Space Alignment Matters: The Missing Piece for Inducing Neural Collapse in Long-Tailed Learning

Jinping Wang, Zhiqiang Gao*, Zhiwu Xie

College of Science, Mathematics and Technology, Wenzhou-Kean University
{1306325, zgao, zxie} @wku.edu.cn

Abstract

Recent studies on Neural Collapse (NC) reveal that, under class-balanced conditions, the class feature means and classifier weights spontaneously align into a simplex equiangular tight frame (ETF). In long-tailed regimes, however, severe sample imbalance tends to prevent the emergence of the NC phenomenon, resulting in poor generalization performance. Current efforts predominantly seek to recover the ETF geometry by imposing constraints on features or classifier weights, yet overlook a critical problem: There is a pronounced misalignment between the feature and the classifier weight spaces. In this paper, we theoretically quantify the harm of such misalignment through an optimal error exponent analysis. Built on this insight, we propose three explicit alignment strategies that plug-and-play into existing long-tail methods without architectural change. Extensive experiments on the CIFAR-10-LT, CIFAR-100-LT, and ImageNet-LT datasets consistently boost examined baselines and achieve the state-of-the-art performances.

Introduction

Long-tailed learning refers to scenarios where a few head classes dominate the dataset, while numerous tail classes have scarce examples. This distribution poses a significant challenge for neural networks, often resulting in suboptimal feature learning, biased predictions toward head classes, and poor generalization on minority classes (Fang et al. 2021). In contrast, when trained on balanced data, as shown in Figure 1(b), deep models exhibit the phenomenon of *Neural Collapse* (NC) (Papyan, Han, and Donoho 2020), where the learned features and classifier converge to a highly symmetric structure (Figure 1(b)). Specifically, NC includes four key properties: (NC1) features from the same class collapse to their class mean; (NC2) class means are maximally separated and form a Simplex Equiangular Tight Frame (ETF); (NC3) feature means and classifier weights are mutually aligned as mirror images; and (NC4) classification reduces to a nearest-center decision rule. This structure is theoretically optimal for linear separability, minimizing intra-class variance and maximizing inter-class margin, and empirically correlates with low test error under balanced settings.

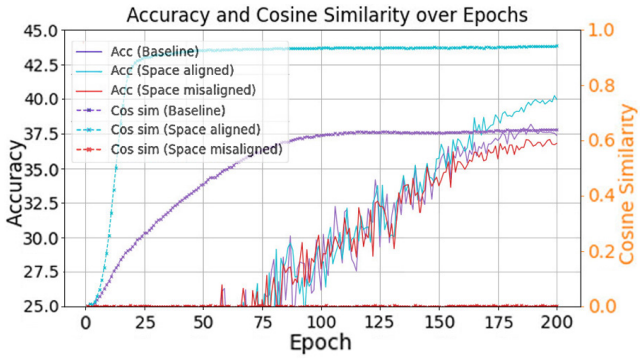
*Corresponding author: zgao@wku.edu.cn.

However, under long-tail distributions, the NC structure is disrupted, leading to a common failure mode termed *Minority Collapse*, where classifier weights for tail classes degenerate to nearly identical directions, inducing severe misclassification. As highlighted by (Yang et al. 2022), this is due to the imbalance in gradient contributions: majority classes dominate both attraction (intra-class compactness) and repulsion (inter-class separation) terms, while minority classes contribute negligibly. Consequently, classifier weights of tail classes are overwhelmed by repulsion from head classes, and their updates deviate from the NC geometry. The resulting biased decision boundaries favor majority classes and suppress minority accuracy.

To mitigate this, recent studies have attempted to reconstruct the ETF structure (NC1 and NC2), either by manually producing classifier weights (Yang et al. 2022), class prototypes (Zhu et al. 2022), handcrafted ETF structure (Gao et al. 2024), or reshaping the feature space into a more stable configuration (Peifeng et al. 2023; Xie et al. 2023; Liu et al. 2023). However, these approaches overlook an important question: *To what extent existing methods achieve feature-classifier alignment (NC3), and whether explicitly promoting this alignment is beneficial?* Intuitively, enforcing this alignment is helpful for further inducing NC in long-tail learning, which will encourage a better representation learning and provide a extra performance improvement.

To answer this, we study the space misalignment between feature and classifier vectors in long-tailed settings. As visualized in Figure 1(c), models trained with cross-entropy loss fail to achieve alignment, in contrast to the balanced case in Figure 1(b). We further observe that the similarity between these spaces correlates strongly with performance: Compared to baseline (purple curve), applying a constraint that further reduces this alignment (red curve) leads to consistent performance degradation. This suggests that space misalignment is not merely a by-product, but a core obstacle to robust representation learning under imbalance.

To formally analyze this, we introduce a geometric framework based on the Optimal Error Exponent (OEE) which is a classical information-theoretic measure that quantifies how quickly misclassification probability decays as the noise level decreases. We show that angular misalignment between feature and classifier directions provably slows convergence and deteriorates generalization. This analysis pro-



(a)

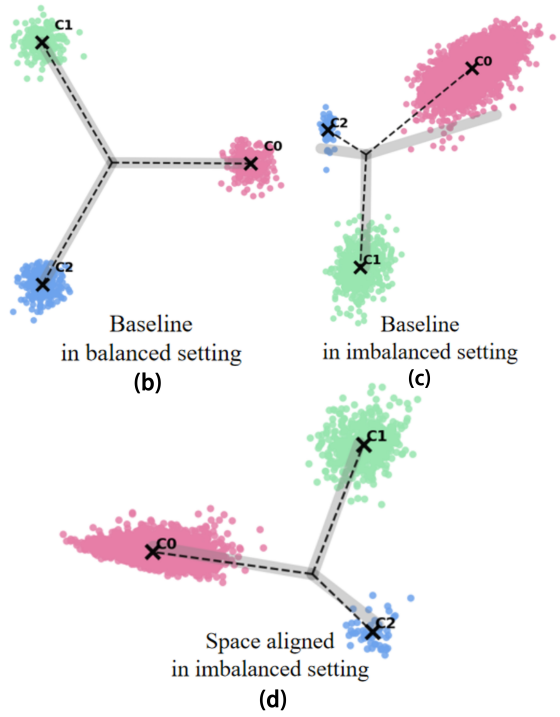


Figure 1: Space misalignment issue. (a) shows ResNet-32 model performances and corresponding cosine similarities between class feature means and classifier weights on the CIFAR-100 dataset. The baseline, aligned, and misaligned models are obtained by training with cross-entropy (CE), SpA-Reg (proposed method for alignment), and negative SpA-Reg loss, respectively. (b), (c) and (d) are toy examples with 2-dimensional features and 3 classes, which illustrate different geometric structures of the class feature means (black crosses) and the classifier weights (gray lines).

vides theoretical insight suggesting that even moderate misalignment can significantly deteriorate the generalization performance of the model. Motivated by this, we propose three plug-and-play strategies to reduce space misalignment in standard long-tailed learning setups. As shown in Figure 1(d), our regularization leads to high alignment, and consistently improves both space similarity and classification accuracy (blue line in Figure 1(a)). Through extensive experiments on benchmarks, our approach achieves state-of-the-art performance while recovering stronger NC properties in both feature and decision spaces.

Preliminaries

The training set consists of C classes, where the given dataset is balanced; each class contains n samples and can be denoted as $\{(x_{i,c}, y_{i,c})\}$. Here, $x_{i,c} \in \mathbb{R}^d$ denotes the i_{th} input sample of class c and $y_{i,c} = c$ denotes its real label. The model is composed of a deep neural network. We can consider the layers before the classifier, which act as a feature extractor as a mapping h: $\mathbb{R}^d \to \mathbb{R}^p$ that outputs a p-dimensional feature vector $h(x)$. Following this, a linear classifier with weight matrix $\mathbf{W} \in \mathbb{R}^{C\times p}$ and biases $b \in \mathbb{R}^C$ takes the last-layer features as inputs and then outputs the class label. In detail, through classification scores via $f(x) = \mathbf{W}h(x) + b$, the predicted label is then given by $argmax_{c'}\langle w_{c'}, h\rangle + b_{c'}$, where $w_{c'}$ denotes the classifier weight for a specific class. Furthermore, we denote the class mean $\mu_c = \frac{1}{n}\sum_{i=1}^{n} h(x_{i,c})$ and the global mean $\mu_G = \frac{1}{C}\sum_{c=1}^{C} \mu_c$.

Simplex Equiangular Tight Frame (Simplex ETF)

A general Simplex Equiangular Tight Frame (Simplex ETF) matrix $\mathbf{M} \in \mathbb{R}^{p\times C}$ is a collection of points in $\mathbb{R}^p$ specified by the columns of:

$$\mathbf{M} = \sqrt{\frac{C}{C-1}}\,\mathbf{R}\left(\mathbf{I} - \frac{1}{C-1}\mathbf{1}_C\mathbf{1}_C^\top\right), \tag{1}$$

where $\mathbf{I} \in \mathbb{R}^{C\times C}$ is the identity matrix and $\mathbf{1}_C \in \mathbb{R}^{C\times 1}$ is the ones vector, and $\mathbf{R} \in \mathbb{R}^{p\times C}(p \geq C)$ is a rotation orthogonal matrix ($\mathbf{R}^\top\mathbf{R} = \mathbf{I}$). $\mathbf{M} := \{m_1, m_2, m_3..., m_C\} \in \mathbb{R}^{p\times C}$includes C classes with the m_c weight.

Neural Collapse

In the terminal phase of training on balanced datasets, as shown in Figure 1(b), it can be observed that the last-layer features will converge to class means, which in turn align with classifier weights, all forming the vertices of a symmetric simplex ETF (Papyan, Han, and Donoho 2020). Specifically, Neural Collapse (NC) can be formally described within 4 phases as follows:

(NC1) Within-class variability collapse As training progresses, the features belonging to the same class collapse to their class means. Mathematically, this means the within-class covariance matrix Σ_W approaches zero, that is:

$$\Sigma_W = \frac{1}{Cn}\sum_{c=1}^{C}\sum_{i=1}^{n}\left(h(x_{i,c}) - \mu_c\right)\left(h(x_{i,c}) - \mu_c\right)^\top,$$
$$\Sigma_W \to 0, \tag{2}$$

(NC2) Convergence to a simplex ETF The mean vectors of each class converge to a simplex ETF:

$$\|\mu_c - \mu_G\|_2 - \|\mu_{c'} - \mu_G\|_2 \to 0, \quad \forall c, c'$$
$$\langle \hat{\mu}_c, \hat{\mu}_{c'}\rangle \to \frac{C}{C-1}\delta_{c,c'} - \frac{1}{C-1}, \quad \forall c, c' \tag{3}$$

where $\hat{\mu_c} = (\mu_c - \mu_G)/\|\mu_c - \mu_G\|_2$ denotes the renormalized class means and $\delta_{c,c'}$ is the Kronecker delta symbol.

(NC3) Convergence to self-duality The classifier weight vectors w_c become aligned with the centered class mean vectors $(\mu_c - \mu_G)$, reinforcing the global simplex ETF structure in both feature and decision spaces:

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

where $\dot{M} = [\mu_c - \mu_G, c = 1, ..., C] \in \mathbb{R}^{p \times C}$.

(NC4) Simplification to nearest center The neural network classifier converges to a nearest class mean:

$$\arg\max_{c'} \langle w_{c'}, h \rangle + b_{c'} \;\to\; \arg\min_{c'} \|h - \mu_{c'}\|_2. \tag{5}$$

Space Misalignment Under Long-tail Setting

The original theory (Papyan, Han, and Donoho 2020) of NC identified NC1 and NC2 as prerequisites for achieving alignment between the feature space and classifier vector space (NC3). However, in long-tail settings, the imbalanced sample sizes across classes induce skewed gradient signals, causing the phenomenon known as Minority Collapse (Fang et al. 2021), which damages the ETF geometric structure. To address this challenge, existing methods have made notable progress in restoring the simplex structure. Nevertheless, as illustrated in Figure 4, we still observe space misalignment in these studies, i.e., low cosine similarity between feature means and classifier weights. This motivates us to investigate the detrimental effects of space misalignment, even if the decision and feature spaces approximately approach the ideal ETF structure.

Optimal Error Exponent Under Perfect Alignment

Setting To focus on and better quantify the harm that space misalignment might cause, following the settings of (Papyan, Han, and Donoho 2020), we construct our problem under an idealized condition where feature and classifier vector spaces already both converge to the simplex ETF. Assume we are given an observation: $h = \mu_c + z \in \mathbb{R}^C; z \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and $c \sim \{1, ..., C\}$ denotes the unknown class index, independently distributed from z. We use a linear classifier, $\mathbf{W}h(x) + \mathbf{b}$ where weights $\mathbf{W} = [w_c : c = 1, ..., C] \in \mathbb{R}^{C \times C}$ and biases $\mathbf{b} = b_c \in \mathbb{R}^C$. Our decision rule is $\hat{\gamma}(h) = \hat{\gamma}(h; \mathbf{W}, \mathbf{b}) = \arg\max_c \langle w_c, h \rangle + b_c$.

Optimal Error Exponent (OEE) Following (Papyan, Han, and Donoho 2020), to quantify classification performance theoretically, we consider the large-deviations error exponent. This exponent shows how quickly the misclassification probability decays as conditions become ideal (e.g., as noise or classification difficulty is reduced). It is worth noting that in classification tasks, the training process drives the training loss to almost zero. Moreover, according to the research in (Dang et al. 2024), NC1 (within-class variance collapse) occurs in both of these situations. Therefore, here, leveraging the large-deviation theory is appropriate to analyze misclassification error and explain poor generalization performance.

Theorem 1 (Large-Deviations Error Exponent)**.** *If modeling classification with a bit of noise, the error exponent is defined as:*

$$\beta(\mathbf{M}, \mathbf{W}, b) = \lim_{\sigma \to 0} -\sigma^2 \log P_\sigma \{\hat{\gamma}(\mathbf{h}) \neq \gamma\} \tag{6}$$

where σ represents the noise level.

For very small noise, the misclassification probability P_σ typically behaves like $\exp(-\beta/\sigma^2)$; *a larger β means the error probability decreases* ***exponentially faster*** *as the task gets easier* (noise $\sigma \to 0$). The error exponent thus captures the geometric separability of classes: if classes are well separated in feature space, the error of an ideal classifier will decrease rapidly (high β), while if some classes are too close (misaligned), the error decreases slowly (low β). Furthermore, as shown in Theorem 2 proved by (Papyan, Han, and Donoho 2020), the maximum possible error exponent (over all possible arrangements of class feature means in a given dimension) is achieved when the class means are arranged as a centered simplex ETF (which essentially represents the most symmetric configuration).

Theorem 2 (Optimal Error Exponent (OEE))**.** *When both class means and classifier weights form a renormalized Simplex ETF, with the classifier bias being* 0*:*

$$\begin{aligned} \beta^* &= \max_{\mathbf{M}, \mathbf{W}, \mathbf{b}} \beta(\mathbf{M}, \mathbf{W}, \mathbf{b}) \quad s.t. \quad \|\mu_c\|_2 \leq 1 \quad \forall c, \\ \beta^* &= \beta(\mathbf{M}^*, \mathbf{M}^*, 0), \\ \beta^* &= \frac{C}{C-1} \cdot \frac{1}{4}, \end{aligned} \tag{7}$$

where $\mathbf{M}^*$ is a $C \times C$ matrix constructed by renormalized class mean $\hat{\mu_c}$ where $\hat{\mu_c} = (\mu_c - \mu_G)/\|\mu_c - \mu_G\|_2$.

Impact of Misalignment

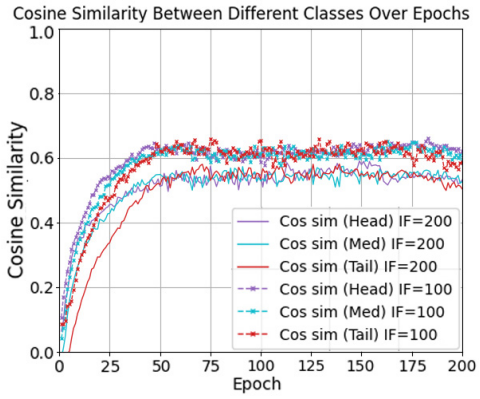


Figure 2: Cosine similarity between the feature space and the decision space across d classes under two imbalance factors (IF=100 and IF=200).

Setting To quantify the influence of space misalignment, we follow the conventional setting in the original NC theory (Papyan, Han, and Donoho 2020), where both the feature and weight vector space converge to the Simplex ETF, i.e.,

$\|\hat{\mu_c}\|_2 = \|\hat{w_c}\|_2 = 1; \langle \hat{\mu_c}, \hat{\mu_{c'}} \rangle = \langle \hat{w_c}, \hat{w_{c'}} \rangle = -\frac{1}{C-1}, (c \neq c')$. Moreover, we consider a simplified but insightful case: *uniform angular misalignment*. This setting assumes that each classifier weight vector w_c forms a fixed angle $\alpha \in [0, \frac{\pi}{2}]$ with its corresponding feature mean μ_c such that

$$\langle \hat{\mu_c}, \hat{w_c} \rangle = \cos\alpha, \forall c.$$

That is, the classifier weights can be obtained by a single isoclinic rotation R ($R = \cos\alpha\, I + \sin\alpha\, A$ with $A^\top = -A$, $A^2 = -I$. We discuss an even feature dimension here to guarantee the existence of a real orthogonal A here; Odd-dimension features can always be zero-padded to satisfy this condition without altering the analysis results.) from the unit-norm feature simplex. Actually, this uniform angular misalignment can also be observed from an empirical observation as shown in Figure 2, where the misalignment angles across head, medium, and tail classes are approximately consistent as training progresses.

Denoting $v = \hat{w_c} - \hat{w_{c'}}; d = \hat{\mu_c} - \hat{\mu_{c'}}$, we have the following Theorem 3 for a unit-norm standard simplex. Theorem 3 provides the distance and inner product expressions corresponding to the symmetric structure among the class centers, offering key constants for the explicit expression of the error exponent in subsequent steps. It can be directly applied to simplify the error comparison in both spatially aligned and unaligned scenarios.

Theorem 3 (Standard Simplex ETF Distance Properties)**.** *Let $\{\mu_c\}_{c=1}^C \subset \mathbb{R}^p$ from a standard Simplex ETF, we have:*

$$d^\top \mu_c = \frac{1}{2}\|d\|^2, \quad \|d\|^2 = \frac{2C}{C-1} \tag{8}$$

Proof Sketch In this proof, we aim to quantify how a uniform misalignment angle α affects the Optimal Error Exponent (OEE). The proof proceeds in the following key steps: **Step 1: Problem Reformulation** We first express the misclassification probability as the chance that a Gaussian perturbation z causes the inequality $\langle \hat{w_c}, h \rangle \leq \langle \hat{w_{c'}}, h \rangle$. Then, applying Lemmas 1 and 2, we convert this probability into a constrained optimization problem and get the closed form of this problem. **Step 2: Pairwise Error Exponent Derivation** We then introduce the misalignment angle α and use the geometric properties of ETF (Lemma 3) to derive the pairwise error exponent under the simplex alignment setting (Theorem 4). **Step 3: OEE Upper Bound** Finally, we derive the upper bound for the OEE under the simplex misalignment setting (Theorem 5), demonstrating that even with perfect ETF geometry, a common rotation degrades the OEE quadratically in $\cos\alpha$. **All the detailed proof contents in this part can be found in the Appendix.**

Problem Reformulation Considering the fundamental tool from the Large-Deviations Theory (Dembo 2009) in Lemma 1, we transform the task of calculating the misclassification into the Minimum-Norm optimization problem.

Lemma 1. *Suppose $\mathcal{K}$ is a closed set and $0 \notin \mathcal{K}$. Then as $\sigma \to 0$:*

$$-\sigma^2 \log P_\sigma\{z \in \mathcal{K}\} \to \min\left\{\frac{1}{2}\|z\|_2^2 : z \in \mathcal{K}\right\} \tag{9}$$

Based on Lemma 1, the probability of misclassification of class c to c' can be transformed into an optimization problem, and then obtain in a closed form.

Lemma 2. *For any two distinct classes $c \neq c'$, following the previous setting $v = \hat{w_c} - \hat{w_{c'}}; d = \hat{\mu_c} - \hat{\mu_{c'}}$, the large-deviations error exponent for misclassifying $c \to c'$ $\beta_{c,c'} = -\lim_{\sigma\to 0} \sigma^2 \log P\{\hat{\gamma}(\mathbf{h}) = c' | \gamma = c\}$ admits the closed form:*

$$\beta_{c,c'} = \frac{(v^\top \hat{\mu_c})^2}{2\|v\|^2} \tag{10}$$

Pairwise Error Exponent Derivation To analyze how space misalignment affects the error exponent, the misalignment angle setting is introduced in Lemma 3.

Lemma 3. *Let $\{\mu_j\}_{j=1}^C \subset \mathbb{R}^D$ be unit vectors forming a regular simplex, i.e.*

$$\|\mu_j\| = 1, \qquad \mu_j^\top \mu_{j'} = \begin{cases} 1, & j = j', \\ -\dfrac{1}{C-1}, & j \neq j'. \end{cases}$$

Apply a common orthogonal transform $R = \cos\alpha\, I + \sin\alpha\, A$ with $A^\top = -A$, $A^2 = -I$ to obtain the "mis-aligned" vectors $w_j = R\mu_j$ and assume the unified misalignment angle $\mu_j^\top w_j = \cos\alpha$ for every j. Then for any fixed class c and all $c' \neq c$:

$$\hat{\mu_c}^\top \hat{w_{c'}} = -\frac{\cos\alpha}{C-1} + \zeta_{c,c'}, \quad \zeta_{c,c'} = \sin\alpha\, \hat{\mu_c}^\top A \hat{\mu_{c'}}. \tag{11}$$

The residuals $\zeta_{c,c'}$ satisfy

$$\sum_{c' \neq c} \zeta_{c,c'} = 0. \tag{12}$$

Based on the result of Lemma 3 that provides the inner product expression between $\hat{\mu_c}$ and $\hat{w_{c'}}$, we can substitute these results into the closed form expression in Lemma 2 to derive the pairwise error exponent with misalignment angle α.

Theorem 4 (Error Exponent under simplex misalignment setting)**.** *For any two distinct classes $c \neq c'$, $v = \hat{w_c} - \hat{w_{c'}}$, $d = \hat{\mu}_c - \hat{\mu}_{c'}$. Then $\|d\|^2 = \|v\|^2 = 2C/(C-1)$ and the Chernoff-type error exponent is*

$$\beta_{c,c'} = \frac{C-1}{4C}\left(\cos\alpha\big(1 + \tfrac{1}{C-1}\big) - \zeta_{c,c'}\right)^2 \tag{13}$$

where $\zeta_{c,c'} = \sin\alpha\, \hat{\mu_c}^\top A \hat{\mu_{c'}}$. When the space misalignment angle is 0 (aligned), the same form of Optimal Error Exponent under perfect alignment in Theorem 2 can be obtained (Papyan, Han, and Donoho 2020).

OEE Upper Bound We aim to quantify how such misalignment influences the large-deviations error exponent β. Under this setup, the decision boundary between any two classes c and c' depends on the angle between their projected feature means, i.e., $\langle \hat{\mu_{c'}}, \hat{w_c} \rangle$. Because the classifier is misaligned, the effective projection of the feature mean onto its weight vector is shrunk by a factor of $\cos\alpha$. Therefore, the

margin between any two classes scales by $\cos\alpha$. As such, we connect the OEE between the perfect alignment and the space misalignment case by using Theorem 5, where the detailed proof can be found in the Supplementary.

Theorem 5 (Optimal Error Exponent Under Simplex Misalignment Setting)**.** *Following (Papyan, Han, and Donoho 2020), the optimal error exponent quantifies the asymptotic rate at which the misclassification probability decays as the noise level becomes relatively small. Given a space misalignment angle α, optimal error exponent is formulated as:*

$$\beta^{*'} = \min_{c' \neq c} \beta_{c,c'} \leq \frac{1}{4}\cos^2\alpha \frac{C}{C-1} = \cos^2\alpha\beta^*. \quad (14)$$

This result illustrates that even if the geometry of the feature space is perfectly optimal (e.g., a simplex ETF), angular misalignment between features and classifiers can substantially impair the theoretical error rate decay. As such, reducing the angle of misalignment is crucial to induce NC under long-tail learning.

Proposed Methods

To verify the importance of space alignment, we propose three methods for explicitly aligning the feature space and the decision space, which enable a convenient integration with current methods.

Similarity Regularization (SpA-Reg) A straightforward way is to improve the cosine similarity between the feature and the classifier vector space by applying a regularization term:

$$\mathcal{L}_{SpA-Reg} = \frac{1}{C}\sum_{c=1}^{C}[1 - cos(\hat{w_c}, \hat{\mu_c})], \quad (15)$$

which can be added to the main loss with a scaling hyperparameter λ:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \lambda \cdot \mathcal{L}_{SpA-Reg}. \quad (16)$$

A toy example is provided in Figure 3 to show the effectiveness of directly correcting the misalignment in standard long-tail learning.

Spherical Linear Interpolation (SpA-SLERP) Inspired by the spherical geometry of NC, this approach rotates the classifier weights towards their corresponding class feature mean vectors, which naturally reduces the misalignment angle. During training, given a predefined threshold τ, when the cosine similarity, $\cos sim = \langle \hat{w_c}, \hat{\mu_c} \rangle$, is below this threshold, an SLERP algorithm will execute in the following form:

$$w_c \leftarrow \|w_c\| \left(\frac{\sin((1-\alpha_t)\theta_c)}{\sin\theta_c}\hat{w}_c + \frac{\sin(\alpha_t\theta_c)}{\sin\theta_c}\hat{\mu}_c \right).$$

Here, we apply a cosine schedule to the interpolation coefficient α to realize a more flexible optimization and reduce its interference in the early stage of training:

$$\alpha_t = \alpha_{max}\frac{1-\cos(\frac{\pi t}{T})}{2},$$

where t denotes the current training epoch and T denotes the total number of epochs.

Gradient Projection (SpA-Proj) To prevent the classifier weights from deviating excessively from the mean values of the corresponding class features, we propose a gradient projection mechanism that selectively adjusts the update direction of classifier weights. After calculating the gradient $g_c = \nabla_{w_c}\mathcal{L}$, we decompose the gradient into a radial and a tangential part:

$$g_{rad} = \langle g_c, \hat{w_c} \rangle \hat{w_c} \quad g_{tan} = g_c - g_{rad}. \quad (17)$$

Then, we project the renormalized class mean $\hat{\mu_c}$ onto the same tangent space:

$$d_c = \hat{\mu_c} - \langle g_c, \hat{w_c} \rangle \hat{w_c}, \quad \hat{d_c} = \frac{d_c}{\|d_c\|}. \quad (18)$$

Moving along d_c is the steepest way to decrease the misalignment angle. Based on this, we can remove the harmful gradient component during training. Denote $p_c = \langle g_{tan}, \hat{d_c} \rangle$, we suppress the gradient component that pushes the classifier weights away from the class means:

$$g_{\tan}^{\text{safe}} = \begin{cases} g_{\tan} - p_c\hat{d}_c, & p_c > 0, \\ g_{\tan}, & p_c \leq 0. \end{cases} \quad (19)$$

Then we mix the safe direction with the original gradient to maintain the general training signal:

$$g_{final} = (1-\gamma)g_{\tan}^{\text{safe}} + \gamma g_c, \quad (20)$$

where γ is a soft projection coefficient.

Experiments

Datasets Following the standard long-tail evaluation process (Du et al. 2023; Xie et al. 2023; Gao et al. 2024; Yang et al. 2022), we use the modified long-tail version of CIFAR-10 (Cui et al. 2019), CIFAR-100 (Cui et al. 2019), and ImageNet (Deng et al. 2009) dataset. In specific, we leverage the imbalance ratio q (defined by the ratio of the samples between the most-frequent class and the rarest class in the whole dataset: $q = N_{max}/N_{min}$) and the exponential decay to create a long-tail dataset with different degrees of imbalance. For CIFAR10-LT and CIFAR-100-LT, we modify each of them with three imbalance factors {200,100,50}. For ImageNet-LT, it has an imbalance ratio of 256 with 1000 different classes. We train each model on the imbalanced training set and evaluate it in the balanced validation/test set.

Implementation All experiments are conducted on NVIDIA GeForce RTX 4090 GPUs using PyTorch. Following (Du et al. 2023; Gao et al. 2024; Xie et al. 2023), ResNet-32 (He et al. 2016) is applied on both CIFAR10-LT and CIFAR100-LT, ResNet-50 (He et al. 2016) and ResNet-50-32x4d (Xie et al. 2017) on ImageNet-LT. All models are optimized by applying the SGD optimizer with a momentum of 0.9. When our methods are incorporated into the baseline methods, all hyperparameters and learning rate scheduling strategies will follow those of the respective baselines. All experiments are repeated with three different random seeds; the reported results are the average over these runs. **More details of hyperparameters in our methods are shown in the Appendix.**

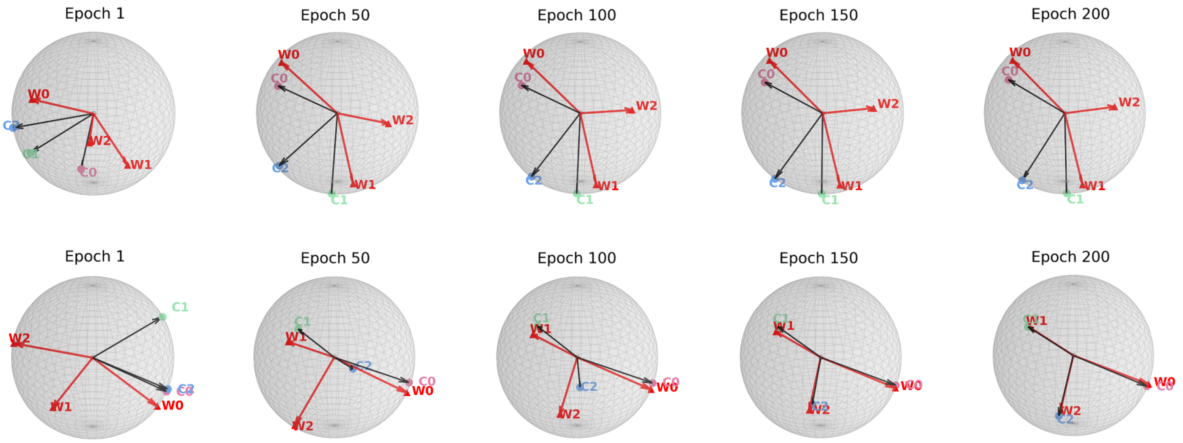


Figure 3: A toy example illustrating the process of space alignment using our proposed SpA-Reg method. Each sphere shows the change of orientations of classifier weights (red arrows) and class feature means (black arrows) as training progressed. The top row visualizes the standard long-tail learning, i.e., training with cross-entropy loss, where significant misalignment between the classifier weights and the feature center persists throughout the whole training process. In contrast, for the bottom row, the classifier weights gradually aligned with the feature mean during training.

Baselines We choose baseline methods with fundamentally different motivations. Except for standard training with cross-entropy (CE) loss, the chosen strategies include contrastive learning and data augmentation-based strategies: KCL (Kang et al. 2020), TSC (Li et al. 2022), HCL (Wang et al. 2021), GLMC (Du et al. 2023). Meanwhile, we also consider two-stage methods: BBN (Zhou et al. 2020), RIDE (Wang et al. 2020), MaxNorm (Alshammari et al. 2022). Moreover, recent approaches motivated by the Neural Collapse (NC) phenomenon, e.g., INC (Liu et al. 2023), fixed classifier as ETF (Yang et al. 2022), RBL (Peifeng et al. 2023), ARB (Xie et al. 2023), DisA (Gao et al. 2024), are also taken into consideration.

Long-tailed Benchmark Results

All results with standard deviation and **accuracies on three splits of the set of classes**: Many-shot, Medium-shot and Few-shot, **are shown in the Appendix**.

CIFAR10-LT and CIFAR100-LT Table 1 reports the accuracy of various methods on CIFAR-10-LT and CIFAR-100-LT with three imbalance ratios: 50, 100, and 200. Our approach achieves an improvement in accuracy ranging from 0.5 to 2.6. Moreover, we can observe that, compared to CIFAR-10-LT, when applying our proposed methods on CIFAR-100-LT, the baseline methods can have a higher performance gain than on CIFAR-10-LT. The main reason is that under a long-tail setting, the increase in the number of classes exacerbates the space misalignment, thereby making the effect of space alignment more significant. As shown in Figure 4, our space alignment strategies lead to high alignment during training, improving the classification performance consistently at the same time. Meanwhile, we can observe that in the later stages of training, as the loss approaches zero and the model enters the small-noise regime, the model with a higher space alignment angle can have a better generalization performance. This observation matches our theoretical analysis: the optimal error exponent β will reduce with respect to $\cos^2 \alpha$ when the model enters the large deviation regime. The experiment results confirm the validity of explicitly aligning the decision space and the feature space during training. Combining space alignment with different types of long-tail based methods can yield consistent performance improvement compared with original baselines and achieve the best performance.

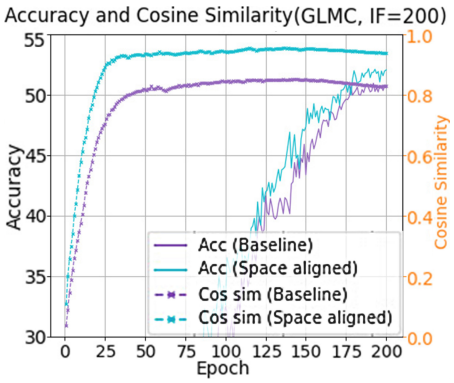


Figure 4: Accuracies and cosine similarities between class feature means and classifier weights on the CIFAR-100 dataset with the imbalance factor of 200. More results for CE, ETF-DR and ARB can be found in the Appendix.

ImageNet-LT We further conduct more experiments with different types of long-tailed classification methods on the ImageNet-LT dataset. To ensure a fair comparison in the experiment, we use the ResNet-50 following (Xie et al. 2023; Yang et al. 2022) and ResNeXt-50 following (Du et al. 2023). As the experiment results shown in Table 2, through aligning the feature space and the decision space, our meth-

Method	CIFAR-10-LT			CIFAR-100-LT		
	200	**100**	**50**	**200**	**100**	**50**
BBN	/	79.9	82.2	/	42.6	47.1
KCL	/	77.6	81.7	/	42.8	46.3
TSC	/	79.7	82.9	/	43.8	47.4
HCL	/	81.4	85.4	/	46.7	51.9
MiSLAS	/	82.1	85.7	/	47.0	52.3
RIDE (3 experts)	/	81.6	84.0	/	48.6	51.4
RBL	81.2	84.7	87.6	48.9	53.1	57.2
INC-DRW	75.8	81.9	82.7	42.5	48.6	51.7
CE*	70.1	75.4	78.3	38.5	42.1	48.1
CE*+**SpAReg**	71.9	76.3	79.3	39.9	44.3	49.2
CE*+**SpASLERP**	71.6	76.5	79.0	39.1	43.1	48.9
CE*+**SpAProj**	71.7	76.6	79.1	39.4	43.3	48.6
ETF-DR	71.9	76.5	81.0	40.9	45.3	50.4
ETF-DR+**DisA**	73.7	78.5	81.4	41.5	45.9	51.1
ETF-DR+**SpAReg**	73.0	79.0	82.3	41.6	46.3	50.9
ARB	79.6	83.3	85.7	44.5*	47.2	52.6
ARB+**SpAReg**	81.2	84.0	86.7	45.6	51.1	55.2
ARB+**SpASLERP**	80.7	83.8	86.6	44.7	49.8	54.2
ARB+**SpAProj**	81.0	84.2	86.4	45.1	49.8	54.1
GLMC	83.4*	87.8	90.2	50.8*	55.9	61.1
GLMC+MaxNorm (two-stage)	/	87.6	90.2	/	57.1	62.3
GLMC+**SpAReg**	83.9	88.5	**91.1**	52.0	58.2	63.5
GLMC+**SpASLERP**	83.9	88.8	90.9	52.1	58.0	63.3
GLMC+**SpAProj**	**84.2**	**88.9**	90.7	**52.2**	**58.3**	**63.6**

Table 1: Long-tailed classification accuracy (%) with ResNet-32 under imbalance ratios {200,100,50} on CIFAR-10-LT and CIFAR-100-LT. The methods or results marked with (*) denote the reproduced result by ourselves.

Method	All	Method	All
Backbone: ResNet-50			
KCL	51.5	LDAM-DRW	47.7
TSC	52.4	LDAM-DRW+DisA	48.5
MiSLAS	52.7	CE-DRW	47.1
CE*	44.3	CE*+**SpAReg**	44.8
CE*+**SpASLERP**	44.7	CE*+**SpAProj**	44.8
ETF-DR	44.7	ETF-DR+**SpAReg**	45.3
ARB	52.8	ARB+**SpAReg**	53.2
ARB+**SpASLERP**	53.1	ARB+**SpAProj**	**53.6**
Backbone: ResNeXt-50			
CE-DRW	46.4	GLMC	56.3
CE-LWS	47.7	GLMC+**SpAReg**	**56.7**
LADE	53.4	GLMC+**SpASLERP**	56.5
RBL	53.5	GLMC+**SpAProj**	**56.7**
INC-DRW	53.0		
INC-DRW-cRT	54.5		

Table 2: Long-tailed classification accuracy (%) on ImageNet-LT. (*) denotes our reproduced results.

ods outperform related long-tail approaches and achieve the highest performance.

Related Work

Recent works inspired by Neural Collapse (NC) have explored how to improve long-tailed learning by promoting NC geometry. One line of methods explicitly induces NC patterns via regularization. For example, (Liu et al. 2023) introduced a feature alignment term to cross-entropy loss to encourage intra-class feature collapse and inter-class orthogonality, effectively recovering NC1 and NC2 even under class imbalance. (Xie et al. 2023) analyzed minority collapse from the gradient perspective and proposed ARB-Loss, which balances attraction and repulsion forces to stabilize classifier gradients and restore the NC structure. From a representation learning angle, (Zhu et al. 2022) proposed Balanced Contrastive Learning (BCL), which uses class-balanced sampling and averaging to prevent head-class dominance in contrastive learning, pushing the feature space toward a simplex ETF.

Another line of work imposes NC-friendly geometry through fixed or structured classifiers. (Yang et al. 2022) argued that since the optimal classifier under NC is a simplex ETF, a learnable classifier may not be necessary. They fixed the classifier to a random ETF and only trained the feature extractor, leading to natural NC emergence even with imbalance. (Peifeng et al. 2023) proposed Rotated Balanced Learning (RBL), adding a learnable rotation to align the feature space with the fixed classifier. (Gao et al. 2024) proposed Distribution Alignment (DisA), using optimal transport to align features with a fixed ETF structure. This lightweight regularization improves compatibility with existing long-tail methods. (Yan et al. 2024) further extended NC by proposing Neural Collapse to Multiple Centers (NCMC), allowing each class to collapse to multiple prototypes, especially benefiting tail classes with enhanced feature diversity.

Conclusion

In this paper, we comprehensively discuss the phenomenon of space misalignment, an often overlooked problem when inducing Neural Collapse into long-tail learning. Through an analytical framework based on the Optimal Error Exponent, we quantified the detrimental effect of space alignment theoretically. Based on this theoretical insight, we proposed three plug-and-play alignment strategies that do not require architectural modifications to origin methods. Extensive experiments verify that these alignment strategies substantially enhance space similarity, achieving the state-of-the-art performances at the same time. Our findings emphasize the important role of space alignment when inducing NC to long-tail learning, offering novel perspectives for addressing data imbalance issues.

Acknowledgements

The work was partially supported by the following: WKU Internal (Faculty/Staff) Start-up Research Grant under No. ISRG2024009, WKU 2025 International Collaborative Research Program under No. ICRPSP2025001.

References

Alshammari, S.; Wang, Y.-X.; Ramanan, D.; and Kong, S. 2022. Long-Tailed Recognition via Weight Balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6897–6907.

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

Dang, H.; Tran, T.; Nguyen, T.; and Ho, N. 2024. Neural collapse for cross-entropy class-imbalanced learning with unconstrained relu feature model. *arXiv preprint arXiv:2401.02058*.

Dembo, A. 2009. *Large deviations techniques and applications*. Springer.

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

Du, F.; Yang, P.; Jia, Q.; Nan, F.; Chen, X.; and Yang, Y. 2023. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15814–15823.

Fang, C.; He, H.; Long, Q.; and Su, W. J. 2021. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43): e2103091118.

Gao, J.; Zhao, H.; dan Guo, D.; and Zha, H. 2024. Distribution alignment optimization through neural collapse for long-tailed classification. In *Forty-first International Conference on Machine Learning*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Li, T.; Cao, P.; Yuan, Y.; Fan, L.; Yang, Y.; Feris, R. S.; Indyk, P.; and Katabi, D. 2022. Targeted Supervised Contrastive Learning for Long-Tailed Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6918–6928.

Liu, X.; Zhang, J.; Hu, T.; Cao, H.; Yao, Y.; and Pan, L. 2023. Inducing neural collapse in deep long-tailed learning. In *International conference on artificial intelligence and statistics*, 11534–11544. PMLR.

Papyan, V.; Han, X.; and Donoho, D. L. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663.

Peifeng, G.; Xu, Q.; Wen, P.; Yang, Z.; Shao, H.; and Huang, Q. 2023. Feature directions matter: Long-tailed learning via rotated balanced representation. In *International Conference on Machine Learning*, 27542–27563. PMLR.

Wang, P.; Han, K.; Wei, X.-S.; Zhang, L.; and Wang, L. 2021. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 943–952.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2020. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. *CoRR*, abs/2010.01809.

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

Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yan, H.; Qian, Y.; Peng, F.; Luo, J.; Li, F.; et al. 2024. Neural collapse to multiple centers for imbalanced data. *Advances in Neural Information Processing Systems*, 37: 65583–65617.

Yang, Y.; Chen, S.; Li, X.; Xie, L.; Lin, Z.; and Tao, D. 2022. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in neural information processing systems*, 35: 37991–38002.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. 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)*.

Zhu, J.; Wang, Z.; Chen, J.; Chen, Y.-P. P.; and Jiang, Y.-G. 2022. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6908–6917.