# MLC-NC: Long-Tailed Multi-Label Image Classification Through the Lens of Neural Collapse

**Zijian Tao[1,2], Shao-Yuan Li[1,2,3*], Wenhai Wan[4], Jinpeng Zheng[1,2], Jia-Yao Chen[1,2], Yuchen Li[5], Sheng-Jun Huang[1,2], Songcan Chen[1,2]**

[1]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
[2]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[3]State Key Laboratory for Novel Software Technology, Nanjing University
[4]School of Computer Science and Technology, Huazhong University of Science and Technology
[5]College of Computer and Software, Hohai University

## Abstract

Long-tailed (LT) data distribution is common in multi-label image classification (MLC) and can significantly impact the performance of classification models. One reason is the challenge of learning unbiased instance representations (i.e. features) for imbalanced datasets. Additionally, the co-occurrence of head/tail classes within the same instance, along with complex label dependencies, introduces further challenges. In this work, we delve into this problem through the lens of neural collapse (*NC*). *NC* refers to a phenomenon where the last-layer features and classifier of a deep neural network model exhibit a simplex Equiangular Tight Frame (ETF) structure during its terminal training phase. This structure creates an optimal linearly separable state. However, this phenomenon typically occurs in balanced datasets but rarely applies to the typical imbalanced problem. To induce NC properties under Long-tailed multi-label classification (LT-MLC) conditions, we propose an approach named MLC-NC, which aims to learn high-quality data representations and improve the model's generalization ability. Specifically, MLC-NC accounts for the fact that different labels correspond to different feature parts located in images. MLC-NC extracts class-wise features from each instance through a cross-attention mechanism. To guide the features toward the ETF structure, we introduce visual-semantic feature alignment with a fixed ETF structured label embedding, which helps to learn evenly distributed class centers. To reduce within-class feature variation, we introduce collapse calibration within a lower-dimensional feature space. To mitigate classification bias, we concatenate features and feed them into a binarized fixed ETF classifier. As an orthogonal approach to existing methods, MLC-NC can be seamlessly integrated into various frameworks. Extensive experiments on widely-used benchmarks demonstrate the effectiveness of our method.

## Introduction

Multi-label image classification involves identifying and predicting a comprehensive set of labels that correspond to the various objects, attributes, or actions present within an image(Zhang and Zhou 2013; Everingham et al. 2015). This field has seen a surge in research on developing advanced algorithms such as probabilistic graphical models(Chen et al. 2019), new deep model architectures(Vaswani 2017), and novel loss functions(Ridnik et al. 2021; Kobayashi 2023). Despite these advancements, the success of these methods is often constrained by the prevailing yet unrealistic assumption: *all categories appear with comparable numbers of instances*. This textbook scenario glosses over the intricate long-tailed data distribution challenge inherent to practical multi-label classification scenarios(Reed 2001; Zhang et al. 2023).

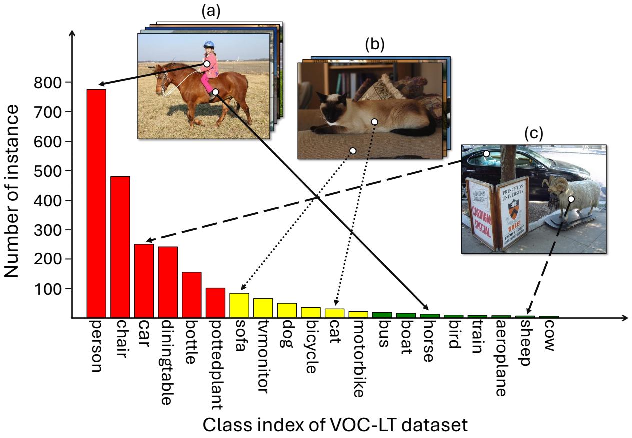

Figure 1: An illustrative example and the challenges of the LT-MLC problem on the VOC-LT dataset(Wu et al. 2020).

Figure 1 plots an illustrative example and the challenges of the LT-MLC problem on the VOC-LT dataset(Wu et al. 2020). It is characterized by: **1) Long-Tailed Class Distribution of Categories**: From the class aspect, we can see a significant disparity in the prevalence of categories, with head classes (e.g., 'person', 'chair', 'car') enjoying abundant instances, in the meanwhile a large number of tail classes (e.g., 'aeroplane', 'sheep', 'cow') are under-represented with rare instances. It's widely acknowledged in the literature that training on such datasets can lead the model to be biased towards overfitting the head classes and underfitting the tail classes, significantly undermining the model's generalization ability(He and Ma 2013; Veit et al.

[*] Corresponding author

2017). **2) Co-occurrence of Head and Tail Classes**: From an instance perspective, each image can be annotated with a diverse array of labels, ranging from frequently assigned head classes to rarely assigned tail classes. For example, images (a) and (c) are tagged with head classes such as 'person' and 'car,' alongside tail classes like 'horse' and 'sheep.' This co-occurrence further complicates the learning of tail classes, as the dominance of head classes can suppress the extraction of features for tail classes. Notably, in VOC-LT, the co-occurrence rate between head and tail classes reaches 100%, while in COCO-LT, it is 99.26%. **3) Complex Correlations of Labels**: Coupled with the reality that individual images may correspond to multiple labels, the labels may have complex dependencies, see {'person' ride on 'horse'} in the image (a), and {'cat' sit on 'sofa'} in the image (b). This layer of complexity necessitates the model to account for both the interdependencies and the individualities of labels.

Due to the presence of multiple classes per instance, the common learning base of extracting one feature vector for each instance in the single-label field is certainly insufficient to capture the complex correlations among labels in multi-label datasets(Cao et al. 2019; Zhou et al. 2020). Additionally, strategies such as resampling and reweighting the instances, and those advanced representation learning and novel loss functions that rely on them are inevitably influenced by the co-occurrence of head and tail classes, leading models to focusing more on head classes and neglect learning the tail classes(Zhang et al. 2017; Hou et al. 2023; Wu et al. 2020; Guo and Wang 2021; Zhang et al. 2023). How to improve the discriminative ability of the learning model on LT-MLC without the interference of multiple label co-occurrence is of considerable significance.

Fortunately, recent studies on the neural collapse (NC) phenomenon have opened a chance for us. The NC phenomenon was first observed on *balanced* single-label multi-class datasets during the terminal phase of training deep classification models(Papyan, Han, and Donoho 2020): (i) the variability of features within every class collapses to zero, (ii) the set of feature means form a simplex equiangular tight frame (ETF), and (iii) the last layer classifiers collapse to the feature means and forms the same simplex ETF upon some scaling. These properties in turn facilitate an optimal linear separable state for classification. More characteristics include the global optimal property(Zhu et al. 2021) and generalization ability(Galanti, György, and Hutter 2021) were subsequently found for NC.

Whereas this phenomenon holds only for balanced datasets. On imbalanced datasets, it was identified that as the imbalance level increases, the learned representations and the classifier weights on minority classes will become indistinguishable(Fang et al. 2021). The absence of NC property is believed to explain the model performance degradation on imbalanced data partially. Subsequently, a few promising attempts to induce NC properties in the single-label imbalanced field were made. (Yang et al. 2022; Li et al. 2023b) proposed training a neural network with a fixed ETF classifier and proved that it naturally leads to optimal ETF structured features under certain assumptions. (Liu et al. 2023) explicitly proposed two feature regularization terms to learn high-quality representation.

Building on the above neural collapse insights, we propose to introduce the NC properties to LT-MLC to improve model generalization ability. With the reasonable assumption that different labels correspond to distinct locations in each image, we extract class-wise features on each instance and leverage the semantic information of labels to guide feature learning. For each instance, we first use label embeddings with a fixed ETF structure to extract class-wise features for each category. By aligning these class-wise features with the corresponding label embeddings, we ensure that features from different classes on each instance are maximally distinguishable, ultimately forming an ETF structure. We further project these features into a lower-dimensional space and perform collapse calibration to reduce the within-class feature variance. Finally, the resulting features are concatenated and fed forward into a fixed ETF classifier to obtain the model's predictions. Extensive experiments provide strong evidence for the effectiveness of our method.

It's worth noting that recent work(Li et al. 2023a) generalized the concept of NC to multi-label learning. It proved a generalized NC phenomenon, i.e., the means of features for instance with multiple labels are the scaled averages of means for their single-label counterparts. However, it requires *balanced* training instances within the same class multiplicity. In summary, our key contributions are:

- To the best of our knowledge, this is the first work to explore LT-MLC with a neural-collapse-inspired approach. It opens the chance to fundamentally solve the LT-MLC problem without the interference of the label co-occurrence and interdependency challenges.
- We propose a novel MLC-NC approach to learn ETF structured class-wise features. It aligns the features with a fixed ETF label embedding to enforce evenly distributed class centers, and reduces within-class feature variation through collapse calibration in a lower dimension space.
- MLC-NC outperforms strong baselines and achieves state-of-the-art results on the extensive benchmark datasets including COCO-LT, VOC-LT and VG200.

## Related Works

**Long-tailed Classification** Methods in literature mitigate long-tail by following aspects: resampling techniques(Chawla et al. 2002; Shen, Lin, and Huang 2016), re-weighted loss functions(Zhang et al. 2017; Cui et al. 2019) and specialized architectures(Liu et al. 2019; Zhou et al. 2020). MiSLAS(Zhong et al. 2021) explored methods to enhance model calibration by addressing the overconfidence issue typically seen in imbalanced data distributions. CSA(Shi et al. 2023) analyzed the effectiveness of re-sampling techniques in addressing class imbalance in long-tailed learning. SBCL(Hou et al. 2023) introduced a subclass-balancing approach to contrastive learning, significantly improving the representation of minority classes.

**MLC** ML-GCN(Chen et al. 2019) is a graph-based model that constructs relationships based on multi-label associations. Focal Loss(Lin et al. 2017) addresses the issue of class imbalance by assigning different weights to instances, thereby focusing the model on learning from difficult examples. ASL(Ridnik et al. 2021) dealt with the imbalance between positive and negative instances within labels by setting thresholds to reject incorrect label annotations. TW(Kobayashi 2023) proposed a novel loss function to cope with intra-class imbalance.

**Long-tailed MLC** DB(Wu et al. 2020) firstly addresses long-tailed MLC by introducing a co-occurence based loss function. It assigned balanced weights to instances to address the imbalance caused by label co-occurrence and incorporated a regularization term to mitigate the overemphasis on negative labels. URS(Guo and Wang 2021) proposed a dual-branch network architecture using uniform and rebalanced sampling as inputs for the branches and introduced a consistency loss to ensure effective learning. PLM(Duarte, Rawat, and Shah 2021) proposed randomly masking certain labels during loss computation to balance the classes. While mask training can alleviate class imbalance, it also results in the loss of useful learning information, preventing optimal performance. MFM(Zhang et al. 2023) decoupled the $\gamma$ in focal loss and proposes a Multi-Focal Modifier to increase the model's attention to tail positive instances.

**Neural Collapse** (Papyan, Han, and Donoho 2020) first investigated the NC phenomenon that occurs during the terminal phase of training deep neural networks. Subsequent research(Tirer and Bruna 2022; Ji et al. 2021) focused on uncovering the underlying mechanisms of NC and identifying the conditions under which it occurs.(Yang et al. 2022) questioned the necessity of learning a linear classifier at the end of a deep neural network when its ETF structure is known. It proposed training a neural network with a fixed ETF classifier and demonstrated that it naturally leads to NC for feature representations.(Liu et al. 2023) proposed two explicit feature regularization terms to learn high-quality representation for class-imbalanced data.(Li et al. 2023b) applied NC to federated learning, and utilized a synthetic and fixed ETF classifier to address the data heterogeneity across different clients. These studies are for sing-label tasks. Recently, (Li et al. 2023a) generalized the concept of NC to the special balanced multi-label classification scenario, where the training instances within the same multiplicity are required to be balanced. It proved that a generalized NC phenomenon holds with the "pick-all-label" formulation where the means of features for instance with multiple labels are the scaled averages of means for their single-label counterparts.

As we discussed in the introduced section, mostly relying on some implicit instance-level resampling/reweighting strategies, exiting LT-MLC learning techniques were inevitably interfered with by the label co-occurrence and interdependency. In this paper, we fundamentally avoid this problem through the lens of NC by learning class-wise features with the optimal ETF structure.

## Preliminaries

We introduce a multi-label dataset with long-tailed distribution $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, $N$ denotes the number of instances, $x_i \in \mathbb{R}^{W \times H \times 3}$ is the $i$-th input image, and $y_i$ represents the label vector corresponding to the $i$-th image. $y_i = [y_i^1, y_i^2, \dots, y_i^C] \in [0, 1]$ represents the label vector of the $i$-th instance. $C$ is the number of classes in the dataset. $y_i^c = 1$ indicates that the $i$-th instance includes class $c$. Otherwise, it does not include class $c$. Let $n_c = \sum_{i=1}^{N} y_i^c$ represent the number of training instances that include class $C$. In the long-tailed distribution, the number of instances for head classes is significantly larger than those of tail classes.

We first give the concept of a simplex ETF in the context of neural collapse.

**Definition 1.** ***Simplex Equiangular Tight Frame (ETF)*** *$\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_C] \in \mathbb{R}^{d \times C}$ is composed of a set of vectors $\mathbf{v}_i \in \mathbb{R}^d$, for $i \in [C]$ and $d \geq C - 1$. $\mathbf{V}$ is called a simplex equiangular tight frame if:*

$$\mathbf{V} = \sqrt{\frac{C}{C-1}} \mathbf{U} \left( \mathbf{I}_C - \frac{1}{C} \mathbf{J}_C \right). \tag{1}$$

Here $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_C] \in \mathbb{R}^{d \times C}$ is an orthogonal matrix satisfying $\mathbf{U}^T \mathbf{U} = \mathbf{I}_C$, $\mathbf{I}_C \in \mathbb{R}^{C \times C}$ is the identity matrix, and $\mathbf{J}_C \in \mathbb{R}^{C \times C}$ is an all-ones matrix. All vectors in a simplex ETF have equal $\ell_2$ norm and identical pair-wise angles, with a cosine value of $-\frac{1}{C-1}$. Hence,

$$\mathbf{v}_i^T \mathbf{v}_j = \begin{cases} -\frac{1}{C-1} & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases} \tag{2}$$

The pair-wise angle $-\frac{1}{C-1}$ is the maximal equiangular separation of $C$ vectors in $\mathbb{R}^d$.

Then, we specify three fundamental characteristics of the **neural collapse (NC)** phenomenon below.

- **Variability Collapse (NC1)**: The variability within the last-layer activations for instances within the same class collapses to zero, meaning the activations converge to their class means.
- **ETF Convergence (NC2)**: The class means collapse to the vertices of a simplex ETF, which is a highly symmetric geometric structure i.e. $\tilde{\mathbf{f}}_i \cdot \tilde{\mathbf{f}}_j \rightarrow -\frac{1}{C-1},\ \forall i, j \in [C],\ i \neq j,\ \tilde{\mathbf{f}}_c$ is the feature prototype of class $c$.
- **Self-Duality (NC3)**: Up to rescaling, the last-layer classifiers also collapse to the class means, leading to a self-dual configuration where classifiers align with the class means which means that the classifier vectors collapse to the same simplex ETF i.e. $\tilde{\mathbf{v}}_i \cdot \tilde{\mathbf{v}}_j \rightarrow -\frac{1}{C-1},\ \forall i, j \in [C],\ i \neq j$ where $\mathbf{v}_c = \frac{\mathbf{v}_c}{\|\mathbf{v}_c\|}, \forall c \in [C]$, $\mathbf{v}_c$ is the classifier vector of ETF classifier.

## Method

NC tells us the optimal structure (i.e. simplex ETF) of classifiers and feature prototypes in a perfect training setting. It inspires us to induce the NC properties in LT-MLC to improve the discriminative ability of the learning model. Concretely, we propose MLC-NC. As elaborated in Figure 2, MLC-NC consists of three major components: ETF Label Embedding Guided Feature Learning to learn distinct

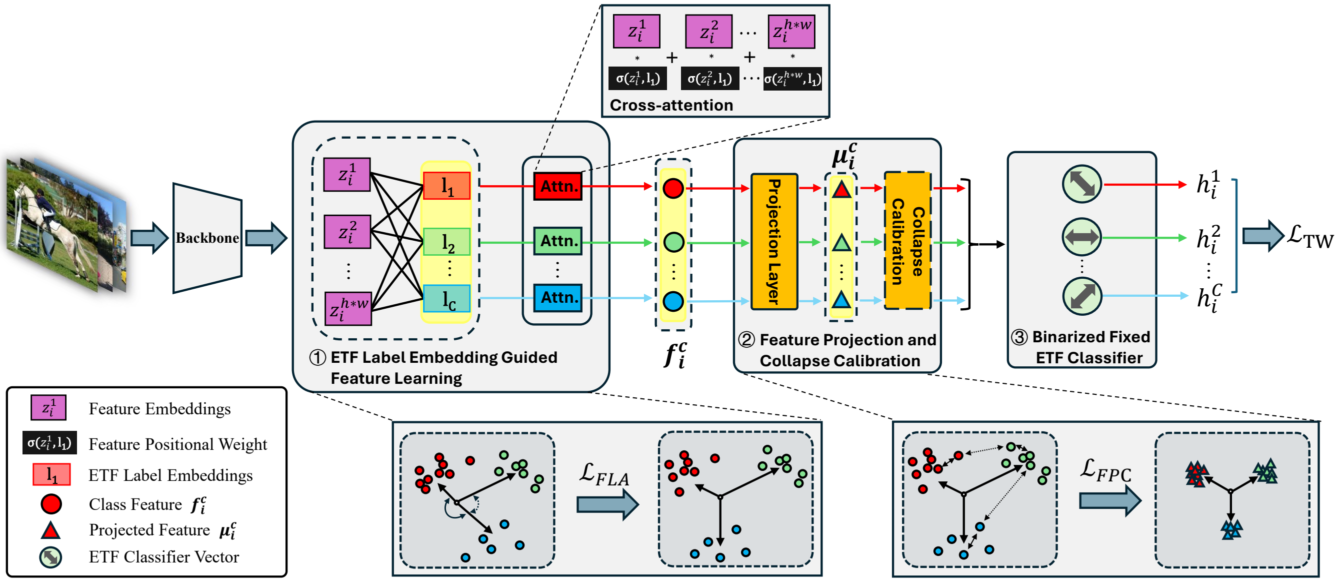

Figure 2: Overall structure of MLC-NC. MLC-NC consists of 3 major components: ETF Label Embedding Guided Feature Learning, Feature Projection and Collapse Calibration, and Binarized Fixed ETF Classifier.

between-class features, feature projection and collapse calibration to reduce within-class feature variation, and binarized fixed ETF classifier to maximally separate the pairwise angles of all classes. In the following, we elaborate on the details.

## ETF Label Embedding Guided Feature Learning

In the long-tail multi-label domain, different labels correspond to features located in different parts of the instance. Therefore, it is crucial to consider spatial information during feature extraction, as different classes emphasize different locations. Hence, we need to extract corresponding features for each class, as detailed below.

Firstly, for an input instance $x_i$, we utilize a feature extractor backbone $G(\cdot)$, e.g., ResNet50, to extract features $z_i = G(x_i)$, where $z_i \in \mathbb{R}^{w\times h\times ch}$. $w$, $h$, and $ch$ represent the width, height, and number of channels of the features, respectively. To obtain features from different spatial positions, as shown in Figure 2, we divide $z_i$ spatially into $\{z_i^1, z_i^2, \dots z_i^{w\times h}\}$, where $z_i^k \in \mathbb{R}^{1\times 1\times ch}$.

Secondly, accounting for the different spatial information of features for various labels, we employ label embedding and feature embedding to compute similarity scores for reweighting the features. We synthesize a simplex ETF label embedding $\mathbf{l} = \{\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_C\} \in \mathbb{R}^{d\times C}$ by Eq. (1), where $d$ is the dimension of the label embedding and $d = ch$. $C$ is the number of labels. The feature positional weight of label embedding $\mathbf{l}_c$ on feature embedding $\mathbf{z}_i^k$ is computed as:

$$\sigma(\mathbf{z}_i^k, \mathbf{l}_c) = \frac{\exp(\mathrm{sim}(\mathbf{z}_i^k, \mathbf{l}_c))}{\sum_{j=1}^{h\times w} \exp(\mathrm{sim}(\mathbf{z}_i^j, \mathbf{l}_c))}, \tag{3}$$

with

$$\mathrm{sim}(\mathbf{z}_i^k, \mathbf{l}_c) = \frac{\mathbf{z}_i^k \cdot \mathbf{l}_c}{\|\mathbf{z}_i^k\| \|\mathbf{l}_c\|}. \tag{4}$$

$\mathrm{sim}(\mathbf{z}_i^k, \mathbf{l}_c)$is the cosine similarity function. The higher the value of $\sigma(\cdot)$, the more attention the label pays to the features at that position. Therefore, the feature corresponding to label $c$ in the $i$-th instance $\mathbf{f}_i^c$ is:

$$\mathbf{f}_i^c = \sum_{k=1}^{h\times w} \sigma(\mathbf{z}_i^k, \mathbf{l}_c) \cdot \mathbf{z}_i^k. \tag{5}$$

Finally, to align the class-wise features with the label embeddings and ensure that the class-wise features also exhibit the ETF structure, we propose the visual-semantic feature alignment loss $\mathcal{L}_{\text{FLA}}$:

$$\mathcal{L}_{\text{FLA}} = -\frac{1}{N\cdot C}\sum_{i=1}^{N}\sum_{c=1}^{C}\left(y_i^c \log\left(\frac{1+\mathrm{sim}(\mathbf{f}_i^c, \mathbf{l}_c)}{2}\right) + \right. \\ \left. (1-y_i^c)\log\left(1 - \frac{C-1}{C}\left|\frac{1}{C-1} + \mathrm{sim}(\mathbf{f}_i^c, \mathbf{l}_c)\right|\right)\right). \tag{6}$$

When $y_i^c = 1$, through $\mathcal{L}_{\text{FLA}}$, we align $\mathbf{f}_i^c$ and $\mathbf{l}_c$ to maximize their cosine similarity. Conversely, when $y_i^c = 0$, we aim to minimize the similarity between the instance's feature and the current label embedding, making the cosine value reflect the ETF structure as $-\frac{1}{C-1}$. For example, we assume that instance $i$ contains label $c_1$, and instance $j$ contains label $c_2$ and $c_1 \neq c_2$. Through $\mathcal{L}_{\text{FLA}}$, $\mathrm{sim}(\mathbf{f}_i^{c_1}, \mathbf{l}_{c_1}) \approx 1$ and $\mathrm{sim}(\mathbf{f}_j^{c_2}, \mathbf{l}_{c_2}) \approx 1$. Therefore, $\frac{\mathbf{f}_i^{c1}}{\|\mathbf{f}_i^{c1}\|} \approx \frac{\mathbf{l}_{c_1}}{\|\mathbf{l}_{c_1}\|}$ and $\frac{\mathbf{f}_j^{c2}}{\|\mathbf{f}_j^{c2}\|} \approx \frac{\mathbf{l}_{c_2}}{\|\mathbf{l}_{c_2}\|}$. Given $\frac{\mathbf{l}_{c_1}}{\|\mathbf{l}_{c_1}\|} \cdot \frac{\mathbf{l}_{c_2}}{\|\mathbf{l}_{c_2}\|} = -\frac{1}{C-1}$, it follows that $\frac{\mathbf{f}_i^{c1}}{\|\mathbf{f}_i^{c1}\|} \cdot \frac{\mathbf{f}_j^{c2}}{\|\mathbf{f}_j^{c2}\|} \approx -\frac{1}{C-1}$. In other words, in each instance, the class feature for $c_1$ and $c_2$ forms an ETF structure.

## Feature Projection and Collapse Calibration

According to NC1, in an ideal training scenario, within-class features should collapse to the corresponding class centers. However, in real-world scenarios, such as in the VOC and COCO datasets, the features of the same label in different instances may vary. For example, instances labeled "bird" may contain seagulls in some cases and sparrows in others. Consequently, it becomes challenging to unify the extracted features $\mathbf{f}_i^c$ in high-dimensional scenarios. Collapsing features based on NC1 in such cases may interfere with the feature extractor. Therefore, it is necessary to project the features into a lower dimension and normalize them:

$$\hat{\mu}_i^c = g(\mathbf{p_c}; \mathbf{f_i^c}), \quad \mu_i^c = \frac{\hat{\mu}_i^c}{\|\hat{\mu}_i^c\|_2}. \tag{7}$$

Here $g(\cdot)$ is the linear projection function with $\mathbf{p_c}$ denoting the parameters of the projection layer for the feature of class c. $\mu_i^c \in \mathbb{R}^{p \times C}$ is the normalized result of the low-dimensional feature vector $\hat{\mu}_i^c$ obtained after projection. $p$ is the dimension of the feature vector after projection.

We note that the projection layer is essential in collapse calibration for the following reasons: (i) If the last layer of the feature extractor employs non-linear activation, e.g., ReLU, the raw feature $\mathbf{f}_c$ will be sparse with zeros. This leads to features easily orthogonal to each other, making it difficult to collapse into the ETF structure. (ii) High-dimensional features of the same label may contain different information due to the difference between samples. By projecting them into a lower dimension, we refine the features and reduce their variability. This allows us to better leverage the principles of NC1 to optimize the features.

In the previous subsection, although we extract class-wise features for each class from an instance, these features all originate from the same instance. This results in a correlation between the class-wise features extracted from the same instance and this correlation limits our ability to collapse $\mu_i^c$ to the class center $\mu^c$. To address this, we adopt a contrastive learning approach, allowing each $\mu_i^c$ to collapse to the class feature center while distinguishing it from the feature vectors of other class centers. In the $t$-th epoch, we define the class center feature of class $c$ as $\mu_t^c$:

$$\mu_t^c = \frac{1}{n_c}\sum_{i=1}^{N} \mathbb{1}(y_i^c = 1)\mu_{i,t}^c. \tag{8}$$

$n_c$ is the number of instances that contain label $c$. $\mu_{i,t}^c$ represents $\mu_i^c$ in $t$-th epoch. During training, we calculate the mean feature for each class in every epoch and use it as the class center for the next epoch. Inspired by contrastive learning, to encourage features to collapse into their respective class prototypes while remaining distant from the prototypes of other classes, we propose a loss function $\mathcal{L}_{\text{FPC}}$ based on **f**eature **p**rojection *and* **c**ollapse calibration. At the $t$-th epoch, $\mathcal{L}_{\text{FPC}}$ is defined as:

$$\mathcal{L}_{\text{FPC}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \mathbb{1}(y_i^c = 1)\log\left(\frac{\exp(S_i^{c,c}/\tau_1)}{\sum_{k=1}^{C}\exp(S_i^{c,k}/\tau_2)}\right). \tag{9}$$

Here $S_i^{c,k} = \text{sim}(\mu_{i,t}^c, \mu_{t-1}^k)$ is the cosine similarity between the feature of class $c$ for instance $i$ in current t-th epoch and the class center feature of class $k$ computed in last epoch. $\tau_1$ and $\tau_2$ are temperature parameters used to control the degree of collapse, with a larger(smaller) value results in a higher degree of collapse. As shown in Figure 2, through $\mathcal{L}_{\text{FPC}}$, we can learn more robust feature information for different classes and maximize the distinction of different classes.

## Binarized Fixed ETF Classifier

To minimize the impact of imbalance on the classifier, we employ the binarized fixed ETF classifier for multi-label datasets. We first synthesize a simplex ETF classifier $V_{\text{ETF}} = \{v_1, v_2, \ldots, v_C\} \in \mathbb{R}^{\mathcal{D} \times C}$ by Eq. (1), where $\mathcal{D} = p \times C$. Extensive previous research has demonstrated the effectiveness of the ETF classifier when dealing with imbalanced datasets(Zhu et al. 2021; Galanti, György, and Hutter 2021). To capture the complete information of image features and avoid the classifier bias influenced by the bias in feature extraction across different classes, we concatenate the projected class-wise features $\mu_i^c$ into $\mu_i$ as the final feature. Then, we compute the inner product of the projected feature vector and the classifier vector to obtain the logit:

$$h_i^c = \gamma \mu_i \cdot v_c, \tag{10}$$

where $\mu_i = concat(\mu_i^1, \mu_i^2, \ldots, \mu_i^C)$, $h_i^c$ is the logit of class c of i-th instance. We also introduce a learnable temperature scalar $\gamma$ to scale the results. Subsequently, the logit $h_i^c$ for each class is fed into the sigmoid function to obtain the binary classification prediction.

Recently, (Kobayashi 2023) proposed a novel loss function to address the issue of imbalance between positive and negative labels in MLC. It approaches the problem from both an instance-wise way and a class-wise way, aiming to increase the margin between the logits of positive and negative labels. This enhancement improves the model's attention to positive labels and suppression of negative labels. The specific loss function is as follows:

$$\ell_i = \text{softplus}\left[\log \sum_{c|y_i^c=0} e^{h_i^c} + T \log \sum_{c|y_i^c=1} e^{-\frac{h_i^c}{T}}\right], \tag{11}$$

$$\ell^c = \text{softplus}\left[\log \sum_{i|y_i^c=0} e^{h_i^c} + T \log \sum_{j|y_j^c=1} e^{-\frac{h_j^c}{T}}\right]. \tag{12}$$

The final two-way multi-label loss function $\mathcal{L}_{TW}$ is defined as follows:

$$\mathcal{L}_{TW} = \frac{1}{M}\sum_{i=1}^{M} \ell_i\left(\{x_i, y_i^c\}_{c=1}^{C}\right) + \frac{1}{C}\sum_{c=1}^{C} \ell^c\left(\{x_i, y_i^c\}_{i=1}^{M}\right). \tag{13}$$

Here $M$ is the number of instances in a batch and $T$ is the temperature parameter. In this paper, we use the two-way loss to handle the imbalance between positive and negative instances within labels.Thus, the final loss function $\mathcal{L}$ is:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{FLA}} + \beta\mathcal{L}_{\text{FPC}} + \mathcal{L}_{\text{TW}}. \tag{14}$$

The pseudocode for MLC-NC is provided in Appendix A.

| Category | Methods | COCO-LT |  |  |  | VOC-LT |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  | Total | Head | Medium | Tail | Total | Head | Medium | Tail |
| MLC | ML-GCN | 44.24 | 44.04 | 48.36 | 38.96 | 68.92 | 70.14 | 76.41 | 62.39 |
|  | Focal Loss | 49.46 | 49.80 | 54.77 | 42.14 | 73.88 | 69.41 | 81.43 | 71.56 |
|  | ASL | 54.35 | 50.59 | 58.76 | 51.82 | 78.31 | 71.12 | 84.95 | 78.71 |
| LT-SLC | ERM | 41.27 | 48.48 | 49.06 | 24.25 | 70.86 | 68.91 | 80.20 | 65.31 |
|  | RS | 46.97 | 47.58 | 50.55 | 41.70 | 75.38 | 70.95 | 82.94 | 73.05 |
|  | RW | 42.27 | 48.62 | 45.80 | 32.02 | 74.70 | 67.58 | 82.81 | 73.96 |
|  | OLTR | 45.83 | 47.45 | 50.63 | 38.05 | 71.02 | 70.31 | 79.80 | 64.96 |
|  | LDAM | 40.53 | 48.77 | 48.38 | 22.92 | 70.33 | 68.73 | 80.38 | 69.09 |
|  | CB Focal | 49.06 | 47.91 | 53.01 | 44.85 | 75.24 | 70.30 | 83.53 | 72.74 |
|  | BBN | 50.00 | 49.79 | 53.99 | 44.91 | 73.37 | 71.31 | 81.76 | 68.62 |
| LT-MLC | DB | 52.53 | 50.25 | 56.33 | 49.54 | 78.65 | 73.16 | 84.11 | 78.66 |
|  | DB-Focal | 53.55 | 51.13 | 57.05 | 51.06 | 78.94 | 73.22 | 84.18 | 79.30 |
|  | URS | 56.90 | **54.13** | 60.59 | 54.47 | 81.44 | **75.68** | 85.53 | 82.69 |
|  | MFM | 55.25 | 48.71 | 58.24 | 57.08 | 79.64 | 66.32 | 84.69 | 85.83 |
|  | **MLC-NC** | **60.52** | 49.69 | **64.94** | **64.21** | **84.37** | 72.75 | **88.15** | **90.31** |

Table 1: Performance (mAP%) comparison on COCO-LT, VOC-LT. The best and second-best performances are highlighted in **bold** and underline notes.

| Methods | Total | G-mAP | Head | Medium | Tail |
|---|---|---|---|---|---|
| BCE | 24.74 | 38.39 | 39.73 | 22.70 | 19.26 |
| DB | 30.83 | 44.69 | 47.07 | 29.07 | 22.80 |
| DB-Focal | 30.51 | 44.46 | 47.65 | 28.58 | 22.36 |
| ASL | 26.75 | 42.99 | 42.54 | 24.73 | 20.38 |
| URS | 10.23 | 17.8 | 24.28 | 8.73 | 3.17 |
| MFM | 29.43 | 44.51 | 44.12 | 27.54 | 23.52 |
| **MLC-NC** | **36.87** | **50.07** | **49.78** | **35.13** | **32.09** |

Table 2: Performance (mAP%) comparison on imbalanced VG200.

## Experiment

**Dataset**: We analyze and conduct experiments on two artificially created long-tailed multi-label image classification datasets VOC-LT and COCO-LT following(Wu et al. 2020). Besides, we verify the universality of the proposed approach on one real-world multi-label dataset VG200 with milder imbalance distribution(Krishna et al. 2017).

**Comparison Methods:** To objectively evaluate MLC-NC, we compare it against methods from three scenarios. **Classical Deep Multi-Label Methods**: We employ ML-GCN, Focal Loss, and ASL. **Classical Long-Tailed Single-Label Methods:** We employ Empirical Risk Minimization(ERM) / Re-Sampling (RS) / Re-Weighting (RW)(Shen, Lin, and Huang 2016), OLTR(Liu et al. 2019), LDAM(Cao et al. 2019), CB Focal(Cui et al. 2019), BBN(Zhou et al. 2020). **Long-Tailed Multi-Label Methods:** We employ DB/DB-Focal(Wu et al. 2020), URS(Guo and Wang 2021) and MFM(Zhang et al. 2023).

**Training Setup**: We use a ResNet50 pre-trained on ImageNet as the feature extractor. In $\mathcal{L}_{\text{FPC}}$, $\tau_1$ and $\tau_2$ are set to 0.5. $\alpha$ and $\beta$ are set to 0.5 and 0.2, The dimension $d$ of the feature projection is set to 20. We evaluate mean average precision (mAP) across all classes, averaging the results over three runs for all methods.

### Results

Tables 1 and 2 present the experimental results on COCO-LT, VOC-LT, and VG200. Our MLC-NC demonstrates significant performance improvements over the second-best baseline, especially in the **medium** and **tail** classes, and achieves the best overall **total** performance. For instance, on COCO-LT, our method achieves a **medium** mAP of 64.94 compared to 60.59 (URS), a **tail** mAP of 64.21 compared to 57.08 (MFM), and a **total** mAP of 60.52 compared to 56.90 (URS). Similarly, on VOC-LT, our method achieves a **medium** mAP of 88.15 compared to 85.53 (URS), a **tail** mAP of 90.31 compared to 85.85 (MFM), and a **total** mAP of 84.37 compared to 81.44 (URS).

**URS** enhances head class learning by enforcing consistency between resampled and non-resampled samples, leading to better performance in head classes, but highly inferior on medium and tail classes. **LT-SLC** methods emphasize tail class learning in single-label tasks but neglect label dependency and head-tail co-occurrence in long-tail multi-label scenarios, resulting in weaker performance.

Methods such as **MFM** and **DB** rely on co-occurrence matrices to shift the model's focus towards tail classes, which inherently unable to improve the model's learning of tail class features. In contrast, our approach learns class-wise optimal features with the guidance of ETF structured label embeddings, which fundamentally allows for a clearer representation of tail class features without the interference by the head-tail label co-occurrence.

On VG200, we additionally analyze the Global mAP (G-

| EGFL | FPC | BEC | Total | Head | Medium | Tail |
|---|---|---|---|---|---|---|
| - | ✓ | ✓ | 82.13 | 72.47 | 87.99 | 84.98 |
| ✓ | - | ✓ | 83.43 | 71.12 | 87.88 | 89.34 |
| ✓ | ✓ | - | 79.59 | 66.40 | 85.64 | 84.96 |
| ✓ | ✓ | ✓ | **84.37** | **72.75** | **88.15** | **90.31** |

Table 3: Ablation study on VOC-LT dataset.

| EGFL | Total | Head | Medium | Tail |
|---|---|---|---|---|
| Trainable embedding | 83.92 | 71.99 | 87.57 | 90.14 |
| Glove embedding | 83.74 | 71.89 | 87.74 | 89.65 |

Table 4: Performance (mAP%) comparison of EGFL using different embedding methods on VOC-LT dataset.

mAP) from an instance-wise perspective. MLC-NC demonstrates superior performance under all scenarios, especially in tail classes, showing our robustness to milder imbalanced datasets.

## Ablation Studies

We first conduct ablation studies on the three key components of our method: **E**TF label embedding **G**uided **F**eature **L**earning (EGFL) through not using label embedding to guide feature learning(w/o $\mathcal{L}_{\text{FLA}}$); **F**eature **P**rojection and **C**ollapse Calibration (FPC) by not projecting and calibrating the feature(w/o $\mathcal{L}_{\text{FPC}}$); **B**inarized Fixed **E**TF **C**lassifier (BEC) by using randomly initialized trainable classifier and not concatenating features. As illustrated in Table 3, the performances on tail classes drop sharply without EGFL, which plays an important role in robust feature extraction. FPCC helps to refine class features. BEC significantly mitigates the impact of head and tail classes on the classification bias.

We then ablate the EGFL by using two other types of label embeddings: randomly initialized trainable label embedding, and fixed Glove label embedding(Pennington, Socher, and Manning 2014), as shown in table 4. The Glove embedding performs the worst for at least two reasons: firstly, it requires strict adherence to pre-trained Glove initialization, which does not cover all VOC labels, e.g., ”diningtable” and ”pottedplant”; besides, its fixed embedding dimension (300) constrains the feature dimensions to be the same. Compared with trainable embedding, our fixed ETF label embedding is optimal in both least training complexity and performance.

## Futher In-depth Analysis

**Geometric structure of features** In Figure 3, we plot the pair-wise angles of the centered feature means of different classes on VOC-LT guided by randomly initialized trainable label embedding and our fixed ETF label embedding. The larger the angle value, the more spread out the features, which makes it easier for the model to distinguish different class features. Our fixed ETF label embedding achieves much larger pair-wise angles, around 78° compared with the 43° of trainable label embedding.

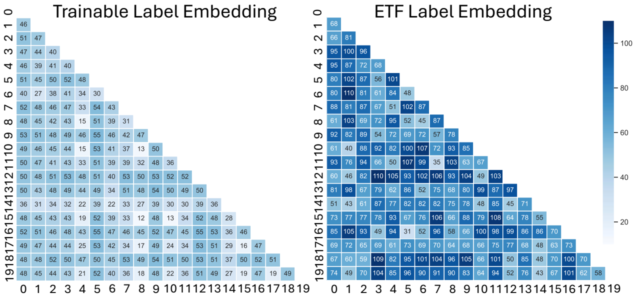

Figure 3: Pair-wise angle degree on VOC-LT of different classes feature center. The bigger the angle is, the easier for the model to differentiate classes. The optimal ETF pairwise angle for 20 classes is $\arccos\left(\frac{-1}{19}\right) \approx 93°$.

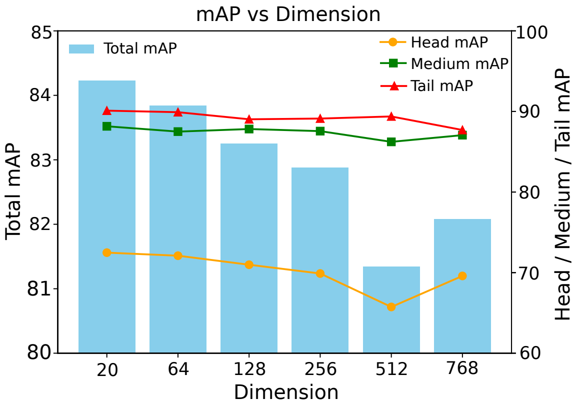

Figure 4: The effect of projection dimension on VOC-LT.

**Effect of projection dimension** Figure 4 illustrates the mAP performance across different projected feature dimensions on VOC-LT. The chart has two y-axes: the left y-axis represents the Total mAP, while the right y-axis shows the mAP for the Head, Medium, and Tail categories. The x-axis indicates the feature projection dimensions, ranging from 20 to 768, with 768 representing the original, unprojected features. It can be seen as the projection dimension increases, the overall mAP of the model tends to decline. The head classes which consist of instances with greater within-class diversity are more impacted, for which reducing features to lower dimensions is more helpful to enhance the model's ability to aggregate features of same classes.

## Conclusion

We address the LT-MLC problem by introducing neural collapse (NC). Our method, MLC-NC, uses fixed Equiangular Tight Frame (ETF) label embeddings and collapse calibration to optimize class-wise feature learning, enhancing discrimination across all classes while handling head-tail label co-occurrence and inter-dependency. Additionally, we employ Binarized Fixed ETF Classifier and concatenated features to mitigate classification bias. Extensive experiments have confirmed MLC-NC's effectiveness.

## Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities (NS2024059), Open Project Funds for the Joint Laboratory of Spatial Intelligent Perception and Large Model Application (SIPLMA-2024-YB-05), and the National Natural Science Foundation of China (62376126).

## References

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

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5177–5186.

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

Duarte, K.; Rawat, Y.; and Shah, M. 2021. Plm: Partial label masking for imbalanced multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2739–2748.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

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

Galanti, T.; György, A.; and Hutter, M. 2021. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121*.

Guo, H.; and Wang, S. 2021. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15089–15098.

He, H.; and Ma, Y. 2013. Imbalanced learning: foundations, algorithms, and applications.

Hou, C.; Zhang, J.; Wang, H.; and Zhou, T. 2023. Subclass-balancing contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5395–5407.

Ji, W.; Lu, Y.; Zhang, Y.; Deng, Z.; and Su, W. J. 2021. An unconstrained layer-peeled perspective on neural collapse. *arXiv preprint arXiv:2110.02796*.

Kobayashi, T. 2023. Two-way multi-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7476–7485.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Li, P.; Wang, Y.; Li, X.; and Qu, Q. 2023a. Neural collapse in multi-label learning with pick-all-label loss. *arXiv preprint arXiv:2310.15903*.

Li, Z.; Shang, X.; He, R.; Lin, T.; and Wu, C. 2023b. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5319–5329.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

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

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2537–2546.

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

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Reed, W. J. 2001. The Pareto, Zipf and other power laws. *Economics letters*, 74(1): 15–19.

Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 82–91.

Shen, L.; Lin, Z.; and Huang, Q. 2016. Relay backpropagation for effective learning of deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, 467–482. Springer.

Shi, J.-X.; Wei, T.; Xiang, Y.; and Li, Y.-F. 2023. How re-sampling helps for long-tail learning? *Advances in Neural Information Processing Systems*, 36.

Tirer, T.; and Bruna, J. 2022. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning*, 21478–21505. PMLR.

Vaswani, A. 2017. Attention is All You Need. *arXiv preprint arXiv:1706.03762*.

Veit, A.; Alldrin, N.; Chechik, G.; Krasin, I.; Gupta, A.; and Belongie, S. 2017. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 839–847.

Wu, T.; Huang, Q.; Liu, Z.; Wang, Y.; and Lin, D. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 162–178. Springer.

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

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, M.-L.; and Zhou, Z.-H. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8): 1819–1837.

Zhang, W.; Liu, C.; Zeng, L.; Ooi, B.; Tang, S.; and Zhuang, Y. 2023. Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1423–1432.

Zhong, Z.; Cui, J.; Liu, S.; and Jia, J. 2021. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16489–16498.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9719–9728.

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