**	31.6	23.3	16.3



(a) vanilla (b) w/ RS (c) w/ RW (d) w/ $\mathcal{L}_W$ and $\mathcal{L}_B$

Figure 2: Pair-wise angle degree between centered class means trained on CIFAR10-LT. Note that the optimal pair-wise angle for 10 classes is $\arccos \frac{-1}{10-1} \approx 96.4°$.

Table 7: **Ablation studies** on the effectiveness of each regularization term on CIFAR10/100-LT. Note that we apply DRW for all experiments here.

Method	CIFAR10-LT		CIFAR100-LT	
imbalance ratio	100	10	100	10
CE	75.1	86.4	42.4	56.2
+Centor Loss	78.7	89.1	46.3	61.2
+$\mathcal{L}_W$	79.1	88.1	46.9	61.3
+$\mathcal{L}_B$	80.1	88.6	47.6	61.7
+Centor Loss &$\mathcal{L}_B$	77.5	89.2	46.5	61.4
+$\mathcal{L}_W$&$\mathcal{L}_B$	**81.9**	**89.8**	**48.6**	**63.1**

(a) CIFAR10(-LT) (b) CIFAR100(-LT)

Figure 3: The norm of centered class means on a balanced dataset, long-tailed dataset w/ and w/o inducing NC is represented in different colors. Note that the class index is inversely sorted by the sample size.

follows a two-stage training scheme that decouples the representation and classifier learning (Kang et al., 2019; Zhong et al., 2021; Li et al., 2022; Kang et al., 2020; Zhu et al., 2022). Kang et al. (2019) observed that a high-quality representation requires fully utilizing the training instances equally, while a re-balancing technique is crucial for an unbiased classifier. On the other hand, some works take advantage of the superior representation learning ability of contrastive loss to extract the feature for deep long-tailed learning; then train a classifier upon the feature extractor with cost-sensitive loss or class-balanced sampling (Kang et al., 2020; Li et al., 2022; Zhu et al., 2022). Supervised contrastive learning shows superiority in representation learning under imbalanced distribution and achieves SOTA for long-tailed recognition tasks (Li et al., 2022; Zhu et al., 2022; Cui et al., 2022). However, these methods usually converge slowly and require complex network structures compared to traditional supervised learning.

Researchers also explored methods based on the ensemble. They usually utilize multiple models over different data distributions (Wang et al., 2021) or perform representation learning and classifier training with separate branches (Zhou et al., 2020; Zhu et al., 2022). This kind of approach is generally considered to be orthogonal to the single-model approach described above.

5.2 Neural collapse

A recent study (Papyan et al., 2020) discovered the phenomenon named *Neural Collapse* ($\mathcal{NC}$), stating that the last-layer embedding and classifiers will converge to a sym-

metric geometry named simplex *Equiangular Tight Frame* (ETF) for deep classifiers trained on balanced data. A more precise description of the $\mathcal{NC}$ phenomena is delivered in Section 2.1. Subsequent studies indicate that $\mathcal{NC}$ will eventually occur, independent of the loss function, the optimizer, batch-normalization, and regularization, as long as the training data exhibits a balanced distribution (Zhu et al., 2021; Han et al., 2022; Kothapalli et al., 2022). Meanwhile, the intrinsic merit of $\mathcal{NC}$ has also been revealed, including ensuring global optimality, stronger generalization and robustness, and transferability (Papyan et al., 2020; Zhu et al., 2021; Galanti et al., 2021).

The investigation of $\mathcal{NC}$ has also been carried over to the imbalanced data case, where different phenomena are uncovered. Fang et al. (2021) demonstrated that the minority classifiers have smaller pair-wise angles than the majority ones and will even merge together as the imbalance level increases, named *Minority Collapse*. This phenomenon provides some reason of the performance drop. Thrampoulidis et al. (2022) provides a general frame that is equivalent to ETF for balanced data, and reveals an asymmetric geometry of the last-layer feature and classifiers for imbalanced distribution. Furthermore, the perfect alignment between the class feature means and classifiers vanished under the imbalanced distribution. However, Thrampoulidis et al. (2022) illustrates the general geometry with a special encoding framework and does not discuss whether this geometry with an imbalanced dataset has merit or defect.

Inspired by the $\mathcal{NC}$ phenomenon, some researchers have attempted to improve the model's classification ability encountering imbalanced distribution by eliminating *Minority Collapse*, including fixing the classifier as an ETF (Yang et al., 2022) and adjusting the CE loss (Xie et al., 2023). Distinct from these works, our work analyzes that obtaining high-quality features is the key to the improvement and thus proposes regularization to guide learning representations.

6 CONCLUSIONS

In this paper, we argue that the existence of $\mathcal{NC}$ is crucial for long-tailed recognition and propose two simple but effective regularization terms to induce the appearance of $\mathcal{NC}$. We empirically show that under the imbalanced data distribution, the class centers of minority classes are close to the majority ones, leading to the overlap among different classes over the feature space and confusion of the classifier. With our method, the deep classification models are able to learn *compact within-class* and maximally *distinct between-class* features. Extensive experiments confirm that our method can enhance the generalization power of the deep classification model, especially when the training set is imbalanced. Our method is more efficient than contrastive loss based methods, and we set new state-of-the-art performance for single model based methods on widely used benchmarks.

Our proposed regularization guides the representation learning to be of 'optimal' geometry for classification, which is particularly beneficial for training sets with imbalanced labels. However, the learned geometry is validated empirically and lacks complete theoretical guarantees, leading to manually tuning the related hyperparameters. In the future, we plan to formally analyze the geometry obtained with our regularization and provide some theoretical justification for the choice of hyperparameters.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China / Research Grants Council (NSFC/RGC) Joint Research Scheme Grant N_HKUST635/20, Hong Kong Research Grant Council (HKRGC) Grant 16308321, ITF UIM/390, as well as awards from Smale Institute of Mathematics of Computation. This research made use of the computing resources of the X-GPU cluster supported by the HKRGC Collaborative Research Fund C6021-19EF.

References

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

Ando, S. and Huang, C. Y. Deep over-sampling framework for classifying imbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 770–785. Springer, 2017.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

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

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

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

Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. Parametric contrastive learning. In *Proceedings of the IEEE/CVF*

international conference on computer vision, pp. 715–724, 2021.

Cui, J., Zhong, Z., Tian, Z., Liu, S., Yu, B., and Jia, J. Generalized parametric contrastive learning. *arXiv preprint arXiv:2209.12400*, 2022.

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

Demirkaya, A., Chen, J., and Oymak, S. Exploring the role of loss functions in multiclass classification. In *2020 54th annual conference on information sciences and systems (ciss)*, pp. 1–5. IEEE, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Endres, D. M. and Schindelin, J. E. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

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(43):e2103091118, 2021.

Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Galanti, T., György, A., and Hutter, M. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2021.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. S. Domain-adversarial training of neural networks. In Csurka, G. (ed.), *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pp. 189–209. Springer, 2017. doi: 10.1007/978-3-319-58347-1_10.

Graf, F., Hofer, C., Niethammer, M., and Kwitt, R. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pp. 3821–3830. PMLR, 2021.

Han, X., Papyan, V., and Donoho, D. L. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022.

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*, pp. 770–778, 2016.

Hu, T., Wang, J., Wang, W., and Li, Z. Understanding square loss in training overparametrized neural network classifiers. *arXiv preprint arXiv:2112.03657*, 2021.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2019.

Kang, B., Li, Y., Xie, S., Yuan, Z., and Feng, J. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020.

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

Kothapalli, V., Rasromani, E., and Awatramani, V. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*, 2022.

Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R. S., 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*, pp. 6918–6928, 2022.

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*, pp. 2980–2988, 2017.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 97–105, Lille, France, 07–09 Jul 2015. PMLR.

Papyan, V., Han, 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(40):24652–24663, 2020.

Shelke, M. S., Deshmukh, P. R., and Shandilya, V. K. A review on imbalanced data handling using undersampling and oversampling technique. *Int. J. Recent Trends Eng. Res*, 3(4):444–449, 2017.

Thrampoulidis, C., Kini, G. R., Vakilian, V., and Behnia, T. Imbalance trouble: Revisiting neural-collapse geometry. *arXiv preprint arXiv:2208.05512*, 2022.

Van Horn, G. and Perona, P. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

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

Wang, Z., Xiang, C., Zou, W., and Xu, C. Mma regularization: Decorrelating weights of neural networks by maximizing the minimal angles. *Advances in Neural Information Processing Systems*, 33:19099–19110, 2020.

Webb, A. R. and Lowe, D. The optimised internal representation of multilayer classifier networks performs nonlinear discriminant analysis. *Neural Networks*, 3(4): 367–375, 1990.

Weiss, G. M., McCarthy, K., and Zabar, B. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? In *Proceedings of the 2007 International Conference on Data Mining (DMIN)*, volume 7, pp. 35–41, Las Vegas, Nevada, USA, 2007.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.

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

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Yang, Y., Xie, L., Chen, S., Li, X., Lin, Z., and Tao, D. Do we really need a learnable classifier at the end of deep neural network? *arXiv preprint arXiv:2203.09081*, 2022.

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

Zhao, H., Des Combes, R. T., Zhang, K., and Gordon, G. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pp. 7523–7532. PMLR, 2019.

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*, pp. 16489–16498, 2021.

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*, pp. 9719–9728, 2020.

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

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.