南京航空航天大學

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]

AAAI 2025

Introduction

解决问题：Long-tailed (LT) data distribution multilabel image classification (MLC)
面临挑战：
1. learning unbiased instance representations (i.e. features) for imbalanced datasets.
2. the co-occurrence of head/tail classes within the same instance
3. complex label dependencies

创新点：
1. 不同的标签对应于图像中不同位置的特征部分。MLC-NC引入**交叉注意力机制。**
2. 采用neural collapse引导特征向ETF结构发展，引入了**具有固定ETF结构标签嵌入的视觉-语义特征对齐。**
3. 为了减少类内特征变化，在**低维特征空间内引入崩溃校准。**
4. 为了减轻分类偏差，将特征连接起来，并将其输入到**二值化的固定ETF分类器**中。

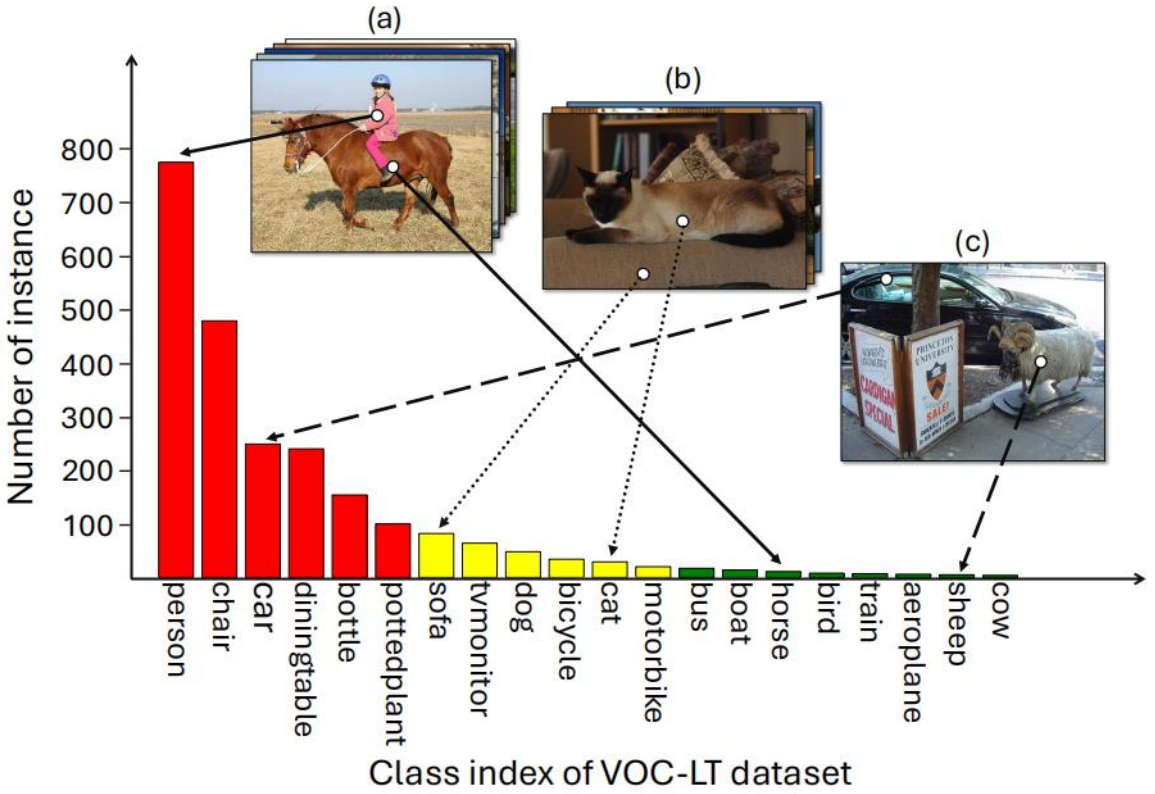
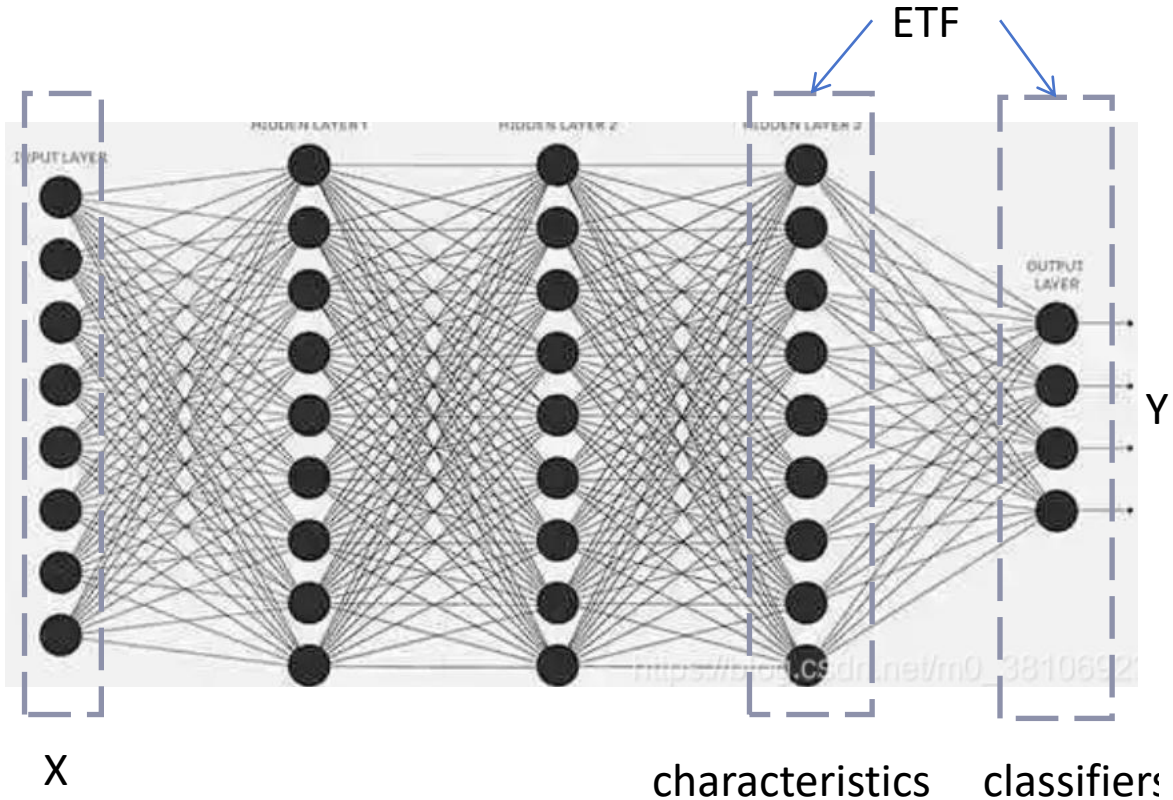


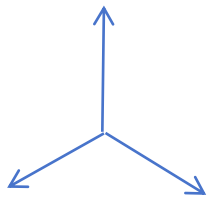
Figure 1: An illustrative example and the challenges of the LT-MLC problem on the VOC-LT dataset(Wu et al. 2020).

Neural Collapse (NC)

南京航空航天大學

深度神经网络 Y=XW+b





Simplex Equiangular Tight Frame (ETF)

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. $\tilde{\mathbf{f}}_i \cdot \tilde{\mathbf{f}}_j \to -\frac{1}{C-1}, \ \forall i, j \in [C], \ i \neq j, \ \tilde{\mathbf{f}}_c$ 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. $\tilde{\mathbf{v}}_i \cdot \tilde{\mathbf{v}}_j \to -\frac{1}{C-1}, \ \forall i, j \in [C], \ i \neq j$ 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

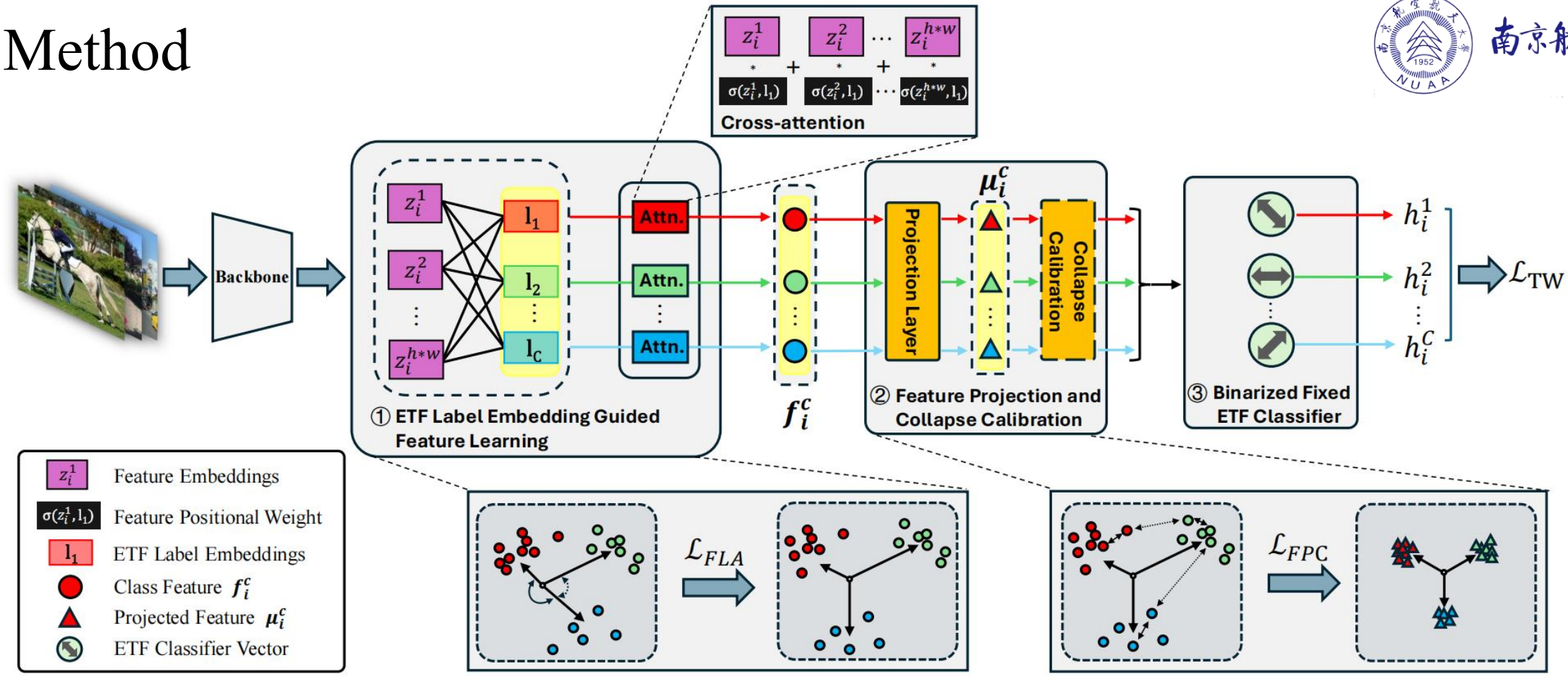


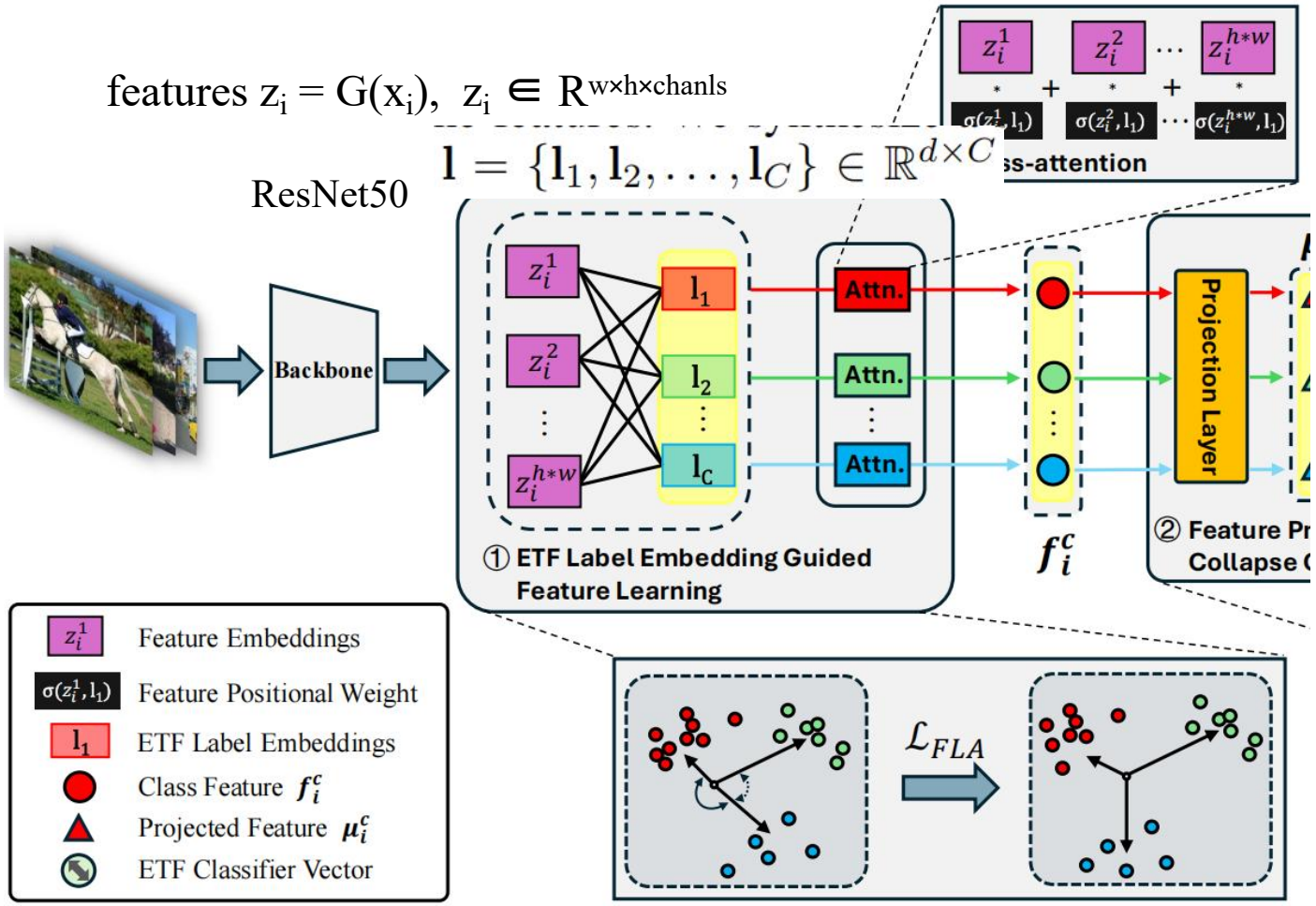
Figure 2: Overall structure of MLC-NC. MLC-NC consists of 3 major components: ETF Label Embedding Guided Feature Learning, Feature Projection and Collapse Calibration, and Binarized Fixed ETF Classifier.

MLC-NC consists of three major components: **ETF Label Embedding Guided Feature Learning (ETF标签嵌入引导特征学习)** to learn distinct between-class features, **feature projection and collapse calibration** （特征投影与塌陷校准） to reduce within-class feature variation, and **binarized fixed ETF classifier (二值化固定ETF分类器)** to maximally separate the pair-wise angles of all classes. In the following, we elaborate on the details.

Method (ETF Label Embedding Guided Feature Learning ETF标签嵌入引导特征学习)

features $z_i = G(x_i)$, $z_i \in R^{w \times h \times chanls}$

$$\mathbf{l} = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_C\} \in \mathbb{R}^{d \times C}$$



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.

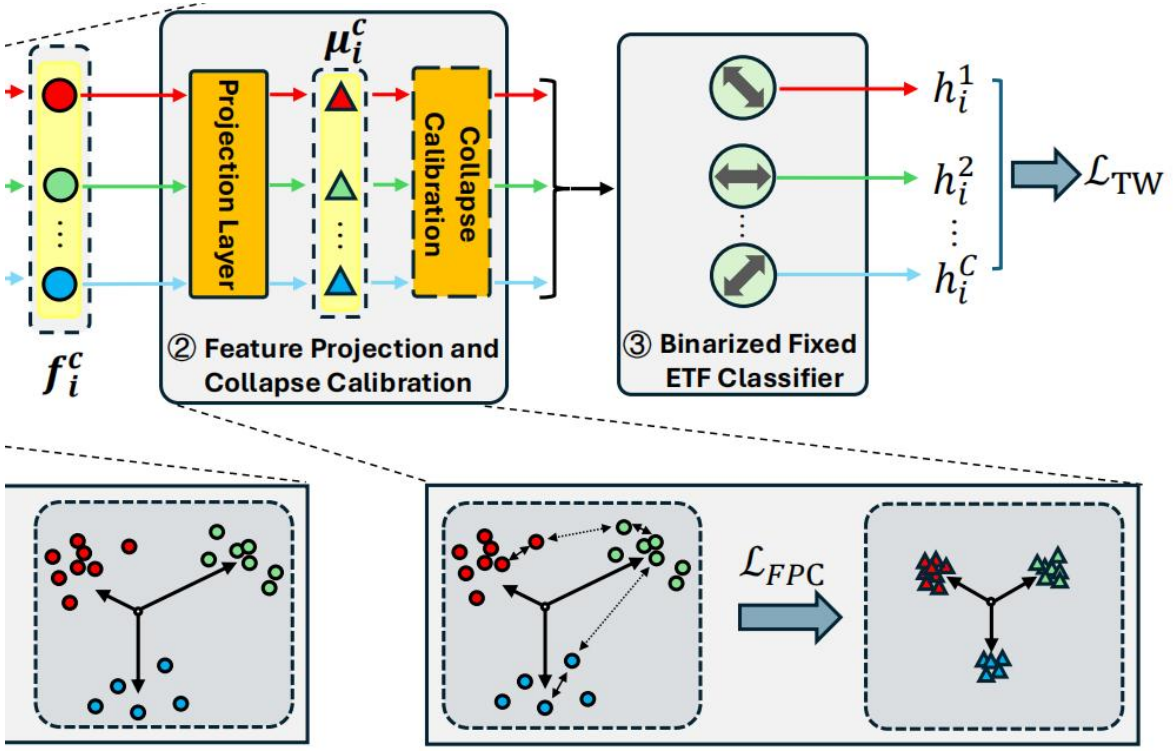
$$\mathbf{V} = \sqrt{\frac{C}{C-1}}\mathbf{U}\left(\mathbf{I}_C - \frac{1}{C}\mathbf{J}_C\right)$$

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

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

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

Method (Feature Projection and Collapse Calibration特征投影与塌陷校准)



同一标签在不同实例中的特征可能有所不同，在高维场景中统一提取的特征fc i变得具有挑战性，有必要将特征投影到低维并对其进行归一化。

$$\hat{\mu}_i^c = g(\mathbf{p_c}; \mathbf{f_i^c}), \quad \mu_i^c = \frac{\hat{\mu}_i^c}{\|\hat{\mu}_i^c\|_2}. \qquad \mu_i^c \in \mathbb{R}^{p \times C}$$

（i）如果特征提取器的最后一层使用非线性激活，例如ReLU，则原始特征fc将包含零，变得稀疏。这导致特征之间容易正交，使得它们难以折叠成ETF结构。

（ii）由于样本之间的差异，同一标签的高维特征可能包含不同的信息。通过将它们投影到较低维度，我们细化特征并减少它们的可变性。这使我们能够更好地利用NC1的原则来优化特征。

$$S_i^{c,k} = \mathrm{sim}(\mu_{i,t}^c, \mu_{t-1}^k)$$

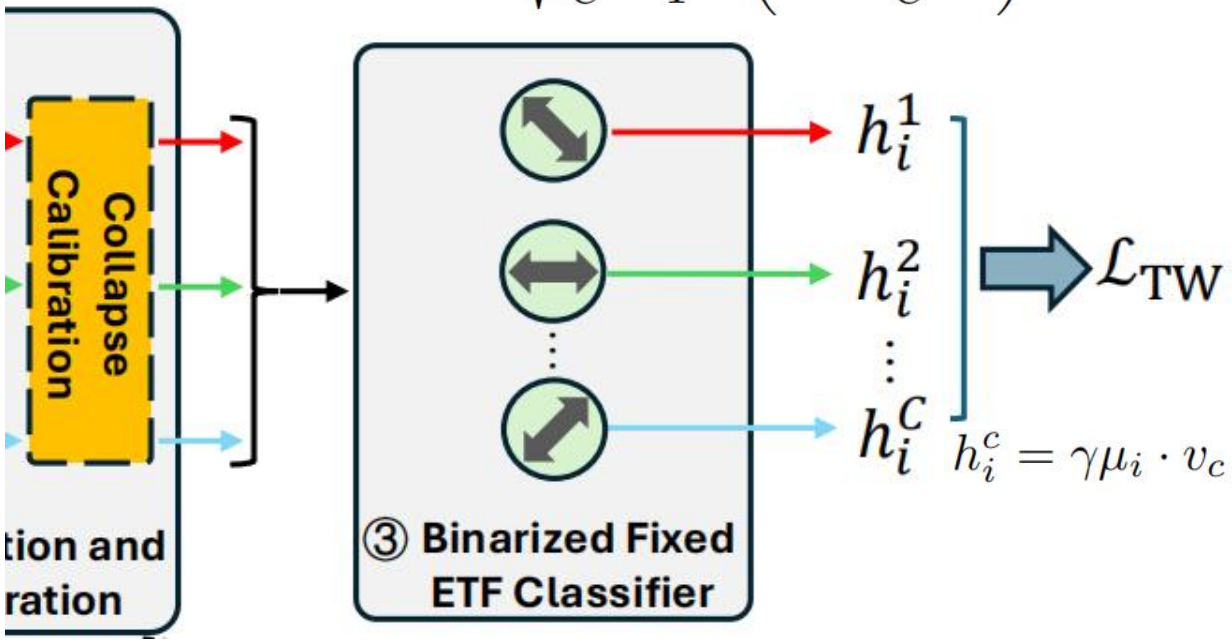
$$\mu_t^c = \frac{1}{n_c} \sum_{i=1}^{N} \mathbb{1}(y_i^c = 1) \mu_{i,t}^c$$

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

$\mu^c_{i,t}$ 表示第 t 个 epoch的类别中心。，我们计算每个类，并将其用作下一

Method (Binarized Fixed ETF Classifier二值化固定ETF分类器)

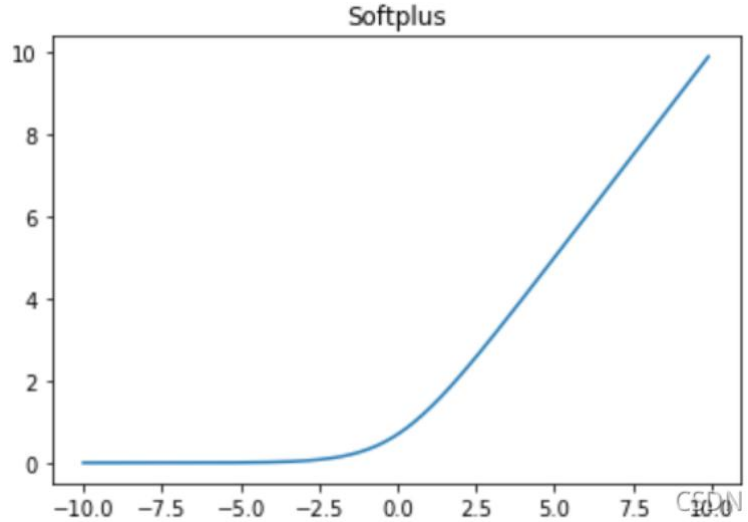
$$\mathbf{V} = \sqrt{\frac{C}{C-1}}\mathbf{U}\left(\mathbf{I}_C - \frac{1}{C}\mathbf{J}_C\right)$$



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 $\mathcal{D} = p \times C$.

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

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



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

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

Result

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	<u>51.13</u>	57.05	51.06	78.94	<u>73.22</u>	84.18	79.30
	URS	<u>56.90</u>	**54.13**	<u>60.59</u>	54.47	<u>81.44</u>	**75.68**	<u>85.53</u>	82.69
	MFM	55.25	48.71	58.24	<u>57.08</u>	79.64	66.32	84.69	<u>85.83</u>
	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 <u>underline</u> notes.