Fixed Non-negative Orthogonal Classifier: Inducing Zero-mean Neural Collapse with Feature Dimension Separation - Supplementary Material -

A Related Work

We provided an elaborate treatment of related work.

A.1 Analyses and Extensions of Neural Collapse

Extensive research has examined the neural collapse from various perspectives, including loss function (Lu & Steinerberger, 2020; Wojtowytsch et al., 2020), the use of mse loss (Han et al., 2021), feature normalization (Yaras et al., 2022), label smoothing (Zhou et al., 2022), fine-grained structure (Yang et al., 2023), generalization (Hui et al., 2022), out-of-distribution with normalization (Haas et al., 2022), imbalanced learning (Dang et al., 2023; Zhong et al., 2023), transfer learning (Galanti et al., 2021; Kornblith et al., 2021), federated learning (Li et al., 2023) or complex of them (Xu et al., 2023). Additionally, some studies have considered extensions of neural collapse such as intermediate layer collapse (Rangamani et al., 2023), generalized neural collapse (Liu et al., 2023), a novel metric (Xu & Liu, 2023). Naturally, geometric analyses have also arisen (Zhu et al., 2021; Thrampoulidis et al., 2022; Tirer et al., 2023).

A.2 Advantages of Orthogonality

Orthogonality has also shown advantages in various environments, such as regularization in multi-task learning (He et al., 2020), vision applications (Cao et al., 2021), features normalization (Lu et al., 2023), optimization (Chiley et al., 2019; Hu et al., 2020; Wu et al., 2020), regularization-better training stability- robustness-generalization (Liu et al., 2021; Achour et al., 2022; Xu et al., 2022; Wang et al., 2020; Li et al., 2019; Ranasinghe et al., 2021; Trockman & Kolter, 2021), explainability (Zhang et al., 2022a; Yang et al., 2020b), continual learning (Pernici et al., 2021; Hersche et al., 2022), reduction on computation load (Yang et al., 2020a), quantization (Ma et al., 2023), graph neural networks (Bodnar et al., 2022; Yang et al., 2022a), transformer (Huang et al., 2022; Kong et al., 2022), representation learning (Medini et al., 2020; Tiao et al., 2023), continual learning (Chaudhry et al., 2020; Ramasesh et al., 2021; Saha et al., 2021; Farajtabar et al., 2020), disentanglement (Sarhan et al., 2020; Liu et al., 2020; Cha & Thiyagalingam, 2023).

B Zero-mean Neural Collapse

Class Mean. In the common neural collapse, the trained class mean is as Eq. 1.

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \boldsymbol{h}_{i,k} \tag{1}$$

where K is the number of classes and N_k is the number of input samples included in the k-th class. $\boldsymbol{h}_{i,k} \in \mathbb{R}^D$ means the feature vector of i-th input sample in the k-th class.

The zero class mean $\boldsymbol{\mu}_k^0$ is calculated in the same way with the trained class mean Eq. 2.

$$\boldsymbol{\mu}_k^0 = \boldsymbol{\mu}_k \tag{2}$$

Global Mean. In the common neural collapse, the trained global mean $\boldsymbol{\mu}_G$ is as Eq. 3.

$$\boldsymbol{\mu}_G = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\mu}_k \tag{3}$$

where $\boldsymbol{\mu}_k$ is the k-th class mean.

This global mean will not be trained and fixed to zero-mean vector like Eq. 4.

$$\boldsymbol{\mu}_G^0 = \mathbf{0}_D \tag{4}$$

where $\boldsymbol{\mu}_G^0$ denotes the zero global mean and $\mathbf{0}_D \in \mathbb{R}^D$ is the zeros vector.

Total Covariance. In the common neural collapse, the trained total covariance $\mathbf{\Sigma}_T \in \mathbb{R}^{D\times D}$ is as Eq. 5.

$$\mathbf{\Sigma}_T = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N_k}\sum_{i=1}^{N_k}\left(\boldsymbol{h}_{i,k}-\boldsymbol{\mu}_G\right)\left(\boldsymbol{h}_{i,k}-\boldsymbol{\mu}_G\right)^{\intercal}\right) \tag{5}$$

This total covariance will be changed like Eq. 6.

$$\mathbf{\Sigma}_T^0 = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N_k}\sum_{i=1}^{N_k}\boldsymbol{h}_{i,k}\boldsymbol{h}_{i,k}^{\intercal}\right) \quad \left(\because \boldsymbol{\mu}_G^0 = \mathbf{0}_D\right) \tag{6}$$

where $\mathbf{\Sigma}_T^0 \in \mathbb{R}^{D\times D}$ denotes the total covariance in the zero-mean NC.

Between-class Covariance. In the common neural collapse, the between-class covariance $\mathbf{\Sigma}_B \in \mathbb{R}^{D\times D}$ is as Eq. 7.

$$\mathbf{\Sigma}_B = \frac{1}{K}\sum_{k=1}^{K}\left(\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}_G\right)\left(\boldsymbol{\mu}_k-\boldsymbol{\mu}_G\right)^{\intercal}\right) \tag{7}$$

This between-class covariance will be changed like Eq. 8.

$$\mathbf{\Sigma}_B^0 = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\intercal} \quad \left(\because \boldsymbol{\mu}_G^0 = \mathbf{0}_D\right) \tag{8}$$

where $\mathbf{\Sigma}_B^0 \in \mathbb{R}^{D\times D}$ denotes the between-class covariance in the zero-mean NC.

Within-class Covariance. In the common neural collapse, the within-class covariance $\mathbf{\Sigma}_W \in \mathbb{R}^{D\times D}$ is as Eq. 9.

$$\mathbf{\Sigma}_W = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N_k}\sum_{i=1}^{N_k}\left(\boldsymbol{h}_{i,k}-\boldsymbol{\mu}_k\right)\left(\boldsymbol{h}_{i,k}-\boldsymbol{\mu}_k\right)^{\intercal}\right) \tag{9}$$

The zero within-class covariance $\mathbf{\Sigma}_W^0$ is calculated in the same way with the within-class covariance (Eq. 10.

$$\mathbf{\Sigma}_W^0 = \mathbf{\Sigma}_W \tag{10}$$

where $\mathbf{\Sigma}_W^0 \in \mathbb{R}^{D\times D}$ denotes the within-class covariance in the zero-mean NC.

(NC1) Variability collapse: $\mathbf{\Sigma}_W \to \mathbf{0}_D$

(ZNC1) Variability collapse: $\mathbf{\Sigma}_W^0 \to \mathbf{0}_D$

(NC2) Convergence to simplex ETF:

$$\left|\left\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_G\right\|_2 - \left\|\boldsymbol{\mu}_{k'}-\boldsymbol{\mu}_G\right\|_2\right| \to 0 \quad \forall_{k,k'}$$

$$\left\langle\tilde{\boldsymbol{\mu}_k}, \tilde{\boldsymbol{\mu}_{k'}}\right\rangle \to \frac{K}{K-1}\delta_{k,k'} - \frac{K}{K-1} \quad \forall_{k,k'}$$

where $\tilde{\boldsymbol{\mu}}_k = \left(\boldsymbol{\mu}_k-\boldsymbol{\mu}_G\right)/\left\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_G\right\|_2$ denotes the normalized k-th class mean and $\delta_{k,k'}$ is the Kronecker delta symbol.

(ZNC2) Convergence to orthogonal matrix:

$$\left|\left\|\boldsymbol{\mu}_k\right\|_2 - \left\|\boldsymbol{\mu}_{k'}\right\|_2\right| \to 0 \quad \forall_{k,k'}$$

$$\left\langle\tilde{\boldsymbol{\mu}_k}, \tilde{\boldsymbol{\mu}_{k'}}\right\rangle \to 0 \quad \forall_{k,k'}$$

where $\tilde{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_k/\left|\boldsymbol{\mu}_k\right|_2$ denotes the normalized k-th class mean.

Table 1: Test Accuracy (%) at zero error and last epoch.

Arch \ Dataset	MNIST		Fashion MNIST		SVHN		CIFAR10		CIFAR100	
	Zero	Last	Zero	Last	Zero	Last	Zero	Last	Zero	Last
VGG	99.40	99.56	92.92	93.31	93.82	94.53	87.85	88.65	63.03	63.85
ResNet	99.32	99.71	93.29	93.64	94.64	95.70	88.72	89.44	66.19	66.21
DenseNet	99.65	99.70	94.18	94.35	95.87	95.93	91.14	91.19	77.19	76.56
VGG†	99.49	99.56	93.25	93.56	93.91	94.51	87.54	88.44	63.14	63.95
ResNet†	99.52	99.64	93.96	93.77	92.45	95.21	87.49	89.45	66.03	65.11
DenseNet†	99.64	99.61	93.92	94.21	95.21	95.41	88.20	88.43	70.66	70.95
VGG+FNO†	99.49	99.49	92.98	93.31	93.89	94.16	88.14	89.10	62.59	64.17
ResNet+FNO†	99.40	99.67	93.90	94.28	92.30	95.41	87.61	89.59	67.15	66.20
DenseNet+FNO†	99.62	99.60	93.91	94.08	95.11	95.22	88.64	88.76	72.23	72.54

(NC3) Convergence to self-duality:

$$\left\| \frac{\boldsymbol{W}^\intercal}{\|\boldsymbol{W}\|_F} - \frac{\dot{\boldsymbol{M}}}{\|\dot{\boldsymbol{M}}\|_F} \right\|_F \to 0$$

where $\dot{\boldsymbol{M}} = [\boldsymbol{\mu}_k - \boldsymbol{\mu}_G, 1 \leq k \leq K] \in \mathbb{R}^{D \times K}$ denotes the matrix obtained by concatenating the class means into the columns of a matrix.

(ZNC3) Convergence to self-duality:

$$\left\| \frac{\boldsymbol{W}^\intercal}{\|\boldsymbol{W}\|_F} - \frac{\dot{\boldsymbol{M}}}{\|\dot{\boldsymbol{M}}\|_F} \right\|_F \to 0$$

where $\dot{\boldsymbol{M}} = [\boldsymbol{\mu}_k, 1 \leq k \leq K] \in \mathbb{R}^{D \times K}$.

(NC4) Simplification to NCC:

$$\arg\max_{k'} \langle \boldsymbol{w}_{k'}, \boldsymbol{h} \rangle + b_{k'} \to \arg\min_{k'} \|\boldsymbol{h} - \boldsymbol{\mu}_{k'}\|_2$$

(ZNC4) Simplification to NCC:

$$\arg\max_{k'} \langle \boldsymbol{w}_{k'}, \boldsymbol{h} \rangle \to \arg\min_{k'} \|\boldsymbol{h} - \boldsymbol{\mu}_{k'}\|_2$$

B.1 Zero-mean Neural Collapse in Image Classification Benchmarks

Datasets. To reproduce the analyses environment for zero-mean neural collapse, we follow (Papyan et al., 2020). We utilize the MNIST (Deng, 2012), Fashion MNIST (Xiao et al., 2017), SVHN (Netzer et al., 2011), CIFAR10, and CIFAR100 datasets (Krizhevsky et al., 2009). To eliminate imbalance factor in the datasets, we subsamples 5,000 samples per class for MNIST and 4,600 samples per class for SVHN. The remaining datasets are already balanced: 6,000 examples for Fashion MNIST, 5,000 examples for CIFAR10, and 500 examples for CIFAR100. There was no data augmentation except normalization.

Architectures and Implementation Details. We train three types of convolutional networks: VGG (Simonyan & Zisserman, 2014), ResNet (He et al., 2015), and DenseNet (Huang et al., 2017). Following (Papyan et al., 2020), we minimize the cross-entropy loss using SGD with momentum 0.9 and the weight decay is set to $5e^{-4}$. The batch size and the number of epochs are set to 256 and 350, respectively. The initial learning is set differently for datasets and architectures and it is decayed by 10 at 1/3 and 2/3. We itemize architectures for each dataset with the initial learning rate:

- **MNIST**: VGG11 (0.06786), ResNet18 (0.013296), and DenseNet40 (0.094015)
- **Fashion MNIST**: VGG11 (0.009597), ResNet18 (0.13025), and DenseNet250 (0.009597)
- **SVHN**: VGG11 (0.094015), ResNet18 (0.009597), and DenseNet40 (0.06786)
- **CIFAR10**: VGG13 (0.048982), ResNet18 (0.06786), and DenseNet40 (0.094015)
- **CIFAR100**: VGG13 (0.180451), ResNet50 (0.13025), and DenseNet250 (0.13025)

Results and Analyses. We visualize the observations of various experiments. Best descriptions in captions of each figure.

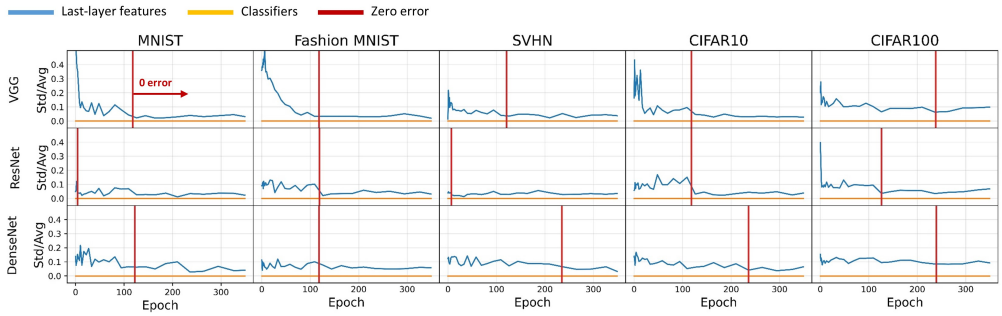

Figure 1: Class mean vectors become equinorm. The y-axis in each cell displays the coefficient of variation of averaged normalization of class means and class weight vectors. The blue lines represent $\mathrm{Std}_k\left(\|\boldsymbol{\mu}_k\|\right)/\mathrm{Avg}_k\left(\|\boldsymbol{\mu}_k\|\right)$ where $\boldsymbol{\mu}_k,\ 1 \leq k \leq K$ denotes the class means of the last-layer features of the training input samples. The orange lines show $\mathrm{Std}_k\left(\|\boldsymbol{w}_k\|\right)/\mathrm{Avg}_k\left(\|\boldsymbol{w}_k\|\right)$ where $\boldsymbol{w}_k$ is the class weight vector of the k-th class. As training advances, the coefficients of variation for both class means and class weight vectors decrease.

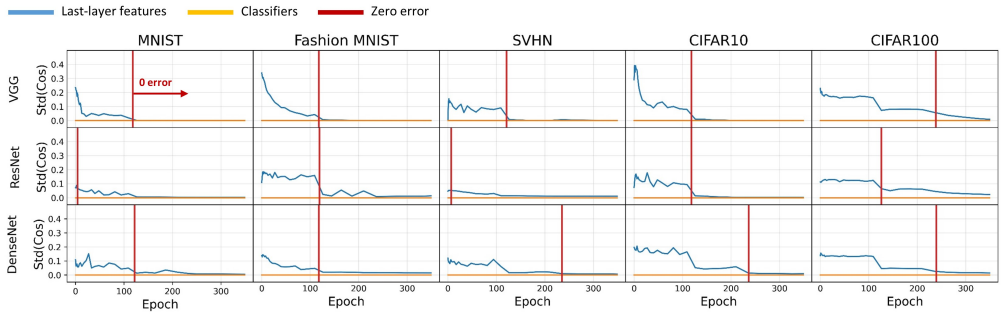

Figure 2: Class means converge to equiangularity. The y-axis in each cell displays the standard deviation of the cosine similarity between pairs of centered class means and class weight vectors across all pairs of classes k and k', $\forall\ k \neq k'$. $\boldsymbol{w}_k$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\mu}_G$ are as in Figure 1. The blue lines represent $\cos_{\boldsymbol{\mu}}\left(k,k'\right) = \left(\boldsymbol{\mu}_k - \boldsymbol{\mu}_G\right)\left(\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_G\right)^{\mathsf{T}} / \left(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_G\|_2 \|\boldsymbol{\mu}_{k'} - \boldsymbol{\mu}_G\|_2\right)\ \forall\ k \neq k'$. The orange lines show $\cos_{\boldsymbol{w}}\left(k,k'\right) = \boldsymbol{w}_k \boldsymbol{w}_{k'}^{\mathsf{T}} / \left(\|\boldsymbol{w}_k\|_2 \|\boldsymbol{w}_{k'}\|_2\right)\ \forall\ k \neq k'$. As training advances, the standard deviations of $\cos_{\boldsymbol{\mu}}\left(k,k'\right)$ and $\cos_{\boldsymbol{w}}\left(k,k'\right)$ converge to zero, signifying equiangularity.

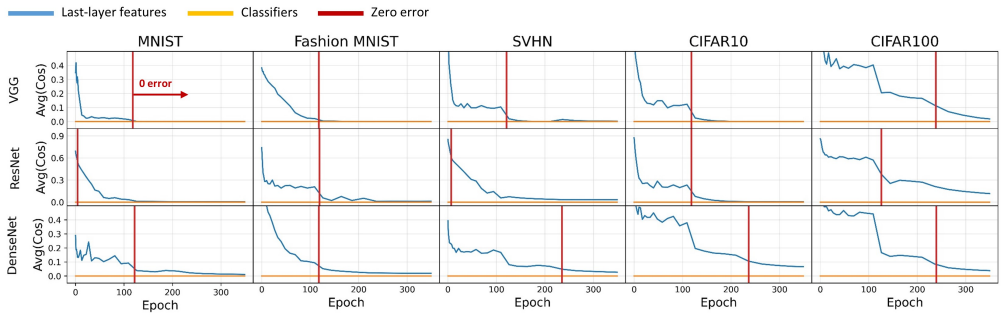

Figure 3: Class means converge to maximal-angle equiangularity. The y-axis in each cell displays the average of cosine similarity of between pairs of centered class means and between pairs of class weight vectors across all pairs of classes k and k', $\forall\ k \neq k'$. The blue lines represent $\mathrm{Avg}_{k,k'}\cos_{\boldsymbol{\mu}}\left(k,k'\right),\ \forall k \neq k'$. The orange lines show $\mathrm{Avg}_{k,k'}\cos_{\boldsymbol{w}}\left(k,k'\right),\ \forall k \neq k'$. This result represents the maximum separation achievable among globally centered, equiangular vectors.

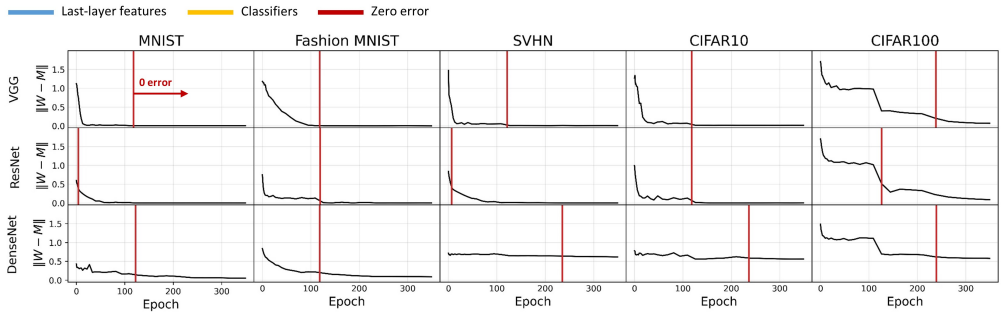

Figure 4: Class means approach to class weight vectors. The y-axis in each cell displays the distance between the class weight vectors and the normalized, centered class means. The lines represent the quantities of $\|\boldsymbol{W} - \tilde{\boldsymbol{M}}\|_F^2$. $\boldsymbol{W}$ is the class weight matrix. $\tilde{\boldsymbol{M}} = \dot{\boldsymbol{M}}/\|\dot{\boldsymbol{M}}\|_F$ where $\dot{\boldsymbol{M}} = (\boldsymbol{\mu}_k)_{1 \leq k \leq K} \in \mathbb{R}^{D \times K}$ is the matrix whose columns consist of the centered class means. This distance decreases as training advances. This result indicates that the centered class means are proportionally related to the class weight vectors like a self-dual manner.

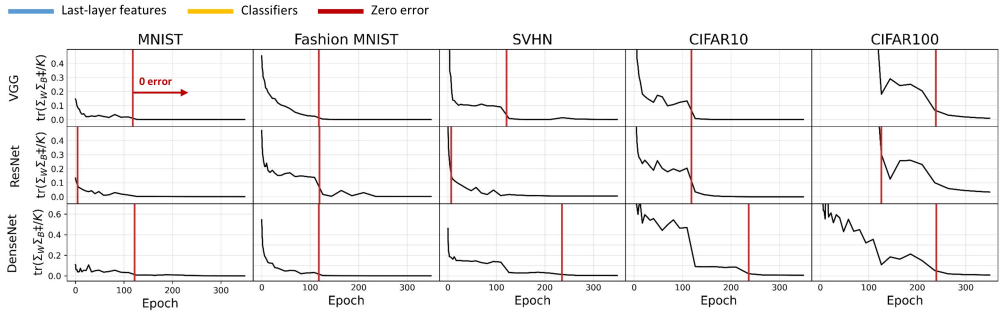

Figure 5: Within-class variation collapses. The y-axis in each cell displays the magnitude of the between-class covariance compared with the within-class covariance of the last-layer features. The lines represent $\mathrm{tr}(\Sigma_W \Sigma_B^{\ddagger}/K)$, where $\mathrm{tr}(\cdot)$ denotes the trace operator, $[\cdot]^{\ddagger}$ indicates Moore-Penrose pseudoinverse, and K is the number of classes. Σ_W and Σ_B are as in Table 1 of the main paper. This magnitude decreases as training advances. This result indicates that collapse of within-class variations occurs.

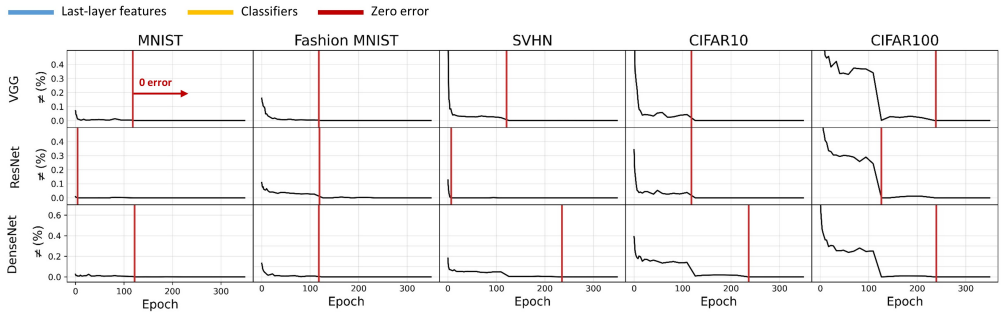

Figure 6: Classifier works in a similar way to NCC. The y-axis in each cell displays the percentage of input samples in the training set where there is a mismatch between the classifier;s output and the result that would have been obtained by selecting $\arg\min_k \|\boldsymbol{h} - \boldsymbol{\mu}_k\|_2$ where $\boldsymbol{h}$ is a last-layer feature and $\boldsymbol{\mu}_k$, $\forall\, 1 \leq k \leq K$ are the class means of last-layer features. The proportion of disagreement decreases as training advances. This result indicates that the classifier's decision is gradually simplified to the nearest class mean decision rule.

C PROOF FOR THEOREM 1

Note that the constraint constrained optimization problem of OLPM is reduced to an entangled constraint. We consider the k-th ($1 \leq k \leq K$ and $K \geq 2$) problem as:

$$\begin{aligned} \min_{H} \ & \frac{1}{N}\sum_{i=1}^{n_k}\mathcal{L}_{ce}\left(\mathbf{h}_{k,i}, \mathbf{Q}^*\right), \\ s.t. \ & \|\mathbf{h}_{k,i}\|^2 - 2\sum_{j\neq k}^{K}\mathbf{h}_{k,i}^{\intercal}\mathbf{w}_j \leq E_{HW}, \ 1 \leq i \leq n_k, \end{aligned} \tag{11}$$

where $\mathbf{Q}^*$ is the fixed non-negative orthogonal classifier. The problem above is convex as the objective is a sum of affine functions and log-sum-exp functions with convex constraints. We have the Lagrange function as:

$$\tilde{L} = \frac{1}{N}\sum_{i=1}^{n_k} -\log\frac{\exp(\mathbf{h}_{k,i}^{\intercal}\mathbf{q}_k^*)}{\sum_{j=1}^{K}\exp(\mathbf{h}_{k,i}^{\intercal}\mathbf{q}_j^*)} + \sum_{i=1}^{n_k}\lambda_i\left(\|\mathbf{h}_{k,i}\|^2 - 2\sum_{j\neq k}^{K}\mathbf{h}_{k,i}^{\intercal}\mathbf{q}_j - E_{HW}\right), \tag{12}$$

where λ_i is the Langrange multiplier. We have its gradient with respect to $\mathbf{h}_{k,i}$ as:

$$\frac{\partial\tilde{L}}{\partial\mathbf{h}_{k,i}} = -\frac{(1-p_k)\,\mathbf{q}_k^*}{N} + \frac{1}{N}\sum_{j\neq k}^{K}p_j\mathbf{q}_j^* + 2\lambda_i\left(\mathbf{h}_{k,i} - \sum_{j\neq k}^{K}\mathbf{q}_j^*\right), 1 \leq i \leq n_k. \tag{13}$$

First we consider the case when $\lambda_i = 0$. $\partial\tilde{L}/\partial\mathbf{h}_{k,i} = 0$ gices the following equation:

$$\begin{aligned} \frac{(1-p_k)\,\mathbf{q}_k^*}{N} &= \frac{1}{N}\sum_{j\neq k}^{K}p_j\mathbf{q}_j^* \\ \sum_{j\neq k}^{K}p_j\mathbf{q}_k^* &= \sum_{j\neq k}^{K}p_j\mathbf{q}_j^*. \end{aligned} \tag{14}$$

Multiplying $\mathbf{q}_k^*$ by both sides of the equation, we should have:

$$\sum_{j\neq k}^{K}p_j = 0 \quad \left(\because \ \mathbf{q}_k^{*\intercal}\mathbf{q}_{k'}^* = \delta_{k,k'}, \ \forall k, k' \in [1, K]\right), \tag{15}$$

which contradicts with $p_j > 0, \forall 1 \leq i \leq K$ when the ℓ_2 norm of $\mathbf{h}_{k,i}$ is constrained and $\mathbf{Q}^*$ has a fixed ℓ_2 norm. So we have $\lambda_i > 0$ and according to the KKT condition:

$$\|\mathbf{h}_{k,i}\|^2 - 2\sum_{j\neq k}^{K}\mathbf{h}_{k,i}^{\intercal}\mathbf{q}_j = E_{HW}. \tag{16}$$

Then we have the equation:

$$\frac{\partial\tilde{L}}{\partial\mathbf{h}_{k,i}^*} = \frac{1}{N}\sum_{j\neq k}^{K}p_j\left(\mathbf{q}_j^* - \mathbf{q}_k^*\right) + 2\lambda_i\left(\mathbf{h}_{k,i}^* - \sum_{j\neq k}^{K}\mathbf{q}_j^*\right) = 0, \tag{17}$$

where $\mathbf{h}_{k,i}^*$ is the optimal solution of $\mathbf{h}_{k,i}$. Multiplying $\mathbf{q}_{j'}^*$ ($j' \neq k$) by both sides of Eq. 17, we get:

$$\frac{p_{j'}}{N} + 2\lambda_i\left(\mathbf{h}_{k,i}^{*\intercal}\mathbf{q}_j^* - 1\right) = 0. \tag{18}$$

Since $p_{j'} > 0$, we have $\mathbf{h}_{k,i}^{*\intercal}\mathbf{q}_j^* < 1$. Then for any pair $j, j' \neq k$, we have:

$$\frac{p_j}{p_{j'}} = \frac{\exp\left(\mathbf{h}_{k,i}^{*\intercal}\mathbf{q}_j^*\right)}{\exp\left(\mathbf{h}_{k,i}^{*\intercal}\mathbf{q}_{j'}^*\right)} = \frac{\exp\left(\mathbf{h}_{k,i}^{*\intercal}\mathbf{q}_j^* - 1\right)}{\exp\left(\mathbf{h}_{k,i}^{*\intercal}\mathbf{q}_{j'}^* - 1\right)} = \frac{\mathbf{h}_{k,i}^{*\intercal}\mathbf{q}_j^* - 1}{\mathbf{h}_{k,i}^{*\intercal}\mathbf{q}_{j'}^* - 1}. \tag{19}$$

Considering that the function $f(x) = x/\exp(x)$ is monotonically increasing when $x < 0$, we have:

$$\begin{aligned} \frac{{\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_j - 1}{\exp\left({\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_j - 1\right)} &= \frac{{\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_{j'} - 1}{\exp\left({\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_{j'} - 1\right)} \\ {\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_j = {\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_{j'} = C, p_j = p_{j'} = p,\ &\forall j, j' \neq k, \end{aligned} \tag{20}$$

where C amd p are constants. From Eq. 18, we have:

$$p = -2N\lambda_i(C-1), \tag{21}$$

and

$$1 - p_k = (K-1)p = -2N\lambda_i(K-1)(C-1). \tag{22}$$

From Eq. 17, we have:

$$\begin{aligned} \mathbf{h}^*_{k,i} &= \frac{1}{2N\lambda_i}\left((1-p_k)\,\mathbf{q}^*_k - \sum_{j\neq k}^{K} p_j \mathbf{q}^*_j\right) + \sum_{j\neq k}^{K} \mathbf{q}^*_j \\ &= \frac{1}{2N\lambda_i}\left(-2N\lambda_i(K-1)(C-1)\mathbf{q}^*_k + 2N\lambda_i(C-1)\sum_{j\neq k}^{K}\mathbf{q}^*_j\right) + \sum_{j\neq k}^{K}\mathbf{q}^*_j \\ &= -(K-1)(C-1)\mathbf{q}^*_k + C\sum_{j\neq k}^{K}\mathbf{q}^*_j, \end{aligned} \tag{23}$$

From the theorem in (Ji et al., 2022), the margin of a single feature $\mathbf{h}^*_{k,i}$ is defined:

$$\mathcal{M}_{k,i} := {\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_k - \max_{j\neq k} {\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_j. \tag{24}$$

Multiplying $\mathbf{q}^*_k$ by both sides of Eq. 23, we should have:

$${\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_k = -(K-1)(C-1) \leq K-1 \quad (\because 0 \leq C < 1)\,, \tag{25}$$

and

$${\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_j = {\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_{j'} = C \geq 0,\ \ \forall j, j' \neq k \tag{26}$$

When the equality holds, we have:

$$\begin{aligned} {\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_k &= K-1 \\ {\mathbf{h}^*_{k,i}}^\intercal \mathbf{q}^*_j &= 0,\ \ \forall j \neq k, \end{aligned} \tag{27}$$

which is equivalent to Eq. 8 of the main paper and concludes the proof.

D ALGORITHMS FOR SECTION 6.1 AND 6.2

In this section, we provide algorithms that are more detailed and may help with implementation for *masked and weighted softmax with FNO classifier in a task T_t* in section 6.1 and for *arc-mixup with FNO classifier and feature masking in a mini-batch $\mathbb{B}$* in section 6.2. These algorithms represent the whole process in training classification models in continual learning and imbalanced learning, respectively.

Algorithm 1 Masked and Weighted Softmax with FNO classifier in a Task T_t

Input: $(\boldsymbol{X}_t, \boldsymbol{Y}_t)$, $\boldsymbol{Q}$
Output: $\mathbf{P} \in \mathbb{R}^{N_t \times K}$

1: $\mathbf{H} \leftarrow \text{RELU}(f_\theta(\boldsymbol{X}_t))$ # get features from $\boldsymbol{X}_t$ as Eq. 1 of the main paper.
2: $\mathbb{K} = \{c_i \mid c_i \text{ of } \boldsymbol{X}_t\}$ # initialize a set of class labels in the task T_t.
3: $M^{(\text{-}\infty)} = (m_{i,j})_{1 \le i \le N_t, 1 \le j \le K}$, where $m_{i,j} = \text{-}\infty$ if $j \notin \mathbb{K}$ otherwise 1 # initialize $M^{(\text{-}\infty)}$
4: $\mathbf{P} =$W-$\text{SOFTMAX}(M^{(\text{-}\infty)} \odot \text{MATMUL}(\boldsymbol{Q}, \mathbf{H}))$ # get the confidence of $\mathbf{H}$

Algorithm 2 Arc-mixup with FNO classifier and feature masking in a mini-batch $\mathbb{B}$

Input: $(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{B}$, $\boldsymbol{Q}$
Output: $\mathbf{P} \in \mathbb{R}^{|\mathbb{B}| \times K}$

1: $(\hat{\boldsymbol{X}}, \hat{\boldsymbol{Q}}) \leftarrow ArcMixup(\boldsymbol{X}, \boldsymbol{Q})$ # mixup input samples and class vectors as Eq. 11 of the main paper
2: $\mathbb{K} = \{c_i \mid c_i \in \mathbb{B},\ \forall 1 \le i \le |\mathbb{B}|\}$ # initialize a set of class labels in the mini-batch
3: $\hat{\mathbb{J}} = \bigcup_{i=1}^{|\mathbb{B}|} \mathbb{J}_i$ # initialize an index set including all index sets of $\boldsymbol{Q}$
4: $M^{(0)} = (m_{i,j})_{1 \le i \le |\mathbb{B}|, 1 \le j \le D}$, where $m_{i,j} = \mathbf{1}_{j \in \hat{\mathbb{J}}}$ # initialize a zero mask $M^{(0)}$
5: $\hat{\mathbf{H}} \leftarrow \text{RELU}(f_\theta(\hat{\boldsymbol{X}}))$ # get features from $\hat{\boldsymbol{X}}$ as Eq. 1 of the main paper.
6: $\mathbf{P} =\text{MATMUL}(\hat{\boldsymbol{Q}}, \text{LAYERNORM}(M^{(0)} \odot \hat{\mathbf{H}}))$ # get the confidence of $\hat{\mathbf{H}}$

E ADDITIONAL IMPLEMENTATION DETAILS FOR EXPERIMENTS IN SECTION 7

E.1 CONTINUAL LEARNING IN SPLIT DATASETS

Dataset Explanation. Following (Hsu et al., 2018; Van de Ven & Tolias, 2019; Buzzega et al., 2020; Kim et al., 2023), the split datasets of the MNIST, CIFAR10, CIFAR100, and Tiny-ImageNet are described as:

- **S-MNIST, S-CIFAR10**: T_1(0-1), T_2(2-3), T_3(4-5), T_4(6-7), T_5(8-9)
- **S-CIFAR100**: T_1(0-9), T_2(10-19), T_3(20-29), T_4(30-39), T_5(40-49), T_6(50-59), T_7(60-69), T_8(70-79), T_9(80-89), T_{10}(90-99)
- **S-Tiny-ImageNet**: T_1(0-9), T_2(10-19), T_3(20-29), T_4(30-39), T_5(40-49), T_6(50-59), T_7(60-69), T_8(70-79), T_9(80-89), T_{10}(90-99) T_{11}(100-109), T_{12}(110-119), T_{13}(120-129), T_{14}(130-139), T_{15}(140-149), T_{16}(150-159), T_{17}(160-169), T_{18}(170-179), T_{19}(180-189), T_{20}(190-199)

where $T_t(i\text{-}j)$ indicates that the t-th task has class labels from i to j and all classes are sequentially divided to each task. In CIFAR10, for instance, we can convert the class index in each task as below:

- T_1(0-1): {airplane, autombile}
- T_2(2-3): {bird, cat}
- T_3(4-5): {deer, dog}
- T_4(6-7): {frog, horse}
- T_5(8-9): {ship, truck}

Table 2: Hyperparameter settings on all experiments in Continual Learning Benchmarks. (Table credit: (Kim et al., 2023))

Method	Buffer	S-MNIST	S-CIFAR10	S-CIFAR100	S-Tiny-ImageNet
SGD	-	lr: 0.03	lr: 0.1	lr: 0.03	lr: 0.03
ER	200	lr: 0.01	lr: 0.1	lr: 0.1	lr: 0.1
	500	lr: 0.1	lr: 0.1	lr: 0.1	lr: 0.03
	5120	lr: 0.1	lr: 0.1	lr: 0.1	lr: 0.1
DER++	200	lr: 0.03 α: 0.2 β: 1.0	lr: 0.03 α: 0.1 β: 0.5	lr: 0.03 α: 0.1 β: 0.5	lr: 0.03 α: 0.1 β: 1.0
	500	lr: 0.03 α: 1.0 β: 0.5	lr: 0.03 α: 0.2 β: 0.5	lr: 0.03 α: 0.1 β: 0.5	lr: 0.03 α: 0.2 β: 0.5
	5120	lr: 0.1 α: 0.2 β: 0.5	lr: 0.03 α: 0.1 β: 1.0	lr: 0.03 α: 0.1 β: 0.5	lr: 0.03 α: 0.1 β: 0.5

Hyperparameter Settings. We have the same settings to (Buzzega et al., 2020; Kim et al., 2023). To summarize them, we borrow the table of hyperparameter settings in (Kim et al., 2023) (Please refer to Table 2).

Task-Incremental Learning (Task-IL) and Class-Incremental Learning (Class-IL. As described in (Hsu et al., 2018; Van de Ven & Tolias, 2019), the sequence of tasks can be modeled in two ways: Task-IL and Class-IL. Both divide datasets and train models for tasks in the same way. The only one difference lies in the evaluation system. To be more specific, Task-IL provides task information the task information during prediction, so the models classify test samples into the target task's classes. In contrast, Class-IL does not provide task information, so the models must predict the test samples as one of the total classes, regardless of the target task. We trained and validated our models in continual learning benchmarks using these training strategies.

E.2 Imbalanced Learning in Long-Tailed Datasets

Dataset Explanation. Following (Zhong et al., 2021) and (Zhou et al., 2020), the long-tailed datasets of CIFAR10, CIFAR100, ImageNet (Russakovsky et al., 2015), and Places365 (Zhou et al., 2017) are described as:

- **CIFAR10-LT** comprises ten imbalanced classes, subsampled at an exponentially decreasing rate from the initial class of CIFAR10 as mentioned in (Zhong et al., 2021).
- **CIFAR100-LT** includes one hundred imbalanced classes, constructed in the same way as CIFAR10-LT.
- **ImageNet-LT** is a long-tailed dataset for large-scale object classification, derived from the ImageNet. The sampling is based on Pareto distribution with a power value $\alpha = 5$ and the classes have varying cardinality from 5 to 1,280. Therefore, it contains 115.8K images sorted into 1,000 classes.
- **Places-LT** is an extended version of the large-scale scene classification dataset Places. The classes differ in their cardinality, ranging from 5 to 4,980, and therefore, it contains 184.5K images from 365 classes.

Architectures. We utilize a ResNet32 (Zhong et al., 2021) consisting of three residual blocks, with each output dimensions of 16, 32, and 64, respectively, for the CIFAR10-LT dataset. For the CIFAR100-LT dataset, each output dimension is twice that of the CIFIAR10-LT dataset. Unlike the ResNet architecture for ImageNet, the kernel size, stride, and padding are set to 3, 1, and 1, respectively for the first convolutional layer. ResNet50 and 152 are same to He et al. (2015).

Hyperparameters Settings. We employed ResNet32 for CIFAR10/100-LT and trained the model with 128 mini-batches, utilizing SGD with momentum of 0.9 and weight decay of 2e-4, for 200 epochs. The learning rate was warmed up from 0.02 to the initial learning rate in a linear fashion, divided by 0.1 at epochs 160 and 180. For the other datasets, we utilized ResNet50 and 152, trained the models using SGD with momentum of 0.9 and weight decay of 5e-4, and updated the learning rate using a cosine annealing scheduler. Additionally, we differently set mixup alpha of mixup and arc-mixup as the datasets: $\alpha = 1.0$ for both in CIFAR10/100-LT, $\alpha = 0.2$ for mixup and $\alpha = 5.0$ for arc-mixup in ImageNet, and $\alpha = 0.2$ for both in Places-LT.

F ADDITIONAL EXPERIMENTAL RESULTS.

F.1 CONTINUAL LEARNING IN SPLIT DATASETS

Table 3: Classification results for standard continual learning benchmarks including the performance of commonly used methods in these benchmarks. Best in bold for each buffer setting. (Final Average Accuracy ↑ (%): $mean_{std}$)

Buffer	Method	S-MNIST		S-CIFAR-10		S-Tiny-ImageNet	
		Class-IL	*Task-IL*	*Class-IL*	*Task-IL*	*Class-IL*	*Task-IL*
-	JOINT	$95.57_{0.24}$	$99.51_{0.07}$	$92.20_{0.15}$	$98.31_{0.12}$	$59.99_{0.19}$	$82.04_{0.10}$
	SGD	$19.60_{0.04}$	$94.94_{2.18}$	$19.62_{0.05}$	$61.02_{3.33}$	$7.92_{0.26}$	$18.31_{0.68}$
-	oEWC (Schwarz et al., 2018)	$20.46_{1.01}$	$98.39_{0.48}$	$19.49_{0.12}$	$68.29_{3.92}$	$7.58_{0.10}$	$19.20_{0.31}$
	SI (Zenke et al., 2017)	$19.27_{0.30}$	$96.00_{2.04}$	$19.48_{0.17}$	$68.05_{5.91}$	$6.58_{0.31}$	$36.32_{0.13}$
	LwF (Li & Hoiem, 2017)	$19.62_{0.01}$	$94.11_{3.01}$	$19.61_{0.05}$	$63.29_{2.35}$	$8.46_{0.22}$	$15.85_{0.58}$
	PNN (Rusu et al., 2016)		$99.23_{0.20}$		$95.13_{0.72}$		$67.84_{0.29}$
200	ER (Riemer et al., 2018)	$80.43_{1.89}$	$97.86_{0.35}$	$44.79_{1.86}$	$91.19_{0.94}$	$8.49_{0.16}$	$38.17_{2.00}$
	GEM (Lopez-Paz & Ranzato, 2017)	$80.11_{1.54}$	$97.78_{0.25}$	$25.54_{0.76}$	$90.44_{0.94}$		
	A-GEM (Chaudhry et al., 2018)	$45.72_{4.26}$	$98.61_{0.24}$	$20.04_{0.34}$	$83.88_{1.49}$	$8.07_{0.08}$	$22.77_{0.03}$
	iCaRL (Rebuffi et al., 2017)	$70.51_{0.53}$	$98.28_{0.09}$	$49.02_{3.20}$	$88.99_{2.13}$	$7.53_{0.79}$	$28.19_{1.47}$
	FDR (Benjamin et al., 2018)	$79.43_{3.26}$	$97.66_{0.18}$	$30.91_{2.74}$	$91.01_{0.68}$	$8.70_{0.19}$	$40.36_{0.68}$
	GSS (Aljundi et al., 2019)	$38.90_{2.49}$	$95.02_{1.85}$	$39.07_{5.59}$	$88.80_{2.89}$		
	HAL (Chaudhry et al., 2021)	$84.70_{0.87}$	$97.96_{0.21}$	$32.36_{2.70}$	$82.51_{3.20}$		
	DER (Buzzega et al., 2020)	$84.55_{1.64}$	$98.80_{0.15}$	$61.93_{1.79}$	$91.40_{0.92}$	$11.87_{0.78}$	$40.22_{0.67}$
	DER++ (Buzzega et al., 2020)	$85.61_{1.40}$	$98.76_{0.28}$	$64.88_{1.17}$	$91.92_{0.60}$	$10.96_{1.17}$	$40.87_{1.16}$
	ER†	$78.27_{1.37}$	$97.73_{0.26}$	$49.38_{2.15}$	$91.54_{0.81}$	$8.58_{0.19}$	$38.39_{0.72}$
	FNOER†	$78.59_{1.27}$	$98.03_{0.22}$	$50.56_{1.52}$	$91.79_{0.64}$	$9.21_{0.28}$	$40.46_{1.15}$
	ERMR†	$82.98_{1.03}$	$98.13_{0.16}$	$61.75_{6.07}$	$91.39_{2.13}$	$15.47_{0.67}$	$44.11_{0.50}$
	FNOERMR†	$84.26_{1.16}$	$98.45_{0.19}$	$63.84_{1.47}$	$92.03_{0.52}$	$17.31_{0.74}$	$44.76_{0.90}$
	DER++†	$85.64_{1.02}$	$98.84_{0.11}$	$63.67_{1.01}$	$91.61_{0.73}$	$11.59_{1.07}$	$41.00_{0.88}$
	FNODER++†	$83.29_{1.64}$	$97.67_{0.33}$	$64.15_{1.50}$	$92.52_{0.55}$	$13.75_{1.28}$	$45.52_{1.29}$
	DERMR++†	$84.45_{0.88}$	$99.03_{0.09}$	$66.35_{1.52}$	$93.17_{0.54}$	$13.21_{0.56}$	$49.75_{0.99}$
	FNODERMR++†	$\mathbf{86.27}_{0.88}$	$\mathbf{99.11}_{0.08}$	$\mathbf{67.53}_{1.25}$	$\mathbf{93.98}_{0.39}$	$\mathbf{18.44}_{0.94}$	$\mathbf{53.06}_{0.67}$
500	ER (Riemer et al., 2018)	$86.12_{1.89}$	$99.04_{0.18}$	$57.74_{0.27}$	$93.61_{0.27}$	$9.99_{0.29}$	$48.64_{0.46}$
	GEM (Lopez-Paz & Ranzato, 2017)	$85.99_{1.35}$	$98.71_{0.20}$	$26.20_{1.26}$	$92.16_{0.69}$		
	A-GEM (Chaudhry et al., 2018)	$46.66_{5.85}$	$98.93_{0.21}$	$22.67_{0.57}$	$89.48_{1.45}$	$8.06_{0.04}$	$25.33_{0.49}$
	iCaRL (Rebuffi et al., 2017)	$70.10_{1.08}$	$98.32_{0.07}$	$47.55_{3.95}$	$88.22_{2.62}$	$9.38_{1.53}$	$31.55_{3.27}$
	FDR (Benjamin et al., 2018)	$85.87_{4.04}$	$97.54_{1.90}$	$28.71_{3.23}$	$93.29_{0.59}$	$10.54_{0.21}$	$49.88_{0.71}$
	GSS (Aljundi et al., 2019)	$49.76_{4.73}$	$97.71_{0.53}$	$49.73_{4.78}$	$91.02_{1.57}$		
	HAL (Chaudhry et al., 2021)	$87.21_{0.49}$	$98.03_{0.22}$	$41.79_{4.46}$	$84.54_{2.36}$		
	DER (Buzzega et al., 2020)	$90.54_{1.18}$	$98.84_{0.13}$	$70.51_{1.67}$	$93.40_{0.39}$	$17.75_{1.14}$	$51.78_{0.88}$
	DER++ (Buzzega et al., 2020)	$91.00_{1.49}$	$98.94_{0.27}$	$72.70_{1.36}$	$93.88_{0.50}$	$19.38_{1.41}$	$51.91_{0.68}$
	ER†	$85.99_{1.52}$	$99.14_{0.07}$	$62.38_{1.40}$	$94.12_{0.31}$	$10.12_{0.22}$	$48.06_{0.80}$
	FNOER†	$85.92_{1.76}$	$99.10_{0.15}$	$63.41_{1.36}$	$93.99_{0.45}$	$11.07_{0.41}$	$45.77_{0.46}$
	ERMR†	$89.35_{0.59}$	$\mathbf{99.20}_{0.16}$	$70.64_{1.28}$	$94.22_{0.41}$	$20.43_{0.38}$	$53.21_{0.84}$
	FNOERMR†	$89.42_{0.72}$	$99.16_{0.17}$	$71.43_{0.95}$	$94.38_{0.43}$	$22.41_{0.57}$	$52.60_{0.58}$
	DER++†	$\mathbf{91.01}_{0.46}$	$98.95_{0.07}$	$73.15_{0.80}$	$94.07_{0.39}$	$19.82_{0.87}$	$52.24_{0.94}$
	FNODER++†	$90.56_{0.64}$	$98.06_{0.16}$	$72.70_{1.21}$	$94.37_{0.67}$	$19.42_{0.72}$	$53.35_{0.95}$
	DERMR++†	$83.10_{1.22}$	$99.08_{0.09}$	$71.85_{3.76}$	$94.28_{1.49}$	$17.71_{0.58}$	$59.86_{1.08}$
	FNODERMR++†	$86.75_{0.75}$	$99.00_{0.10}$	$\mathbf{74.77}_{0.66}$	$\mathbf{95.56}_{0.16}$	$\mathbf{22.45}_{0.36}$	$\mathbf{59.87}_{1.91}$
5120	ER (Riemer et al., 2018)	$93.40_{1.29}$	$99.33_{0.22}$	$82.47_{0.52}$	$96.98_{0.17}$	$27.40_{0.31}$	$67.29_{0.23}$
	GEM (Lopez-Paz & Ranzato, 2017)	$95.11_{0.87}$	$99.44_{0.12}$	$25.26_{3.46}$	$95.55_{0.02}$		
	A-GEM (Chaudhry et al., 2018)	$54.24_{6.49}$	$98.93_{0.20}$	$21.99_{2.29}$	$90.10_{2.09}$	$7.96_{0.13}$	$26.22_{0.65}$
	iCaRL (Rebuffi et al., 2017)	$70.60_{1.03}$	$98.32_{0.11}$	$55.07_{1.55}$	$92.23_{0.84}$	$14.08_{1.92}$	$40.83_{3.11}$
	FDR (Benjamin et al., 2018)	$87.47_{3.15}$	$97.79_{1.33}$	$19.70_{0.07}$	$94.32_{0.97}$	$28.97_{0.41}$	$68.01_{0.42}$
	GSS (Aljundi et al., 2019)	$89.39_{0.75}$	$98.33_{0.17}$	$67.27_{4.27}$	$94.19_{1.15}$		
	HAL (Chaudhry et al., 2021)	$89.52_{0.96}$	$98.35_{0.17}$	$59.12_{4.41}$	$88.51_{3.32}$		
	DER (Buzzega et al., 2020)	$94.90_{0.57}$	$99.29_{0.11}$	$83.81_{0.33}$	$95.43_{0.33}$	$36.73_{0.64}$	$69.50_{0.26}$
	DER++ (Buzzega et al., 2020)	$95.30_{1.20}$	$99.47_{0.07}$	$85.24_{0.49}$	$96.12_{0.21}$	$39.02_{0.97}$	$69.84_{0.63}$
	ER†	$93.42_{1.08}$	$99.41_{0.15}$	$84.31_{0.38}$	$97.02_{0.26}$	$27.30_{0.51}$	$67.69_{0.33}$
	FNOER†	$92.95_{1.57}$	$99.41_{0.10}$	$84.33_{1.47}$	$96.90_{0.49}$	$28.38_{0.46}$	$65.05_{0.81}$
	ERMR†	$93.51_{0.60}$	$99.38_{0.12}$	$82.63_{1.34}$	$96.45_{0.27}$	$35.73_{0.41}$	$67.50_{0.53}$
	FNOERMR†	$93.98_{0.39}$	$99.47_{0.09}$	$82.88_{1.35}$	$96.79_{0.38}$	$36.90_{0.41}$	$66.86_{0.32}$
	DER++†	$\mathbf{95.09}_{0.56}$	$99.50_{0.08}$	$85.56_{0.38}$	$96.30_{0.22}$	$\mathbf{39.66}_{0.89}$	$69.95_{0.32}$
	FNODER++†	$94.99_{0.74}$	$99.40_{0.08}$	$\mathbf{85.83}_{0.36}$	$96.68_{0.14}$	$31.72_{1.31}$	$65.76_{0.78}$
	DERMR++†	$93.75_{0.23}$	$\mathbf{99.62}_{0.05}$	$84.71_{0.65}$	$96.78_{0.16}$	$34.72_{0.46}$	$72.40_{0.25}$
	FNODERMR++†	$94.26_{0.24}$	$99.59_{0.05}$	$85.65_{0.38}$	$\mathbf{97.20}_{0.13}$	$38.95_{0.71}$	$\mathbf{72.70}_{0.27}$

Evaluation Metrics. (Kumari et al., 2022) Final Average Accuracy (A_T) and Forgetting (A_F) where $a_{j,t}$ denotes the test accuracy on the t-task after the model has trained all task up to j.

$$A_T = \frac{1}{T}\sum_{t=1}^{T} a_{T,t} \quad F_T = \frac{1}{T-1}\sum_{t=1}^{T} \max_{j\in\{1,\dots T-1\}} (a_{j,t} - a_{T,t})$$

Table 4: Classification results for standard CL benchmarks. The experiment on S-Tiny-ImageNet was conducted using 5 trials with random seeds while other experiments were conducted using 10 trials with random seeds. Best in bold for each buffer setting. (Final Average Forgetting ↓ (%): $mean_{std}$)

Buffer	Method	S-MNIST		S-CIFAR-10	
		Class-IL	*Task-IL*	*Class-IL*	*Task-IL*
-	SGD	$99.10_{0.55}$	$5.15_{2.74}$	$96.39_{0.12}$	$46.24_{2.12}$
-	oEWC (Schwarz et al., 2018)	$97.79_{1.24}$	$0.44_{0.16}$	$91.64_{3.07}$	$29.33_{3.84}$
	SI (Zenke et al., 2017)	$98.89_{0.86}$	$5.15_{2.74}$	$95.78_{0.64}$	$38.76_{0.89}$
	LwF (Li & Hoiem, 2017)	$99.30_{0.11}$	$5.15_{2.74}$	$96.69_{0.25}$	$32.56_{0.56}$
	PNN (Rusu et al., 2016)		$0.00_{0.00}$		$0.00_{0.00}$
200	ER (Riemer et al., 2018)	$21.36_{2.46}$	$0.84_{0.41}$	$61.24_{2.62}$	$7.08_{0.64}$
	GEM (Lopez-Paz & Ranzato, 2017)	$22.32_{2.04}$	$1.19_{0.38}$	$82.61_{1.60}$	$9.27_{2.07}$
	A-GEM (Chaudhry et al., 2018)	$66.15_{6.84}$	$0.96_{0.28}$	$95.73_{0.20}$	$16.39_{0.86}$
	iCaRL (Rebuffi et al., 2017)	$11.73_{0.73}$	$0.28_{0.08}$	$28.72_{0.49}$	$2.63_{3.48}$
	FDR (Benjamin et al., 2018)	$21.15_{4.18}$	$0.52_{0.18}$	$86.40_{2.67}$	$7.36_{0.03}$
	GSS (Aljundi et al., 2019)	$74.10_{3.03}$	$4.30_{2.31}$	$75.25_{4.07}$	$8.56_{1.78}$
	HAL (Chaudhry et al., 2021)	$14.54_{1.49}$	$0.53_{0.19}$	$69.11_{4.21}$	$12.26_{0.02}$
	DER (Buzzega et al., 2020)	$17.66_{2.10}$	$0.57_{0.18}$	$40.76_{0.42}$	$6.57_{0.20}$
	DER++ (Buzzega et al., 2020)	$16.27_{1.73}$	$0.66_{0.28}$	$32.59_{2.32}$	$5.16_{0.21}$
	ER†	$24.09_{1.61}$	$0.89_{0.24}$	$58.97_{2.70}$	$6.49_{0.90}$
	FNOER†	$24.28_{1.58}$	$0.84_{0.16}$	$57.78_{1.94}$	$6.41_{0.83}$
	ERMR†	$10.22_{1.20}$	$0.69_{0.12}$	$26.30_{6.18}$	$6.18_{1.56}$
	FNOERMR†	$9.88_{1.79}$	$0.50_{0.16}$	$\mathbf{21.93}_{3.75}$	$5.91_{0.69}$
	DER++†	$16.34_{1.23}$	$0.53_{0.12}$	$34.80_{1.68}$	$6.45_{1.07}$
	FNODER++†	$19.61_{2.07}$	$2.13_{0.43}$	$35.58_{2.18}$	$5.40_{0.81}$
	DERMR++†	$\mathbf{4.29}_{0.75}$	$0.32_{0.09}$	$23.10_{1.46}$	$4.82_{0.71}$
	FNODERMR++†	$5.59_{0.76}$	$\mathbf{0.25}_{0.05}$	$27.00_{2.20}$	$\mathbf{3.83}_{0.68}$
500	ER (Riemer et al., 2018)	$15.97_{2.46}$	$0.39_{0.20}$	$45.35_{0.07}$	$3.54_{0.35}$
	GEM (Lopez-Paz & Ranzato, 2017)	$15.57_{1.77}$	$0.54_{0.15}$	$74.31_{4.62}$	$9.12_{0.21}$
	A-GEM (Chaudhry et al., 2018)	$65.84_{7.24}$	$0.64_{0.20}$	$94.01_{1.16}$	$14.26_{4.18}$
	iCaRL (Rebuffi et al., 2017)	$11.84_{0.73}$	$0.30_{0.09}$	$25.71_{1.10}$	$2.66_{2.47}$
	FDR (Benjamin et al., 2018)	$13.90_{5.19}$	$1.35_{2.40}$	$85.62_{0.36}$	$4.80_{0.00}$
	GSS (Aljundi et al., 2019)	$60.35_{6.03}$	$0.89_{0.40}$	$62.88_{2.67}$	$7.73_{3.99}$
	HAL (Chaudhry et al., 2021)	$9.97_{1.62}$	$0.35_{0.21}$	$62.21_{4.34}$	$5.41_{1.10}$
	DER (Buzzega et al., 2020)	$9.58_{1.52}$	$0.45_{0.13}$	$26.74_{0.15}$	$4.56_{0.45}$
	DER++ (Buzzega et al., 2020)	$8.85_{1.86}$	$0.35_{0.15}$	$22.38_{4.41}$	$4.66_{1.15}$
	ER†	$16.28_{1.99}$	$0.38_{0.11}$	$42.83_{1.89}$	$3.51_{0.40}$
	FNOER†	$16.47_{2.30}$	$0.48_{0.16}$	$41.52_{1.94}$	$3.79_{0.65}$
	ERMR†	$7.18_{1.37}$	$0.39_{0.17}$	$17.31_{1.58}$	$3.09_{0.36}$
	FNOERMR†	$6.79_{1.80}$	$0.46_{0.20}$	$\mathbf{15.60}_{2.61}$	$3.20_{0.48}$
	DER++†	$9.00_{0.50}$	$0.41_{0.05}$	$23.12_{1.88}$	$3.56_{0.38}$
	FNODER++†	$9.86_{0.84}$	$0.88_{0.17}$	$24.12_{2.00}$	$3.22_{0.82}$
	DERMR++†	$\mathbf{2.07}_{0.89}$	$\mathbf{0.23}_{0.09}$	$16.95_{3.77}$	$3.03_{0.76}$
	FNODERMR++†	$3.81_{0.85}$	$0.85_{0.28}$	$17.34_{1.35}$	$\mathbf{1.97}_{0.37}$
5120	ER (Riemer et al., 2018)	$6.08_{1.84}$	$0.25_{0.23}$	$13.99_{0.12}$	$0.27_{0.06}$
	GEM (Lopez-Paz & Ranzato, 2017)	$4.30_{1.16}$	$0.16_{0.09}$	$75.27_{4.41}$	$6.91_{2.33}$
	A-GEM (Chaudhry et al., 2018)	$55.10_{10.79}$	$0.63_{0.21}$	$84.49_{3.08}$	$11.36_{1.68}$
	iCaRL (Rebuffi et al., 2017)	$11.64_{0.72}$	$0.26_{0.06}$	$24.94_{0.14}$	$1.59_{0.57}$
	FDR (Benjamin et al., 2018)	$11.58_{3.97}$	$0.95_{1.61}$	$96.64_{0.19}$	$1.93_{0.48}$
	GSS (Aljundi et al., 2019)	$7.90_{1.21}$	$0.18_{0.11}$	$58.11_{9.12}$	$7.71_{2.31}$
	HAL (Chaudhry et al., 2021)	$6.55_{1.63}$	$0.13_{0.07}$	$27.19_{7.53}$	$5.21_{0.50}$
	DER (Buzzega et al., 2020)	$4.53_{0.83}$	$0.32_{0.08}$	$10.12_{0.80}$	$2.59_{0.08}$
	DER++ (Buzzega et al., 2020)	$4.19_{1.63}$	$0.23_{0.06}$	$7.27_{0.84}$	$1.18_{0.19}$
	ER†	$6.26_{1.56}$	$0.21_{0.10}$	$14.44_{0.65}$	$0.46_{0.21}$
	FNOER†	$6.92_{2.32}$	$0.20_{0.11}$	$14.34_{2.02}$	$0.70_{0.63}$
	ERMR†	$3.12_{1.10}$	$0.24_{0.10}$	$7.57_{1.31}$	$0.81_{0.28}$
	FNOERMR†	$2.44_{0.60}$	$0.16_{0.13}$	$6.16_{1.03}$	$0.56_{0.20}$
	DER++†	$4.59_{0.79}$	$0.24_{0.10}$	$7.50_{0.80}$	$1.05_{0.34}$
	FNODER++†	$4.72_{0.97}$	$0.32_{0.06}$	$7.68_{0.71}$	$0.79_{0.18}$
	DERMR++†	$\mathbf{0.90}_{0.18}$	$\mathbf{0.09}_{0.06}$	$\mathbf{5.02}_{0.69}$	$0.71_{0.27}$
	FNODERMR++†	$1.47_{0.25}$	$0.11_{0.04}$	$5.56_{0.37}$	$\mathbf{0.43}_{0.13}$

F.2 Imbalanced Learning in Long-Tailed Datasets

We present the top-1 test accuracy for three class divisions: *Head-Many* (more than 100 images), *Medium* (20 to 100 images), and *Tail-Few* (less than 20 images). The imbalance factor ω in

CIFAR10/100-LT datasets means the ratio of the number of a head class n_{max} to the number of a tail class n_{min}, *i.e.*, $\omega = n_{max}/n_{min}$.

Table 5: Results of Imbalanced Learning with other methods on CIFAR10/100-LT. Each column of Method indicates the augmentation method, the type of classifier, and the type of loss function, respectively. Best in bold. (Accuracy (%): *mean*$_{std}$)

Method			Reference	CIFAR10-LT			CIFAR100-LT		
Augmentation	Classifier	Loss		100	50	10	100	50	10
-	FC	CE	(Yang et al., 2022b)	$72.10_{0.30}$	$77.60_{0.30}$	$87.40_{0.30}$	-	-	-
-	ETF	CE	(Yang et al., 2022b)	$72.90_{0.30}$	$79.50_{0.20}$	$87.20_{0.10}$	-	-	-
-	ETF	DR	(Yang et al., 2022b)	$73.00_{0.20}$	$78.40_{0.30}$	$86.90_{0.20}$	-	-	-
mixup	FC	CE	(Yang et al., 2022b)	$73.90_{0.30}$	$79.30_{0.20}$	$87.80_{0.10}$	43.00	48.10	-
B-mixup	FC	CE	(Zhang et al., 2022b)	78.70	-	**89.60**	-	-	-
mixup	ETF	CE	(Yang et al., 2022b)	$67.00_{0.40}$	$77.20_{0.30}$	$87.00_{0.20}$	-	-	-
mixup	ETF	DR	(Yang et al., 2022b)	$76.50_{0.30}$	$81.00_{0.20}$	$87.70_{0.20}$	45.30	50.40	
-	FC	CE	(reproduced.)†	$71.86_{0.65}$	$77.58_{0.48}$	$87.42_{0.30}$	$41.65_{0.41}$	$46.89_{0.38}$	$60.48_{0.34}$
mixup	FC	CE	(reproduced.)†	$74.24_{0.44}$	$80.00_{0.54}$	$89.08_{0.32}$	$43.80_{0.42}$	$49.57_{0.37}$	$\mathbf{63.90}_{0.33}$
arc-mixup	FNO	CE	(reproduced.)†	$\mathbf{82.59}_{0.26}$	$\mathbf{85.13}_{0.25}$	$89.50_{0.14}$	$\mathbf{49.26}_{2.82}$	$\mathbf{54.44}_{2.32}$	$63.14_{3.82}$

Table 6: Detailed Results of Imbalanced Learning on CIFAR10/100-LT. Best in bold. (Aug: the augmentation method, Clf: the type of classifier, $\mathcal{L}$: the type of loss function) (Accuracy (%): *mean*$_{std}$)

	Method			CIFAR10-LT				CIFAR100-LT			
	Aug	Clf	$\mathcal{L}$	Many	Median	Few	All	Many	Median	Few	All
imb 100	-	FC	CE†	$91.99_{3.02}$	$71.72_{1.67}$	$51.92_{5.34}$	$71.86_{0.65}$	$68.41_{0.55}$	$40.41_{0.75}$	$9.94_{0.55}$	$41.65_{0.41}$
	mixup	FC	CE†	$\mathbf{94.72}_{0.43}$	$75.70_{1.03}$	$51.82_{1.61}$	$74.24_{0.44}$	$\mathbf{73.25}_{0.52}$	$42.56_{0.81}$	$8.73_{0.49}$	$43.80_{0.42}$
	arc-mixup	FNO	CE†	$84.60_{0.78}$	$\mathbf{80.19}_{0.71}$	$\mathbf{83.79}_{0.86}$	$\mathbf{82.59}_{0.26}$	$63.46_{3.55}$	$\mathbf{53.04}_{2.93}$	$\mathbf{27.05}_{2.08}$	$\mathbf{49.26}_{2.82}$
imb 50	-	FC	CE†	$93.30_{1.26}$	$76.74_{1.00}$	$62.98_{1.91}$	$77.58_{0.48}$	$69.95_{0.61}$	$47.05_{0.41}$	$18.08_{0.63}$	$46.89_{0.38}$
	mixup	FC	CE†	$\mathbf{94.98}_{0.30}$	$79.98_{0.70}$	$65.04_{1.58}$	$80.00_{0.54}$	$\mathbf{74.53}_{0.42}$	$50.26_{0.75}$	$17.75_{0.66}$	$49.57_{0.37}$
	arc-mixup	FNO	CE†	$86.00_{0.69}$	$\mathbf{82.80}_{0.56}$	$\mathbf{87.38}_{0.81}$	$\mathbf{85.13}_{0.25}$	$64.38_{2.97}$	$\mathbf{57.81}_{1.77}$	$\mathbf{38.01}_{2.49}$	$\mathbf{54.44}_{2.32}$
imb 10	-	FC	CE†	$94.00_{0.40}$	$84.77_{0.45}$	$84.37_{0.64}$	$87.42_{0.30}$	$72.53_{0.37}$	$61.12_{0.53}$	$44.73_{0.76}$	$60.48_{0.34}$
	mixup	FC	CE†	$\mathbf{95.37}_{0.29}$	$\mathbf{86.81}_{0.47}$	$85.80_{0.89}$	$89.08_{0.32}$	$\mathbf{76.89}_{0.43}$	$64.65_{0.45}$	$46.88_{0.68}$	$\mathbf{63.90}_{0.33}$
	arc-mixup	FNO	CE†	$89.94_{0.55}$	$86.59_{0.53}$	$\mathbf{92.95}_{0.38}$	$\mathbf{89.50}_{0.14}$	$66.26_{5.32}$	$\mathbf{64.70}_{3.06}$	$\mathbf{57.39}_{3.07}$	$63.14_{3.82}$

Table 7: Detailed Results of Imbalanced Learning on Places-LT. All experiments were conducted using 3 trials with random seeds. Best in bold. (Aug: the augmentation method, Clf: the type of classifier, $\mathcal{L}$: the type of loss function) (Accuracy (%): *mean*$_{std}$)

Method			ResNet152				ResNet152 (FT)			
Aug	Clf	$\mathcal{L}$	Many	Median	Few	All	Many	Median	Few	All
-	FC	CE†	$40.63_{0.21}$	$17.69_{0.64}$	$2.04_{0.44}$	$22.62_{0.29}$	$40.77_{0.08}$	$19.99_{0.67}$	$4.41_{0.55}$	$24.16_{0.41}$
mixup	FC	CE†	$\mathbf{42.10}_{0.31}$	$15.82_{0.60}$	$0.86_{0.18}$	$22.10_{0.27}$	$\mathbf{43.14}_{0.63}$	$20.10_{1.61}$	$3.32_{1.33}$	$24.83_{1.19}$
arc-mixup	FNO	CE†	$35.82_{0.32}$	$\mathbf{31.74}_{0.64}$	$\mathbf{16.89}_{0.23}$	$\mathbf{30.07}_{0.40}$	$40.31_{0.18}$	$\mathbf{35.31}_{0.19}$	$\mathbf{20.89}_{0.33}$	$\mathbf{34.06}_{0.10}$

REFERENCES

El Mehdi Achour, François Malgouyres, and Franck Mamalet. Existence, stability and scalability of orthogonal convolutional neural networks. *The Journal of Machine Learning Research*, 23(1): 15743–15798, 2022.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.

Ari S Benjamin, David Rolnick, and Konrad Kording. Measuring and regularizing networks in function space. *arXiv preprint arXiv:1805.08289*, 2018.

Cristian Bodnar, Francesco Di Giovanni, Benjamin Chamberlain, Pietro Liò, and Michael Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. *Advances in Neural Information Processing Systems*, 35:18527–18541, 2022.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pp. 1188–1197, 2021.

Jaehoon Cha and Jeyan Thiyagalingam. Orthogonality-enforced latent space in autoencoders: An approach to learning disentangled representations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3913–3948. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/cha23b.html`.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

Arslan Chaudhry, Naeemullah Khan, Puneet Dokania, and Philip Torr. Continual learning in low-rank orthogonal subspaces. *Advances in Neural Information Processing Systems*, 33:9900–9911, 2020.

Arslan Chaudhry, Albert Gordo, Puneet Dokania, Philip Torr, and David Lopez-Paz. Using hindsight to anchor past knowledge in continual learning. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 6993–7001, 2021.

Vitaliy Chiley, Ilya Sharapov, Atli Kosson, Urs Koster, Ryan Reece, Sofia Samaniego de la Fuente, Vishal Subbiah, and Michael James. Online normalization for training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Hien Dang, Tan Nguyen, Tho Tran, Hung Tran, and Nhat Ho. Neural collapse in deep linear network: From balanced to imbalanced data. *arXiv preprint arXiv:2301.00437*, 2023.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3762–3773. PMLR, 2020.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121*, 2021.

Jarrod Haas, William Yolland, and Bernhard T Rabus. Linking neural collapse and l2 normalization with improved out-of-distribution detection in deep neural networks. *Transactions on Machine Learning Research*, 2022.

XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

Guiqing He, Yincheng Huo, Mingyao He, Haixi Zhang, and Jianping Fan. A novel orthogonality loss for deep hierarchical multi-task learning. *IEEE Access*, 8:67735–67744, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9057–9067, 2022.

Yen-Chang Hsu, Yen-Cheng Liu, Anita Ramasamy, and Zsolt Kira. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*, 2018.

Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. *arXiv preprint arXiv:2001.05992*, 2020.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pp. 2261–2269. IEEE Computer Society, 2017. ISBN 978-1-5386-0457-1. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2017.html#HuangLMW17.

Huaibo Huang, Xiaoqiang Zhou, and Ran He. Orthogonal transformer: An efficient vision transformer backbone with token orthogonalization. *Advances in Neural Information Processing Systems*, 35: 14596–14607, 2022.

Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J. Su. An unconstrained layer-peeled perspective on neural collapse, 2022.

Hoyong Kim, Minchan Kwon, and Kangil Kim. Revisiting softmax masking for stability in continual learning, 2023.

Lingkai Kong, Yuqing Wang, and Molei Tao. Momentum stiefel optimizer, with applications to suitably-orthogonal attention, and optimal transport. *arXiv preprint arXiv:2205.14173*, 2022.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems*, 34:28648–28662, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Lilly Kumari, Shengjie Wang, Tianyi Zhou, and Jeff A Bilmes. Retrospective adversarial replay for continual learning. *Advances in Neural Information Processing Systems*, 35:28530–28544, 2022.

Shuai Li, Kui Jia, Yuxin Wen, Tongliang Liu, and Dacheng Tao. Orthogonal deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1352–1368, 2019.

Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. *arXiv preprint arXiv:2303.10058*, 2023.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Bingchen Liu, Yizhe Zhu, Zuohui Fu, Gerard De Melo, and Ahmed Elgammal. Oogan: Disentangling gan with one-hot sampling and orthogonal regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4836–4843, 2020.

Sheng Liu, Xiao Li, Yuexiang Zhai, Chong You, Zhihui Zhu, Carlos Fernandez-Granda, and Qing Qu. Convolutional normalization: Improving deep convolutional network robustness and training. *Advances in neural information processing systems*, 34:28919–28928, 2021.

Weiyang Liu, Longhui Yu, Adrian Weller, and Bernhard Schölkopf. Generalizing and decoupling neural collapse via hyperspherical uniformity gap. *arXiv preprint arXiv:2303.06484*, 2023.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Jianfeng Lu and Stefan Steinerberger. Neural collapse with cross-entropy loss. *arXiv preprint arXiv:2012.08465*, 2020.

Yao Lu, Stephen Gould, and Thalaiyasingam Ajanthan. Bidirectionally self-normalizing neural networks. *Neural Networks*, 167:283–291, 2023.

Yuexiao Ma, Taisong Jin, Xiawu Zheng, Yan Wang, Huixia Li, Yongjian Wu, Guannan Jiang, Wei Zhang, and Rongrong Ji. Ompq: Orthogonal mixed precision quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9029–9037, 2023.

Tharun Medini, Beidi Chen, and Anshumali Shrivastava. Solar: Sparse orthogonal learned and random embeddings. *arXiv preprint arXiv:2008.13225*, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117.

Federico Pernici, Matteo Bruni, Claudio Baecchi, Francesco Turchini, and Alberto Del Bimbo. Class-incremental learning with pre-allocated fixed classifiers. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6259–6266. IEEE, 2021.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*, 2021.

Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Orthogonal projection loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12333–12343, 2021.

Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A Poggio. Feature learning in deep classifiers through intermediate neural collapse. In *International Conference on Machine Learning*, pp. 28729–28745. PMLR, 2023.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pp. 746–761. Springer, 2020.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pp. 4528–4537. PMLR, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL http://arxiv.org/abs/1409.1556.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

Louis C Tiao, Vincent Dutordoir, and Victor Picheny. Spherical inducing features for orthogonally-decoupled gaussian processes. *arXiv preprint arXiv:2304.14034*, 2023.

Tom Tirer, Haoxiang Huang, and Jonathan Niles-Weed. Perturbation analysis of neural collapse. In *International Conference on Machine Learning*, pp. 34301–34329. PMLR, 2023.

Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. *arXiv preprint arXiv:2104.07167*, 2021.

Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X Yu. Orthogonal convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11505–11515, 2020.

Stephan Wojtowytsch et al. On the emergence of simplex symmetry in the final and penultimate layers of neural network classifiers. *arXiv preprint arXiv:2012.05420*, 2020.

Xia Wu, Xueyuan Xu, Jianhong Liu, Hailing Wang, Bin Hu, and Feiping Nie. Supervised feature selection with orthogonal regression and feature weighting. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):1831–1838, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL `http://arxiv.org/abs/1708.07747`. cite arxiv:1708.07747Comment: Dataset is freely available at https://github.com/zalandoresearch/fashion-mnist Benchmark is available at http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/.

Jing Xu and Haoxiong Liu. Quantifying the variability collapse of neural networks. *arXiv preprint arXiv:2306.03440*, 2023.

Mengjia Xu, Akshay Rangamani, Qianli Liao, Tomer Galanti, and Tomaso Poggio. Dynamics in deep classifiers trained with the square loss: Normalization, low rank, neural collapse, and generalization bounds. *Research*, 6:0024, 2023.

Xiaojun Xu, Linyi Li, and Bo Li. Lot: Layer-wise orthogonal training on improving l2 certified robustness. *Advances in Neural Information Processing Systems*, 35:18904–18915, 2022.

Huanrui Yang, Minxue Tang, Wei Wen, Feng Yan, Daniel Hu, Ang Li, Hai Li, and Yiran Chen. Learning low-rank deep neural networks via singular vector orthogonality regularization and singular value sparsification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 678–679, 2020a.

Liang Yang, Lina Kang, Qiuliang Zhang, Mengzhe Li, Dongxiao He, Zhen Wang, Chuan Wang, Xiaochun Cao, Yuanfang Guo, et al. Open: Orthogonal propagation with ego-network modeling. *Advances in Neural Information Processing Systems*, 35:9249–9261, 2022a.

Yibo Yang, Liang Xie, Shixiang Chen, Xiangtai Li, Zhouchen Lin, and Dacheng Tao. Do we really need a learnable classifier at the end of deep neural network? *arXiv preprint arXiv:2203.09081*, 2022b.

Yongyi Yang, Jacob Steinhardt, and Wei Hu. Are neurons actually collapsed? on the fine-grained structure in neural representations. 2023.

Zebin Yang, Aijun Zhang, and Agus Sudjianto. Enhancing explainability of neural networks through architecture constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6): 2610–2621, 2020b.

Can Yaras, Peng Wang, Zhihui Zhu, Laura Balzano, and Qing Qu. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. *Advances in neural information processing systems*, 35:11547–11560, 2022.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. PMLR, 2017.

Borui Zhang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Bort: Towards explainable neural networks with bounded orthogonal constraint. *arXiv preprint arXiv:2212.09062*, 2022a.

Shaoyu Zhang, Chen Chen, Xiujuan Zhang, and Silong Peng. Label-occurrence-balanced mixup for long-tailed recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3224–3228. IEEE, 2022b.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19550–19560, 2023.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. pp. 1–8, 2020.

Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. *Advances in Neural Information Processing Systems*, 35:31697–31710, 2022.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. 2021.