南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Long-Tailed Out-of-Distribution Detection: Prioritizing Attention to Tail

Yina He[1], Lei Peng[1], Yongcun Zhang[1], Juanjuan Weng[2*], Shaozi Li[1,3], Zhiming Luo[1,3*]

[1]Department of Artificial Intelligence, Xiamen University, Xiamen, China
[2]College of Information Science and Technology, Jinan University, Guangzhou, China
[3]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University

AAAI 2025

Current **out-of-distribution (OOD) detection** methods typically assume balanced **in-distribution (ID) data**, while most real-world data follow a **long-tailed distribution**.

Outlier Exposure (OE)
(Hendrycks, Mazeika, and Dietterich 2018)

$$\mathcal{L}_{OE} = \mathbb{E}_{x,y\sim\mathcal{D}_{in}}[\ell(f(x), y] + \gamma\mathbb{E}_{x\sim\mathcal{D}_{out}}[\ell(f(x), u)], \quad (1)$$

Supervised Contrastive Learning (SCL)
(Khosla et al. 2020)

$$\mathcal{L}_{scl}^{in}(\boldsymbol{z}_i, y) = -\log\left\{\frac{1}{|B_y|}\sum_{p\in B_y}\frac{e^{\boldsymbol{z}_i\cdot\boldsymbol{z}_p/\tau}}{\sum_{j=1}^{K}\sum_{a\in B_j}e^{\boldsymbol{z}_i\cdot\boldsymbol{z}_a/\tau}}\right\}, \quad (2)$$

Logit Adjustment (LA)
(Menon et al. 2021)

$$\mathcal{L}_{\text{la}}(\boldsymbol{z}_i, y) = -\log\frac{\pi_y e^{\varphi_y(\boldsymbol{z}_i)}}{\sum_{y'\in\mathcal{Y}}\pi_{y'}e^{\varphi_{y'}(\boldsymbol{z}_i)}}, \quad (3)$$

Framework

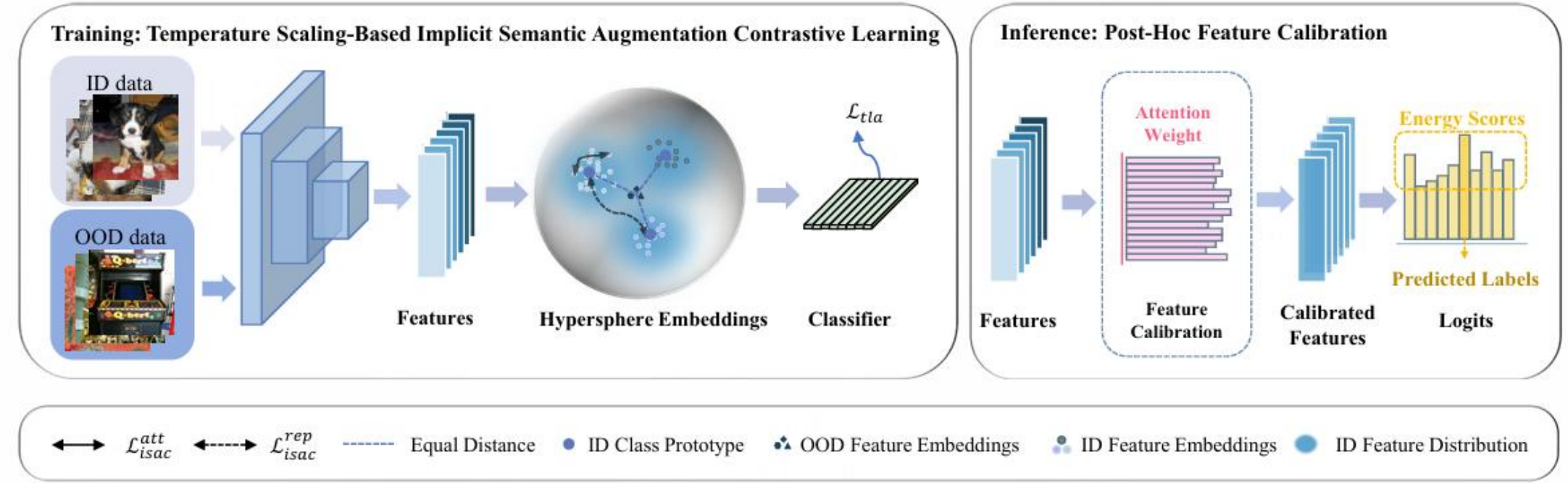
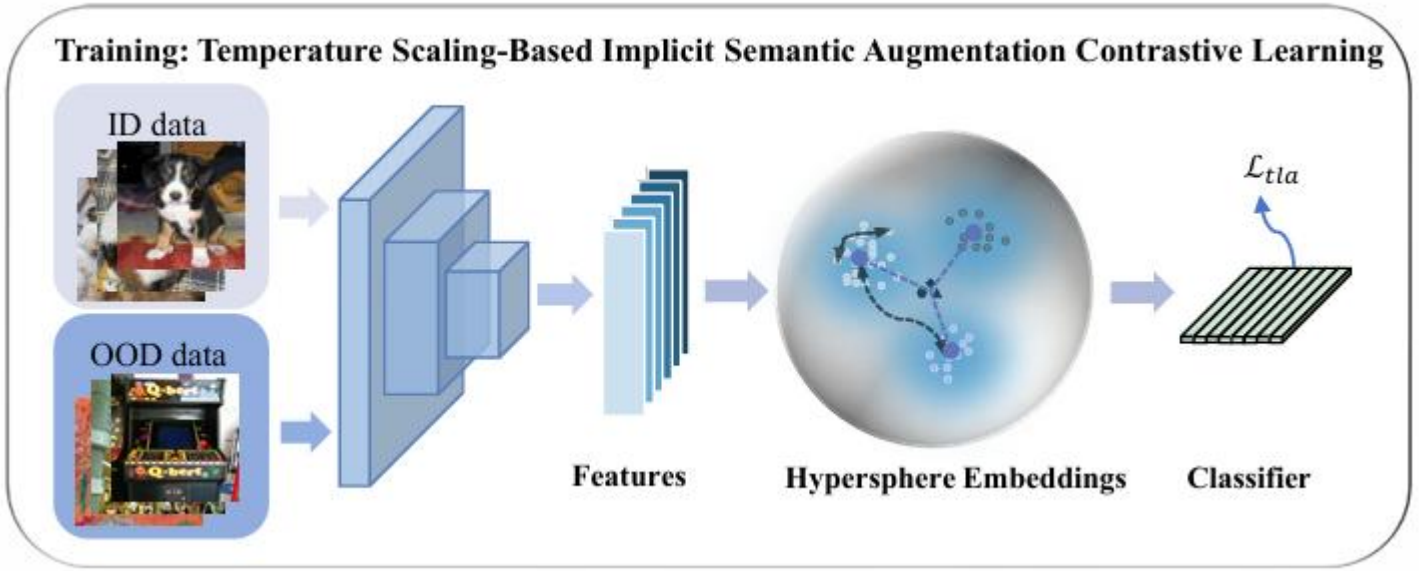


Figure 2: Overview of the proposed framework. The framework consists of a temperature scaling-based implicit semantic augmentation training phase and a feature calibration inference phase. We jointly optimize two complementary terms to encourage desirable hypersphere embeddings: an implicit semantic augmentation contrastive loss to encourage a balanced feature encoder and a temperature scaling-based logit adjustment loss to encourage a balanced high-confidence classifier. Feature calibration fine-tunes features during the inference phase by using an attention weight extracted from the training set, thereby achieving desirable ID classification and OOD detection results.

Training: von Mises-Fisher (vMF) Distributions



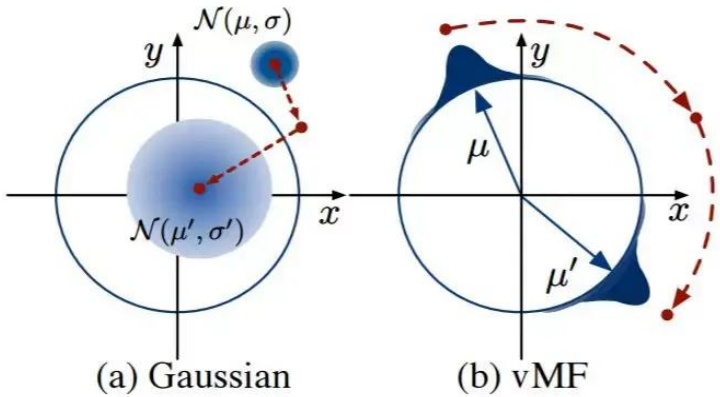
a mixture of vMF distributions

$$P_d(\boldsymbol{z}) = \sum_{y=1}^{K} P_d(y) P_d(\boldsymbol{z}|y) = \sum_{y=1}^{K} \pi_y Z_d(\kappa) e^{\kappa \boldsymbol{\mu}_y^\top \boldsymbol{z}}, \quad (5)$$

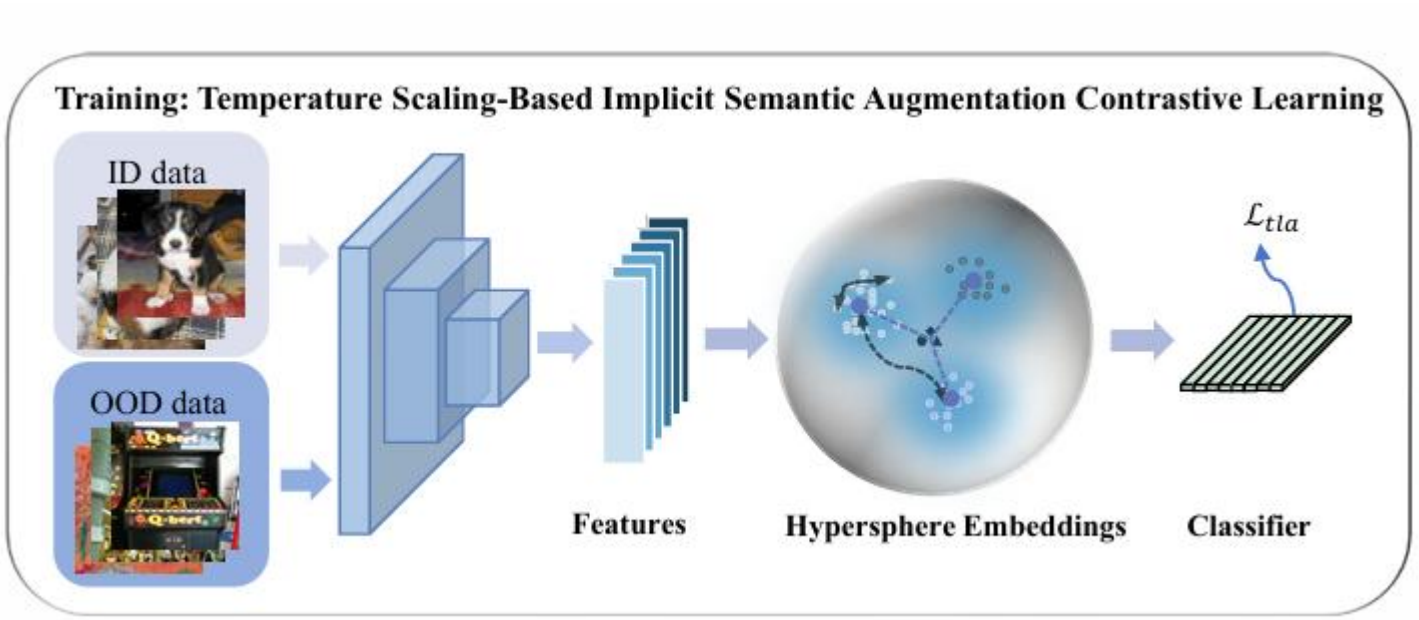
von Mises-Fisher (vMF) distributions

$$P_d(\boldsymbol{z}; \boldsymbol{\mu}_y, \kappa_y) = Z_d(\kappa_y) e^{\kappa_y \boldsymbol{\mu}_y^\top \boldsymbol{z}},$$

- $\boldsymbol{z} \in \mathbb{R}^d$ is a unit vector
- $\boldsymbol{\mu}_y$ is the class prototype
- $\kappa_y \geq 0$ indicates the concentration of the distribution
- $Z_d(\kappa_y)$ serving as the normalization factor.



Training: Implicit Semantic Augmentation Contrastive Learning



Temperature Scaling-Based Logit Adjustment

ICLR2021

$$\mathcal{L}_{\text{tla}}(\boldsymbol{z}_i, y) = -\log \frac{\pi_y e^{\varphi_y(\boldsymbol{z}_i)/\varepsilon}}{\sum_{y' \in \mathcal{Y}} \pi_{y'} e^{\varphi_{y'}(\boldsymbol{z}_i)/\varepsilon}}. \quad (7)$$

- hyperparameter ε

通过数学推导直接计算无限采样下的期望损失:

$$\mathcal{L}_{\text{isac}}(\boldsymbol{z}_i, y) = \log \left\{ \sum_{j=1}^{K} \frac{\pi_j Z_d(\tilde{\kappa}_y) Z_d(\kappa_j)}{\pi_y Z_d(\kappa_y) Z_d(\tilde{\kappa}_j)} \right\}, \quad (6)$$

- $\tilde{\kappa}_j = ||\kappa_j \boldsymbol{\mu}_j + \boldsymbol{z}_i/\tau||_2$，类别 j 的原分布参数，i样本的方向向量
- $\boldsymbol{z}_j \sim \text{vMF}(\boldsymbol{\mu}_j, \kappa_j)$
- τ is a temperature parameter

Finally, the overall loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{isac}} + \alpha \mathcal{L}_{\text{tla}} + \beta \mathcal{L}_{\text{out}},$$

- α and β are hyperparameters

Inference: Attention Weight

The **penultimate feature layer** is more correlated with the final classification.

- The attention weight extracted from **a class-balanced ID dataset** and **an OOD dataset**

$$\mathcal{X}^{cb} = \mathcal{D}_{in}^{cb} \cup \mathcal{D}_{out}, \quad x_i \in \mathcal{X}^{cb}$$

- d-dimensional feature embedding

$$z_i = f(x_i, \theta) = [z_i^1, z_i^2, \ldots, z_i^d]^T$$

- The score of z_i being predicted as class y_i is

$$S_{y_i}(z_i) = \varphi(z_i)$$

- The importance of k-th dimension of z_i is defined as $I^k(z_i) = \frac{\partial S_{y_i}(z_i)}{\partial z_i^k} \cdot z_i^k$

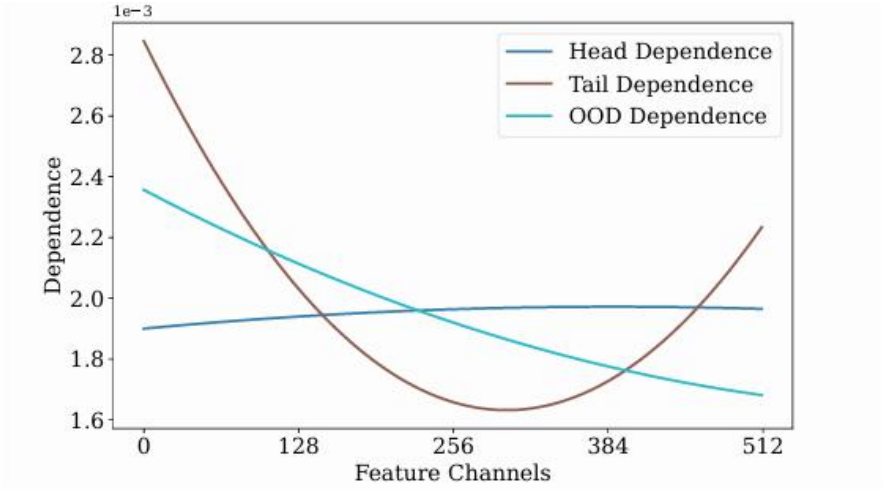


Figure 3: It visualizes the dependence of OOD, head class, and tail class samples on feature channels, showing that these three types of samples rely on different feature channels.

$$A = \frac{1}{|N|} \sum_{j=1}^{K} \left\{ \sum_{i \in N_j^{in}} \frac{I(z_i^{in})}{\pi_j} - \sum_{i \in N_j^{out}} \frac{I(z_i^{out})}{\pi_j} \right\}, \quad (9)$$

They scale A to the range [0, 2], where values between [0,1] are attenuated and values between [1,2] are enhanced.

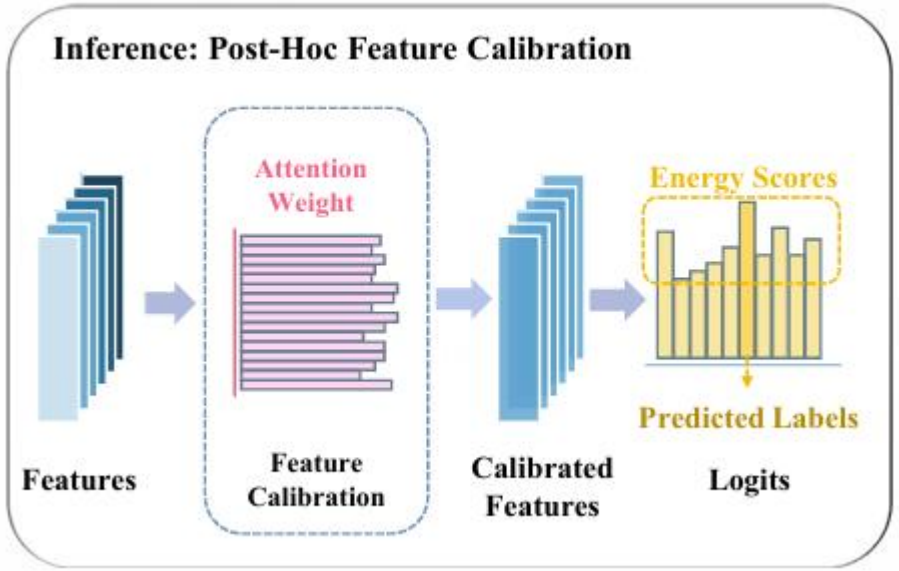
Inference: OOD Score

$$A = \frac{1}{|N|}\sum_{j=1}^{K}\left\{\sum_{i\in N_j^{in}}\frac{I(z_i^{in})}{\pi_j} - \sum_{i\in N_j^{out}}\frac{I(z_i^{out})}{\pi_j}\right\}, \quad (9)$$

$$z^{cal} = z \odot A^{scale}$$

- Instead of using maximum softmax probability(MSP) as OOD scores, we are inspired by energyOE(Liu etal.2020) and use energy scores as OOD scores in the inference phase, which is defined as follows:

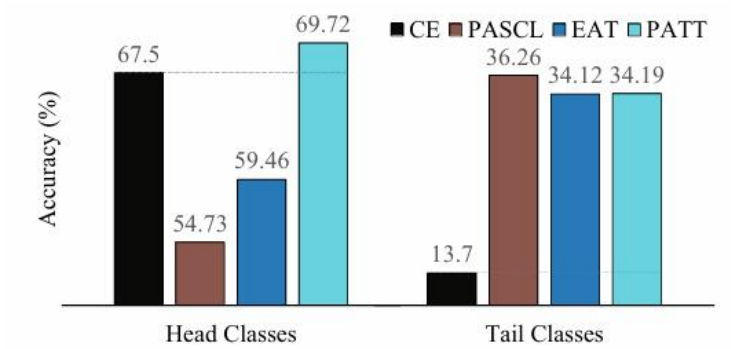
$$\mathcal{S}_{\text{ood}}(\boldsymbol{z_i}) = \log\sum_{j=1}^{K} e^{\varphi_j(\boldsymbol{z}_i^{cal})}.$$



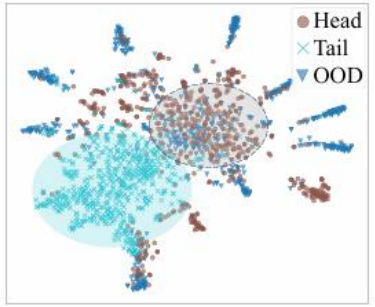
Results

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

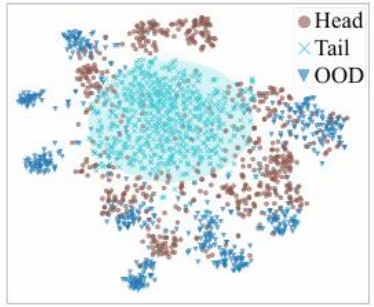
- They conduct experiments on widely used datasets, i.e., CIFAR10-LT,CIFAR100-LT (Caoetal.2019), and ImageNet-LT (Liuetal.2019) as ID training sets (D_{in}) .



(a) Separate ACC for head and tail on ImageNet-LT.



(b) PASCL's feature distribution



(c) PATT's feature distribution

$\mathcal{D}_{out}^{test}$	Method	AUROC↑	AUPR-in↑	AUPR-out↑	FPR95↓
Texture	OE	$92.59_{\pm 0.4}$	$96.01_{\pm 1.4}$	$83.32_{\pm 1.7}$	$25.10_{\pm 1.1}$
	PASCL	$93.16_{\pm 0.4}$	$96.57_{\pm 1.2}$	$84.80_{\pm 1.5}$	$\mathbf{23.26}_{\pm 0.9}$
	Ours	$\mathbf{93.96}_{\pm 0.6}$	$\mathbf{97.69}_{\pm 0.8}$	$\mathbf{86.49}_{\pm 0.8}$	$26.65_{\pm 1.6}$
SVHN	OE	$95.10_{\pm 1.0}$	$91.59_{\pm 0.5}$	$97.14_{\pm 0.8}$	$16.15_{\pm 1.5}$
	PASCL	$96.63_{\pm 0.9}$	$92.89_{\pm 0.5}$	$98.06_{\pm 0.6}$	$12.18_{\pm 3.3}$
	Ours	$\mathbf{98.21}_{\pm 0.7}$	$\mathbf{97.19}_{\pm 0.6}$	$\mathbf{98.50}_{\pm 0.5}$	$\mathbf{5.73}_{\pm 2.0}$
CIFAR100	OE	$83.40_{\pm 0.3}$	$84.06_{\pm 0.3}$	$80.93_{\pm 0.6}$	$56.96_{\pm 0.9}$
	PASCL	$84.43_{\pm 0.2}$	$85.32_{\pm 0.5}$	$82.99_{\pm 0.5}$	$57.27_{\pm 0.9}$
	Ours	$\mathbf{85.36}_{\pm 0.2}$	$\mathbf{86.01}_{\pm 0.3}$	$\mathbf{83.29}_{\pm 0.5}$	$\mathbf{51.12}_{\pm 0.9}$
Tiny ImageNet	OE	$86.14_{\pm 0.3}$	$89.88_{\pm 0.7}$	$79.33_{\pm 0.7}$	$47.78_{\pm 0.7}$
	PASCL	$87.14_{\pm 0.2}$	$90.22_{\pm 0.5}$	$81.54_{\pm 0.4}$	$47.69_{\pm 0.6}$
	Ours	$\mathbf{88.62}_{\pm 0.2}$	$\mathbf{90.82}_{\pm 0.7}$	$\mathbf{84.54}_{\pm 0.1}$	$\mathbf{41.30}_{\pm 1.7}$
LSUN	OE	$91.35_{\pm 0.2}$	$93.06_{\pm 0.3}$	$87.62_{\pm 0.8}$	$27.86_{\pm 0.7}$
	PASCL	$\mathbf{93.17}_{\pm 0.15}$	$82.59_{\pm 0.3}$	$91.76_{\pm 0.5}$	$26.40_{\pm 1.0}$
	Ours	$91.64_{\pm 0.3}$	$\mathbf{93.16}_{\pm 0.6}$	$\mathbf{92.27}_{\pm 0.1}$	$\mathbf{24.41}_{\pm 1.5}$
Place365	OE	$90.07_{\pm 0.3}$	$82.09_{\pm 0.4}$	$95.15_{\pm 0.2}$	$34.04_{\pm 0.9}$
	PASCL	$91.43_{\pm 0.2}$	$82.59_{\pm 0.2}$	$96.28_{\pm 0.1}$	$33.40_{\pm 0.9}$
	Ours	$\mathbf{91.95}_{\pm 0.6}$	$\mathbf{82.63}_{\pm 0.3}$	$\mathbf{97.81}_{\pm 0.2}$	$\mathbf{30.15}_{\pm 1.9}$
Average	OE	$89.77_{\pm 0.3}$	$89.45_{\pm 0.4}$	$87.25_{\pm 0.6}$	$34.65_{\pm 0.5}$
	PASCL	$90.99_{\pm 0.2}$	$90.18_{\pm 0.4}$	$89.24_{\pm 0.3}$	$33.36_{\pm 0.8}$
	Ours	$\mathbf{91.62}_{\pm 0.5}$	$\mathbf{91.25}_{\pm 0.5}$	$\mathbf{90.48}_{\pm 0.3}$	$\mathbf{29.89}_{\pm 1.4}$

(a) Comparison of PATT to PASCL and OE on six OOD datasets.

Table 1: Comparison results on CIFAR10-LT. The best results are shown in bold, and the second-best results are underlined.

Ablation Results

ID Dataset	ISAC	TLA	FC	AUROC↑	AUPR-in↑	AUPR-out↑	FPR95↓	ACC↑	ACC-t↑
CIFAR10-LT	✗	✗	✗	71.21	74.84	64.37	58.09	78.10	63.00
	✓	✗	✗	74.10	78.70	66.43	49.09	79.55	66.47
	✗	✓	✗	84.70	83.68	81.12	48.18	81.11	70.77
	✗	✗	✓	79.75	80.48	74.38	52.17	79.55	67.50
	✓	✓	✗	91.06	90.90	88.92	32.35	82.09	70.00
	PATT			**91.62**	**91.25**	**90.48**	**29.89**	**84.77**	**79.67**
CIFAR100-LT	✗	✗	✗	71.70	72.77	65.95	72.22	44.68	10.97
	✓	✗	✗	74.43	76.80	67.50	64.91	45.49	17.00
	✗	✓	✗	72.10	73.88	66.35	69.80	49.01	27.12
	✗	✗	✓	72.95	74.43	66.51	70.79	44.90	12.72
	✓	✓	✗	75.48	77.19	68.87	64.25	49.56	27.12
	PATT			**76.25**	**77.84**	**70.37**	**62.94**	**50.07**	**31.03**
ImageNet-LT	✗	✗	✗	67.87	44.28	69.74	88.76	45.13	9.24
	✓	✗	✗	68.39	45.31	79.39	83.39	48.77	13.79
	✗	✓	✗	72.74	51.17	86.25	82.08	50.27	31.58
	✗	✗	✓	72.07	47.84	82.07	82.36	46.17	14.01
	✓	✓	✗	73.77	50.10	87.32	81.92	55.06	33.17
	PATT			**74.13**	**51.41**	**87.43**	**80.57**	**55.14**	**34.19**

Table 4: Ablation results of three key modules for PATT on CIFAR10-LT, CIFAR100-LT and ImageNet-LT.

The **penultimate feature layer** is more correlated with the final classification.

- The attention weight extracted from **a class-balanced ID dataset** and **an OOD dataset**

$$\mathcal{X}^{cb} = \mathcal{D}_{in}^{cb} \cup \mathcal{D}_{out}, \quad x_i \in \mathcal{X}^{cb}$$

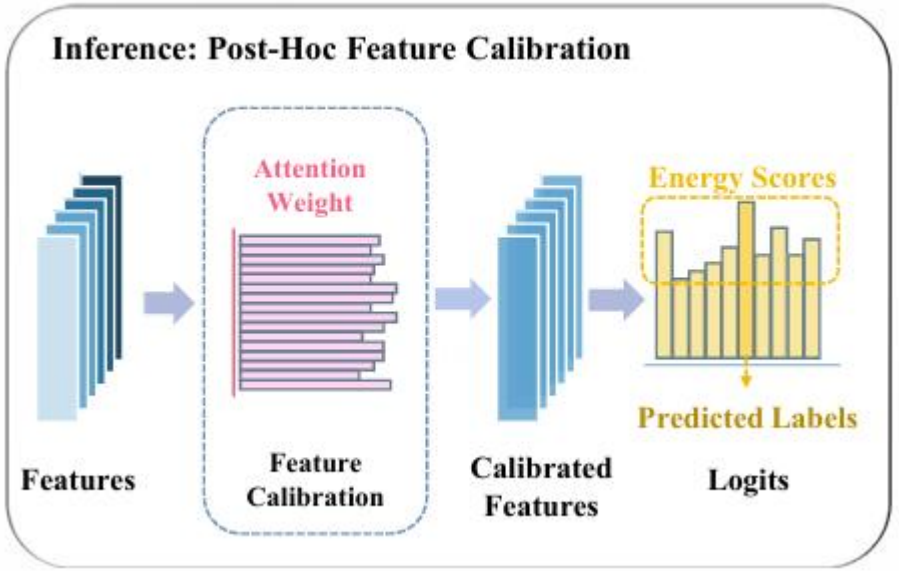
- d-dimensional feature embedding

$$z_i = f(x_i, \theta) = [z_i^1, z_i^2, \ldots, z_i^d]^T$$

- The score of z_i being predicted as class y_i is $S_{y_i}(z_i) = \varphi(z_i)$

- The importance of k-th dimension of z_i is defined as $I^k(z_i) = \frac{\partial S_{y_i}(z_i)}{\partial z_i^k} \cdot z_i^k$

Class-Balanced?



Thanks