Out-of-Distribution Detection in Long-Tailed Recognition with Calibrated Outlier Class Learning

Wenjun Miao[1], Guansong Pang[2*], Xiao Bai[1,3], Tianqi Li[1], Jin Zheng[1, 4*]

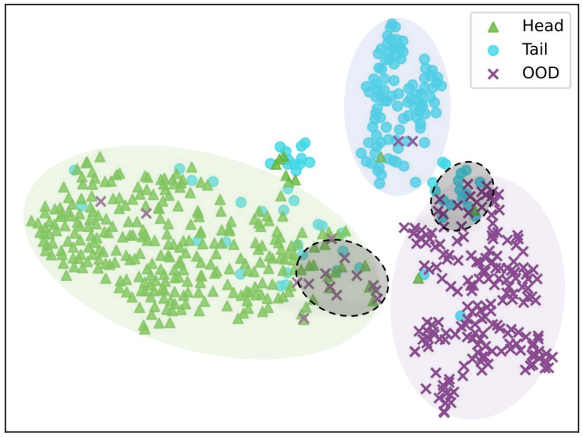
[1]School of Computer Science and Engineering, Beihang University
[2]School of Computing and Information Systems, Singapore Management University
[3]State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University
[4]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
{miaowenjun, jinzheng, baixiao, tianqili}@buaa.edu.cn, gspang@smu.edu.sg

Abstract

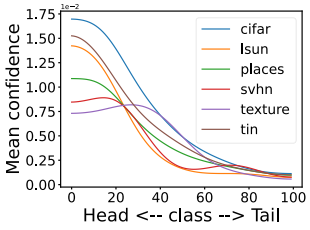
Existing out-of-distribution (OOD) methods have shown great success on balanced datasets but become ineffective in long-tailed recognition (LTR) scenarios where 1) OOD samples are often wrongly classified into head classes and/or 2) tail-class samples are treated as OOD samples. To address these issues, current studies fit a prior distribution of auxiliary/pseudo OOD data to the long-tailed in-distribution (ID) data. However, it is difficult to obtain such an accurate prior distribution given the unknowingness of real OOD samples and heavy class imbalance in LTR. A straightforward solution to avoid the requirement of this prior is to learn an outlier class to encapsulate the OOD samples. The main challenge is then to tackle the aforementioned confusion between OOD samples and head/tail-class samples when learning the outlier class. To this end, we introduce a novel calibrated outlier class learning (COCL) approach, in which 1) a debiased large margin learning method is introduced in the outlier class learning to distinguish OOD samples from both head and tail classes in the representation space and 2) an outlier-class-aware logit calibration method is defined to enhance the long-tailed classification confidence. Extensive empirical results on three popular benchmarks CIFAR10-LT, CIFAR100-LT, and ImageNet-LT demonstrate that COCL substantially outperforms state-of-the-art OOD detection methods in LTR while being able to improve the classification accuracy on ID data. Code is available at https://github.com/mala-lab/COCL.

Introduction

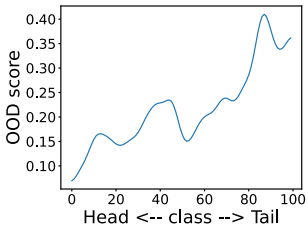
Deep neural networks (DNNs) have achieved remarkable success in various fields (Russakovsky et al. 2015; Krizhevsky, Sutskever, and Hinton 2017). However, their application in real-world scenarios, such as autonomous driving (Kendall and Gal 2017) and medical diagnosis (Leibig et al. 2017), remains challenging due to the presence of long-tailed distribution and unknown classes (Huang and Li 2021; Wang et al. 2020b). In particular, DNNs often have high confidence predictions that classify out-of-distribution (OOD) samples from unknown classes as one of the known classes. This issue is further amplified when the in-distribution (ID) data has a class-imbalanced/long-tailed distribution (Zhu et al. 2023; Li et al. 2022; Li, Cheung, and Lu 2022; Wang et al. 2022). This is because, as illustrated in Fig. 1a, DNNs trained on long-tailed data can be heavily biased towards head classes (the majority classes) due to the overwhelm-



(a) Feature representations of CIFAR100-LT test data



(b) Prediction confidence



(c) OOD score for ID samples

Figure 1: Visualization and qualitative results on test data of CIFAR100-LT using an LTR model augmented with an outlier learning module (see Eq. 2) for OOD detection. (a) Feature representations of samples randomly selected from head class, tail class, and OOD samples. The gray areas highlight obscure regions between head/tail samples and OOD samples. (b) The mean prediction confidence of the model classifying six OOD datasets into one of the ID classes. (c) The mean OOD score (i.e., the softmax probability of the outlier class) of samples from each ID class.

*Corresponding authors: G. Pang and J. Zheng

ing presence of samples from these classes, and as a result, long-tailed recognition (LTR) models often misclassify OOD samples into head classes with high confidence (see Fig. 1b); further, the LTR models tend to treat tail samples as part of OOD samples due to the rareness of tail samples in the training data, i.e., the tail samples often have a much higher OOD score than the head samples (see Fig. 1c).

Compared to OOD detection on balanced ID datasets, significantly less work has been done in the LTR scenarios. Recent studies (Wang et al. 2022; Wei et al. 2022a; Jiang et al. 2023; Choi, Jeong, and Choi 2023) are among the seminal works exploring OOD detection in LTR. Current methods in this line focus on distinguishing OOD samples from ID samples using an approach called outlier exposure (OE) (Hendrycks, Mazeika, and Dietterich 2018) that fits auxiliary/pseudo OOD data to a prior distribution (e.g., uniform distribution) of ID data. However, unlike balanced ID datasets, LTR datasets heavily skewed the distribution of ID data, so using the commonly-adopted uniform distribution as the prior becomes ineffective. Estimating this prior from the sample size of ID classes is a simple solution to alleviate this issue, but it can intensify the LTR models' bias toward head classes. Another line of approach is focused on learning discriminative representations to separate OOD samples from tail samples. However, the lack of sufficient samples in the tail classes renders this approach less effective, furthermore, it often fails to distinguish head and OOD samples.

In this work, we aim to synthesize both approaches and introduce a novel approach, namely calibrated outlier class learning (COCL). Intuitively, a straightforward solution to avoid the requirement of this prior in the OE-based approach is to learn an outlier class to encapsulate the OOD samples. The main challenge is then mainly about tackling the aforementioned confusion between OOD samples and head/tail-class samples when learning the outlier class. To address this challenge, we introduce a debiased large margin learning method, which is jointly optimized with the outlier class learning to distinguish OOD samples from both head and tail classes in the representation space. We further introduce an outlier-class-aware logit calibration method that takes into account the outlier class when calibrating the ID prediction probability. This helps enhance long-tailed classification confidence while improving OOD detection performance. In summary, our main contributions are as follows:

- We show that outlier class learning is generally more effective for OOD detection in LTR than fitting to a prior distribution when auxiliary OOD data is available.
- We then introduce a novel calibrated outlier class learning (COCL) approach that learns an accurate LTR model with a strong OOD detector that effectively mitigates the biases towards head and OOD samples. To this end, we introduce two components, including the debiased large margin learning and the outlier-class-aware logit calibration, which work in the respective training and inference stages, enabling substantially improved OOD detection and long-tailed classification performance.
- Extensive empirical results on three popular benchmarks CIFAR10-LT, CIFAR100-LT, and ImageNet-LT demonstrate that COCL substantially outperforms state-of-the-art OOD detection methods in LTR while improving the classification accuracy of ID data.

Related Work

OOD Detection The objective of this task is to determine whether a given input sample belongs to known classes (in-distribution) or unknown classes (out-of-distribution). In recent years, OOD detection has been extensively developed, including post hoc strategies (Sun, Guo, and Li 2021; Wang et al. 2023; Zhang and Xiang 2023) and training-time strategies (Liu et al. 2020; Wei et al. 2022b; Tian et al. 2022; Yu et al. 2023; Li et al. 2023; Liu et al. 2023). The post hoc methods focus on devising new OOD scoring functions in the inference phase. The training-time methods focus on separating OOD samples from ID samples by utilizing auxiliary data during training. Outlier exposure (OE) (Hendrycks, Mazeika, and Dietterich 2018) is arguably the most popular approach in this line that utilizes the OOD data by enforcing a uniform distribution of its prediction probability to ID classes. EnergyOE (Liu et al. 2020) improves OE and maximizes the free energy of OOD samples instead. However, all these methods are focused on cases with balanced ID training data, which fail to work well on imbalanced ID datasets.

Long-Tailed Recognition (LTR) LTR aims to improve the accuracy of the tail classes with the least influence on the head classes. Re-sampling (Wang et al. 2020a; Tang et al. 2022; Bai et al. 2023) and re-weighting (Tan et al. 2020; Alshammari et al. 2022; Gou et al. 2023; Hong et al. 2023) that focus on balancing the ratio between head and tail classes are the most straightforward solutions for LTR. Additionally, logit adjustment (LA) (Menon et al. 2020) emerges as an effective statistical framework that can be applied in both the training and inference phases to further enhance ID recognition performance. Although these LTR methods show effective performance in the long-tailed classification of ID samples, they do not have an explicit design to handle OOD samples.

OOD Detection in LTR PASCL (Wang et al. 2022) formulates the OOD detection problem in LTR and reveals the difficulty that simple combinations of existing OOD detection and LTR methods do not work well. In particular, PASCL evaluates different baseline methods in the SC-OOD benchmark (Yang et al. 2021) to establish performance benchmarks for OOD detection in LTR. OS (Wei et al. 2022a) finds that leveraging equivalent noisy labels does not harm training, so it introduces a noisy labels assignment method for utilizing unlabeled auxiliary OOD data to enhance the robustness of OOD detection and improve ID classification accuracy. Recent studies (Choi, Jeong, and Choi 2023; Jiang et al. 2023) find that fitting the prediction probability of OOD data to a long-tailed distribution in either the scratch or fine-tuning approach is more effective than using a uniform distribution. They specify this prior distribution based on the number of samples in ID classes or a pre-trained ID model to learn the OOD detection model. However, it is difficult to obtain such an accurate prior dis-

OOD Method		LTR Method	CIFAR10-LT					CIFAR100-LT				
			AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
OE	+	None	89.76	89.45	87.22	53.19	73.59	73.52	75.06	67.27	86.30	39.42
		Re-weight	89.34	88.63	86.39	56.24	70.35	73.08	73.86	66.05	87.22	39.45
		τ-norm	89.58	88.21	85.88	52.84	73.33	73.62	74.67	66.59	86.02	40.87
		LA	89.46	88.74	86.39	53.38	73.93	73.44	74.33	66.48	86.13	42.06
OCL	+	None	89.91	88.15	90.38	41.13	74.48	73.56	74.12	69.65	81.93	41.54
		Re-weight	90.45	89.12	90.58	38.86	74.84	74.23	74.29	70.68	79.45	42.06
		τ-norm	90.95	89.59	91.11	37.91	75.14	74.57	75.12	70.76	81.27	44.21
		LA	91.56	90.52	91.51	36.50	76.67	74.77	75.15	71.13	80.33	43.02
Our method COCL			**93.28**	**92.24**	**92.89**	**30.88**	**81.56**	**78.25**	**79.37**	**73.58**	**74.09**	**46.41**

Table 1: Comparison of outlier exposure (OE) and outlier class learning (OCL) approaches when combined with three LTR methods. All methods are trained on CIFAR10/100-LT using ResNet18. Reported are the average performance across six different OOD test sets (including CIFAR, Texture, SVHN, LSUN, Places365, and TinyImagenet) in the commonly-used SC-OOD detection benchmark (Yang et al. 2021) (See the Experiments section for the description of evaluation measures).

tribution of OOD data in LTR. We instead utilize the outlier class learning to eliminate the need for such a prior.

Outlier Class Learning vs. Outlier Exposure

Problem Statement Let $\mathcal{X} = X^{in} \cup X^{out}$ denote the input space and $Y^{in} = \{1, 2, \ldots, k\}$ be the set of k imbalanced ID classes in the label space. OOD detection in LTR is to learn a classifier f that for any test data $x \in \mathcal{X}$: if x drawn from X^{in} (from either head or tail classes), then f can classify x into the correct ID class; and if x is drawn from X^{out}, then f can detect x as OOD data. It is normally assumed that genuine OOD data X^{out} is not available during training since OOD samples are unknown instances. On the other hand, auxiliary samples that are not X^{out} but are drawn from a different distribution other than X^{in} are often available. These auxiliary samples can be used as pseudo OOD samples to fine-tune/re-train the LTR models.

Outlier Exposure (OE) OE is a popular OOD detection approach that uses auxiliary data as outliers to train ID classifiers for separating ID and OOD samples. Specifically, given ID data $\mathcal{D}_{in} = (X_{in}, Y_{in})$ and auxiliary data $\mathcal{D}_{out} = (X_{out}, u)$ for training, where u is a uniform distribution-based pseudo label for OOD data, OE then minimizes:

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

where γ denotes a hyper-parameter, and ℓ is a cross entropy loss. During inference, it uses the maximum softmax probability (MSP) over the ID classes as an OOD score.

Outlier Class Learning (OCL) Outlier class learning (OCL) aims at learning a new (outlier) class that encapsulates OOD samples, rather than enforcing a uniform prediction probability distribution as in the second term of Eq. 1. Specifically, for a k-class classification problem, it extends the label space by explicitly adding a separate class $k+1$ as outlier class, i.e., ID data $\mathcal{D}_{in} = (X_{in}, Y_{in})$ and auxiliary data $\mathcal{D}_{out} = (X_{out}, k+1)$ are used during training, and we then minimize the following loss function:

$$\mathcal{L}_{OCL} = \mathbb{E}_{x,y\sim\mathcal{D}_{in}}[\ell(f(x), y] + \gamma\mathbb{E}_{x\sim\mathcal{D}_{out}}[\ell(f(x), \tilde{y}], \quad (2)$$

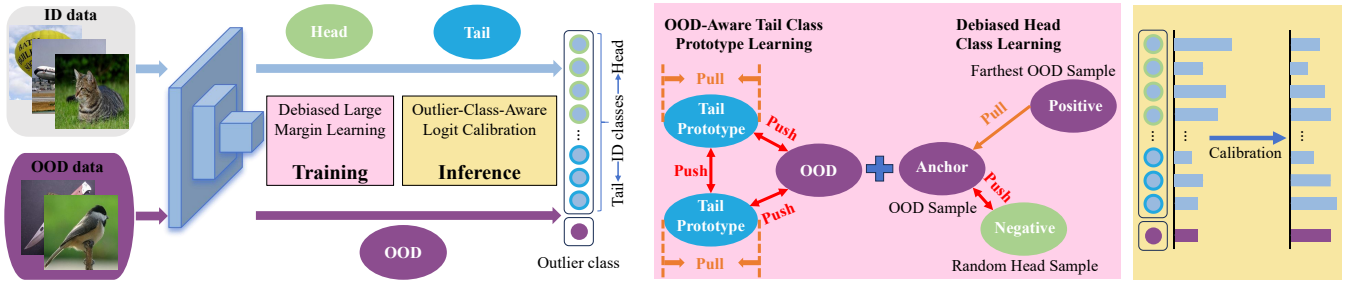
where $\tilde{y} = k+1$ and γ denotes a hyper-parameter. The softmax probability from the $k+1$ class is used as an OOD score during inference.

OCL Aligns Better with LTR Than OE We find empirically that OE achieves promising performance in general OOD detection scenarios, but works less effectively when applied to LTR settings. It is mainly because the uniform prediction probability prior in Eq. 1 does not hold in LTR. OCL helps eliminate this prior input and learns the outlier class that separates the OOD samples from the ID samples in the representation space. In Table 1, we compare the performance of OE and OCL when combining with three widely used LTR methods: Re-weight (Cui et al. 2019), τ-norm (Kang et al. 2019), and logit adjustment (LA) (Menon et al. 2020). The results show that OCL largely improves not only ID classification accuracy but also OOD detection performance on both datasets, substantially outperforming the OE method. Motivated by the large performance gap, we promote the use of OCL for LTR instead. However, there are two main challenges in the OCL approach: 1) OOD samples can often be wrongly classified into head classes and/or 2) tail-class samples are often misclassified as OOD samples. Our approach calibrated OCL (COCL) is focused on addressing these two challenges, and as shown in Table 1, it can help address the challenges and achieve largely improved classification and detection performance over the general OCL baselines.

Approach

Overview of Our Proposed Approach COCL

We introduce a novel COCL approach to tackle the aforementioned two issues for OOD detection in LTR. COCL consists of two components, namely *debiased large margin learning* and *outlier-class-aware logit calibration*, as shown in Fig. 2. The debiased large margin learning, as shown in Fig. 2b, is designed to reduce the bias towards head classes (leading to the misclassification of OOD samples into head classes) as well as the bias towards OOD samples (leading to the misclassification of tail samples as OOD samples) during training. The outlier-class-aware logit calibration component, as shown in Fig. 2c, is devised to utilize the logit of the outlier class for calibration to enhance OOD detection and the confidence of long-tailed classification during inference. Below we introduce each component in detail.



(a) Pipeline (b) Debiased large margin learning (c) Outlier-class-aware logit calibration

Figure 2: Overview of our approach COCL. (a) presents a high-level pipeline of our two components in our approach COCL, (b) illustrates the key idea of debiased large margin learning which includes OOD-aware tail class prototype learning and Debiased Head Class Learning to reduce biases towards OOD samples and head classes respectively, and (c) shows the outlier-class-aware logit calibration that utilizes the logit of the outlier class to calibrate the prediction results during inference.

Debiased Large Margin Learning

The debiased large margin learning component includes two modules, namely OOD-aware tail class prototype learning and Debiased Head Class Learning, to respectively reduce the model bias towards OOD samples and head classes. Below we elaborate on how these two modules can help reduce the two types of model bias.

OOD-Aware Tail Class Prototype Learning Since tail class samples are rare in the training data, the LTR models lack confidence in classifying them. As a result, they tend to exhibit high OOD scores during LTR inference, i.e., the LTR models' bias towards OOD samples when classifying tail class samples. The seminal work PASCL (Wang et al. 2022) attempts to utilize diverse augmentations to push tail samples away from OOD samples, but it often learns non-discriminative representations between OOD samples and tail samples due to the limited size of tail classes. To address this issue, we utilize a learnable prototype of one tail class as positive sample to pull tail samples closer to their prototype, with OOD samples and other tail class prototypes as negative samples to push the samples and prototype of the positive tail class away from OOD samples and other tail prototypes. This strategy harnesses the tail prototypes to increase the presence of representations for tail classes, helping reduce the model bias towards the OOD samples. Formally, let $\mathcal{M} \in \mathbb{R}^{N \times D}$ be the learnable parameters of N tail prototypes, with each prototype representation spanned in a D-dimensional space, our tail class prototypes are learned by minimizing the following loss:

$$\mathcal{L}_t = \mathbb{E}_{x \sim \mathcal{D}_{tail}}[\mathcal{L}_t(x, \mathcal{M})], \tag{3}$$

where $\mathcal{D}_{tail}$ is all tail samples in $\mathcal{D}_{in}$, and $\mathcal{L}_t(x, \mathcal{M})$ is defined as:

$$\mathcal{L}_t(x, \mathcal{M}) = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} log \frac{exp(z(x) m_x^\intercal / t)}{\sum_{m \in \mathcal{M}} exp(z(x) m^\intercal / t) + P(x)}, \tag{4}$$

where $\mathcal{B}$ is a training sample batch, $z(\cdot)$ is the output of a non-linear projection on the model's penultimate layer, i.e., the learned feature representation of x, $P(x) = \sum_{\hat{x} \in \mathcal{O}} exp(z(x) z(\hat{x})^\intercal / t)$ with $\mathcal{O}$ being a batch of OOD samples from $\mathcal{D}_{out}$, $\intercal$ is a transpose operation, m is a tail prototype in $\mathcal{M}$, m_x is the tail prototype corresponding to the tail class of sample x, and t is the scaling temperature.

As illustrated on the left in Fig. 2b, we only apply this tail class prototype learning to the tail-class in-distribution data and OOD data, as it is specifically designed to tackle the problem that tail samples exhibit high OOD scores. As for the head samples, they are normally easily distinguished from the OOD samples as there are sufficient head class samples in the training set. Note that we exclusively calculate the loss only when taking tail samples as input; the loss is not calculated for auxiliary OOD samples, since we have already pulled OOD samples together in the joint LTR and outlier class learning in Eq. 2. Thus, this module introduces only minor computation overheads to the general OCL.

Debiased Head Class Learning Due to the overwhelming presence of head class samples, LTR models demonstrate a strong bias towards head classes when performing OOD detection, i.e., OOD samples are often misclassified as one of the head classes. To address this issue, we introduce the debiased head class learning module that performs a one-class learning of OOD samples, where we aim to learn a large outlier-class description region for OOD samples to alleviate the dominant influence of head samples in the feature space. To this end, as illustrated on the right in Fig. 2b, we use only the OOD samples as anchors, with randomly sampled head samples as negative samples and the OOD samples that are distant from the anchors in the feature space as positive samples, and then we perform a semi-supervised one-class learning for OOD samples by minimizing the following loss:

$$\mathcal{L}_h = \mathbb{E}_{x \sim \mathcal{D}_{out}}[\mathcal{L}_h(x)], \tag{5}$$

where $\mathcal{L}_h(x)$ is defined as:

$$\mathcal{L}_h(x) = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} max(0, \|z(x) - z(x^p)\|_2^2 - \|z(x) - z(x^n)\|_2^2 + margin), \tag{6}$$

where $\mathcal{B}$ is a batch of OOD training samples from $\mathcal{D}_{out}$, x^p is a positive sample that is set to the most distance OOD sample from the anchor sample x in $\mathcal{B}$, x^n is a randomly head sample in the same batch, and $margin$ is a user-defined hyperparameter that specifies the margin between the one-class OOD description region and the head samples. Note that since popular contrastive learning is a two-way learning method, the model would be reinforced to bias towards the head class if the original contrastive learning is directly applied. Our design in Eq. 6 is to explicitly correct this bias and refine the learning of the outlier class.

Lastly, the overall objective of our debiased large margin learning is as follows:

$$\begin{aligned}\mathcal{L}_{total} &= \mathcal{L}_{OCL} + \alpha\mathcal{L}_t + \beta\mathcal{L}_h \\ &= \mathbb{E}_{x,y\sim\mathcal{D}_{in}}[\ell(f(x),y] + \gamma\mathbb{E}_{x\sim\mathcal{D}_{out}}[\ell(f(x),\tilde{y}] \\ &+ \alpha\mathbb{E}_{x\sim\mathcal{D}_{tail}}[\mathcal{L}_t(x,\mathcal{M})] + \beta\mathbb{E}_{x\sim\mathcal{D}_{out}}[\mathcal{L}_h(x)],\end{aligned} \tag{7}$$

where $\mathcal{L}_{OCL}$ is the same as the outlier class learning in Eq. (2), $\mathcal{L}_t$ is as defined in Eq. (4), and $\mathcal{L}_h$ is as defined in Eq. (6). α and β denote two hyperparameters to control the reduction of biases towards the OOD data and head classes.

Outlier-Class-Aware Logit Calibration

Our LTR model is then equipped with an OOD detector by minimizing Eq. 7 on the training data. However, due to the inherent class imbalance in the training data, the LTR model often tends to have a higher confidence on the prediction of head samples than both the tail samples and the OOD samples. To avoid this issue, we propose the outlier-class-aware logit calibration component that calibrates the predictions using the model logits and prior probability of both ID and outlier classes in the inference stage. This is different from existing LTR calibration methods that focus on ID classes only. Specifically, given a test sample x, we calibrate its posterior probability via:

$$P(y=i|x) = \frac{e^{f_i(x)-\tau\cdot\log n_i}}{\sum_{j=1}^{k+1} e^{f_j(x)-\tau\cdot\log n_j}}, \tag{8}$$

where $f_i(x)$ denotes the predicted model logit of x belonging to the class i, τ is a hyperparameter to balance how much we want to bring in the prior of the outlier class, and n_i is a prior probability for the class i and it is estimated by

$$n_i = \frac{N_i}{N_1 + N_2 + \cdots + N_k}, \tag{9}$$

where N_i is the training sample size for class i. We do not have genuine OOD samples during training, so their prior probability can not be estimated in the same way as ID classes. Motivated by the fact that detecting OOD samples should be as important as ID classification, we set $n_{k+1} = 1$ for the outlier class, which equals the summation of the prior probabilities of all ID classes. In doing so, our model's prediction is calibrated to decrease the probability of head classes and increase that of tail classes, while taking into account the influence of OOD samples on the prediction. Thus, this calibration is beneficial both for ID classification and OOD detection. This effect cannot be achieved using the general logit calibration used in LTR.

OOD	Method	AUC↑	AP-in↑	AP-out↑	FPR↓
Texture	OE	92.30	96.01	82.57	48.65
	OCL	93.71	95.95	91.07	27.22
	COCL	**96.81**	**98.21**	**93.86**	**14.65**
SVHN	OE	94.86	91.59	97.00	29.11
	OCL	95.14	90.88	97.73	25.47
	COCL	**96.98**	**93.25**	**98.61**	**12.59**
CIFAR100	OE	83.32	84.06	80.83	65.82
	OCL	82.04	82.52	81.92	63.35
	COCL	**86.63**	**86.66**	**86.28**	**52.21**
Tiny ImageNet	OE	86.35	89.88	79.30	64.50
	OCL	85.90	88.98	82.17	57.46
	COCL	**90.43**	**92.52**	**87.03**	**46.12**
LSUN	OE	91.57	93.06	88.37	53.99
	OCL	92.75	92.69	93.10	30.95
	COCL	**94.85**	**95.43**	**93.98**	**27.48**
Place365	OE	90.20	82.09	95.24	57.06
	OCL	89.91	77.91	96.28	42.33
	COCL	**93.97**	**87.36**	**97.56**	**32.25**

(a) Comparison of COCL with OE and OCL on six OOD datasets.

Method	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
MSP	74.33	73.96	72.14	85.33	72.17
OE	89.76	89.45	87.22	53.19	73.59
EnergyOE	91.92	91.03	91.97	33.80	74.57
OCL	89.91	88.15	90.38	41.13	74.48
PASCL	90.99	90.56	89.24	42.90	77.08
OS	91.94	91.08	89.35	36.92	75.78
Class Prior	92.08	91.17	90.86	34.42	74.33
BERL	92.56	91.41	91.94	32.83	81.37
COCL	**93.28**	**92.24**	**92.89**	**30.88**	**81.56**

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

Table 2: Comparison results on CIFAR10-LT.

Experiments

Experiment Settings

Datasets We use three popular long-tailed image classification datasets as ID data, including CIFAR10-LT (Cao et al. 2019), CIFAR100-LT (Cao et al. 2019), and ImageNet-LT (Liu et al. 2019). Following (Wang et al. 2022; Choi, Jeong, and Choi 2023), TinyImages 80M (Torralba, Fergus, and Freeman 2008) dataset is used for auxiliary OOD data to CIFAR10-LT and CIFAR100-LT, and ImageNet-Extra (Wang et al. 2022) is used for auxiliary OOD data to ImageNet-LT. The default imbalance ratio is set to $\rho = 100$ on CIFAR10-LT and CIFAR100-LT as (Wang et al. 2022). For OOD test set, we use six datasets CIFAR (Krizhevsky, Hinton et al. 2009), Texture (Cimpoi et al. 2014), SVHN (Netzer et al. 2011), LSUN (Yu et al. 2015), Places365 (Zhou et al. 2017), and TinyImagenet (Le and Yang 2015) introduced in the SC-OOD benchmark (Yang et al. 2021) for the LTR task on CIFAR10-LT and CIFAR100-LT. Following (Wang et al. 2022), we use ImageNet-1k-OOD (Wang et al. 2022) as the OOD test set for ImageNet-LT.

Evaluation Measures Following (Yang et al. 2021; Wang et al. 2022), we use the below common metrics for OOD detection and ID classification: (1) FPR is the false posi-

OOD	Method	AUC↑	AP-in↑	AP-out↑	FPR↓
Texture	OE	76.01	85.28	57.47	87.45
	OCL	75.92	82.99	66.48	70.01
	COCL	**81.99**	**88.05**	**74.38**	**59.79**
SVHN	OE	81.82	73.25	89.10	80.98
	OCL	78.64	69.21	86.26	86.38
	COCL	**89.20**	**81.57**	**94.21**	**54.46**
CIFAR10	OE	**62.60**	**66.16**	**57.77**	**93.53**
	OCL	60.29	63.21	55.71	94.22
	COCL	62.05	66.14	56.82	93.88
Tiny ImageNet	OE	68.22	79.36	51.82	88.54
	OCL	69.56	79.97	54.47	85.91
	COCL	**71.87**	**81.89**	**57.12**	**83.93**
LSUN	OE	76.81	85.33	60.94	83.79
	OCL	79.14	86.56	66.58	75.07
	COCL	**84.10**	**89.89**	**69.80**	**74.67**
Place365	OE	75.68	60.99	86.51	83.55
	OCL	77.81	62.80	88.39	79.97
	COCL	**80.30**	**68.65**	**89.16**	**77.83**

(a) Comparison of COCL to OE and OCL on six OOD datasets.

Method	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
MSP	63.93	64.71	60.76	89.71	40.51
OE	73.52	75.06	67.27	86.30	39.42
EnergyOE	76.40	77.32	72.24	76.33	41.32
OCL	73.56	74.12	69.65	81.93	41.54
PASCL	73.32	74.84	67.18	79.38	43.10
OS	74.37	75.80	70.42	78.18	40.87
Class Prior	76.03	77.31	72.26	76.43	40.77
BERL	77.75	78.61	73.10	74.86	45.88
COCL	**78.25**	**79.37**	**73.58**	**74.09**	**46.41**

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

Table 3: Comparison results on CIFAR100-LT.

tive rate of OOD examples when the true positive rate of ID examples is at 95% (as is typically done in previous OOD detection studies (Huang and Li 2021; Yang et al. 2022; Zhang and Xiang 2023)), (2) AUC computes the area under the receiver operating characteristic curve of detecting OOD samples, (3) AP measures the area under the precision-recall curve. Depending on the selection of the positive class, AP contains AP-in which ID class samples are treated as positive, as well as AP-out where the OOD samples are regarded as positive, and (4) ACC calculates the classification accuracy of the ID data. The reported results are averaged over six runs with different random seeds by default.

Implementation Details We compared our approach COCL with several existing OOD detection methods on long-tailed training sets, including classical methods MSP (Hendrycks and Gimpel 2016), OE (Hendrycks, Mazeika, and Dietterich 2018), EnergyOE (Liu et al. 2020), and very recently published methods PASCL (Wang et al. 2022), OS (Wei et al. 2022a), Class Prior (Jiang et al. 2023), and BERL (Choi, Jeong, and Choi 2023). The OCL method in our results is a baseline that is trained based on Eq. 2 only. Following PASCL (Wang et al. 2022) and BERL (Choi, Jeong, and Choi 2023), we use ResNet18 as our backbone on CIFAR10-LT and CIFAR100-LT, and use ResNet50 on ImageNet-LT.

Method	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
MSP	55.78	35.60	74.18	94.01	45.36
OE	68.33	43.87	82.54	90.98	44.00
EnergyOE	69.43	45.12	84.75	76.89	44.42
OCL	68.67	43.11	84.15	77.46	44.77
PASCL	68.00	43.32	82.69	82.28	47.29
OS	69.23	44.21	84.12	79.37	45.73
Class Prior	70.43	45.26	84.82	77.63	46.83
BERL	71.16	45.97	85.63	76.98	50.42
COCL	**71.85**	**46.76**	**86.21**	**75.60**	**51.11**

Table 4: Comparison results on ImageNet-LT with ImageNet-1k-OOD as OOD test dataset.

Metric	CIFAR10-LT			CIFAR100-LT		
	OE	OCL	COCL	OE	OCL	COCL
AUC↑	82.60	84.84	**91.91**	64.08	66.11	**74.85**
AP-in↑	60.47	61.56	**76.98**	34.07	34.97	**47.76**
AP-out↑	92.28	94.75	**97.15**	83.19	85.74	**87.59**
FPR↓	72.10	52.73	**34.30**	92.48	82.53	**77.01**

(a) On separating tail samples from OOD data.

Metric	CIFAR10-LT			CIFAR100-LT		
	OE	OCL	COCL	OE	OCL	COCL
AUC↑	95.97	95.79	**96.34**	84.42	83.85	**87.73**
AP-in↑	91.09	88.72	**93.34**	70.16	68.44	**73.84**
AP-out↑	98.17	98.54	**98.67**	92.85	92.83	**93.94**
FPR↓	20.57	22.67	**19.59**	70.17	67.94	**66.01**

(b) On separating head samples from OOD data.

Table 5: Comparison results on separating tail/head samples from OOD samples. The results are averaged over six OOD test datasets in the SC-OOD benchmark.

Main Results

Table. 2a and Table. 3a presents the comparison of our COCL with the baseline OE and OCL on CIFAR10/100-LT using six commonly used OOD test datasets. COCL substantially outperforms OE and OCL on both datasets across six OOD datasets except CIFAR100-LT with the CIFAR10 OOD test set where OE performs slightly better than COCL due to the difficulty of learning the outlier class given the similarity between these two datasets, which slightly drags down the performance of COCL in this case. OCL generally achieves better performance than OE, especially in FPR, indicating that OCL can detect OOD samples better with less influence on ID classification accuracy. Our COCL improves OCL further through the three components we introduced.

Table. 2b and Table. 3b show the comparison of COCL to state-of-the-art OOD detectors in LTR on CIFAR10-LT and CIFAR100-LT. To demonstrate the scalability of COCL, we also perform experiments on the large-scale ID dataset ImageNet-LT. The empirical results are presented in Table 4. COCL can improve not only OOD detection performance but also ID classification accuracy, and achieves the SOTA performance in both scenarios.

To show the effectiveness of COCL in improving the capability of distinguishing OOD data from head and tail samples, we perform two particular OOD detection settings: one

ID Dataset	TCPL	DHCL	OLC	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑	ACC-t↑
CIFAR10-LT	Baseline (OE)			89.76	89.45	87.22	53.19	73.59	55.91
	✗	✗	✗	89.91	88.15	90.38	41.13	74.48	56.52
	✓	✗	✗	91.23	89.47	91.51	34.27	74.58	57.10
	✗	✓	✗	91.08	89.40	91.10	35.28	74.61	56.92
	✗	✗	✓	92.06	91.29	91.78	34.41	79.40	76.57
	✓	✓	✗	91.74	89.91	92.04	33.85	75.20	57.30
	✓	✓	✓	**93.28**	**92.24**	**92.89**	**30.88**	**81.56**	**77.90**
CIFAR100-LT	Baseline (OE)			73.52	75.06	67.27	86.30	39.42	12.59
	✗	✗	✗	73.56	74.12	69.65	81.93	41.54	12.06
	✓	✗	✗	75.14	75.74	71.25	78.39	41.93	13.53
	✗	✓	✗	74.70	75.36	70.63	78.96	42.42	13.33
	✗	✗	✓	75.51	75.83	71.66	77.57	45.62	28.44
	✓	✓	✗	76.09	76.59	71.92	76.20	42.46	13.89
	✓	✓	✓	**78.25**	**79.37**	**73.58**	**74.09**	**46.41**	**29.44**
ImageNet-LT	Baseline (OE)			68.33	43.87	82.54	90.98	44.00	7.65
	✗	✗	✗	68.67	43.11	84.15	77.46	44.77	8.02
	✓	✗	✗	70.08	44.68	85.04	76.61	44.59	8.49
	✗	✓	✗	69.64	44.11	84.83	76.62	45.00	8.43
	✗	✗	✓	70.37	45.07	85.35	76.31	50.16	26.03
	✓	✓	✗	70.78	45.19	85.61	76.26	45.24	9.92
	✓	✓	✓	**71.85**	**46.76**	**86.21**	**75.60**	**51.11**	**28.05**

Table 6: Ablation study results on CIFAR10-LT, CIFAR100-LT and ImageNet-LT.

with only tail samples and OOD samples, and another one with only head samples and OOD samples. The empirical results are shown in Table. 5a and Table. 5b respectively. It can be observed that (1) differentiating tail and OOD samples is often more difficult than differentiating head and OOD samples, as indicated by the AUC performance, which applies to both COCL and the two baselines, and (2) COCL does a better job than the two baselines in both scenarios, resulting in significantly enhanced OOD performance.

Ablation Study Our COCL consists of OOD-Aware Tail Class Prototype Learning (TCPL), Debiased Head Class Learning (DHCL), and Outlier-Class-Aware Logit Calibration (OLC), as elaborated in the Approach section. Table 6 presents the results of the ablation study on these three components on all three ID datasets to show the importance of each component, with OE used as a baseline. The method immediately below OE is another baseline OCL based on Eq. 2. The results show that (1) TCPL can largely reduce FPR, while at the same time increasing ACC-t, indicating improved performance in handling tail classes, (2) DHCL also largely reduces FPR while having similar ACC and ACC-t as OCL, indicating its effect mainly on handling head and OOD samples, (3) combining TCPL and DHCL helps leverage the strengths of both components, (4) adding the OLC component consistently improves not only the classification accuracy but also the OOD detection performance.

Qualitative Analysis Fig. 3 presents a qualitative analysis of the prediction confidence of our method COCL on OOD samples belonging to each ID class in CIFAR10-LT (**Left**), and the mean OOD scores for each ID class (**Right**), with the results of OCL as the comparison baseline. It shows on the left panel that OCL has high confidence in predicting OOD samples as head classes, while COCL can significantly reduce these over-confident predictions. On the right panel, it is clear that our COCL can largely decrease the OOD scores for all ID class samples, particularly for tail class samples. These results justify that the aforementioned biases towards head classes and OOD samples are effectively reduced in our model COCL, significantly enhancing COCL in distinguishing OOD samples from both head and tail classes.

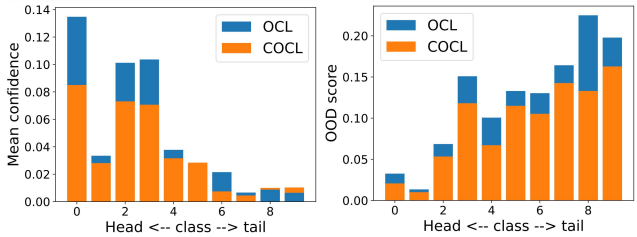


Figure 3: Results on CIFAR10-LT. (Left) The mean prediction confidence of six OOD datasets belonging to each ID class. (Right) The mean OOD score for each ID class.

Conclusion

To address the OOD detection problem in LTR, we propose a novel approach, calibrated outlier class learning (COCL), to discriminate OOD samples from long-tailed ID samples. COCL equips the general OCL with debiased large margin learning to reduce the model biases towards head classes and OOD samples. It also introduces outlier-class-aware logit calibration to guarantee the long-tailed classification performance when presented with OOD samples. Extensive experiments show that COCL significantly enhances the performance of both OOD detection and long-tailed classification on three popular LTR and OOD detection benchmarks.

Acknowledgments

W. Miao, T. Li, X. Bai, and J. Zheng were supported by National Natural Science Foundation of China (No. 62372029 and No. 62276016). We thank the anonymous reviewers for their insightful comments that help largely enhance our work. Due to page limitation, we provide additional settings, empirical results and their analyses at https://arxiv.org/abs/2312.10686 to address some of the reviewers' concerns.

References

Alshammari, S.; Wang, Y.-X.; Ramanan, D.; and Kong, S. 2022. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6897–6907.

Bai, J.; Liu, Z.; Wang, H.; Hao, J.; Feng, Y.; Chu, H.; and Hu, H. 2023. On the Effectiveness of Out-of-Distribution Data in Self-Supervised Long-Tail Learning. *arXiv preprint arXiv:2306.04934*.

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.

Choi, H.; Jeong, H.; and Choi, J. Y. 2023. Balanced Energy Regularization Loss for Out-of-distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15691–15700.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9268–9277.

Gou, Y.; Hu, P.; Lv, J.; Zhu, H.; and Peng, X. 2023. Rethinking Image Super Resolution From Long-Tailed Distribution Learning Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14327–14336.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Hong, F.; Yao, J.; Zhou, Z.; Zhang, Y.; and Wang, Y. 2023. Long-tailed partial label learning via dynamic rebalancing. *arXiv preprint arXiv:2302.05080*.

Huang, R.; and Li, Y. 2021. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8710–8719.

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 *International Conference on Machine Learning*. PMLR.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

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

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Leibig, C.; Allken, V.; Ayhan, M. S.; Berens, P.; and Wahl, S. 2017. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1): 17816.

Li, J.; Chen, P.; He, Z.; Yu, S.; Liu, S.; and Jia, J. 2023. Rethinking Out-of-distribution (OOD) Detection: Masked Image Modeling is All You Need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11578–11589.

Li, M.; Cheung, Y.-m.; and Lu, Y. 2022. Long-tailed visual recognition via gaussian clouded logit adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6929–6938.

Li, T.; Cao, P.; Yuan, Y.; Fan, L.; Yang, Y.; Feris, R. S.; Indyk, P.; and Katabi, D. 2022. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6918–6928.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475.

Liu, Y.; Ding, C.; Tian, Y.; Pang, G.; Belagiannis, V.; Reid, I.; and Carneiro, G. 2023. Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1151–1161.

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.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Sun, Y.; Guo, C.; and Li, Y. 2021. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34: 144–157.

Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; and Yan, J. 2020. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11662–11671.

Tang, K.; Tao, M.; Qi, J.; Liu, Z.; and Zhang, H. 2022. Invariant feature learning for generalized long-tailed classification. In *European Conference on Computer Vision*, 709–726. Springer.

Tian, Y.; Liu, Y.; Pang, G.; Liu, F.; Chen, Y.; and Carneiro, G. 2022. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *European Conference on Computer Vision*, 246–263. Springer.

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

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 *International Conference on Machine Learning*, 23446–23458. PMLR.

Wang, T.; Li, Y.; Kang, B.; Li, J.; Liew, J.; Tang, S.; Hoi, S.; and Feng, J. 2020a. The devil is in classification: A simple framework for long-tail instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 728–744. Springer.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2020b. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*.

Wang, Z.; Li, Y.; Chen, X.; Lim, S.-N.; Torralba, A.; Zhao, H.; and Wang, S. 2023. Detecting everything in the open world: Towards universal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11433–11443.

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

Wei, H.; Xie, R.; Cheng, H.; Feng, L.; An, B.; and Li, Y. 2022b. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, 23631–23644. PMLR.

Yang, J.; Wang, H.; Feng, L.; Yan, X.; Zheng, H.; Zhang, W.; and Liu, Z. 2021. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8301–8309.

Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; et al. 2022. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35: 32598–32611.

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*.

Yu, Y.; Shin, S.; Lee, S.; Jun, C.; and Lee, K. 2023. Block Selection Method for Using Feature Norm in Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15701–15711.

Zhang, Z.; and Xiang, X. 2023. Decoupling MaxLogit for Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3388–3397.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.

Zhu, F.; Cheng, Z.; Zhang, X.-Y.; and Liu, C.-L. 2023. OpenMix: Exploring Outlier Samples for Misclassification Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12074–12083.