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

Abstract

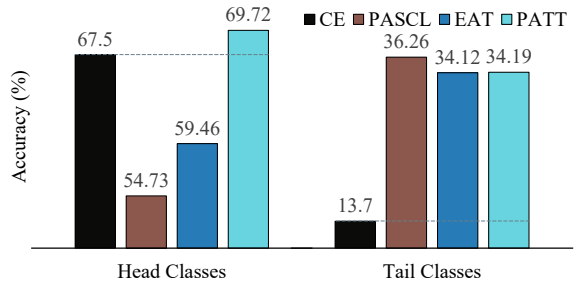
Current out-of-distribution (OOD) detection methods typically assume balanced in-distribution (ID) data, while most real-world data follow a long-tailed distribution. Previous approaches to long-tailed OOD detection often involve balancing the ID data by reducing the semantics of head classes. However, this reduction can severely affect the classification accuracy of ID data. The main challenge of this task lies in the severe lack of features for tail classes, leading to confusion with OOD data. To tackle this issue, we introduce a novel Prioritizing Attention to Tail (PATT) method using augmentation instead of reduction. Our main intuition involves using a mixture of von Mises-Fisher (vMF) distributions to model the ID data and a temperature scaling module to boost the confidence of ID data. This enables us to generate infinite contrastive pairs, implicitly enhancing the semantics of ID classes while promoting differentiation between ID and OOD data. To further strengthen the detection of OOD data without compromising the classification performance of ID data, we propose feature calibration during the inference phase. By extracting an attention weight from the training set that prioritizes the tail classes and reduces the confidence in OOD data, we improve the OOD detection capability. Extensive experiments verified that our method outperforms the current state-of-the-art methods on various benchmarks.

Code — https://github.com/InaR-design/PATT.

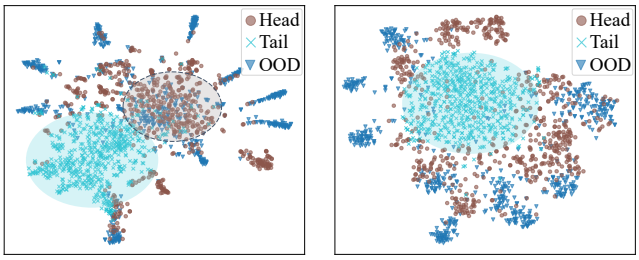
Introduction

When confronted with a sample that does not match any of the known classes, deep neural networks (DNNs) tend to predict this OOD sample as one of the training classes with high confidence (Hendrycks and Gimpel 2017; Hein, Andriushchenko, and Bitterwolf 2019; Hsu et al. 2020). To address this issue, numerous OOD detection methods have been proposed and achieved significant improvements (Hendrycks, Mazeika, and Dietterich 2018; Liu et al. 2020; Mohseni et al. 2020). Current state-of-the-art methods introduce surrogate OOD datasets during training, enabling the model to see beyond the training set (Zhu et al. 2024b; Ming et al. 2023). These methods tackle the OOD detection task by maximizing uncertainty (Hendrycks, Mazeika, and Dietterich 2018) and using informative extrapolation based on surrogate outliers (Zhu et al. 2024b). However, most of these approaches assume a balanced ID data, a condition not typically hold in real-world scenarios characterized by long-tail distributions (e.g., Cybersecurity (Yi et al. 2021) and Autonomous Driving (Kendall and Gal 2017)).

In long-tailed recognition, the data distribution is highly imbalanced, and combining long-tailed recognition with OOD detection methods still fails to achieve optimal results (Wang et al. 2022). Recent methods like PASCL (Wang et al. 2022) and EAT (Wei, Wang, and Zhang 2024) focus on differentiating tail classes from OOD data by handling head and tail classes differently. However, these methods significantly reduce ID classification accuracy compared to long-



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



(b) PASCL's feature distribution (c) PATT's feature distribution

Figure 1: Visualization of the Comparison between PATT and other methods on ImageNet-LT. (a) Comparison of separate accuracy for head and tail classes on ImageNet-LT between PATT and other methods. (b) (c) Visualization of feature distribution across the top ten classes (Head), the bottom ten classes (Tail), and OOD data from PASCL and PATT.

*Corresponding author

tailed learning approaches. To identify the cause of the reduction, Fig. 1 (a) compares the accuracy of head and tail classes on ImageNet-LT (Liu et al. 2019) across different methods, where CE stands for cross-entropy. We can observe that after incorporating the OOD detection task, both PASCL and EAT show a decline in head class accuracy. For further analysis, Fig. 1 (b) shows PASCL's feature distribution, where head classes appear slightly overfitted, with tail classes, head classes, and OOD data intermixed, shedding light on the decline in head class accuracy.

Motivated by these observations, we propose a Prioritizing Attention to Tail (PATT) method for long-tailed OOD detection. The core idea of our design is: address the imbalance in ID data and subsequently enhance the distinction in confidence levels between ID and OOD samples. Specifically, we propose a temperature scaling-based implicit semantic augmentation contrastive learning (TISAC). Compared to traditional supervised contrastive learning (Khosla et al. 2020), TISAC incorporates two key insights tailored for long-tailed OOD detection: (1) *implicit semantic augmentation contrastive learning* effectively balances imbalanced ID data. Particularly, it models the ID data using a mixture of von Mises-Fisher (vMF) distributions, allowing us to sample a large number of contrastive pairs. Yet, sampling sufficient data from the vMF distributions at each training iteration is still inefficient, and we extend the number of samples to infinity and rigorously derive a closed-form for the expected contrastive loss. (2) *temperature scaling*, which ensures high confidence in ID data, thereby increasing the confidence gap between ID and OOD data. By doing so, our method improves accuracy for both head and tail classes as shown in Fig. 1 (a), where CE is a pure classification model, and the other three methods are designed for long-tailed OOD tasks.

To further boost the long-tailed OOD detection, we propose Post-Hoc Feature Calibration, which derives an attention weight from the training set to refine features during the inference phase, yielding a more balanced ID feature distribution and clearly distinguishing ID and OOD data based on confidence. Ultimately, the feature distribution of our method shows that both head and tail classes occupy nearly equal space, and OOD data is similarly distanced from each class as illustrated in Fig. 1 (c). Our key contributions are:

- We identified the potential issues in current long-tailed OOD detection methods, requiring a trade-off where improved OOD detection performance comes at the expense of ID classification accuracy.
- We propose a novel Prioritizing Attention to Tail (PATT) framework, composed of temperature scaling-based implicitly augmented contrastive learning and post-hoc feature calibration, functioning during the training and inference stages respectively. The former ensures a balanced feature extractor and classifier, while the latter calibrates features to focus more on tail classes.
- Extensive experiments were conducted to validate the effectiveness of PATT in improving long-tailed recognition and OOD detection performance, resulting in a 4.29% increase in AUROC and an 8.35% increase in ID classification accuracy on ImageNet-LT.

Related Work

OOD Detection The OOD detection (Nguyen, Yosinski, and Clune 2015) aims to determine whether an input sample belongs to ID classes or OOD classes. Some works enhance performance by generating virtual OOD data, such as DivOE (Zhu et al. 2024a), VOS (Du et al. 2022) and NPOS (Tao et al. 2023). These methods adaptively sample virtual outliers from low-likelihood regions to extrapolate the feature distribution of the OOD data and obtain a more reasonable decision boundary. Other works leverage original OOD data, such as Outlier Exposure (OE) (Hendrycks, Mazeika, and Dietterich 2018) and EnergyOE (Liu et al. 2020). They exploits information from OOD data by enforcing a uniform distribution to its logits or maximizing the free energy of OOD samples. Additionally, there are post-hoc strategies that focus on designing new OOD scoring functions, which are always used in conjunction with the aforementioned methods, such as MSP, EnergyOE, and ODIN. Despite the high performance of existing OOD detectors, they are typically trained on class-balanced ID datasets and cannot be directly applied to long-tailed tasks.

Long-Tailed Recognition (LTR) Early LTR methods primarily involve re-sampling (Buda, Maki, and Mazurowski 2018; Byrd and Lipton 2019; He and Garcia 2009; Wallace et al. 2011) and re-weighting (Huang et al. 2016; Cui et al. 2019), which are effective but limited in addressing intra-class diversity in tail classes, leading to overfitting. To increase the intra-class diversity of tail data, many data augmentation methods have been employed, but explicit data augmentation methods (Bowles et al. 2018; Zhang et al. 2018; Yun et al. 2019) require significant time and resources. ISDA (Wang et al. 2019), an implicit data augmentation method that utilizes a mixture of Gaussian distribution to generate infinite samples, is an excellent solution to the aforementioned issue. Subsequently, RISDA (Chen et al. 2022) and ProCo (Du et al. 2024)emerged, which effectively applies ISDA to long-tailed learning. Although these LTR methods exhibit excellent performance in the long-tailed classification, they lack specific designs for OOD samples.

Long-tailed OOD detection Long-tailed OOD detection has garnered increasing attention, and several methods (Jiang et al. 2023; Wang et al. 2022; Wei et al. 2022; Choi, Jeong, and Choi 2023) have been proposed for this challenging task. Among them, PASCL (Wang et al. 2022) and BERL (Choi, Jeong, and Choi 2023) are OE-based methods. Specifically, PASCL optimizes the contrastive objective between tail class samples and OOD data, pushing them apart in the latent feature space. BERL (Choi, Jeong, and Choi 2023) is a balanced version of EnergyOE (Liu et al. 2020). Recent methods (Wei, Wang, and Zhang 2024; Miao et al. 2024) utilize Outlier Class Learning, which directly learns outlier classes for OOD data. EAT (Wei, Wang, and Zhang 2024) builds on this by learning multiple outlier classes and employing Cutmix (Yun et al. 2019) augmentation on the tail with OOD data, which encourages the model to focus on the foreground. COCL (Miao et al. 2024) uses a

partial contrastive learning approach similar to PASCL and performs logit calibration during inference. However, these methods reduce semantic information for head classes to improve OOD detection performance, while sacrificing the ID classification accuracy.

Preliminaries

Task Definition: Following (Wei, Wang, and Zhang 2024; Wang et al. 2022), let $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ denote the training set, characterized by a long-tailed ID set, and a surrogate OOD set, respectively. The entire task can be seen as a combination of a multi-class classification for ID data and a binary classification for OOD detection. Let $\mathcal{X} = \mathcal{D}_{in} \cup \mathcal{D}_{out}$ and $Y^{in} = \{1, 2, \ldots, K\}$ denote the input and label space of the ID data. OOD detection in LTR aims to learn a encoder $f\colon \mathcal{X} \to \mathcal{Z}$ and a classifier $\varphi\colon \mathcal{Z} \to \mathcal{Y}$ such that for any test data $x \in \mathcal{X}$: if x is drawn from $\mathcal{D}_{in}$, then model can classify x into the correct ID class, otherwise if x is drawn from $\mathcal{D}_{out}$, then model can detect x as OOD data.

Outlier Exposure (OE) OE (Hendrycks, Mazeika, and Dietterich 2018) formulates the training objective for OOD detection as follows:

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

where γ is a hyperparameter, u represents a uniform distribution, and ℓ denotes the cross entropy loss. We define $\mathcal{L}_{out} = \mathbb{E}_{x\sim\mathcal{D}_{out}}[\ell(f(x), u]$. However, CE loss is not an optimal solution for long-tailed recognition, as it treats each sample equally and tends to optimize for those classes that appear more frequently (Liu et al. 2019).

Supervised Contrastive Learning (SCL) SCL (Khosla et al. 2020) is a form of contrastive learning in a supervised setting, achieving classification by repelling instances from different classes and attracting those from the same class, and has recently emerged as a paradigm for long-tailed recognition. For a feature z_i of class y in a batch B, SCL minimize the following loss function:

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

where $\boldsymbol{z}_i$ denotes the feature of x_i drawn from f, B_y is a subset of B that contains all instances with the same label y, and $|B_y|$ is its cardinality. $\tau > 0$ is a scalar temperature hyperparameter that controls tolerance to similar samples; a smaller temperature leads to less tolerance for similar samples (Wang and Liu 2021).

Logit Adjustment (LA) LA (Menon et al. 2021) aims to refine the model's output probabilities to better match the true probabilities observed in the data. In LTR, it is often used in conjunction with contrastive learning (Zhu et al. 2022; Hong et al. 2021; Tan et al. 2020), the former focuses on representation learning, while the latter learns a balanced classifier. Its definition is as follows:

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

Here, $\pi_y = \frac{N_y}{N}$ denotes the class prior of label y, $\varphi_y(\boldsymbol{z}_i)$ is the logits of class y drawn from $\boldsymbol{z}_i$.

The Proposed Method

Framework

The overview of our unified end-to-end model PATT is shown in Fig. 2. It consists of two main components, i.e., *temperature scaling-based implicit semantic augmentation contrastive learning (TISAC)* and *post-hoc feature calibration (FC)*. TISAC is designed to achieve balanced representation learning and classifier for ID data, ensuring high confidence in ID data for better OOD detection capability. FC prioritizes attention to the features of tail classes and reduces the confidence of OOD data, thus benefiting both ID classification and OOD detection. Below, we introduce each component in detail.

Temperature Scaling-Based Implicit Semantic Augmentation Contrastive Learning

Samples from tail classes struggle to represent the diversity of tail data. Additionally, as aforementioned, contrastive learning achieves classification by repelling instances from different classes and attracting same-class ones. Due to the severe imbalance in the training set, the repelling force from the head classes can cause the tail classes to be compressed together, making them difficult to separate. Naive reweighting and re-sampling methods are difficult to achieve ideal results, as these approaches fail to address the semantic deficiency of tail classes. Inspired by implicit semantic data augmentation methods (Wang et al. 2019; Chen et al. 2022; Du et al. 2024), we propose a temperature scaling-based implicit semantic augmentation contrastive learning module (TISAC), which consists of two components: Implicit Semantic Augmentation Contrastive Learning (ISAC) and Temperature Scaling-Based Logit Adjustment (TLA). ISAC is inspired by ISDA, which employs Gaussian distributions to model unconstrained features and obtains an upper bound on the expected loss. However, Our method diverges significantly from ISDA-based methods. Compared to Gaussian distribution, the vMF distribution focuses on feature direction rather than magnitude, which largely reduces model bias between classes. Additionally, the Gaussian distribution requires extensive data to compute covariance matrices, making it impractical to estimate Gaussian parameters for all classes with long-tailed data. In contrast, vMF's mean direction vector and concentration parameter can be effectively calculated across batches. Moreover, we rigorously derived the expected loss instead of an upper bound. TLA is a temperature scaling-based version of logit adjustment, which, when combined with OE, creates a larger separation between ID and OOD data.

Hypersphere Distribution Given the desirable properties of hypersphere embeddings, we choose a mixture of von Mises-Fisher (vMF) distributions (Mardia and Jupp 2009) to model the ID data. The probability density function of the vMF distribution for a unit vector $\boldsymbol{z} \in \mathbb{R}^d$ is defined as:

$$P_d(\boldsymbol{z}; \boldsymbol{\mu_y}, \kappa_y) = Z_d(\kappa_y)e^{\kappa_y\boldsymbol{\mu_y}^\top\boldsymbol{z}}, \quad (4)$$

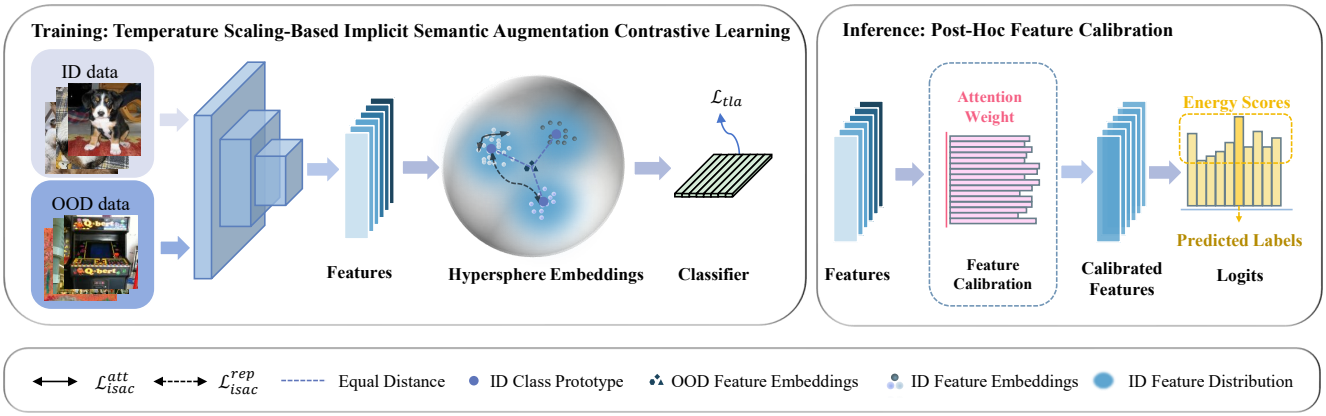


Figure 2: Overview of the proposed framework. The framework consists of a temperature scaling-based implicit semantic augmentation training phase and a feature calibration inference phase. We jointly optimize two complementary terms to encourage desirable hypersphere embeddings: an implicit semantic 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.

where $\boldsymbol{\mu_y}$ is the class prototype with unit norm of class y, $\kappa_y \geq 0$ indicates the concentration of the distribution around the mean direction $\boldsymbol{\mu_y}$, and $Z_d(\kappa_y)$ serving as the normalization factor. A larger κ_y means the distribution is more tightly concentrated around the mean direction. Vice versa when $\kappa_y = 0$, its turned into a uniform distribution. Under this probability model, we use a mixture of vMF distributions to model the feature distribution:

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

Implicit Semantic Augmentation Contrastive Learning Given the above distribution, an intuitive idea is to obtain contrastive learning pairs by infinitely sampling from the distribution. However, we realize that sampling a sufficient amount of data from the vMF distributions at each training iteration is still inefficient. Therefore, we mathematically derive a closed-form expression when considering an infinite set of training sample pairs, akin to the methods proposed in previous works (Wang et al. 2019; Du et al. 2024). Thus, we have the following formula:

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

where $\tilde{\kappa}_j = ||\kappa_j \boldsymbol{\mu}_j + \boldsymbol{z}_i/\tau||_2$, which $\boldsymbol{z}_j \sim \text{vMF}(\boldsymbol{\mu}_j, \kappa_j)$, τ is a temperature parameter. See proof in Appendix B.

Temperature Scaling-Based Logit Adjustment Through Eq. 6, we can achieve balanced representation learning for ID data. Besides, in our task, we further need high confidence in ID data to ensure excellent OOD detection capability. Therefore, we introduce Logit Adjustment (Ding et al. 2021; Joy et al. 2023; Kull et al. 2019) with a temperature scaling hyperparameter ε to achieve this goal, which is defined as follows:

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

Finally, the overall loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{isac}} + \alpha \mathcal{L}_{\text{tla}} + \beta \mathcal{L}_{\text{out}}, \quad (8)$$

where α and β are hyperparameters for each component.

Post-Hoc Feature Calibration

Attention Weight The penultimate feature layer is more correlated with the final classification, and different feature channels within it have different correlations with different classes. We further propose to use an attention weight to balance the influence of different feature channels. Unlike (Tang et al. 2023), which measures the importance of feature channels by introducing a learnable weight matrix, we aim to determine which channels are crucial for tail class classification and OOD detection by an attention weight extracted from a class-balanced ID dataset and an OOD dataset, denoted as $\mathcal{D}_{in}^{cb} \subset \mathcal{D}_{in}$ and $\mathcal{D}_{out}$. Let $\mathcal{X}^{cb} = \mathcal{D}_{in}^{cb} \cup \mathcal{D}_{out}$. Given a sample $x_i \in \mathcal{X}^{cb}$, if it comes from D_{in}^{cb}, its label is naturally y_i. If it comes from $\mathcal{D}_{out}$, its label is determined by the prediction obtained after passing through the network. For convenience, we also refer to this predicted label as y_i. Then, we obtain the corresponding d-dimensional feature embedding $z_i = f(x_i, \theta) = [z_i^1, z_i^2, \ldots, z_i^d]^T$. Then 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 $I^k(z_i) = \frac{\partial S_{y_i}(z_i)}{\partial z_i^k} \cdot z_i^k$. We can find that $I(z_i)$ represents the contribution of all channels of z_i to the correct classification. As shown in Fig. 3, we can obtain

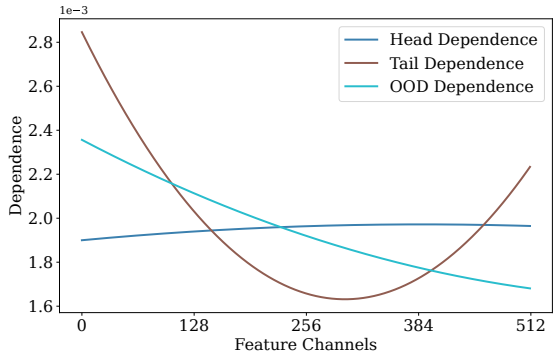


Figure 3: It visualizes the dependence of OOD, head class, and tail class samples on feature channels, showing that these three types of samples rely on different feature channels.

that OOD data, head class, and tail class depend on different feature channels. To achieve balanced results for ID data and satisfactory OOD detection performance, it is essential to derive feature-level representations that favor tail classes and enforce a uniform distribution for OOD data. Specifically, we calculate an attention weight A for each feature in the test set for calibration, which is defined as follows:

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

where $z_i^{in} \in \mathcal{D}_{in}^{cb}$, $z_i^{out} \in \mathcal{D}_{out}$ and its label j is a virtual label comes from the prediction obtained after passing through the network, $|N|$ is the cardinality of the class-balanced set $\mathcal{X}^{cb}$. Besides, the resulting attention weight A has a large variance, then we scale it to the range [0,2] as A^{scale}, where values between [0,1] are attenuated and values between [1,2] are enhanced.

OOD Score During the inference phase, we element-wise multiply this weight A^{scale} with feature z to obtain the final calibrated feature $z^{cal} = z \odot A^{scale}$. By doing so, our model will prioritize attention to the features of tail classes and reduce the confidence of OOD data. Thus, feature calibration benefits both ID classification and OOD detection. Additionally, instead of using maximum softmax probability (MSP) as OOD scores, we are inspired by energyOE (Liu et al. 2020) and use energy scores as OOD scores in the inference phase, which is defined as follows:

$$\mathcal{S}_{\text{ood}}(\boldsymbol{z_i}) = \log\sum_{j=1}^{K} e^{\varphi_j(\boldsymbol{z}_i^{cal})}. \quad (10)$$

Experiments

Experiment Settings

Datasets: We conduct experiments on widely used datasets, i.e., CIFAR10-LT, CIFAR100-LT (Cao et al. 2019), and ImageNet-LT (Liu et al. 2019) as ID training sets ($\mathcal{D}_{in}$).

$\mathcal{D}_{\text{out}}^{\text{test}}$	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.

Method	AUROC	AUPR-in	AUPR-out	FPR95	ACC
MSP	72.28	73.96	70.27	66.07	72.34
EnergyOE	89.31	91.01	88.92	40.88	74.68
OE	$89.77_{\pm 0.3}$	$89.45_{\pm 0.4}$	$87.25_{\pm 0.6}$	$34.65_{\pm 0.5}$	$73.84_{\pm 0.8}$
PASCL	$90.99_{\pm 0.2}$	$90.18_{\pm 0.4}$	$89.24_{\pm 0.3}$	$33.36_{\pm 0.8}$	$77.08_{\pm 1.0}$
EAT	$\mathbf{91.73}_{\pm 0.3}$	$\mathbf{91.46}_{\pm 0.5}$	$\underline{90.35}_{\pm 0.5}$	$\underline{30.30}_{\pm 1.2}$	$\underline{81.31}_{\pm 0.3}$
Ours	$\underline{91.62}_{\pm 0.5}$	$\underline{91.25}_{\pm 0.5}$	$\mathbf{90.48}_{\pm 0.3}$	$\mathbf{29.89}_{\pm 1.4}$	$\mathbf{84.77}_{\pm 0.2}$

(b) Comparison results with different competing methods. The results are averaged over the six OOD test datasets in (a).

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

The standard CIFAR10, CIFAR100, and ImageNet test sets are used as ID test sets (D_{in}^{test}). Following PASCL (Wang et al. 2022), we utilize 300,000 samples from TinyImages80M (Torralba, Fergus, and Freeman 2008) as the surrogate OOD training data for CIFAR10/100-LT and ImageNet-Extra as the surrogate OOD training for ImageNet-LT. We set the default imbalance ratio to 100 for CIFAR10/100-LT. For OOD test data, we use Textures (Cimpoi et al. 2014), SVHN (Netzer et al. 2011), Tiny ImageNet (Le and Yang 2015), LSUN (Yu et al. 2015), and Places365 (Zhou et al. 2017) introduced in the SC-OOD benchmark (Yang et al. 2021) as $\mathcal{D}_{test}^{out}$ for CIFAR10/100-LT. Additionally, for near-OOD experiments (Fort, Ren, and Lakshminarayanan 2021), we use CIFAR-100 as $\mathcal{D}_{test}^{out}$ for CIFAR10-LT and vice versa. We use ImageNet-1k-OOD as $\mathcal{D}_{test}^{out}$ for ImageNet-LT. More information about the datasets can be found in Appendix C.

Evaluation Metrics: Following Wang et al. (2022) and Yang et al. (2021), we use four metrics to evaluate OOD detection and ID classification: (1) **AUROC** (↑) is the area under the receiver operating characteristic curve from OOD scores; (2) **AUPR** (↑) is the area under precision-recall curve. AUPR contains AUPR-in which ID samples are treated as positive and AUPR-out is vice versa; (3) **FPR95** (↓) is the false positive rate (FPR) when 95% OOD samples have been successfully detected; (4) **ACC** (↑) is the classification accuracy of the ID data.

$\mathcal{D}_{\text{out}}^{\text{test}}$	Method	AUROC↑	AUPR-in↑	AUPR-out↑	FPR95↓
Texture	OE	76.71$_{\pm 1.2}$	85.28$_{\pm 1.0}$	58.79$_{\pm 1.4}$	68.28$_{\pm 1.5}$
	PASCL	76.01$_{\pm 0.7}$	85.84$_{\pm 1.0}$	58.12$_{\pm 1.1}$	67.43$_{\pm 1.9}$
	Ours	**76.86**$_{\pm 0.7}$	**85.86**$_{\pm 0.9}$	**59.16**$_{\pm 0.9}$	**66.64**$_{\pm 1.4}$
SVHN	OE	77.61$_{\pm 3.3}$	73.25$_{\pm 1.5}$	86.82$_{\pm 2.5}$	58.04$_{\pm 4.8}$
	PASCL	80.19$_{\pm 2.2}$	67.81$_{\pm 2.4}$	88.49$_{\pm 1.6}$	53.45$_{\pm 3.6}$
	Ours	**90.21**$_{\pm 2.5}$	**84.66**$_{\pm 1.3}$	**92.49**$_{\pm 1.7}$	**32.12**$_{\pm 3.2}$
CIFAR-10	OE	62.23$_{\pm 0.3}$	66.16$_{\pm 0.4}$	57.57$_{\pm 0.3}$	80.64$_{\pm 1.0}$
	PASCL	62.33$_{\pm 0.4}$	67.21$_{\pm 0.2}$	57.14$_{\pm 0.2}$	79.55$_{\pm 0.8}$
	Ours	**63.12**$_{\pm 0.5}$	**67.69**$_{\pm 0.3}$	**60.77**$_{\pm 0.3}$	**78.89**$_{\pm 0.4}$
Tiny ImageNet	OE	68.04$_{\pm 0.4}$	79.36$_{\pm 0.3}$	51.66$_{\pm 0.5}$	76.66$_{\pm 0.5}$
	PASCL	68.20$_{\pm 0.4}$	79.65$_{\pm 0.4}$	51.53$_{\pm 0.4}$	76.11$_{\pm 0.8}$
	Ours	**71.02**$_{\pm 0.2}$	**80.94**$_{\pm 0.6}$	**56.57**$_{\pm 0.7}$	**75.42**$_{\pm 1.4}$
LSUN	OE	77.10$_{\pm 0.6}$	85.33$_{\pm 0.6}$	61.42$_{\pm 1.0}$	63.98$_{\pm 1.4}$
	PASCL	77.19$_{\pm 0.4}$	85.73$_{\pm 0.5}$	61.27$_{\pm 0.7}$	63.31$_{\pm 0.9}$
	Ours	**78.46**$_{\pm 0.9}$	**86.24**$_{\pm 0.5}$	**65.79**$_{\pm 0.9}$	**59.86**$_{\pm 1.2}$
Place365	OE	75.80$_{\pm 0.5}$	60.99$_{\pm 0.6}$	86.68$_{\pm 0.4}$	65.72$_{\pm 0.9}$
	PASCL	76.02$_{\pm 0.2}$	60.84$_{\pm 0.4}$	86.52$_{\pm 0.3}$	64.81$_{\pm 0.3}$
	Ours	**77.85**$_{\pm 0.6}$	**61.65**$_{\pm 0.5}$	**87.45**$_{\pm 0.9}$	**64.70**$_{\pm 1.5}$
Average	OE	72.91$_{\pm 0.7}$	75.06$_{\pm 0.6}$	67.16$_{\pm 0.6}$	68.89$_{\pm 1.1}$
	PASCL	73.56$_{\pm 0.3}$	74.52$_{\pm 0.4}$	67.18$_{\pm 0.1}$	67.44$_{\pm 0.6}$
	Ours	**76.25**$_{\pm 0.9}$	**77.84**$_{\pm 0.5}$	**70.37**$_{\pm 0.9}$	**62.94**$_{\pm 1.4}$

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

Method	AUROC	AUPR-in	AUPR-out	FPR95	ACC
MSP	61.00	64.52	57.54	82.01	40.97
EnergyOE	71.10	75.42	67.23	71.78	39.05
OE	72.91$_{\pm 0.7}$	75.06$_{\pm 0.6}$	67.16$_{\pm 0.6}$	68.89$_{\pm 1.1}$	39.04$_{\pm 0.4}$
PASCL	73.32$_{\pm 0.3}$	74.52$_{\pm 0.4}$	67.18$_{\pm 0.1}$	67.44$_{\pm 0.6}$	43.10$_{\pm 0.5}$
EAT	<u>74.41</u>$_{\pm 0.9}$	<u>76.04</u>$_{\pm 0.7}$	<u>69.57</u>$_{\pm 0.9}$	<u>65.05</u>$_{\pm 1.5}$	<u>46.13</u>$_{\pm 0.3}$
Ours	**76.25**$_{\pm 0.9}$	**77.84**$_{\pm 0.5}$	**70.37**$_{\pm 0.9}$	**62.94**$_{\pm 1.4}$	**50.07**$_{\pm 0.3}$

(b) Comparison results with different competing methods. The results are averaged over the six OOD test datasets in (a).

Table 2: Comparison results on CIFAR100-LT. The best results are shown in bold, and the second-best results are underlined.

Method	AUROC↑	AUPR-in↑	AUPR-out↑	FPR95↓	ACC↑
MSP	53.81	37.68	51.63	90.15	39.65
EnergyOE	64.76	44.63	64.77	87.72	38.50
OE	66.33	43.17	68.29	89.96	40.13
PASCL	68.00	44.32	70.15	87.53	45.49
EAT	69.84	46.67	69.25	87.63	46.79
Ours	**74.13**	**51.41**	**87.43**	**80.57**	**55.14**

Table 3: Comparison results on ImageNet-LT with ImageNet-1k-OOD as OOD test dataset.

Configuration: Following PASCL (Wang et al. 2022), we use ResNet-18 (He et al. 2016) as our backbone and perform the experiments using the Adam optimizer (Kingma and Ba 2014) with an initial learning rate 1×10^{-3} for experiments on CIFAR10/100-LT. For ImageNet-LT, we use ResNet-50 (He et al. 2016) as our backbone and train the model using the SGD optimizer with an initial learning rate of 0.1. All experiments are conducted by training the model for 100 epochs, with a batch size of 128. The reported results for OE, PASCL, and EAT are presented as the mean and standard deviation over six runs. More detailed configuration information is presented in Appendix C, and the full algorithm of our PATT is described in Appendix A.

Comparison with Other Methods

We compare our method with several leading long-tailed OOD detection methods, including classical methods MSP (Hendrycks and Gimpel 2017), OE (Hendrycks, Mazeika, and Dietterich 2018), and its variants EnergyOE (Liu et al. 2020), PASCL (Wang et al. 2022) and EAT (Wei, Wang, and Zhang 2024). Following EAT, we mainly compare the experimental results with OE and PASCL. For a fair comparison, some results in this paper are directly taken from (Wang et al. 2022). For AUPR-in, which is not reported in PASCL, and for some incomplete results, we strictly reproduced them under the same setting. Results on CIFAR10-LT, CIFAR100-LT and ImageNet-LT are presented in Table 1, Table 2 and Table 3.

Table 1a and Table 2a present a comparison between our method with OE and PASCL on CIFAR10-LT and CIFAR100-LT using six OOD datasets as we mentioned before. Our method consistently outperforms OE and PASCL on six OOD datasets, except for FPR95 of Texture on CIFAR10-LT. Both our method and PASCL use OE as a baseline and employ contrastive learning as the primary approach. Our method implicitly enhances the semantic information of tail classes, while PASCL's partial and asymmetric reduce the semantic information of head classes. Thus, our method reduces the average FPR95 by 3.47% on CIFAR10-LT and 4.50% on CIFAR100-LT compared to PASCL, demonstrating a significant improvement.

Table 1b, Table 2b and Table 3 report the comparison of PATT to state-of-the-art Long-tailed OOD methods. We can observe that as the dataset size increases, our method can distinguish itself more from other methods. For instance, in CIFAR10-LT, the OOD detection results of PATT and EAT are similar, and PATT even falls short of EAT in terms of AUROC and AUPR-in. However, in CIFAR100-LT and ImageNet-LT, our method greatly outperforms EAT across all metrics. On CIFAR100-LT, our method improves average AUROC by 1.84% and ID classification accuracy by 3.94% over EAT. The improvement is even more remarkable on ImageNet-LT, with an increase of 4.29% in AUROC and 8.35% in ID classification accuracy. This indicates that our method possesses a more comprehensive semantic representation, making it suitable for real-world scenarios.

Ablation Study

Analysis of key modules in PATT As described in the Method section, our PATT consists of Implicit Semantic Augmentation Contrastive Learning (ISAC), Temperature Scaling-Based Logit Adjustment (TLA), and Feature Calibration (FC). Table 4 shows the effectiveness of these three modules on all three ID datasets, reporting the average performance across six OOD datasets. Rows without the ISAC module use SCL to better highlight the importance of ISAC. From this table, we can infer that (1) All three modules enhance the model's OOD performance and ID classification accuracy across all three datasets; (2) TLA achieves a balanced classifier while ensuring high-confidence classification results, significantly improves the AUROC for OOD detection and the ACC of tail classes; (3) The combination of ISAC and TLA maximizes the effectiveness of both modules, as ISAC ensures a balanced feature extractor and TLA achieves a balanced classifier; (4) Adding FC not only enhances OOD detection capability but also improves classi-

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.

Baseline	Method	AUROC	AUPR-in	AUPR-out	FPR95	ACC
OE +	none	72.91	75.06	67.16	68.89	39.04
	τ-norm	73.14	74.63	66.48	68.76	40.08
	LA	73.05	74.36	66.35	69.24	39.16
	FC	**74.86**	**76.34**	**68.50**	**67.28**	**40.24**
TISAC +	none	75.48	77.19	68.87	64.25	49.56
	τ-norm	74.39	76.29	67.86	65.08	47.83
	LA	73.96	75.63	67.41	66.41	41.18
	FC	**76.25**	**77.84**	**70.37**	**62.94**	**50.07**

Table 5: Ablation study of feature calibration on CIFAR100-LT using ResNet18.

Method	**ACC (↑)**	
	Head classes	Tail classes
OE	54.29	20.90
PASCL	54.73 (+0.44)	36.26 (+15.36)
EAT	59.46 (+5.17)	34.12 (+13.22)
Ours	69.72 (+15.43)	34.19 (+13.29)

Table 6: Separate ACC for head and tail on ImageNet-LT.

fication accuracy; (5) While FC provides substantial benefits when used alone, its gains are less pronounced when the model is already relatively balanced.

On post-hoc feature calibration Post-hoc feature calibration calibrates all features using an attention weight extracted from the training set. In this section, we compare post-hoc feature calibration with post-hoc logit adjustment (Menon et al. 2021) and τ-norm (Kang et al. 2020). The results are shown in Table 5. As can be seen, post-hoc feature calibration, whether combined with the imbalanced feature encoder OE or the balanced encoder TISAC, consistently enhances the OOD detection and ID classification. In contrast, LA and τ-norm only provide slight improvements with the imbalanced encoder and can even lead to worse results when combined with a balanced encoder. A reasonable explanation is that the attention weight obtained from the training set comprehensively summarizes the entire dataset, which cannot be achieved simply by using class priors for debiasing or blindly regularizing the classifier.

Improvements on head and tail classes In Table 6, we show the ACC gains of our method compared to OE on both head and tail classes. It can be seen that our method significantly improves the ACC for both head and tail classes, with an increase of 15.43% for head classes and 13.29% for tail classes. Compared to PASCL and EAT, our enhancements are balanced for both head and tail classes. PASCL, in particular, is extremely biased towards tail classes, providing almost no gain for head classes.

Conclusion

We propose an advanced method for long-tailed OOD detection called Prioritizing Attention to Tail (PATT). PATT implicitly enhances ID data using a mixture of vMF distribution, significantly improving the performance of long-tailed classification. Additionally, we introduce temperature scaling-based logit adjustment, which, combined with Outlier Exposure (OE), creates a large confidence margin between ID and OOD data, enabling the model to identify OOD data effectively. To address biases at important feature levels, we propose an attention-based feature calibration, applied during the inference phase. Compared to post-hoc methods that directly apply debiasing at the logit level, our approach is more comprehensive. Extensive experiments demonstrate that our proposed PATT significantly improves both OOD detection and ID classification performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62276221, No. 62376232); the Fujian Provincial Natural Science Foundation of China (No. 2022J01002).

References

Bowles, C.; Chen, L.; Guerrero, R.; Bentley, P.; Gunn, R.; Hammers, A.; Dickie, D. A.; Hernández, M. V.; Wardlaw, J.; and Rueckert, D. 2018. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*.

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106: 249–259.

Byrd, J.; and Lipton, Z. 2019. What is the effect of importance weighting in deep learning? In *ICML*, 872–881.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 32.

Chen, X.; Zhou, Y.; Wu, D.; Zhang, W.; Zhou, Y.; Li, B.; and Wang, W. 2022. Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification. In *AAAI*, 356–364.

Choi, H.; Jeong, H.; and Choi, J. Y. 2023. Balanced energy regularization loss for out-of-distribution detection. In *CVPR*, 15691–15700.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *CVPR*, 3606–3613.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *CVPR*, 9268–9277.

Ding, Z.; Han, X.; Liu, P.; and Niethammer, M. 2021. Local temperature scaling for probability calibration. In *CVPR*, 6889–6899.

Du, C.; Wang, Y.; Song, S.; and Huang, G. 2024. Probabilistic Contrastive Learning for Long-Tailed Visual Recognition. *TPAMI*, 1–17.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. In *ICLR*.

Fort, S.; Ren, J.; and Lakshminarayanan, B. 2021. Exploring the limits of out-of-distribution detection. *NeurIPS*, 34.

He, H.; and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9): 1263–1284.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hein, M.; Andriushchenko, M.; and Bitterwolf, J. 2019. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, 41–50.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018. Deep Anomaly Detection with Outlier Exposure. In *ICLR*.

Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; and Chang, B. 2021. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, 6626–6636.

Hsu, Y.-C.; Shen, Y.; Jin, H.; and Kira, Z. 2020. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, 10951–10960.

Huang, C.; Li, Y.; Loy, C. C.; and Tang, X. 2016. Learning deep representation for imbalanced classification. In *CVPR*, 5375–5384.

Jiang, X.; Liu, F.; Fang, Z.; Chen, H.; Liu, T.; Zheng, F.; and Han, B. 2023. Detecting out-of-distribution data through in-distribution class prior. In *ICML*, 15067–15088.

Joy, T.; Pinto, F.; Lim, S.-N.; Torr, P. H.; and Dokania, P. K. 2023. Sample-dependent adaptive temperature scaling for improved calibration. In *AAAI*, 14919–14926.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling Representation and Classifier for Long-Tailed Recognition. In *ICLR*.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 30.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *NeurIPS*, 33: 18661–18673.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kull, M.; Perello Nieto, M.; Kängsepp, M.; Silva Filho, T.; Song, H.; and Flach, P. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *NeurIPS*, 32.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. In *NeurIPS*, volume 33, 21464–21475.

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 *CVPR*, 2537–2546.

Mardia, K. V.; and Jupp, P. E. 2009. *Directional statistics*. John Wiley & Sons.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *ICLR*.

Miao, W.; Pang, G.; Bai, X.; Li, T.; and Zheng, J. 2024. Out-of-distribution detection in long-tailed recognition with calibrated outlier class learning. In *AAAI*, 4216–4224.

Ming, Y.; Sun, Y.; Dia, O.; and Li, Y. 2023. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *ICLR*.

Mohseni, S.; Pitale, M.; Yadawa, J.; and Wang, Z. 2020. Self-supervised learning for generalizable out-of-distribution detection. In *AAAI*, 5216–5223.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NeurIPS workshop on deep learning and unsupervised feature learning*.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 427–436.

Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; and Yan, J. 2020. Equalization loss for long-tailed object recognition. In *CVPR*, 11662–11671.

Tang, C.; Zheng, X.; Zhang, W.; Liu, X.; Zhu, X.; and Zhu, E. 2023. Unsupervised feature selection via multiple graph fusion and feature weight learning. *Science China Information Sciences*, 66(5): 152101.

Tao, L.; Du, X.; Zhu, J.; and Li, Y. 2023. Non-parametric Outlier Synthesis. In *ICLR*.

Torralba, A.; Fergus, R.; and Freeman, W. T. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *TPAMI*, 30(11): 1958–1970.

Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2011. Class imbalance, redux. In *2011 IEEE 11th international conference on data mining*, 754–763.

Wang, F.; and Liu, H. 2021. Understanding the behaviour of contrastive loss. In *CVPR*, 2495–2504.

Wang, H.; Zhang, A.; Zhu, Y.; Zheng, S.; Li, M.; Smola, A. J.; and Wang, Z. 2022. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *ICML*, 23446–23458.

Wang, Y.; Pan, X.; Song, S.; Zhang, H.; Huang, G.; and Wu, C. 2019. Implicit semantic data augmentation for deep networks. *NeurIPS*, 32.

Wei, H.; Tao, L.; Xie, R.; Feng, L.; and An, B. 2022. Open-sampling: Exploring out-of-distribution data for rebalancing long-tailed datasets. In *ICML*, 23615–23630.

Wei, T.; Wang, B.-L.; and Zhang, M.-L. 2024. EAT: Towards Long-Tailed Out-of-Distribution Detection. In *AAAI*, 15787–15795.

Yang, J.; Wang, H.; Feng, L.; Yan, X.; Zheng, H.; Zhang, W.; and Liu, Z. 2021. Semantically coherent out-of-distribution detection. In *ICCV*, 8301–8309.

Yi, M.; Hou, L.; Sun, J.; Shang, L.; Jiang, X.; Liu, Q.; and Ma, Z. 2021. Improved ood generalization via adversarial training and pretraing. In *ICML*, 11987–11997.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 6023–6032.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*, 1–13.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *TPAMI*, 40(6): 1452–1464.

Zhu, J.; Geng, Y.; Yao, J.; Liu, T.; Niu, G.; Sugiyama, M.; and Han, B. 2024a. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. *NeurIPS*, 36.

Zhu, J.; Wang, Z.; Chen, J.; Chen, Y.-P. P.; and Jiang, Y.-G. 2022. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, 6908–6917.

Zhu, J.; Yu, G.; Yao, J.; Liu, T.; Niu, G.; Sugiyama, M.; and Han, B. 2024b. Diversified Outlier Exposure for Out-of-Distribution Detection via Informative Extrapolation. In *NeurIPS*.