This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Understanding Imbalanced Semantic Segmentation Through Neural Collapse

Zhisheng Zhong[1,*] Jiequan Cui[1,*] Yibo Yang[2,*] Xiaoyang Wu[3] Xiaojuan Qi[3] Xiangyu Zhang[4] Jiaya Jia[1,5]
CUHK[1] JD Explore Academy[2] HKU[3] MEGVII[4] SmartMore[5]

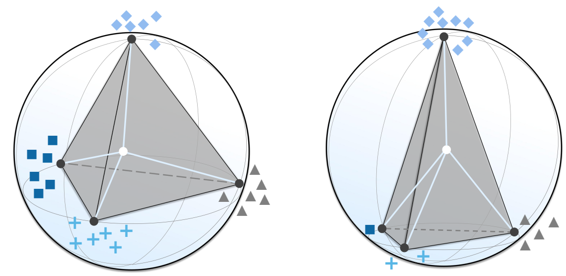
Code: https://github.com/dvlab-research/Imbalanced-Learning

Abstract

A recent study has shown a phenomenon called neural collapse in that the within-class means of features and the classifier weight vectors converge to the vertices of a simplex equiangular tight frame at the terminal phase of training for classification. In this paper, we explore the corresponding structures of the last-layer feature centers and classifiers in semantic segmentation. Based on our empirical and theoretical analysis, we point out that semantic segmentation naturally brings contextual correlation and imbalanced distribution among classes, which breaks the equiangular and maximally separated structure of neural collapse for both feature centers and classifiers. However, such a symmetric structure is beneficial to discrimination for the minor classes. To preserve these advantages, we introduce a regularizer on feature centers to encourage the network to learn features closer to the appealing structure in imbalanced semantic segmentation. Experimental results show that our method can bring significant improvements on both 2D and 3D semantic segmentation benchmarks. Moreover, our method ranks **1**st *and sets a new record (**+6.8%** mIoU) on the ScanNet200 test leaderboard.*

1. Introduction

The solution structures of the last-layer representation and classifier provide a geometric perspective to delve into the learning behaviors in a deep neural network. The neural collapse phenomenon discovered by Papyan et al. [49] reveals that as a classification model is trained towards convergence on a balanced dataset, the last-layer feature centers of all classes will be located on a hyper-sphere with maximal equiangular separation, as known as a simplex equiangular tight frame (ETF), which means that any two centers have an equal cosine similarity, as shown in Fig. 1a. The final classifiers will be formed as the same structure and aligned with the feature centers. The following studies try to theoretically explain this elegant phenomenon, showing



(a) (b)

Figure 1. Illustration of equiangular separation (a) and non-equiangular separation (b) in a 3D space. Neural collapse reveals the structure in (a), where features are collapsed into their within-class centers with ***maximal equiangular separation*** as a simplex ETF, and classifiers are aligned with the same structure. We observe that in semantic segmentation the feature centers and classifiers do not satisfy such a structure, as illustrated in (b) for an example. As some minor class features and classifier vectors lie in a close position, the discriminate ability of the network degrades.

that neural collapse is the global optimality under the cross-entropy (CE) and mean squared error (MSE) loss functions in an approximated model [21, 22, 27, 32, 44, 47, 51, 64, 66, 80]. However, all the current studies on neural collapse focus on the training in image recognition, which performs classification for each image. Semantic segmentation as an important pixel-wise classification problem receives no attention from the neural collapse perspective yet.

In this paper, we explore the solution structures of feature centers and classifiers in semantic segmentation. Surprisingly, it is observed that the symmetric equiangular separation as instructed by the neural collapse phenomenon in image recognition does not hold in semantic segmentation for both feature centers and classifiers. An example of non-equiangular separation is illustrated in Fig. 1b. We point out two reasons that may explain the difference.

First, classification benchmark datasets usually have low correlation among classes. In contrast, different classes in the semantic segmentation task are contextually related. In this case, the classifier needs to be adaptable to class correlation, so does not necessarily equally separate the label space. We conduct a simple experiment to verify it:

* Equal contribution. Part of the work was done in MEGVII.

Classifier	ScanNet200	ADE20K
Learned	**27.8**	**44.5**
Fixed	26.5↓	43.6↓

It is shown that a semantic segmentation model with the classifier fixed as a simplex ETF performs much worse than a learnable classifier. Although using a fixed classifier of the simplex ETF structure has been proven to be effective for image recognition [22, 70, 80], we hold that in semantic segmentation the classifier needs to be learnable and does not have to be equiangular.

Second, the neural collapse phenomenon observed in image recognition highly relies on a balanced class distribution of training samples. It is indicated that neural collapse will be broken when data imbalance emerges, which explains the deteriorated performance of training on imbalanced data [21]. We notice that semantic segmentation naturally suffers from data imbalance because some semantic classes are prone to cover a large area with significantly more points/pixels. Under the point/pixel-wise classification loss, the gradients will be also extremely imbalanced with respect to the backbone parameters, which breaks the equiangular separation structure for feature centers. In this case, the network makes the feature and classifier of minor classes lie in a close position and does not have the ability to discriminate the minor classes. However, the simplex ETF structure in neural collapse renders feature centers ***equiangular separation*** and the ***maximal discriminative*** ability, which is able to effectively improve the performance of minor classes in imbalanced recognition [36, 70, 79].

Inspired by our observations and analyses, we propose to induce the simplex ETF structure for feature centers, but keep a learnable classifier to enable adaptive class correlation for semantic segmentation. To this end, we propose an accompanied center regularization branch that extracts the feature centers of each semantic class. We regularize them by another classifier layer that is fixed as a simplex ETF. The fixed classifier forces feature centers to be aligned with the appealing structure, which enjoys the equiangular separation and the maximal discriminative ability. It in turn helps the feature learning in the original branch to improve the performance of minor classes for better semantic segmentation quality. We also provide theoretical results for a rigorous explanation. Our method can be easily integrated into any segmentation architecture and experimental results also show that our simple method consistently brings improvements on multiple image and point cloud semantic segmentation benchmarks.

Our overall contributions can be listed as follows:

- We are the first to explore neural collapse in semantic segmentation. We show that semantic segmentation naturally brings contextual correlation and imbalanced distribution among classes, which breaks the symmetric structure of neural collapse for both feature centers and classifiers.
- We propose a center collapse regularizer to encourage the network to learn class-equiangular and class-maximally separated structured features for imbalanced semantic segmentation.
- Our method is able to bring significant improvements on both point cloud and image semantic segmentation. Moreover, our method ranks **1**$^{\text{st}}$ and sets a new record (+**6.8** mIoU) on the ScanNet200 test leaderboard.

2. Related Work

Neural collapse. Papyan et al. [49] first discovered the neural collapse phenomenon that at the terminal phase of training, a classification model trained on a balanced dataset will have the last-layer features collapsed into their within class centers. These centers and classifiers will be formed as a simplex equiangular tight frame. Due to its elegant symmetry, later studies try to theoretically unravel such a phenomenon. It is proved that neural collapse is the global optimality of a simplified model with regularization [64, 78, 80], constraint [21, 22, 44, 66], or no explicit constraint [32], under the CE [21, 22, 32, 44, 66, 80] and the MSE loss functions [27, 47, 51, 64, 78]. Some studies try to induce neural collapse in imbalanced learning for better accuracy of minor classes [62, 68, 70]. However, all these studies on neural collapse are limited in recognition. In contrast, we discover the solution structures of the last-layer feature centers and classifiers in semantic segmentation and propose to better induce the equiangular separation state accordingly.

Semantic segmentation. Substantial progress was made for **2D semantic segmentation** with the introduction of FCN [43], which formulates the semantic segmentation task as per-point/pixel classification. Subsequently, many advanced methods have been introduced. Many approaches [2, 4, 38, 52, 58] combine up-sampled high-level feature maps and low-level feature maps to capture global information and recover sharp object boundaries. A large receptive field also plays an important role in semantic segmentation. To capture better global information, many studies [11–14, 63, 71] adopted spatial pyramid pooling to capture multi-scale and larger contextual information. On the other hand, with the introduction of large-scale annotated real-world 3D datasets [3, 6, 20], **3D semantic segmentation** has seen significant focus in recent years. Approaches for point cloud segmentation can be grouped into two categories, *i.e.*, voxel-based and point-based methods. Voxel-based solutions [15–17, 23, 25] first divide the 3D space into regular voxels, and then apply sparse convolutions upon them. Point-based methods [35, 37, 53, 54, 74] directly adopt the point features and positions as inputs, thus keeping the position information intact. However, all these studies on 2D & 3D semantic segmentation mainly focus on network architecture and module design and ignore the impact of

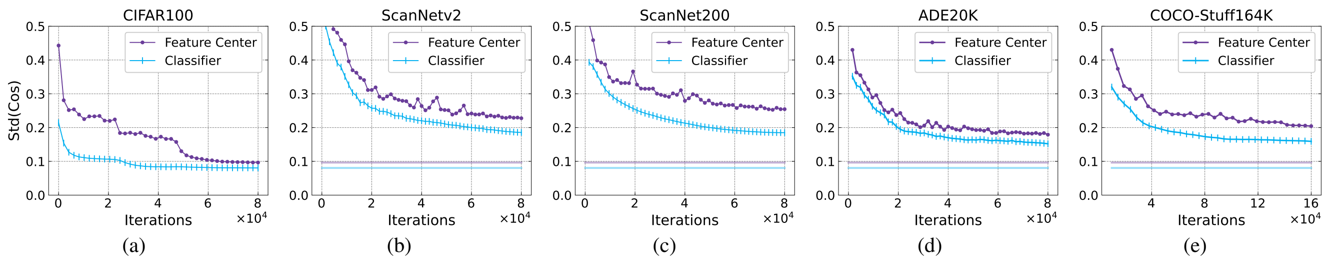


Figure 2. Classifiers and train class-means approach equiangularity for (a) *recognition*, and worse equiangularity for (b-e) *semantic segmentation*. The vertical axis shows the standard deviation of the cosines between pairs of centered class means (**purple** lines) and classifiers (**blue** lines) across all distinct pairs of classes k and k'. Mathematically, denote $\mathrm{Std}_{k\neq k'}(\cos(\hat{\mathbf{z}}_k, \hat{\mathbf{z}}_{k'}))$ and $\mathrm{Std}_{k\neq k'}(\cos(\hat{\mathbf{w}}_k, \hat{\mathbf{w}}_{k'}))$. As training progresses, the standard deviations of the cosines approach zero indicating equiangularity.

data distribution. As the datasets of semantic segmentation usually and naturally follow a heavily imbalanced distribution among classes, neural networks perform poorly when training on them [18, 33, 45, 57, 76]. With the available large-scale imbalanced segmentation datasets, such as ScanNet200 [59], ADE20K [77], COCO [26, 40], exploring segmentation from an **imbalanced learning** view draws more attention recently.

3. Neural Collapse Observations

3.1. Neural Collapse in Recognition

We consider a dataset, having N training samples in total and the annotations $\mathbf{y} \in \mathbb{R}^N$. More concretely, it has K classes and n_k examples in k-th class. We refer to $\mathbf{y}_i \in \{1, ..., K\}$ as the label, $\mathbf{z}_i \in \mathbb{R}^d$ as the last-layer d-dimensional feature of the i-th sample. We define that $\bar{\mathbf{z}}_k = \mathrm{Avg}_{\mathbf{y}_i=k}\{\mathbf{z}_i\}$ is the within-class mean of the last-layer features in the k-th class. The linear classifier is specified by weights $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_K] \in \mathbb{R}^{d\times K}$.

Papyan et al. [49] revealed the neural collapse phenomenon for image recognition. It states that the last-layer features will converge to their within-class means, and the within-class means together with the classifiers will collapse to the vertices of a simplex equiangular tight frame at the terminal phase of training (after 0 training error rate).

Definition 1 (Simplex Equiangular Tight Frame) *A collection of vectors $\mathbf{m}_k \in \mathbb{R}^d$, $k = 1, 2, ..., K$, $d \geq K$, is said to be a simplex equiangular tight frame if:*

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

where $\mathbf{M} = [\mathbf{m}_1, ..., \mathbf{m}_K] \in \mathbb{R}^{d\times K}$, $\mathbf{U} \in \mathbb{R}^{d\times K}$ allows a rotation and satisfies $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_K$, $\mathbf{I}_K$ is the identity matrix, and $\mathbf{1}_K$ is an all-ones vector.

All vectors in a simplex ETF have an equal ℓ_2 norm and the same pair-wise angle, *i.e.,*

$$\mathbf{m}_i^\top\mathbf{m}_j = \frac{K}{K-1}\delta_{i,j} - \frac{1}{K-1}, \forall i, j \in \{1, ..., K\}, \qquad (2)$$

where $\delta_{i,j}$ equals to 1 when $i = j$ and 0 otherwise. The pair-wise angle $-\frac{1}{K-1}$ is the maximal equiangular separation of K vectors in $\mathbb{R}^d$, $d \geq K-1$ [21, 50, 70].

Then the two important geometric properties instructed by neural collapse can be formally described as: (1) The normalized within-class centers converge to a simplex ETF, *i.e.,* $\hat{\mathbf{z}}_k = (\bar{\mathbf{z}}_k - \mathbf{z}_G)/||\bar{\mathbf{z}}_k - \mathbf{z}_G||$ satisfies Eq. (2), where $\mathbf{z}_G = \mathrm{Avg}_i\{\mathbf{z}_i\}$ is the global mean of the last-layer features for all samples; (2) The normalized classifier vectors converge to the same simplex ETF as feature centers, *i.e.,* $\hat{\mathbf{w}}_k = \mathbf{w}_k/||\mathbf{w}_k|| = \hat{\mathbf{z}}_k$ and satisfies Eq. (2), where $\mathbf{w}_k$ is the classifier of the k-th class.

3.2. Neural Collapse in Semantic Segmentation

The elegant phenomenon has only been discovered and studied in image recognition, which performs image-wise classification. In this paper, we explore the corresponding structures of the last-layer feature centers and classifiers in image and point cloud semantic segmentation, which performs pixel- and point-wise classification, respectively.

Following [49], we calculate statistics during training to show the neural collapse convergence in semantic segmentation. We first compare the standard deviations of two cosine similarities, $\cos(\hat{\mathbf{z}}_k, \hat{\mathbf{z}}_{k'})$ and $\cos(\hat{\mathbf{w}}_k, \hat{\mathbf{w}}_{k'})$, for all pairs of different classes $k \neq k'$. The standard deviations of the cosines approach zero indicating equiangularity of the feature centers and classifier weights. As shown in Fig. 2, their standard deviations are much larger on semantic segmentation (Fig. 2b to 2e, light-colored lines for CIFAR100 results), compared with them on image recognition (Fig. 2a) as observed by Papyan et al. [49]. It indicates that the equiangular structure of the feature centers and classifiers in semantic segmentation is not as valid as that in image recognition. Similarly, the feature centers and classifiers approach the maximal-angle structure more closely in classification than in semantic segmentation. More analysis and experiments about the maximal separation structure of the feature centers and classifiers are shown in *Appendix* A.

We point out that semantic segmentation naturally brings contextual correlation and imbalanced distribution among

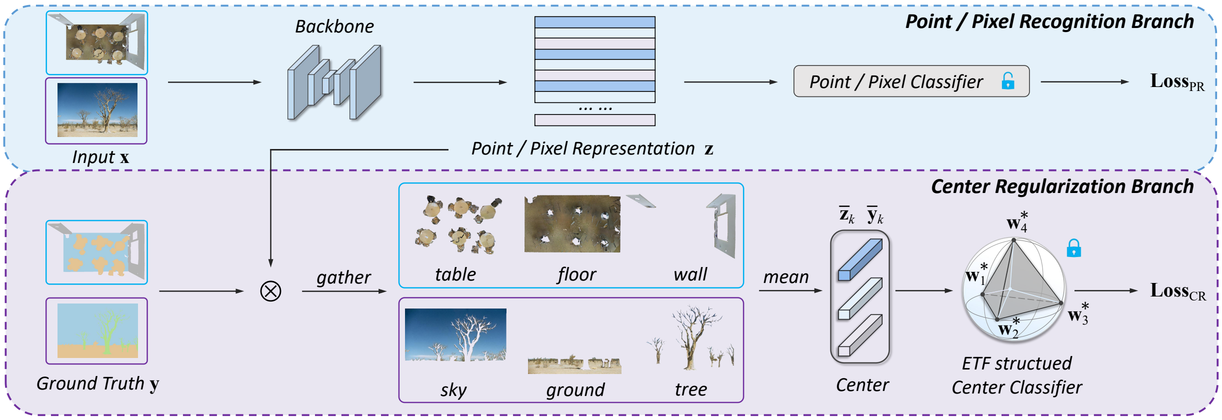


Figure 3. Framework of our method. The **blue** square part and **purple** square part represent 3D scene input illustration and 2D image input illustration, respectively. The point/pixel recognition branch is similar to the conventional segmentation model. We introduce a center regularization branch to make the feature center collapse to an ETF structure during training and remove it during evaluation.

classes, which may break the symmetric structure of neural collapse for both feature centers and classifiers. A learnable classifier will lead to a non-equiangular structure to be *adaptable* to class contextual correlation. But the class imbalance will cause imbalanced gradients. As a result, the feature centers for minor classes will be closed after training, which impedes their performance. A rigorous analysis is conducted in Sec. 4.4.

4. Main Approach

4.1. Motivation

As said in Sec. 1, although the classifier does not have to be equiangular, we note that the ETF structure in Eq. (1) is appealing for feature centers in a classification problem, especially when the learning is imbalanced. First, such a structure has a balanced separation among all classes and the separation is maximally enlarged. We formally state this property in Lemma 1. Second, as suggested by prior studies, inducing neural collapse in imbalanced learning is able to improve the performance of minor classes [62, 68, 70].

Lemma 1 ***(Equiangular & Maximal Separated Property)*** *For any normalized matrix,* $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_K] \in \mathbb{R}^{d\times K}, d \geq K, \mathbf{w}_k^\top \mathbf{w}_k = 1, \forall k = 1, ..., K$*, the maximal separation value is* $-\frac{1}{K-1}$*, i.e.,* $\max_{k\neq k'} \cos(\mathbf{w}_k, \mathbf{w}_{k'}) \geq -\frac{1}{K-1}$*. Equality holds if and only if the matrix is a simplex ETF, i.e.,* $\mathbf{W}$ *satisfies Eq. (1) and it enjoys the equiangular property, i.e.,* $\forall k \neq k', \cos(\mathbf{w}_k, \mathbf{w}_{k'})$ *is a constant.*

Proof 1 *Please refer to our Appendix B for proof.* □

Based on our observations and analyses, we propose to regularize the feature centers in a segmentation model into the simplex ETF structure in Eq. (1) to relieve the imbalance dilemma for semantic segmentation.

4.2. Center Collapse Regularizer

To take advantage of the equiangular and maximum separation properties for better performance on minor classes, we propose a **Ce**nter **Co**llapse Regularizer (**CeCo**) for imbalanced semantic segmentation problems. The overview of our method is shown in Fig. 3. The whole framework can be divided into two branches: the point/pixel recognition branch (upper part) and the center regularization branch (bottom part). The point/pixel recognition branch refers to a point cloud or image segmentation model, dealing with the semantic segmentation task in a point/pixel level manner. Therefore, our CeCo can be easily integrated into any off-the-shelf segmentation architecture.

For simplicity, we use a 3D scene (**blue** square part) and an image (**purple** square part) in Fig. 3 as an example to illustrate the point cloud and image semantic segmentation, respectively. The input is represented as $\mathbf{x} \in \mathbb{R}^{N\times s}$, where $N = HW$ is the number of pixels and $s = 3$ denotes the color dimension for the image case. For point cloud, N is the number of points, and $s = 6$ is for both the color and the position dimension. As shown in the point/pixel recognition branch, we can get the feature representation $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_N]^\top \in \mathbb{R}^{N\times d}$ after the backbone feature extraction. The loss $\mathcal{L}_{\text{PR}}$ for the point/pixel recognition branch is a point/pixel-wise CE for supervised learning.

In the center regularization branch, we first gather $\mathbf{z}_i$ of the same class, compute the feature centers $\bar{\mathbf{z}}_k$, and generate center labels $\bar{\mathbf{y}}_k$ of all classes based on the ground truth $\mathbf{y}$:

$$\bar{\mathbf{z}}_k = \frac{1}{n_k} \sum_{\mathbf{y}_i=k}^{n_k} \mathbf{z}_i, \;\; \bar{\mathbf{y}}_k = \mathbf{y}_i = k, \tag{3}$$

where n_k is the number of samples in $\mathbf{Z}$ belonging to the k-th class. We concatenate feature centers $\bar{\mathbf{Z}} = [\bar{\mathbf{z}}_1, ..., \bar{\mathbf{z}}_K] \in$

$\mathbb{R}^{d\times K}$ and center labels $\bar{\mathbf{y}}=[1,...,K]\in\mathbb{R}^K$ for the center regularization branch. To achieve center collapse, we introduce an ETF structured classifier for this branch. Another CE-based center collapse regularization loss $\mathcal{L}_{\text{CR}}$ upon feature centers $\bar{\mathbf{Z}}$ and their center labels $\bar{\mathbf{y}}$ is used to measure the degree of feature center collapse. Concretely, we initialize the classifier $\mathbf{W}^*$ as a random simplex ETF by Eq. (1). During training, the classifier is fixed to make the maximal equiangular separation property satisfied all the time, *i.e.*,

$$\mathbf{w}_k^{*\top}\mathbf{w}_{k'}^* = \alpha^2\left(\frac{K\delta_{k,k'}}{K-1}-\frac{1}{K-1}\right),\ \forall k,\ k'\in\{1,...,K\},$$

where α is a hyper-parameter of weight scaling, and $\delta_{k,k'}$ equals to 1 when $k=k'$ and 0 otherwise. Then, a CE type of $\mathcal{L}_{\text{CR}}$ loss can be written as follows:

$$\mathcal{L}_{\text{CR}}(\bar{\mathbf{Z}},\mathbf{W}^*) = -\sum_{k=1}^{K}\log\left(\frac{\exp\left(\bar{\mathbf{z}}_k^\top\mathbf{w}_k^*\right)}{\sum_{k'=1}^{K}\exp\left(\bar{\mathbf{z}}_{k'}^\top\mathbf{w}_{k'}^*\right)}\right). \quad (4)$$

We define the total loss as the combination of two branches:

$$\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{PR}}(\mathbf{Z},\mathbf{y})+\lambda\mathcal{L}_{\text{CR}}(\bar{\mathbf{Z}},\mathbf{W}^*), \quad (5)$$

where λ is a loss weight hyper-parameter.

Additionally, in the evaluation of our method, we only preserve the point/pixel recognition branch. It means that just the point/pixel classifier is preserved, while the center regularization branch is discarded. Therefore, the evaluation of CeCo is very efficient: It is consistent with a conventional backbone like the vanilla ResNet [28], without any additional computations.

4.3. Empirical Support

In this subsection, we give some empirical evidence to show that our CeCo can do better rebalance on semantic segmentation. Due to limited pages, the detailed results and evidence are shown in *Appendix* C. We list the main conclusions in the following:

First, the center regularization branch in CeCo transforms each point/pixel pair $(\mathbf{z}_i,\mathbf{y}_i)$ of the k-th class to a center pair $(\bar{\mathbf{z}}_k,\bar{\mathbf{y}}_k)$, which greatly decreases the imbalanced factor [19, 42]: (The imbalanced factor is defined as $\frac{n_{\max}}{n_{\min}}$, where $n_{\max}$ and $n_{\min}$ are the maximal and minimal numbers of samples in all classes, respectively.)

ScanNetv2	ScanNet200	ADE20K	COCO-Stuff
(point) 116	(point) 37256	(pixel) 827	(pixel) 2612
(center) **11**$_\downarrow$	(center) **597**$_\downarrow$	(center) **282**$_\downarrow$	(center) **528**$_\downarrow$

As the imbalance severity of the dataset is significantly reduced, it becomes more friendly to the minor classes.

Second, in long-tailed classification, the class accuracy is *positively correlated* with the class image number of the training dataset. However, in both point cloud and image semantic segmentation, the class accuracy has *weak correlations* with class point/pixel numbers due to correlations among neighboring. The widely used correlation measurement, Pearson correlation coefficients [7] between the class accuracy and the sample numbers in each class:

CIFAR100-LT-100	ScanNet200	ADE20K
(image) 0.76	(point) 0.32	(pixel) 0.31
–	(center) **0.51**$_\uparrow$	(center) **0.49**$_\uparrow$

However, most of the imbalanced recognition methods are using the sample numbers of classes for guidance, *e.g.*, reweighting [19, 30, 31, 61] and loss adjustment [10, 45, 57, 76]. Due to the low correlations for the point and pixel cases, it is probably not feasible for imbalanced semantic segmentation. In contrast, CeCo regularizes imbalance in a feature center space. Feature centers are more global representation and greatly eliminate the effects of correlations among neighboring. Thus, it has a *better correlation* with class accuracy than the point/pixel frequency, further indicating its suitability for semantic segmentation rebalancing.

4.4. Theoretical Support

In this part, we rethink the CE-based center collapse loss $\mathcal{L}_{\text{CR}}$ from the perspective of gradients to analyze its imbalanced learning behaviors.

Gradient *w.r.t* the center classifier. Recalling the center computation Eq. (3), we first analyze the gradient of $\mathcal{L}_{\text{CR}}$ *w.r.t* the center classifier $\mathbf{W}=[\mathbf{w}_1,\cdots,\mathbf{w}_K]\in\mathbb{R}^{d\times K}$:

$$\begin{aligned}\frac{\partial\mathcal{L}_{\text{CR}}}{\partial\mathbf{w}_k} &= \left(p_k\left(\bar{\mathbf{z}}_k\right)-1\right)\bar{\mathbf{z}}_k+\sum_{k'\neq k}^{K-1}p_k\left(\bar{\mathbf{z}}_{k'}\right)\bar{\mathbf{z}}_{k'}, \quad (6)\\ &= \underbrace{\sum_{\mathbf{y}_i=k}^{n_k}\left(p_k\left(\bar{\mathbf{z}}_k\right)-1\right)\frac{\mathbf{z}_i}{n_k}}_{\text{within-class}}+\underbrace{\sum_{k'\neq k}^{K-1}\sum_{\mathbf{y}_j=k'}^{n_k}p_k\left(\bar{\mathbf{z}}_{k'}\right)\frac{\mathbf{z}_j}{n_{k'}}}_{\text{between-class}},\end{aligned}$$

where $p_k(\bar{\mathbf{z}})$ is the predicted probability that $\bar{\mathbf{z}}$ belongs to the k-th class. It is calculated by the softmax transformation and takes the following form in the CE loss:

$$p_k(\bar{\mathbf{z}})=\frac{\exp\left(\bar{\mathbf{z}}^\top\mathbf{w}_k\right)}{\sum_{k'=1}^{K}\exp\left(\bar{\mathbf{z}}^\top\mathbf{w}_{k'}\right)},\ k=1,2,...,K. \quad (7)$$

It reveals that the gradient *w.r.t* $\mathbf{w}_k$ is also *imbalanced* and can be decomposed into two parts. The "within-class" part contains n_k terms and pulls $\mathbf{w}_k$ towards the directions of the same class feature center, *i.e.*, $\bar{\mathbf{z}}_k$. While the "between-class" part contains $\sum_{k'\neq k}n_{k'}$ terms and pushes $\mathbf{w}_k$ away from the directions of the features of the other classes. Thus, the gradients of some minor classes can be more likely swallowed by the other classes, which is unfriendly to the decision boundaries of minor classes. If we fix the center classifier weights, it can avoid the imbalance gradient updating and benefit the minor classes discrimination.

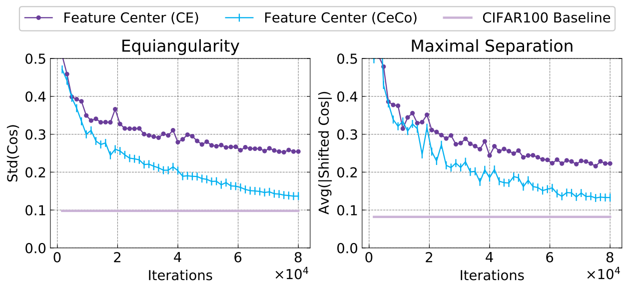


Figure 4. Ablation on the center equiangularity (a) and maximal separation (b) on ScanNet200. Adopting CeCo, the feature centers can obtain a better equiangular and maximally separated structure. It is better to look together with Fig. 2c.

Gradient *w.r.t* point/pixel features. Since the center collapse loss $\mathcal{L}_{\text{CR}}$ is a regularization upon the point/pixel feature $\mathbf{z}_i$, we analyze how the center collapse loss influences the point/pixel feature also from the gradient view. Suppose the class label $\mathbf{y}_i = k$, $\mathbf{z}_i$ contributes to $\bar{\mathbf{z}}_k$. The gradient of $\mathcal{L}_{\text{CR}}$ with respect to $\mathbf{z}_i$ is:

$$\frac{\partial \mathcal{L}_{\text{CR}}}{\partial \mathbf{z}_i} = \frac{\partial \mathcal{L}_{\text{CR}}}{\partial \bar{\mathbf{z}}_k} \cdot \frac{\partial \bar{\mathbf{z}}_k}{\partial \mathbf{z}_i} = \sum_{k'=1}^{K} p_{k'}(\bar{\mathbf{z}}_k)(\mathbf{w}_{k'} - \mathbf{w}_k) \cdot \frac{1}{n_k}. \quad (8)$$

For Eq.(8), we mainly consider the medium term, *i.e.*, $\mathbf{w}_{k'} - \mathbf{w}_k$. In [21], the authors describe a minority collapse phenomenon that the classifier weight vectors of minor classes converge in a similar direction under the extreme imbalance case, *i.e.*, $\lim_{\frac{n_{\max}}{n_k} \to \infty, \frac{n_{\max}}{n_{k'}} \to \infty} \mathbf{w}_{k'} - \mathbf{w}_k = \mathbf{0}_d$. If the classifier is fixed ETF structured, $\|\mathbf{w}_{k'} - \mathbf{w}_k\| = \frac{2K}{K-1}, \forall k \neq k'$, is always satisfied during training, according to the ETF property Eq. (2). It means that ETF structured classifier enables the gradient transfer between minor classes and obtains *better discrimination among minor classes*. Thus it avoids the features of minor classes converging in a close position (see Fig. 1b).

5. Experiments

5.1. Datasets and Implementation Details

To evaluate the effectiveness of our method, we conduct experiments on the most popular benchmarks in semantic segmentation: *i.e.*, ScanNet200 [59] for point cloud semantic segmentation, ADE20K [77] and COCO-Stuff164K [9] for image semantic segmentation. All these three datasets contain more than *150 classes*, which is more suitable to validate the imbalanced performance of the proposed method under severe imbalance cases. Implementation details and dataset descriptions are available in *Appendix* D.

5.2. Ablation Study

In this subsection, we conduct ablation experiments to examine the effectiveness of CeCo. The ablation results are reported on the ScanNet200 validation dataset with the MinkowskiNet [15] backbone for 3D semantic segmentation, and on ADE20K (single-scale inference mIoU) with the Swin-T [41] backbone for 2D semantic segmentation.

PC	CC	ScanNet200	ADE20K
Learned	–	27.8	44.5
Fixed	Fixed	28.3	44.8
Fixed	Learned	27.2	44.0
Learned	**Fixed**	**30.3** (+2.5)	**45.7** (+1.2)
Learned	Learned	26.9	43.7

Table 1. Ablation on the fixed ETF structured or learned classifier. PC: point/pixel classifier. CC: feature center classifier.

Loss Weight	ScanNet200	ADE20K
λ = 0.0 (Baseline)	27.8	44.5
λ = 0.1	28.6	44.8
λ = 0.2	29.4	45.0
λ = 0.3	30.1	45.2
λ = 0.4	**30.3** (+2.5)	45.3
λ = 0.5	29.7	**45.7** (+1.2)
λ = 0.6	28.7	44.9

Table 2. Ablation on the loss weight hyper-parameter λ.

Equiangularity and maximal separation analysis. We first analyze the equiangularity property of the feature centers. Following [49] and Sec. 3.2, we calculate the standard deviation of the cosines between pairs of centered class means across all distinct pairs of classes k and k'. Mathematically, we measure $\text{Std}_{k \neq k'}(\cos(\hat{\mathbf{z}}_k, \hat{\mathbf{z}}_k))$ and $\text{Avg}_{k \neq k'}|\cos(\hat{\mathbf{z}}_k, \hat{\mathbf{z}}_k) + 1/(K-1)|$. Fig. 4a and 4b show the change of standard deviation and average at different iterations. As training progresses, the standard deviations of the cosines approach zero indicating equiangularity, and the shifted averages of the cosines approach zero indicating maximal angle. CeCo (**blue** lines) introduces a center collapse regularization and achieves a much *smaller* standard deviation and shifted average values than the vanilla CE model (**purple** lines), and are *closer* to the classification neural collapse results (**light-colored** lines for reference).

Ablation on the fixed ETF structured center classifier. As discussed in Sec. 4, a fixed ETF structured center classifier can bring many benefits under severely imbalanced cases. In this part, we conduct experiments to answer the following question: For both the point/pixel classifier and center classifier, should we make them fixed or learnable? We list the performance results for four kinds of variants in Table 1. The fixed ETF structured center classifier induces a strong regularization for feature center distribution (equiangular and maximal separated) and

Loss Type	ScanNet200	ADE20K
CE (Baseline)	27.8	44.5
+ Dice [46]	28.8	44.8
+ Dice + **CeCo**	**30.6** (+1.8)	**45.8** (+1.0)
+ Lovász [8]	30.0	45.0
+ Lovász + **CeCo**	**32.0** (+2.0)	**46.5** (+1.5)

Table 3. Ablations on orthogonality to the Dice and Lovász loss.

Method	Head	Comm.	Tail	**All**
DLV3P (R50)	67.7	48.3	36.4	44.9
+ DisAlign	67.7	48.6	37.8	45.7
+ **CeCo**	67.7	**48.7** (+0.1)	**39.0** (+1.2)	**46.4** (+0.7)
DLV3P (R101)	68.7	49.0	38.4	46.4
+ DisAlign	68.7	49.4	39.6	47.1
+ **CeCo**	**68.8** (+0.1)	49.4	**40.9** (+1.3)	**48.0** (+0.9)

Table 4. Imbalanced performance comparison of our method and the SOTA long-tailed segmentation framework DisAlign [73] on ADE20K. All compared methods are based on DLV3P [14] with the ResNet-50 (R50) and ResNet-101 (R101) backbone.

Method	Head	Comm.	Tail	**All**
Minkowski. [15]	48.3	19.1	7.9	25.1
Ins. Samp. [59]	48.2	18.9	9.2	25.4
C-Focal [19,39]	48.1	20.2	9.3	25.8
SupCon [34]	48.5	19.1	10.3	26.0
CSC [29]	49.4	19.5	10.3	26.5
LG (CLIP) [59]	50.4	22.8	10.1	27.7
LG [59]	51.5	22.7	12.5	28.9
CeCo	51.2	22.9 (+0.2)	17.1 (+4.6)	30.3 (+1.4)
CeCo (Lovász)	**52.4** (+0.9)	**26.2** (+3.5)	**17.9** (+5.4)	**32.0** (+3.1)

Table 5. mIoU comparison on the ScanNet200 validation sets.

Method	Head	Comm.	Tail	**All**
Minkowski. [15]	46.3	15.4	10.2	25.3
CSC [29]	45.5	17.1	7.9	24.9
LG [59]	48.5	18.4	10.6	27.2
CeCo (Lovász)	52.1 (+3.6)	23.6 (+5.2)	15.2 (+4.6)	31.7 (+4.5)
CeCo*(Lovász)	**55.1** (+6.6)	**24.7** (+6.3)	**18.1** (+7.5)	**34.0** (+6.8)

Table 6. Comparison with other methods on the ScanNet200 test set. All numbers are from the benchmark on 11$^{\text{th}}$ November 2022.

improves the mIoU performance by 2.6%, 1.2% on ScanNet200 and ADE20K compared with the based model, respectively. Consistent with our analysis in Sec. 1 and similar to the conventional semantic segmentation models, a learnable point/pixel classifier can achieve better results. it is *adaptive* and dynamically updated and hence can better handle a finer point/pixel-level classification.

Ablation on the loss weight hyper-parameter λ. In Eq. (5), we introduce the hyper-parameter λ for the weighting between the loss function of the point/pixel recognition branch $\mathcal{L}_{\rm PR}$ and the loss function of the center regularization branch $\mathcal{L}_{\rm CR}$. To show the sensitivity of our CeCo to different λ values, we conduct an ablation on ScanNet200 and ADE20K. Table 2 lists the experimental results, showing that the performance can be consistently improved with the value of λ in a wide range. According to Table 2, $\lambda = 0.4$ can achieve the best mIoU results among others on ScanNet200, and $\lambda = 0.5$ can achieve the best mIoU results among others on ADE20K.

Imbalanced performance. To evaluate the imbalanced performance of our method, We also follow [59,73] to split the categories of ScanNet200 and ADE20K into three subsets and report the average IoU and accuracy in these three subsets: head-shot, medium-shot, and few-shot, which are also called the *Head*, *Common* and *Tail* categories, respectively. We list the detailed results of ADE20K in Table 4 and ScanNet200 in Table 5. For ADE20K, we mainly compare our method with the plain DeepLabV3+ [14] (DLV3P) model and the state-of-the-art (SOTA) long-tailed segmentation framework DisAlign [73]. In Table 4, our method further outperforms the baseline and DisAlign on both the ResNet-50 (0.7% mIoU improvement), and ResNet-101 (0.9% mIoU improvement) backbone. For ScanNet200, our CeCo achieves **3.1%** mIoU improvement in mIoU using MinkowskiNet-34 [15] compared with previous methods like Contrastive Scene Contexts (CSC) [29] and CLIP [55] based language grounded model (LG) [59]. The performance of the common (**3.6%** mIoU improvement) and tail (**5.4%** mIoU improvement) are improved significantly.

Orthogonality to segmentation losses. By now, many segmentation losses, *e.g.*, Focal [39], Dice [46], SoftIOU [56], SoftTversky [60], and Lovász [8], have been proposed for improving the segmentation results. As our method is a novel regularization for neural collapse convergence, in this part, we mainly verify its orthogonality to three famous segmentation losses, Focal loss, Dice loss, and Lovász loss. As shown in Table 3 and Table 5 (the third row), CeCo can be jointly trained with these three segmentation losses and further improve their performance by **1.0-3.0%**. The above experiments clearly verify the orthogonality of CeCo to these wide-used segmentation losses.

5.3. Main Results

Comparison on ScanNet200. To show the powerful rebalance performance of our method, we compare with a SOTA point cloud pre-training approaches CSC [29] and Supervised Contrastive Learning (SupCon) [34], along with SOTA LG [59] in Table 5 and 6. Following [59], we use the same 3D MinkowskiNet [15] backbone and training setting for a fair comparison. For both the validation dataset

Method	Backbone	mIoU (s.s.)	mIoU (m.s.)
OCRNet	HRNet-W18	39.3	40.8
+ **CeCo**	HRNet-W18	**41.8** (+2.5)	**43.5** (+2.7)
DLV3P	ResNet-50	43.9	44.9
+ **CeCo**	ResNet-50	**45.0** (+1.1)	**46.4** (+1.5)
OCRNet	HRNet-W48	43.2	44.9
+ **CeCo**	HRNet-W48	**44.5** (+1.3)	**46.1** (+1.2)
UperNet	ResNet-101	43.8	44.8
+ **CeCo**	ResNet-101	**44.8** (+1.0)	**46.1** (+1.3)
DLV3P	ResNet-101	45.5	46.4
+ **CeCo**	ResNet-101	**46.7** (+1.2)	**48.0** (+1.6)
UperNet	Swin-T	44.5	45.8
+ **CeCo**	Swin-T	**45.7** (+1.2)	**47.6** (+1.8)
UperNet	Swin-B	50.0	51.7
+ **CeCo**	Swin-B	**51.2** (+1.2)	**52.9** (+1.2)
UperNet	BEiT-L	56.7	57.0
+ **CeCo**	BEiT-L	**57.3** (+0.6)	**57.7** (+0.7)

Table 7. Comparison on ADE20K.

Method	Backbone	mIoU (s.s.)	mIoU (m.s.)
OCRNet	HRNet-W18	31.6	32.4
+ **CeCo**	HRNet-W18	**38.0** (+6.4)	**38.8** (+6.4)
UPerNet	ResNet-50	39.9	40.3
+ **CeCo**	ResNet-50	**41.2** (+1.3)	**41.7** (+1.4)
DLV3P	ResNet-50	40.9	41.5
+ **CeCo**	ResNet-50	**42.7** (+1.8)	**43.5** (+2.0)
OCRNet	HRNet-W48	40.4	41.7
+ **CeCo**	HRNet-W48	**41.8** (+1.4)	**43.3** (+1.6)
UperNet	ResNet-101	41.2	41.5
+ **CeCo**	ResNet-101	**42.3** (+1.1)	**42.8** (+1.3)
DLV3P	ResNet-101	42.4	43.0
+ **CeCo**	ResNet-101	**43.9** (+1.5)	**44.6** (+1.6)
UperNet	Swin-T	43.8	44.6
+ **CeCo**	Swin-T	**44.5** (+0.7)	**45.2** (+0.6)
UperNet	Swin-B	47.7	48.6
+ **CeCo**	Swin-B	**48.2** (+0.5)	**49.2** (+0.6)

Table 8. Comparison on COCOStuff-164K.

and the test dataset, our CeCo consistently surpasses previous methods by a large margin on all split IoU measurements. Concretely, CeCo outperforms the previous best by 0.9%, **3.6%**, **5.4%**, and **3.1%** on the ScanNet200 validation dataset, **3.6%**, **5.2%**, **4.6%**, and **4.5%** on the ScanNet200 test dataset, under the head, common, tail, and all measurements, respectively. Following Mix3D [48], the SOTA 3D model on ScanNetv2, we also report the ensemble results of three CeCo models (denoted as CeCo*) for the test set. Our CeCo yields the highest mIoU and ranks **1**$^{\text{st}}$ on the ScanNet200 test leaderboard. Because CeCo enables the network to learn equiangular and more separated feature distribution, which brings great improvements over baselines and across common and tail categories.

Comparison on ADE20K. To show the flexibility of CeCo, we experiment on ADE20K with various semantic segmentation head methods, *e.g*., UperNet [67], OCRNet [72], and DLV3P, and different backbones, *e.g*., ResNet [28], HRNet [65]. We report the performance of both single-scale (s.s.) inference and multi-scale (m.s.) inference. As shown in Table 7, plugging our CeCo into those methods leads to significant improvements. Specifically, for ResNet-101 or HRNet-W48, after training with CeCo, there are 1.3%, 1.2%, and 1.6% gains for UperNet, OCRNet, and DLV3P respectively. We also experiment with our method on awesome transformers like Swin [41] and BEiT [5]. For CNN-based models, we achieve 48.0% mIoU with ResNet-101, surpassing the baseline by 1.6. With BEiT, the performance is improved from 57.0% mIoU to 57.7% mIoU.

Comparison on COCO-Stuff164K. On the large-scale COCO-Stuff164K dataset, we again demonstrate the flexibility of our CeCo. The experimental results are summarized in Table 8. Equipped with CeCo in training, CNN-based models, *i.e*., ResNets and HRNets, surpass their baselines by a large margin. Specifically, with HRNet-18 and OCRNet, our trained model outperforms the baseline by **6.4%** mIoU. With the large CNN-based ResNet-101 and DLV3P, our model achieves 44.6% mIoU, surpassing the baseline by 1.6% mIoU. We also verify the effectiveness of CeCo with the Swin transformer on COCO-Stuff164K. Experimental results with Swin-T and Swin-B show clear improvements after adopting our center collapse regularization in the training phase. Both Table 7 and Table 8 demonstrate the effectiveness and generalization of CeCo on various backbones and segmentation head methods.

Visual comparison. More visual comparisons over the above three datasets are included in *Appendix* E. We observe that CeCo obtains more precise semantic segmentation masks for both common and tail classes.

6. Conclusion

In this paper, we explore the neural collapse structures of feature centers and classifiers for semantic segmentation. Semantic segmentation naturally brings contextual correlation and imbalanced distribution among classes. It breaks the equiangular and maximal separated structure of neural collapse for both feature centers and classifiers. To preserve these properties for minor classes, we introduce a center collapse loss to regularize the feature centers. Adopting the new regularization, feature centers become more symmetric and class-separated and the performance of minor classes is greatly improved. We hope that our findings could advance future studies of 2D & 3D imbalanced semantic segmentation. Limitation analysis is provided in *Appendix* F.

References

[1] MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[2] Md Amirul Islam, Mrigank Rochan, Neil DB Bruce, and Yang Wang. Gated feedback refinement network for dense image labeling. In *CVPR*, 2017.

[3] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. In *CVPR*, 2016.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 2017.

[5] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022.

[6] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *ICCV*, 2019.

[7] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[8] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The Lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, pages 4413–4421, 2018.

[9] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018.

[10] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, volume 32, 2019.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015.

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE TPAMI*, 2017.

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[15] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019.

[16] Ruihang Chu, Yukang Chen, Tao Kong, Lu Qi, and Lei Li. ICM-3D: Instantiated category modeling for 3D instance segmentation. *IEEE Robotics and Automation Letters*, 7(1):57–64, 2021.

[17] Ruihang Chu, Xiaoqing Ye, Zhengzhe Liu, Xiao Tan, Xiaojuan Qi, Chi-Wing Fu, and Jiaya Jia. TWIST: Two-way inter-label self-training for semi-supervised 3D instance segmentation. In *CVPR*, pages 1100–1109, 2022.

[18] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, 2021.

[19] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019.

[20] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.

[21] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

[22] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *ICML*, pages 3821–3830. PMLR, 2021.

[23] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018.

[24] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018.

[25] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv:1706.01307*, 2017.

[26] Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

[27] XY Han, Vardan Papyan, and David L Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *ICLR*, 2022.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[29] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3D scene understanding with contrastive scene contexts. In *CVPR*, pages 15587–15597, 2021.

[30] Chen Huang, Yining Li, Change Loy Chen, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE TPAMI*, 2019.

[31] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *CVPR*, 2020.

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

[33] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

[34] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and

Dilip Krishnan. Supervised contrastive learning. In *NIPS*, volume 33, pages 18661–18673, 2020.

[35] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3D point cloud segmentation. In *CVPR*, pages 8500–8509, 2022.

[36] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *CVPR*, pages 6918–6928, 2022.

[37] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. PointCNN: Convolution on x-transformed points. *NeurIPS*, 2018.

[38] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017.

[39] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, pages 2980–2988, 2017.

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[42] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019.

[43] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

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

[45] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2020.

[46] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pages 565–571. IEEE, 2016.

[47] Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

[48] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3D: Out-of-context data augmentation for 3D scenes. In *3DV*, pages 116–125. IEEE, 2021.

[49] Vardan Papyan, XY 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.

[50] Federico Pernici, Matteo Bruni, Claudio Baecchi, and Alberto Del Bimbo. Regular polytope networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[51] Tomaso Poggio and Qianli Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. *arXiv preprint arXiv:2101.00072*, 2020.

[52] Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *CVPR*, 2017.

[53] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017.

[54] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017.

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[56] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*, pages 234–244. Springer, 2016.

[57] Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, pages 4175–4186, 2020.

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[59] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3D semantic segmentation in the wild. In *ECCV*, 2022.

[60] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 379–387. Springer, 2017.

[61] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-Net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019.

[62] Christos Thrampoulidis, Ganesh R Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *arXiv preprint arXiv:2208.05512*, 2022.

[63] Zhuotao Tian, Pengguang Chen, Xin Lai, Li Jiang, Shu Liu, Hengshuang Zhao, Bei Yu, Ming-Chang Yang, and Jiaya Jia. Adaptive perspective distillation for semantic segmentation. *IEEE TPAMI*, 45(2):1372–1387, 2023.

[64] Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *ICML*, 2022.

[65] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 2020.

[66] E Weinan and Stephan Wojtowytsch. On the emergence of tetrahedral symmetry in the final and penultimate layers of neural network classifiers. *arXiv preprint arXiv:2012.05420*, 2020.

[67] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018.

[68] Liang Xie, Yibo Yang, Deng Cai, and Xiaofei He. Neural collapse inspired attraction-repulsion-balanced loss for imbalanced learning. *arXiv preprint arXiv:2204.08735*, 2022.

[69] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *arXiv preprint arXiv:2208.06677*, 2022.

[70] 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? In *NeurIPS*, 2022.

[71] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[72] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020.

[73] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, pages 2361–2370, 2021.

[74] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021.

[75] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[76] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, pages 16489–16498, 2021.

[77] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, pages 633–641, 2017.

[78] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under MSE loss: Global optimality with unconstrained features. *arXiv preprint arXiv:2203.01238*, 2022.

[79] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, pages 6908–6917, 2022.

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