Long-Tailed Out-of-Distribution Detection via Normalized Outlier Distribution Adaptation

Wenjun Miao[*]
Beihang University, China
miaowenjun@buaa.edu.cn

Guansong Pang[†]
Singapore Management University, Singapore
gspang@smu.edu.sg

Jin Zheng[*†]
Beihang University, China
jinzheng@buaa.edu.cn

Xiao Bai[*]
Beihang University, China
baixiao@buaa.edu.cn

Abstract

One key challenge in Out-of-Distribution (OOD) detection is the absence of ground-truth OOD samples during training. One principled approach to address this issue is to use samples from external datasets as outliers (*i.e.*, pseudo OOD samples) to train OOD detectors. However, we find empirically that the outlier samples often present a distribution shift compared to the true OOD samples, especially in Long-Tailed Recognition (LTR) scenarios, where ID classes are heavily imbalanced, *i.e.*, the true OOD samples exhibit very different probability distribution to the head and tailed ID classes from the outliers. In this work, we propose a novel approach, namely *normalized outlier distribution adaptation* (AdaptOD), to tackle this distribution shift problem. One of its key components is *dynamic outlier distribution adaptation* that effectively adapts a vanilla outlier distribution based on the outlier samples to the true OOD distribution by utilizing the OOD knowledge in the predicted OOD samples during inference. Further, to obtain a more reliable set of predicted OOD samples on long-tailed ID data, a novel *dual-normalized energy loss* is introduced in AdaptOD, which leverages class- and sample-wise normalized energy to enforce a more balanced prediction energy on imbalanced ID samples. This helps avoid bias toward the head samples and learn a substantially better vanilla outlier distribution than existing energy losses during training. It also eliminates the need of manually tuning the sensitive margin hyperparameters in energy losses. Empirical results on three popular benchmarks for OOD detection in LTR show the superior performance of AdaptOD over state-of-the-art methods. Code is available at https://github.com/mala-lab/AdaptOD.

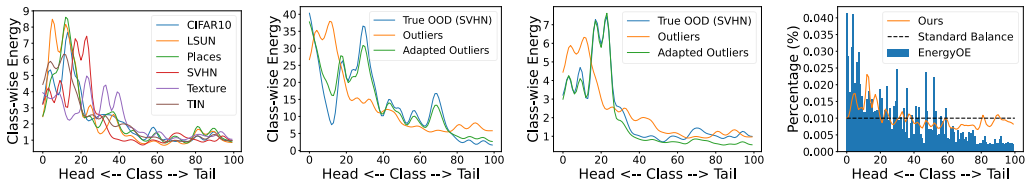
1 Introduction

Deep neural networks (DNNs) are widely known to be overconfident about what they do not know when applying them to real-world scenarios in open environments [15,42], such as autonomous driving [18] and medical diagnosis [21]. Consequently, the high-confidence predictions can misclassify out-of-distribution (OOD) samples from unknown classes as one of the known or in-distribution (ID) classes [24, 48]. This issue is further amplified when ID samples exhibit a class-imbalanced distribution in Long-Tailed Recognition (LTR) scenarios. This is because head samples often receive similarly high-confident prediction as OOD samples, while the tail samples receive substantially

[*]Beihang University and Beihang Jiangxi Research Institute
[†]Corresponding authors: G. Pang and J. Zheng

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-4199



(a) (b) (c) (d)

Figure 1: **(a)** Mean energy distribution on six OOD datasets with CIFAR100-LT [3] as ID data. **(b)** The results of the energy distribution of the OOD dataset SVHN [31] using our proposed dynamic outlier distribution adaptation (DODA) and an existing energy loss EnergyOE [24], where the outlier data is taken from TinyImages80M [38]. **(c)** The results of using both of our proposed DODA and dual-normalized energy loss (DNE). **(d)** The ratio of the energy of each ID class to the aggregated energy of all ID classes.

lower-confident prediction, leading to an indistinguishability between OOD and head samples and the tendency of wrongly detecting tail samples as OOD samples [30, 40, 59]. We address the problem of *long-tailed OOD detection*, aiming at ensuring LTR accuracy while rejecting unknown samples.

One notorious challenge in OOD detection is the lack of ground-truth information on OOD samples, as they can be drawn from any unknown distribution. One popular solution to tackle this challenge is to use samples from external datasets as outliers (*i.e.*, samples that do not overlap with ID and OOD samples, also known as pseudo OOD samples) to train OOD detectors [4, 12, 17, 24, 30]. This approach can be implemented by fitting the prediction probability of the outlier data to a prior distribution over the ID classes [17] or a margin-based global energy function [4]. Despite showing good performance on various benchmarks, all of these methods assume that the distribution of the outliers is well aligned with that of the true OOD samples in the target data. However, the outliers often present a distribution shift compared to the true OOD samples, especially in LTR scenarios [49, 56], *i.e.*, the true OOD samples exhibit very different probability distribution to the head and tailed ID classes from the outliers. Due to the bias toward head classes, the distribution shift is particularly severe w.r.t. the head samples. For example, as shown in Fig. 1a, the energy distribution of six popular OOD datasets differs significantly from each other, where CIFAR100-LT [3] is used as the ID dataset. This implies that any of these OOD datasets used as outlier data source can largely mismatch the distribution of the true OOD data if the other five datasets are used as the true OOD data. Such a distribution shift can largely mislead the training of detection models, leading to downgraded detection performance.

To tackle this problem, in this work, we propose a novel approach for OOD detection in LTR, namely Normalized Outlier Distribution Adaptation (**AdaptOD**). Dynamic Outlier Distribution Adaptation (**DODA**) is a key component of AdaptOD. Given a vanilla outlier distribution, DODA performs test-time adaptation (TTA) to dynamically adapt the outlier distribution to the true OOD distribution by utilizing the OOD knowledge embedded in the predicted OOD samples. This reduces the distribution gap between the outlier and the OOD distributions, enabling a more accurate estimation of OOD scores. As illustrated in Fig. 1b, a large gap exists between the energy distribution of the outlier data (TinyImages80M [38]) and the true OOD data (SVHN [31]). By contrast, our adapted outlier distribution is better aligned to the OOD distribution. Importantly, the ground truth of the test data is assumed to be unavailable during TTA, and as we will show in the experiments (see Table 5), DODA based on the predicted OOD samples can well approximate the upper-bound performance obtained when DODA can get access to the ground truth of the test data to perform TTA (*i.e.*, an oracle model). There have been a few TTA methods for OOD detection, but they require online model retraining [49] or feature memory augmentation [56]. By contrast, DODA focuses on the calibration of the outlier distribution, effectively eliminating the retraining or memory overheads.

On the other hand, training OOD detectors using energy loss functions [4, 24] is a principled approach to learn the vanilla outlier distribution. However, existing energy losses can underestimate the tail class distribution and involve sensitive hyperparameters on energy margins. As a result, the vanilla outlier distribution learned by using these losses often misclassifies tail samples as OOD samples during TTA. This can significantly affect the distribution adaptation in DODA, leading to still a relatively large gap between the adapted outlier distribution and the OOD distribution, as shown in Fig. 1b. Therefore, AdaptOD introduces a novel Dual-Normalized Energy loss (**DNE**) to

balance energy prediction for imbalanced ID samples and learn a better vanilla outlier distribution for subsequent DODA. Unlike existing energy losses that are focused on sample-wise energy estimation, DNE utilizes both class-wise and sample-wise normalized energy. This helps obtain more balanced prediction energy on the head and tail samples, transferring the energy from the head samples to the tail samples, thereby avoiding the bias toward the head classes (see Fig. 1d). In doing so, DNE is also free of energy margin hyperparameters and enables the learning of a better vanilla outlier distribution. This guarantees a better starting point for the outlier distribution adaptation and the accuracy of the predicted OOD samples at testing time in DODA, and thus yielding substantially better aligned outlier distribution (see Fig. 1c vs. Fig. 1b). Our main contributions are as follows:

- We propose the novel approach AdaptOD for OOD detection in LTR. To our best knowledge, it is the first approach for adapting the outlier distribution to the true OOD distribution from both the training and inference stages.
- In AdaptOD, we introduce two new components, DODA and DNE, to reduce the gap between the learned outlier distribution and the true OOD distribution in the presence of long-tailed ID data. DODA builds upon a vanilla outlier distribution and then dynamically adapts this distribution to the true OOD distribution with the OOD knowledge obtained at testing time. DNE is designed to perform class- and sample-wise normalized energy training, which enforces more balanced prediction energy for imbalanced ID samples, enabling the learning of largely enhanced vanilla outlier distribution for more effective DODA.
- Extensive empirical results on three LTR benchmarks CIFAR10-LT, CIFAR100-LT, and ImageNet-LT using six popular OOD datasets demonstrate that AdaptOD substantially outperforms the state-of-the-art (SOTA) OOD detection methods in various LTR scenarios.

2 Related Work

OOD Detection in Long-Tailed Recognition (LTR). In recent years, OOD detection and LTR have been extensively developed. The former determines whether a given input sample belongs to known classes (in-distribution) or unknown classes (out-of-distribution) [22, 23, 25, 35, 37, 43, 44, 52, 57], while the latter expects to train on class-imbalanced datasets [1, 2, 9, 13, 33, 34, 36]. PASCL [40] reveals the difficulty of the OOD detection problem in LTR, and establishes performance benchmarks for OOD detection in LTR based on the SC-OOD benchmark [48]. This setting is also extended to medical image analysis [28, 50], which utilizes a strong data augmentation to discriminate ID data and OOD data. Recent studies [4, 17] find that fitting the prediction probability of outlier data to a long-tailed distribution is more effective than using a uniform distribution. They specify this distribution based on the number of samples in ID classes or a pre-trained ID model to learn this outlier distribution. However, it is difficult to obtain such an accurate distribution for outliers in LTR. Other studies [30, 45] attempt to learn an extra outlier class to overcome the need for learning long-tailed distributions of outliers. But they need a more complex model design. More importantly, all these methods assume that the outlier samples can well represent the distribution of the true OOD data, but this often does not hold in practice since OOD data can be sampled from highly different unknown distributions in different application scenarios. Our approach tackles this problem by adapting the outlier distribution to that of the true OOD data.

Test-Time Adaptation (TTA) for OOD Detection. Recently, TTA [8, 14, 53] has been introduced to OOD detection, in which unlabeled test data that can be seen only once are used to perform online updating pre-trained DNNs for enhancing task performance and quickly adapting to real-world scenarios. There are two primary approaches for TTA in other tasks: retraining the model based on unsupervised objectives [39, 41, 54] and updating the feature memory for each class [16, 47, 55]. However, unlike these TTA methods that generalize training data to test data and maintain the same label space between them, TTA for OOD detection [7, 49, 56] addresses the challenge of identifying unknown classes in test data. While training data includes ID data and outlier data, test data comprises not only ID data but also true OOD data consisting of unknown classes that do not overlap with the ID and outlier data. In particular, AUTO [49] is a recent method that attempts to assign pseudo labels to unlabeled test data, and then directly uses these pseudo-labels and test data to retrain the model online through Outlier Exposure [12]. AdaOOD [56] utilizes a memory bank to store feature memories of ID data, then updates these memories online during inference, and lastly employs a kNN-based distance method to detect OOD samples. However, they fail to work well in the LTR scenarios due

https://doi.org/10.52202/079017-4199

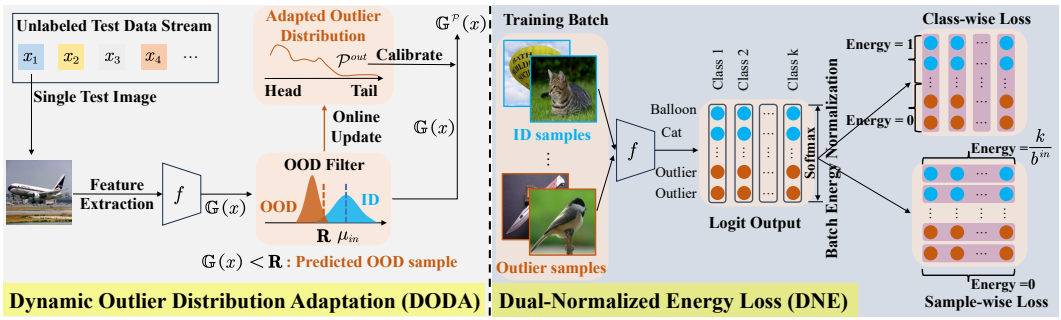


Figure 2: Overview of AdaptOD, which consists of two components, DODA (Left) and DNE (Right). **Left:** Each test sample is assigned a global energy-based OOD score $\mathbb{G}(x)$ to adapt the outlier distribution $\mathcal{P}^{out}$. DODA then uses the adapted outlier distribution $\mathcal{P}^{out}$ to calibrate the global energy score $\mathbb{G}(x)$, obtaining the calibrated global energy score $\mathbb{G}^{\mathcal{P}}(x)$ as the OOD score. **Right:** For each iteration, DNE first applies Batch Energy Normalization on logit output to obtain the normalized energy, and then utilizes this energy to optimize a dual energy loss function at both the class and sample levels.

to the large variation in the heavily imbalanced training ID data. Moreover, these methods require retraining or additional memory overheads. ETLT [7] attempts to calibrate OOD scores by a linear regression of its input feature but requires a batch-wise inference to obtain sufficient test samples for the regression. DODA instead utilizes the dynamically adapted outlier distribution to calibrate the prediction output of test data during inference without any retraining or memory overhead.

3 Approach

Preliminaries. Let X^{in} denote the input space of the ID data and $Y^{in} = \{1, 2, \ldots, k\}$ be the set of k imbalanced ID classes in the label space. We have genuine OOD data X^{true_out} that is different from X^{in}. It is normally assumed that genuine OOD data X^{true_out} are not available during training since OOD samples are unknown instances. However, we can obtain auxiliary OOD data from external datasets, which can be used as outliers X^{aux_out} to act as surrogate OOD data for training/fine-tuning LTR models. That is, X^{aux_out} is still different from X^{true_out}, but both of them are OOD w.r.t. X^{in}. There is no class overlapping among ID data X^{in}, genuine OOD data X^{true_out}, and outlier data X^{aux_out}. Then the training and test sets can be respectively denoted as: $\mathcal{X}^{train} = X^{in} \cup X^{aux\_out}$ and $\mathcal{X}^{test} = X^{in} \cup X^{true\_out}$.

OOD detection in LTR is to learn a classifier f with training data $\mathcal{X}^{train}$ so that for any test data $x \in \mathcal{X}^{test}$, if x is drawn from X^{in} (from either head or tail classes), then f can classify x into the correct ID class, whereas if x is drawn from X^{true_out}, then f can detect x as OOD data.

TTA for OOD detection in LTR is to online update the above pre-trained classifier f with test data $\mathcal{X}^{test}$ during the inference stage, in which for any unlabeled single test sample $x \in \mathcal{X}^{test}$, utilizing pre-trained classifier f to predict whether x belongs to ID or OOD data at the current iteration, then using the predicted label and the test sample x to update the classifier f. At the next iteration, the updated classifier f is used to identify a new test sample and continuously update the classifier f. Notably, each sample can only be seen by f once during inference.

3.1 Overview of AdaptOD

The proposed AdaptOD approach is designed to tackle the aforementioned distribution shift issue for OOD detection in LTR. As shown in Fig. 2, AdaptOD consists of two components, namely Dynamic Outlier Distribution Adaptation (DODA) and Dual-Normalized Energy Loss (DNE). DODA dynamically adapts the learned outlier distribution to the true OOD distribution during inference to reduce the distribution gap between them. DNE is designed to perform both class-wise and sample-wise normalized energy training to obtain more balanced prediction energy on imbalanced ID samples, thereby yielding an enhanced vanilla outlier distribution and enabling better distribution adaptation in DODA. Below we introduce each component in detail.

https://doi.org/10.52202/079017-4199

3.2 DODA: Dynamic Outlier Distribution Adaptation

Previous OOD detection methods in LTR suffer from a distribution shift between outlier data and true OOD data. This issue can largely limit the performance of these OOD detectors. Therefore, we propose to dynamically adapt the outlier distribution to the true OOD distribution and further use it to calibrate the prediction output of test samples at the inference stage.

Dynamic Distribution Adaptation with Predicted OOD Samples. Recently, energy-based methods [4, 24], which use a global energy score over the ID classes as an OOD score for each test sample, have achieved SOTA performance for OOD detection in LTR. Motivated by this success, we learn and adapt the vanilla outlier distribution $\mathcal{P}^{out}$, which is initialized by the global energy from the LTR model predictions on the outlier data, to that of the true OOD data, and then use a $\mathcal{P}^{out}$-calibrated global energy score as the OOD score. Specifically, given a set of k ID classes, for any test sample $x \in \mathcal{X}^{test}$, its global energy score $\mathbb{G}(\cdot)$ is defined as [24]:

$$\mathbb{G}(x) = \sum_{j=1}^{k} e^{f_j(x)}, \tag{1}$$

where $f_j(x)$ is the logit output of sample x in class $j, j \in \{1, 2, \ldots, k\}$. Let $\mathcal{P}^{out} \in \mathbb{R}^k$ be an initial outlier distribution. DODA performs test-time adaptation to dynamically adapt the outlier distribution $\mathcal{P}^{out}$ to the true OOD distribution based on the OOD knowledge from the samples predicted as OOD during inference. To this end, we designed an OOD filter using training data to identify OOD samples. Since it is easy to obtain the distribution of global energy score for training ID samples, we use an offline method to determine a threshold for filtering OOD samples based on this energy distribution. This avoids adverse effects on the adaptation speed during inference. Formally, given ID examples from training data $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, where $\mathbf{x} \in X^{in}$ and n is the number of training ID samples, we estimate the mean μ_{in} and standard deviation σ_{in} of the global energy distribution by:

$$\mu_{in} = \frac{\sum_{i=1}^{n} \mathbb{G}(x_i)}{n}, \sigma_{in} = \sqrt{\frac{\sum_{i=1}^{n} (\mathbb{G}(x_i) - \mu_{in})^2}{n-1}}. \tag{2}$$

We then utilize a Z-score-based method to implement the OOD filter, with the Z-score defined as:

$$R = \mu_{in} - \alpha \times \sigma_{in}, \tag{3}$$

where α is a hyperparameter. $\alpha = 3$ is used by default during the inference stage, and this setting works well throughout our experiments. More discussion about α is described in Appendix D.3.3.

To adapt the outlier distribution $\mathcal{P}^{out}$, DODA utilizes the predicted OOD samples by the OOD filter to perform a momentum update of $\mathcal{P}^{out}$ during inference, so that $\mathcal{P}^{out}$ will represent the mean of energy distribution for the predicted OOD samples. The entries in the vanilla outlier distribution $\mathcal{P}^{out}$ are initialized from the mean energy distribution of the outlier data, and they are updated in an online fashion. Specifically, when the OOD filter detects the t-th test sample x as an OOD sample (*i.e.*, its global energy $\mathbb{G}(x) < R$), DODA performs an update of $\mathcal{P}^{out}$ as follows:

$$\mathcal{P}^{out}(t+1) = \begin{cases} \frac{M * \mathcal{P}^{out}(t) + e^{f(x)}}{M+1}, & \mathbb{G}(x) < R \\ \mathcal{P}^{out}(t), & \mathbb{G}(x) \geq R \end{cases} \tag{4}$$

where DODA only keep the number of predicted OOD samples M and current $\mathcal{P}^{out}$ during inference.

Calibrated OOD Score based on the Adapted Outlier Distribution. After obtaining the adapted $\mathcal{P}^{out}$, we use it to calibrate the global energy score $\mathbb{G}(\cdot)$ and define the OOD score as follows:

$$\mathbb{G}^{\mathcal{P}}(x) = \sum_{j=1}^{k} \frac{e^{f_j(x)}}{1 + \mathcal{P}^{out}_j}, \tag{5}$$

where $x \in \mathcal{X}^{test}$ and $\mathbb{G}^{\mathcal{P}}(\cdot)$ denotes the calibrated global energy score with the adapted outlier distribution. This way helps reduce the energy proportion of the head classes that true OOD distribution leans toward in the original global energy score $\mathbb{G}(\cdot)$. In doing so, the distribution gap between the outliers and true OOD is effectively reduced in final OOD score $\mathbb{G}^{\mathcal{P}}(\cdot)$, enabling more

 https://doi.org/10.52202/079017-4199

accurate estimation of OOD scores in heavily imbalanced ID data without incurring any retraining cost or additional memory expense.

3.3 DNE: Dual-Normalized Energy Loss

When using existing global energy loss to obtain the vanilla outlier distribution $\mathcal{P}^{out}$, the distribution of tail samples is indistinguishable from that of OOD samples due to the underestimating of tail samples. We are also required to manually tune the sensitive hyperparameters on energy margins under complex class imbalance. These can lead an inaccurate OOD filter used in Eq. 3 and subsequently affect the distribution adaptation in DODA. To tackle these issues, we propose a Dual-Normalized Energy Loss (DNE), which consists of two novel components, namely class-wise normalized energy loss (DNE-C) and sample-wise normalized energy loss (DNE-S). DNE-C is a class-wise training loss for balancing the sum of energy on all ID samples for each ID class, whereas DNE-S is a sample-wise training loss for balancing the sum of energy on all ID classes for each ID sample. DNE learns a balanced prediction energy distribution on imbalanced ID samples, which helps further reduce the bias toward the head classes in Eq. 5, thereby improving vanilla outlier distribution for a better OOD filter in Eq. 3 and a better vanilla outlier distribution $\mathcal{P}^{out}$ in Eq. 4. It also provides stable energy margins, eliminating the need of manual tuning of these margins.

Batch Energy Normalization. To this end, we first propose a novel batch energy normalization method, which conducts energy normalization on the logit output of each class for a batch of training samples. In doing so, the energy of each sample is dependent on the energy of other ID samples and OOD samples relative to the same class. This helps transfer the energy knowledge from the head samples to the tail samples, enabling a better estimation for the energy distribution of tail samples.

Formally, let $\mathbf{x}^{in} \in X^{in}$ be one training batch of ID data, with $\mathbf{x}^{in} = \{x_1^{in}, x_2^{in}, ..., x_{b^{in}}^{in}\}$ and b^{in} be its batch size, and $\mathbf{x}^{out} \in X^{aux\_out}$ be a set of outlier data in a training batch, with $\mathbf{x}^{out} = \{x_1^{out}, x_2^{out}, ..., x_{b^{out}}^{out}\}$ whose set size is b^{out}, then the batch energy normalization $F_j(x_i)$ for a sample $x_i \in x^{in} \cup x^{out}$ in class $j \in \{1, 2, ..., k\}$ with classifier f is defined as:

$$F_j(x_i) = \frac{e^{f_j(x_i)}}{e^{f_j(x_1^{in})} + ... + e^{f_j(x_{b^{in}}^{in})} + e^{f_j(x_1^{out})} + ... + e^{f_j(x_{b^{out}}^{out})}}, \tag{6}$$

where $f_j(x)$ is the logit output of sample x in class j. Essentially, we use the logit output of all samples in a training batch on a class j to normalize the energy prediction of sample x. This largely reduces the energy prediction bias toward the head samples. The energy of the outlier data is included as a calibration modulation. Then, those normalized energy scores are used for the dual-normalized energy losses, DNE-C and DNE-S, to better balance the prediction energy of long-tailed ID samples.

Additionally, compared to the current energy-based method [4, 24] for OOD detection in LTR that requires manually designed energy margin hyperparameters, batch energy normalization adjusts the energy of the batch samples on each class to the same scale, so it can provide stable energy margins for the balanced training without relying on the training dataset and/or the class imbalance factor, without the need of manually tuning them.

Class-wise Normalized Energy Loss (DNE-C). DNE-C independently regularizes the energy for each class to enhance the normalized energy of ID samples for more class-wise balanced energy. Formally, let $\mathcal{D}_{in} = (X^{in}, Y^{in})$ and $\mathcal{D}_{out} = X^{aux\_out}$, then we can independently minimize the class energy on each class as follows:

$$\mathcal{L}_C = \sum_{j=1}^{k} (\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{in}}[(max(0, m_{in}^c - \mathbf{C}_j(\mathbf{x})))^2] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{out}}[(max(0, \mathbf{C}_j(\mathbf{x}) - m_{out}^c))^2]), \tag{7}$$

where $m_{in}^c = 1$ and $m_{out}^c = 0$ are the default margin hyperparameter settings without the need of manual tuning on different datasets (see Appendix C for more details). The class-wise normalized energy $\mathbf{C}_j(\mathbf{x}), j \in \{1, 2, ..., k\}$ is defined as:

$$\mathbf{C}_j(\mathbf{x}) = \sum_{i=1}^{b} F_j(x_i), \tag{8}$$

Table 1: Comparison of AdaptOD with EnergyOE and COCL on six OOD datasets.

OOD Dataset	Method	ID Dataset: CIFAR10-LT				ID Dataset: CIFAR100-LT			
		AUC↑	AP-in↑	AP-out↑	FPR↓	AUC↑	AP-in↑	AP-out↑	FPR↓
Texture [5]	EnergyOE [24]	95.53	97.42	92.93	18.44	79.56	86.03	70.88	79.45
	COCL [30]	96.81	98.21	93.86	14.65	81.99	88.05	74.38	59.79
	AdaptOD(Ours)	**98.22**	**98.81**	**94.91**	**11.60**	**83.88**	**89.43**	**76.47**	**58.47**
SVHN [31]	EnergyOE [24]	96.63	92.33	98.46	14.37	86.19	81.42	91.74	34.36
	COCL [30]	96.98	93.25	98.61	12.59	89.20	81.57	94.21	54.46
	AdaptOD(Ours)	**98.13**	**94.34**	**99.11**	**10.33**	**93.09**	**91.32**	**96.86**	**17.63**
CIFAR [19]	EnergyOE [24]	84.44	85.74	84.63	61.73	61.15	67.12	56.66	91.42
	COCL [30]	86.63	86.66	86.28	52.21	62.05	66.14	56.82	93.88
	AdaptOD(Ours)	**89.05**	**89.93**	**88.22**	**45.51**	**72.77**	**76.37**	**70.58**	**86.04**
TIN [20]	EnergyOE [24]	88.40	91.65	84.95	46.23	70.78	79.40	55.90	90.74
	COCL [30]	90.43	92.52	87.03	46.12	71.87	81.89	57.12	83.93
	AdaptOD(Ours)	**91.40**	**93.85**	**88.18**	**42.77**	**72.87**	**82.06**	**58.92**	**88.24**
LSUN [51]	EnergyOE [24]	94.00	94.78	93.70	28.42	81.61	86.57	69.16	80.57
	COCL [30]	94.85	95.43	93.98	27.48	84.10	89.89	69.80	74.67
	AdaptOD(Ours)	**96.16**	**96.84**	**95.86**	**24.12**	**85.70**	**90.55**	**72.70**	**70.20**
Place365 [58]	EnergyOE [24]	92.51	84.26	97.14	33.63	79.12	63.38	89.09	81.43
	COCL [30]	93.97	87.36	97.56	32.25	80.30	68.65	89.16	77.83
	AdaptOD(Ours)	**95.19**	**89.56**	**98.44**	**29.22**	**83.27**	**68.82**	**91.44**	**71.63**

where b is the batch size of the batch $\mathbf{x}$ (if $\mathbf{x}$ is $\mathbf{x}^{in}$ that b is b^{in}, and $\mathbf{x}$ is $\mathbf{x}^{out}$ that b is b^{out}), and x_i is the i-th sample in the batch $\mathbf{x}$. Notably, even if some classes do not have the corresponding ID samples in a certain training batch, this loss also can work well. This is because there is less distribution shift among classes in the ID data compared to the OOD data. Therefore, the output of ID samples on incorrect ID classes should also be higher than the OOD samples. DNE-C balances the sum of energy on all ID samples for each ID class and distinguishes outlier samples from ID samples, especially for the underestimated tail classes in a class-wise manner.

Sample-wise Normalized Energy Loss (DNE-S). DNE-S independently regularizes the energy for each sample to enhance the energy of ID samples for sample-wise balanced energy. Formally, we minimize the global energy over all classes of each sample as follows:

$$\mathcal{L}_S = \mathbb{E}_{(x,y)\sim\mathcal{D}_{in}}[(max(0, m^s_{in} - \mathbf{S}(x)))^2] + \mathbb{E}_{x\sim\mathcal{D}_{out}}[(max(0, \mathbf{S}(x) - m^s_{out}))^2], \tag{9}$$

where $x \in \mathbf{x}^{in} \cup \mathbf{x}^{out}$, $m^s_{in} = \frac{k}{b^{in}}$ and $m^s_{out} = 0$ are the default margin hyperparameter settings that can also work stably regardless of the ID/OOD datasets (see Appendix C). Then the sample-wise normalized energy $\mathbf{S}(x)$ can be defined as:

$$\mathbf{S}(x) = \sum_{j=1}^{k} F_j(x). \tag{10}$$

After doing this, we can regularize the global energy of the ID data, particularly the low global energy for tail samples. DNE-S efficiently balances the energy between head and tail samples. As a result, the combination of DNE-C and DNE-S can learn substantially more balanced prediction energy of ID samples, facilitating DODA to solve the distribution shift problems.

Overall Training Objective. Overall, we utilize the cross-entropy loss, together with our two normalized energy losses, to train our model. The final objective of our DNE training is as follows:

$$\mathcal{L}_{total} = \mathbb{E}_{x,y\sim\mathcal{D}_{in}}[\ell(f(x), y] + \mathcal{L}_{dne}, \tag{11}$$

where ℓ is a cross-entropy loss, along with the two normalized energy losses:

$$\begin{aligned}\mathcal{L}_{dne} =& \mathcal{L}_S + \mathcal{L}_C = \mathbb{E}_{(x,y)\sim\mathcal{D}_{in}}[(max(0, m^s_{in} - \mathbf{S}(x)))^2] + \mathbb{E}_{x\sim\mathcal{D}_{out}}[(max(0, \mathbf{S}(x) - m^s_{out}))^2], \\ &+ \sum_{j=1}^{k}(\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{in}}[(max(0, m^c_{in} - \mathbf{C}_j(\mathbf{x})))^2] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{out}}[(max(0, \mathbf{C}_j(\mathbf{x}) - m^c_{out}))^2])\end{aligned} \tag{12}$$

where $\mathcal{L}_C$ is as defined in Eq. 7 and $\mathcal{L}_S$ is as defined in Eq. 9. The algorithm of AdaptOD described in Appendix B.

https://doi.org/10.52202/079017-4199

Table 2: Comparison to different long-tailed OOD detection methods.

Method	ID Dataset: CIFAR10-LT					ID Dataset: CIFAR100-LT				
	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
OE [12]	89.76	89.45	87.22	53.19	73.59	73.52	75.06	67.27	86.30	39.42
EnergyOE [24]	91.92	91.03	91.97	33.80	74.57	76.40	77.32	72.24	76.33	41.32
PASCL [40]	90.99	90.56	89.24	42.90	77.08	73.32	74.84	67.18	79.38	43.10
EAT [45]	92.87	91.76	92.40	32.42	81.31	75.45	76.02	70.87	77.83	46.23
Class Prior [17]	92.08	91.17	90.86	34.42	74.33	76.03	77.31	72.26	76.43	40.77
BERL [4]	92.56	91.41	91.94	32.83	81.37	77.75	78.61	73.10	74.86	45.88
COCL [30]	93.28	92.24	92.89	30.88	81.56	78.25	79.37	73.58	74.09	46.41
OE [12]+**DODA(Ours)**	91.62	90.55	89.39	49.02	73.59	75.46	77.14	69.88	83.67	39.42
EnergyOE [24]+**DODA(Ours)**	93.36	92.17	92.97	30.82	74.57	79.40	80.89	76.54	72.63	41.32
BERL [4]+**DODA(Ours)**	93.77	92.62	93.15	29.41	81.37	79.45	81.15	75.52	70.51	45.88
COCL [30]+**DODA(Ours)**	93.89	93.06	93.39	29.25	81.56	79.81	81.26	75.93	70.14	46.41
AdaptOD(Ours)	**94.69**	**93.89**	**94.12**	**27.26**	**82.27**	**81.93**	**83.09**	**77.83**	**67.37**	**47.91**

Table 3: Comparison to different TTA-based OOD detection methods.

Training Method	TTA Method	ID Dataset: CIFAR10-LT				ID Dataset: CIFAR100-LT			
		AUC↑	AP-in↑	AP-out↑	FPR↓	AUC↑	AP-in↑	AP-out↑	FPR↓
OE [12]	w/o TTA	89.76$_{\pm 0.27}$	89.45$_{\pm 0.56}$	87.22$_{\pm 0.61}$	53.19$_{\pm 0.42}$	73.52$_{\pm 0.68}$	75.06$_{\pm 0.59}$	67.27$_{\pm 0.57}$	86.30$_{\pm 0.92}$
	AUTO [49]	90.49$_{\pm 0.29}$	89.83$_{\pm 0.52}$	87.45$_{\pm 0.83}$	52.63$_{\pm 0.47}$	73.93$_{\pm 0.89}$	75.98$_{\pm 0.81}$	67.74$_{\pm 0.65}$	85.71$_{\pm 1.00}$
	AdaODD [56]	90.89$_{\pm 0.26}$	90.17$_{\pm 0.51}$	87.88$_{\pm 0.84}$	51.44$_{\pm 0.56}$	74.67$_{\pm 0.92}$	76.53$_{\pm 0.64}$	67.89$_{\pm 0.82}$	85.34$_{\pm 0.94}$
	DODA(Ours)	91.62$_{\pm 0.23}$	90.55$_{\pm 0.45}$	89.39$_{\pm 0.68}$	49.02$_{\pm 0.41}$	75.46$_{\pm 0.77}$	77.14$_{\pm 0.59}$	69.88$_{\pm 0.80}$	83.67$_{\pm 0.88}$
EnergyOE [24]	w/o TTA	91.92$_{\pm 0.30}$	91.03$_{\pm 0.53}$	91.97$_{\pm 0.62}$	33.80$_{\pm 0.56}$	76.40$_{\pm 0.86}$	77.32$_{\pm 0.59}$	72.24$_{\pm 0.62}$	76.33$_{\pm 1.03}$
	AUTO [49]	92.48$_{\pm 0.32}$	91.43$_{\pm 0.55}$	92.44$_{\pm 0.79}$	31.99$_{\pm 0.36}$	77.65$_{\pm 1.01}$	78.11$_{\pm 0.62}$	74.18$_{\pm 0.78}$	74.66$_{\pm 0.99}$
	AdaODD [56]	92.28$_{\pm 0.26}$	91.63$_{\pm 0.56}$	91.73$_{\pm 0.61}$	32.83$_{\pm 0.59}$	77.67$_{\pm 0.82}$	78.47$_{\pm 0.81}$	74.05$_{\pm 0.83}$	74.86$_{\pm 0.98}$
	DODA(Ours)	93.36$_{\pm 0.28}$	92.17$_{\pm 0.53}$	92.97$_{\pm 0.70}$	30.82$_{\pm 0.51}$	79.40$_{\pm 0.98}$	80.89$_{\pm 0.84}$	76.54$_{\pm 0.64}$	72.63$_{\pm 0.94}$
BERL [4]	w/o TTA	92.56$_{\pm 0.40}$	91.41$_{\pm 0.83}$	91.94$_{\pm 0.85}$	32.83$_{\pm 0.38}$	77.75$_{\pm 0.77}$	78.61$_{\pm 0.56}$	73.10$_{\pm 0.73}$	74.86$_{\pm 1.07}$
	AUTO [49]	92.41$_{\pm 0.49}$	91.73$_{\pm 0.56}$	92.42$_{\pm 0.90}$	31.91$_{\pm 0.36}$	77.99$_{\pm 0.75}$	78.50$_{\pm 0.84}$	73.50$_{\pm 0.87}$	74.03$_{\pm 1.00}$
	AdaODD [56]	92.68$_{\pm 0.26}$	91.79$_{\pm 0.54}$	92.20$_{\pm 0.67}$	31.41$_{\pm 0.51}$	78.26$_{\pm 0.97}$	78.94$_{\pm 0.81}$	73.61$_{\pm 0.75}$	73.76$_{\pm 1.12}$
	DODA(Ours)	93.77$_{\pm 0.30}$	92.62$_{\pm 0.51}$	93.15$_{\pm 0.73}$	29.41$_{\pm 0.37}$	79.45$_{\pm 0.83}$	81.15$_{\pm 0.79}$	75.52$_{\pm 0.69}$	70.51$_{\pm 0.91}$
COCL [30]	w/o TTA	93.28$_{\pm 0.30}$	92.24$_{\pm 0.78}$	92.89$_{\pm 0.72}$	30.88$_{\pm 0.63}$	78.25$_{\pm 0.99}$	79.37$_{\pm 0.65}$	73.58$_{\pm 0.76}$	74.09$_{\pm 0.85}$
	AUTO [49]	93.62$_{\pm 0.43}$	92.74$_{\pm 0.83}$	93.10$_{\pm 0.59}$	30.41$_{\pm 0.40}$	78.85$_{\pm 0.97}$	79.99$_{\pm 0.72}$	74.01$_{\pm 0.86}$	72.75$_{\pm 0.95}$
	AdaODD [56]	93.48$_{\pm 0.22}$	92.60$_{\pm 0.66}$	93.05$_{\pm 0.81}$	30.79$_{\pm 0.39}$	79.07$_{\pm 0.70}$	80.00$_{\pm 0.60}$	74.60$_{\pm 0.84}$	73.09$_{\pm 0.91}$
	DODA(Ours)	93.89$_{\pm 0.36}$	93.06$_{\pm 0.56}$	93.39$_{\pm 0.74}$	29.25$_{\pm 0.40}$	79.81$_{\pm 0.96}$	81.26$_{\pm 0.59}$	75.93$_{\pm 0.72}$	70.14$_{\pm 0.98}$
DNE (Ours)	w/o TTA	92.77$_{\pm 0.48}$	92.18$_{\pm 0.71}$	92.62$_{\pm 0.61}$	31.48$_{\pm 0.36}$	77.92$_{\pm 0.75}$	78.97$_{\pm 0.61}$	73.92$_{\pm 0.81}$	74.44$_{\pm 0.99}$
	AUTO [49]	92.89$_{\pm 0.44}$	92.69$_{\pm 0.86}$	92.25$_{\pm 0.60}$	30.85$_{\pm 0.62}$	79.36$_{\pm 0.91}$	80.19$_{\pm 0.63}$	74.81$_{\pm 0.80}$	72.10$_{\pm 1.19}$
	AdaODD [56]	93.39$_{\pm 0.46}$	92.27$_{\pm 0.69}$	92.92$_{\pm 0.59}$	30.78$_{\pm 0.55}$	80.26$_{\pm 0.81}$	81.72$_{\pm 0.68}$	75.62$_{\pm 0.88}$	71.96$_{\pm 0.95}$
	DODA(Ours)	**94.69**$_{\pm 0.22}$	**93.89**$_{\pm 0.68}$	**94.12**$_{\pm 0.58}$	**27.26**$_{\pm 0.49}$	**81.93**$_{\pm 0.71}$	**83.09**$_{\pm 0.64}$	**77.83**$_{\pm 0.76}$	**67.37**$_{\pm 0.93}$

4 Experiments

4.1 Experiment Settings

Datasets. Following [30, 40, 45], we use three popular long-tailed datasets CIFAR10-LT [3], CIFAR100-LT [3] and ImageNet-LT [26] as ID data X^{in}. The default imbalance ratio is set to $\rho = 100$ on CIFAR10/100-LT. TinyImages80M [38] is used as the outlier data X^{aux_out} for CIFAR10/100-LT and ImageNet-Extra [40] is used as outlier data for ImageNet-LT. We use six datasets CIFAR [19], Texture [5], SVHN [31], LSUN [51], Places365 [58] and TinyImageNet [20], all of which are introduced in the SC-OOD benchmark [48] as the OOD test set for CIFAR10/100-LT, and ImageNet-1k-OOD [40] as the OOD test set for ImageNet-LT. More details about the datasets are presented in Appendix A.1.

Implementation Details. Our AdaptOD is compared with seven SOTA OOD detection methods on long-tailed data, including two popular methods: OE [12] and EnergyOE [24], and five recent methods: PASCL [40], EAT [45], Class Prior [17], BERL [4], and COCL [30]. Further, we also compare AdaptOD with two SOTA TTA methods for OOD detection, including AUTO [49] and AdaOOD [56]. We use ResNet18 [11] as our backbone on CIFAR10/100-LT and ResNet50 [11] on ImageNet-LT. Following fine-tuning-based methods OE [12], EnergyOE [24], and BERL [4], our approach AdaptOD employs a similar training strategy to them that we obtain a pre-trained model with only ID data and fine-tune this model with both ID data and outlier data. The reported results are averaged over six independent runs. More details about the implementation details are presented in Appendix A.2.

Table 4: Comparison results on the large-scale ID dataset ImageNet-LT.

Method	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
OE [12]	68.33	43.87	82.54	90.98	44.00
EnergyOE [24]	69.43	45.12	84.75	76.89	44.42
EAT [45]	69.84	43.15	81.32	80.97	46.79
PASCL [40]	68.00	43.32	82.69	82.28	47.29
Class Prior [17]	70.43	45.26	84.82	77.63	46.83
BERL [4]	71.16	45.97	85.63	76.98	50.42
COCL [30]	71.85	46.76	86.21	75.60	51.11
BERL [4]+**DODA**	73.12	47.34	86.95	74.92	50.42
COCL [30]+**DODA**	73.27	47.98	87.77	74.71	51.11
AdaptOD(Ours)	**74.32**	**49.02**	**88.63**	**72.91**	**51.67**

Training	Test	AUC↑	AP-in↑	AP-out↑	FPR↓
BERL [4]	w/o TTA	$71.16_{\pm 0.96}$	$45.97_{\pm 0.85}$	$85.63_{\pm 0.77}$	$76.98_{\pm 1.79}$
	AUTO [49]	$71.66_{\pm 1.20}$	$46.58_{\pm 0.80}$	$86.05_{\pm 0.77}$	$76.09_{\pm 1.63}$
	AdaODD [56]	$71.80_{\pm 1.14}$	$46.47_{\pm 0.63}$	$85.56_{\pm 1.01}$	$77.36_{\pm 1.69}$
	DODA(Ours)	$73.12_{\pm 1.18}$	$47.34_{\pm 0.75}$	$86.95_{\pm 0.76}$	$74.92_{\pm 1.67}$
COCL [30]	w/o TTA	$71.85_{\pm 1.15}$	$46.76_{\pm 1.13}$	$86.21_{\pm 1.11}$	$75.60_{\pm 1.38}$
	AUTO [49]	$71.79_{\pm 1.22}$	$46.84_{\pm 0.81}$	$86.89_{\pm 1.18}$	$75.28_{\pm 1.69}$
	AdaODD [56]	$72.35_{\pm 1.10}$	$47.20_{\pm 1.16}$	$86.89_{\pm 0.96}$	$75.06_{\pm 1.91}$
	DODA(Ours)	$73.27_{\pm 1.19}$	$47.98_{\pm 1.00}$	$87.77_{\pm 0.74}$	$74.71_{\pm 1.55}$
DNE (Ours)	w/o TTA	$72.04_{\pm 1.07}$	$46.53_{\pm 0.72}$	$86.06_{\pm 0.78}$	$75.82_{\pm 1.38}$
	AUTO [49]	$73.31_{\pm 1.26}$	$47.26_{\pm 1.14}$	$87.11_{\pm 1.19}$	$74.60_{\pm 1.27}$
	AdaODD [56]	$73.10_{\pm 0.81}$	$46.83_{\pm 0.76}$	$86.68_{\pm 0.74}$	$74.64_{\pm 1.41}$
	DODA(Ours)	$\mathbf{74.32}_{\pm \mathbf{0.92}}$	$\mathbf{49.02}_{\pm \mathbf{0.70}}$	$\mathbf{88.63}_{\pm \mathbf{0.73}}$	$\mathbf{72.91}_{\pm \mathbf{1.28}}$

Evaluation Measures. Following [30, 48], we use the below common metrics for OOD detection and ID classification: (1) FPR is the false positive rate of OOD examples when the true positive rate of ID examples is at 95%, (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, which can be either AP-in in which ID samples are treated as positive or AP-out in which OOD samples are regarded as positive, and (4) ACC calculates the classification accuracy of the long-tailed ID data. The reported results are averaged over six independent runs with different random seeds by default.

4.2 Empirical Results

AdaptOD vs. Other OOD Detection Methods in LTR. Table 1 presents the comparison of our AdaptOD with two SOTA OOD detectors in LTR (EnergyOE [24], COCL [30]) on CIFAR10/100-LT using six OOD test datasets. These fine-grained results are not available for the other competing methods and thus they are omitted in this table. AdaptOD shows the best performance in all four metrics on each of the six OOD datasets. Table 2 shows the comparison of our AdaptOD with SOTA OOD detectors in LTR on CIFAR10/100-LT, which is the average performance over six OOD test datasets. Following the previous methods [4, 40, 45], we report our accuracy with AdjLogit [29] for a fair comparison. AdaptOD is also the best performers in the averaged results when comparing to all seven competing methods. This consistent improvement and SOTA performance of AdaptOD on both ID and OOD data indicate that the distribution gap between the outlier samples and the true OOD samples is effectively reduced by AdaptOD. Notably, the improvement is large on the near OOD dataset CIFAR [19], which cannot be achieved by previous SOTA methods [4, 30].

DODA as an Enabler to Existing Methods. Table 2 also presents the results of our proposed component DODA in using as a plug-in to tackle the distribution shift problem in four SOTA methods (OE, EnergyOE, BERL, and COCL) on CIFAR10/100-LT. It shows that DODA can consistently enhance the OOD detectors in all four metrics, demonstrating the strong capability of DODA in reducing the learned outlier distribution gap to the distribution of the true OOD data (see Appendix D for more details). The consistent improvement of having DODA as 'plug-and-play' indicates the presence of the distribution shift problem encountered by existing SOTA detectors and the universal effectiveness of DODA in tackling the problem. Note that AdaptOD as a whole achieves consistent and substantial improvement over the four DODA-enabled models, showcasing that the other component of AdaptOD, DNE, helps to learn balanced ID prediction energy and better align the adapted outlier distribution to the true OOD one.

AdaptOD vs. Other TTA Methods for OOD Detection. Table 3 shows the comparison of AdaptOD with two SOTA TTA methods AUTO and AdaODD for OOD detection on CIFAR10/100-LT. To have a straightforward and extensive comparison, we compare DODA with the two TTA methods, all of which are added on top of the same training method. In the experiments, we use five training methods, including four SOTA long-tailed OOD detection methods and our proposed DNE method. It is impressive that our DODA component consistently remains the best performer when the TTA methods are combined with all five different training methods on both ID datasets. DODA achieves better performance in all four OOD detection metrics across five OOD training methods, indicating that DODA is a stronger and more generic TTA method for different OOD detectors. Moreover, the combination of our training method DNE and TTA method DODA, which is our approach AdaptOD as a whole, achieves the best performance across all 20 possible combinations.

https://doi.org/10.52202/079017-4199

Table 5: Ablation study results on CIFAR10-LT, CIFAR100-LT and ImageNet-LT.

DODA	DNE-C	DNE-S	ID Dataset: CIFAR10-LT				ID Dataset: CIFAR100-LT				ID Dataset: ImageNet-LT			
			AUC↑	AP-in↑	AP-out↑	FPR↓	AUC↑	AP-in↑	AP-out↑	FPR↓	AUC↑	AP-in↑	AP-out↑	FPR↓
Baseline (EnergyOE [24])			91.92	91.03	91.97	33.80	76.40	77.32	72.24	76.33	69.43	45.12	84.75	76.89
✗	✗	✗	80.33	81.46	77.02	78.71	67.42	68.29	63.86	85.44	58.33	38.40	77.61	89.73
✓	✗	✗	92.63	92.05	92.46	30.17	78.10	80.22	74.17	71.65	71.71	45.99	86.37	74.31
✗	✓	✗	92.12	91.54	92.33	31.85	76.89	77.94	72.76	74.97	71.11	45.59	85.77	76.83
✗	✗	✓	91.98	91.36	91.92	32.44	76.53	77.46	72.55	74.62	70.55	45.36	84.95	77.02
✗	✓	✓	92.77	92.18	92.62	31.48	77.92	78.97	73.92	74.44	72.04	46.53	86.06	75.82
✓	✓	✗	93.81	93.32	93.53	28.69	80.07	82.13	75.73	68.64	73.14	47.61	87.19	73.67
✓	✗	✓	93.49	92.98	93.02	29.52	79.76	81.89	75.31	69.19	72.76	47.32	86.83	74.48
✓	✓	✓	**94.69**	**93.89**	**94.12**	**27.26**	**81.93**	**83.09**	**77.83**	**67.37**	**74.32**	**49.02**	**88.63**	**72.91**
Oracle Model			95.33	94.75	94.96	25.02	83.60	85.09	78.85	65.37	75.84	50.20	89.97	70.71

Performance on Large-scale ID Data. To demonstrate the scalability of our approach, we also perform experiments on the large-scale ID dataset ImageNet-LT. The empirical results are presented in Table 4, which shows that our approach AdaptOD also achieves the SOTA performance in both the OOD detection performance and the ID classification accuracy.

4.3 Further Analysis of AdaptOD

Ablation Study. The effectiveness of our two proposed components, DODA and DNE, have been justified in Table 3. Here we provide a more fine-grained analysis of DODA and its combination to two improved energy losses used in DNE, $\mathcal{L}_C$ (Eq. 7, denoted as DNE-C) and $\mathcal{L}_S$ (Eq. 9, denoted as DNE-S), in Table 5, with EnergyOE [24] used as baseline. The results show the important contribution of each component to the overall superior performance of the full model AdaptOD. Further, we compare AdaptOD to an oracle model that utilizes the ground true OOD data to update the outlier distribution $\mathcal{P}^{out}$ in DODA. It shows that AdaptOD has only a small performance gap to the oracle model, indicating that AdaptOD can well approximate the true OOD distribution by the predicted labels of the OOD samples, without involving any ground truth during TTA.

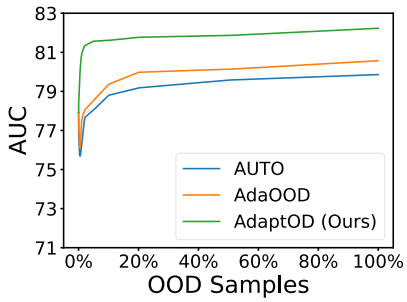


Figure 3: The average performance over six OOD datasets on CIFAR100-LT with an increasing percentage of true OOD samples fed to TTA methods.

OOD Data Exploitation in TTA. To independently evaluate the effectiveness of exploiting OOD data to adapt the outlier distribution, we report the performance of three TTA methods with an increasing number of labeled OOD samples based on our DNE in Fig. 3. All three TTA methods achieve increasing performance for OOD detection in LTR with more and more true OOD data used for the adaptation. However, AUTO and AdaOOD struggle with the difference between training and testing ID data at the early stage of inference, while AdaptOD can utilize the adapted outlier distribution to quickly adapt to the true OOD distribution and achieve significantly improved performance.

5 Conclusion

To address the distribution shift problem in long-tailed OOD detection, we propose a novel approach called AdaptOD. It utilizes a novel normalized energy-based loss – dual-normalized energy loss (DNE) – to learn balanced prediction energy on imbalanced ID samples and enhanced vanilla outlier distribution, then uses a dynamic outlier distribution adaptation (DODA) to adapt the outlier distribution to the true OOD distribution. DODA is shown to be a significantly improved TTA method than existing TTA methods for OOD detection. We also show that DNE can be used to support DODA with its specially designed energy training for better test-time distribution adaptation. Experiments on three popular benchmarks demonstrated that AdaptOD significantly enhances the performance of both OOD detection and long-tailed classification.

https://doi.org/10.52202/079017-4199

Acknowledgments

The participation of W. Miao, J. Zheng, and X. Bai in this work was supported by National Natural Science Foundation of China (No. 62372029 and No. 62276016), while the participation of G. Pang was supported in part by Lee Kong Chian Fellowship.

References

[1] Alshammari, S., Wang, Y.X., Ramanan, D., Kong, S.: Long-tailed recognition via weight balancing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6897–6907 (2022) 3, 15

[2] Bai, J., Liu, Z., Wang, H., Hao, J., Feng, Y., Chu, H., Hu, H.: On the effectiveness of out-of-distribution data in self-supervised long-tail learning. arXiv preprint arXiv:2306.04934 (2023) 3

[3] Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems **32** (2019) 2, 8, 15

[4] Choi, H., Jeong, H., Choi, J.Y.: Balanced energy regularization loss for out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15691–15700 (2023) 2, 3, 5, 6, 8, 9, 15, 20

[5] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014) 7, 8, 15, 18

[6] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 15

[7] Fan, K., Wang, Y., Yu, Q., Li, D., Fu, Y.: A simple test-time method for out-of-distribution detection. arXiv preprint arXiv:2207.08210 (2022) 3, 4

[8] Gao, Z., Yan, S., He, X.: Atta: Anomaly-aware test-time adaptation for out-of-distribution detection in segmentation. Advances in Neural Information Processing Systems **36** (2024) 3

[9] Gou, Y., Hu, P., Lv, J., Zhu, H., Peng, X.: Rethinking image super resolution from long-tailed distribution learning perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14327–14336 (2023) 3

[10] He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering **21**(9), 1263–1284 (2009) 15

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 8, 15

[12] Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606 (2018) 2, 3, 8, 9, 17, 18

[13] Hong, F., Yao, J., Zhou, Z., Zhang, Y., Wang, Y.: Long-tailed partial label learning via dynamic rebalancing. arXiv preprint arXiv:2302.05080 (2023) 3

[14] Hu, X., Uzunbas, G., Chen, S., Wang, R., Shah, A., Nevatia, R., Lim, S.N.: Mixnorm: Test-time adaptation through online normalization estimation. arXiv preprint arXiv:2110.11478 (2021) 3

[15] Huang, R., Li, Y.: Mos: Towards scaling out-of-distribution detection for large semantic space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8710–8719 (2021) 1

[16] Iwasawa, Y., Matsuo, Y.: Test-time classifier adjustment module for model-agnostic domain generalization. Advances in Neural Information Processing Systems **34**, 2427–2440 (2021) 3

[17] Jiang, X., Liu, F., Fang, Z., Chen, H., Liu, T., Zheng, F., Han, B.: Detecting out-of-distribution data through in-distribution class prior. In: International Conference on Machine Learning. PMLR (2023) 2, 3, 8, 9

 https://doi.org/10.52202/079017-4199

[18] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017) 1

[19] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 7, 8, 9, 15, 18

[20] Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015) 7, 8, 15, 18

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

[22] Li, J., Chen, P., He, Z., Yu, S., Liu, S., Jia, J.: 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. pp. 11578–11589 (2023) 3

[23] Li, T., Pang, G., Bai, X., Zheng, J., Zhou, L., Ning, X.: Learning adversarial semantic embeddings for zero-shot recognition in open worlds. Pattern Recognition **149**, 110258 (2024) 3

[24] Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in neural information processing systems **33**, 21464–21475 (2020) 1, 2, 5, 6, 7, 8, 9, 10, 15, 17, 18, 19, 20

[25] Liu, Y., Ding, C., Tian, Y., Pang, G., Belagiannis, V., Reid, I., Carneiro, G.: Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1151–1161 (2023) 3

[26] Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2537–2546 (2019) 8

[27] Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 15

[28] Mehta, D., Gal, Y., Bowling, A., Bonnington, P., Ge, Z.: Out-of-distribution detection for long-tailed and fine-grained skin lesion images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 732–742. Springer (2022) 3

[29] Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314 (2020) 9

[30] Miao, W., Pang, G., Li, T., Bai, X., Zheng, J.: Out-of-distribution detection in long-tailed recognition with calibrated outlier class learning. In: Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence (2024) 2, 3, 7, 8, 9, 15, 17, 18, 19, 20

[31] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011) 2, 7, 8, 15, 18

[32] Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972 (2021) 15

[33] Shi, J.X., Wei, T., Xiang, Y., Li, Y.F.: How re-sampling helps for long-tail learning? Advances in Neural Information Processing Systems **36** (2023) 3

[34] Shi, J.X., Wei, T., Zhou, Z., Shao, J.J., Han, X.Y., Li, Y.F.: Long-tail learning with foundation model: Heavy fine-tuning hurts. In: Forty-first International Conference on Machine Learning (2024) 3

[35] Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. Advances in Neural Information Processing Systems **34**, 144–157 (2021) 3

[36] Tang, K., Tao, M., Qi, J., Liu, Z., Zhang, H.: Invariant feature learning for generalized long-tailed classification. In: European Conference on Computer Vision. pp. 709–726. Springer (2022) 3

[37] Tian, Y., Liu, Y., Pang, G., Liu, F., Chen, Y., Carneiro, G.: Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In: European Conference on Computer Vision. pp. 246–263. Springer (2022) 3

[38] Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for non-parametric object and scene recognition. IEEE transactions on pattern analysis and machine intelligence **30**(11), 1958–1970 (2008) 2, 8, 15

[39] Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020) 3

[40] Wang, H., Zhang, A., Zhu, Y., Zheng, S., Li, M., Smola, A.J., Wang, Z.: Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In: International Conference on Machine Learning. pp. 23446–23458. PMLR (2022) 2, 3, 8, 9, 15

[41] Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7211 (2022) 3

[42] Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S.X.: Long-tailed recognition by routing diverse distribution-aware experts. arXiv preprint arXiv:2010.01809 (2020) 1

[43] Wang, Z., Li, Y., Chen, X., Lim, S.N., Torralba, A., Zhao, H., Wang, S.: Detecting everything in the open world: Towards universal object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11433–11443 (2023) 3

[44] Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: International Conference on Machine Learning. pp. 23631–23644. PMLR (2022) 3

[45] Wei, T., Wang, B.L., Zhang, M.L.: Eat: Towards long-tailed out-of-distribution detection. arXiv preprint arXiv:2312.08939 (2023) 3, 8, 9

[46] Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y.: Large scale incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 374–382 (2019) 15

[47] Xiao, Z., Zhen, X., Shao, L., Snoek, C.G.: Learning to generalize across domains on single test samples. arXiv preprint arXiv:2202.08045 (2022) 3

[48] Yang, J., Wang, H., Feng, L., Yan, X., Zheng, H., Zhang, W., Liu, Z.: Semantically coherent out-of-distribution detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8301–8309 (2021) 1, 3, 8, 9, 15

[49] Yang, P., Liang, J., Cao, J., He, R.: Auto: Adaptive outlier optimization for online test-time ood detection. arXiv preprint arXiv:2303.12267 (2023) 2, 3, 8, 9, 19

[50] Ye, Y., Chen, S., Ni, D., Huang, R.: Triaug: Out-of-distribution detection for robust classification of imbalanced breast lesion in ultrasound. arXiv preprint arXiv:2402.07452 (2024) 3

[51] Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015) 7, 8, 15, 18

[52] Yu, Y., Shin, S., Lee, S., Jun, C., Lee, K.: 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. pp. 15701–15711 (2023) 3

[53] Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. Advances in Neural Information Processing Systems **35**, 38629–38642 (2022) 3

[54] Zhang, Y.F., Zhang, H., Wang, J., Zhang, Z., Yu, B., Wang, L., Tao, D., Xie, X.: Domain-specific risk minimization. arXiv preprint arXiv:2208.08661 (2022) 3

[55] Zhang, Y., Wang, X., Jin, K., Yuan, K., Zhang, Z., Wang, L., Jin, R., Tan, T.: Adanpc: Exploring non-parametric classifier for test-time adaptation. In: International Conference on Machine Learning. pp. 41647–41676. PMLR (2023) 3

[56] Zhang, Y., Wang, X., Zhou, T., Yuan, K., Zhang, Z., Wang, L., Jin, R., Tan, T.: Model-free test time adaptation for out-of-distribution detection. arXiv preprint arXiv:2311.16420 (2023) 2, 3, 8, 9, 19

[57] Zhang, Z., Xiang, X.: Decoupling maxlogit for out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3388–3397 (2023) 3

[58] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017) 7, 8, 15, 18

 https://doi.org/10.52202/079017-4199

[59] Zhu, F., Cheng, Z., Zhang, X.Y., Liu, C.L.: Openmix: Exploring outlier samples for misclassification detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12074–12083 (2023) 2

A More Experiment Settings

A.1 Datasets

For ID datasets, the original version of CIFAR10 [10] and CIFAR100 [10] contains $50,000$ training images and $10,000$ validation images of size 32×32 with 10 and 100 classes, respectively. CIFAR10-LT and CIFAR100-LT are the imbalanced version of them, which reduce the number of training examples per class and keep the validation set unchanged. The imbalance ratio ρ denotes the ratio between sample sizes of the most frequent class and least frequent class. Following [3], we utilize an exponential decay in sample sizes across different classes.

ImageNet-LT [46] is a large-scale dataset in long-tail recognition, which truncates the balanced version ImageNet [6]. ImageNet-LT has $1,000$ classes, which contain $115,846$ training images with the number of per-class training data ranging from 5 to $1,280$, and $20,000$ validation images with a balanced class size.

For outlier data, TinyImages80M [38] contains 80 million images with a size of 3232. We use a subset of random 30K images as the outlier data for CIFAR10-LT and CIFAR100-LT [4, 40]. We use ImageNet-Extra [40] that contains $517,711$ images belonging to 500 classes from ImageNet-22k [6] but having not overlapping with the $1,000$ in-distribution classes in ImageNet-LT [4, 40].

For OOD datasets, we use SC-OOD benchmark [48] as true OOD data for CIFAR10-LT and CIFAR100-LT [4, 40] following [4, 40]. The SC-OOD benchmark contains six datasets: CIFAR [19] with $10,000$ images, Texture [5] with $5,640$ images, SVHN [31] with $26,032$ images, LSUN [51] with $9,998$ images for CIFAR10-LT and $7,571$ images for CIFAR100-LT, Places365 [58] with $35,195$ images for CIFAR10-LT and $33,773$ images for CIFAR100-LT, and TinyImageNet [20] with $8,793$ images for CIFAR10-LT and $7,498$ images for CIFAR100-LT. Following [30, 40], we use ImageNet-1k-OOD [40] that contains $50,000$ images belonging to $1,000$ classes evenly from ImageNet-22k, which have not overlapping with the $1,000$ ID classes in ImageNet-LT and the 500 outlier classes in ImageNet-Extra. A summary of the ID and OOD datasets is presented in Table 6.

Table 6: Key statistics of the ID and OOD datasets used.

Benchmark	CIFAR10-LT			CIFAR100-LT			ImageNet-LT		
	Dataset	Images	Class	Dataset	Images	Class	Dataset	Images	Class
ID data (Training)	CIFAR10-LT	/	10	CIFAR100-LT	/	100	ImageNet-LT	115,846	1,000
ID data (Testing)	CIFAR10-LT	10,000	10	CIFAR100-LT	10,000	100	ImageNet-LT	20,000	1,000
Outlier data	TinyImages80M	30,000	/	TinyImages80M	30,000	/	ImageNet-Extra	517,711	500
OOD data	CIFAR100	10,000	100	CIFAR10	10,000	10	ImageNet-1k-OOD	50,000	1000
	Texture	5,640	47	Texture	5,640	47			
	SVHN	26,032	10	SVHN	26,032	10			
	LSUN	9,998	10	LSUN	7,571	/			
	Places365	35,195	/	Places365	33,773	/			
	TinyImageNet	8,793	/	TinyImageNet	7,498	/			

A.2 Implementation Details

For experiments on CIFAR10-LT [3] and CIFAR100-LT [3], we pre-train our model based on ResNet18 [11] for 320 epochs with an initial learning rate 0.01 [1, 4] using only cross-entropy loss and fine-tune the linear classifier of this model for 20 epochs with an initial learning rate 0.001 [4, 24]. The batch size is 64 for ID data at the pre-training stage, 128 for ID data at the fine-tuning stage, and 256 for outlier data at the fine-tuning stage [4, 30, 40]. Our outlier dataset is a subset of TinyImages80M [38] with 30K images [4, 40].

For large-scale dataset ImageNet-LT, which contains 115,846 images of 1,000 classes, we train our model based on ResNet50 [11] for 100 epochs with an initial learning rate of 0.1 [30] using only cross-entropy loss and also fine-tune the linear classifier of this model for 20 epochs with an initial learning rate 0.01. Our auxiliary dataset is a subset of ImageNet22k [32] with 516K images, following [30, 40].

All experiments use SGD optimizer and decay the learning rate to zero using a cosine annealing learning rate scheduler [27]. All experiences are performed with 8 NVIDIA RTX 3090.

https://doi.org/10.52202/079017-4199

B The AdaptOD Algorithm

The full steps of the training and inference in AdaptOD are given in Algorithm 1 below.

Algorithm 1 : AdaptOD

Training
Input: Pre-trained model f
Data: Training dataset $\mathcal{D}_{in}^{train}$, Auxiliary dataset $\mathcal{D}_{out}^{train}$

1: **for** each iteration **do**
2: Sample a mini-batch of ID training data: $\{(x_i^{in}, y_i)\}_{i=1}^{n}$ from $\mathcal{D}_{in}^{train}$
3: Sample a mini-batch of OOD auxiliary data: $\{(x_i^{out})\}_{i=1}^{n}$ from $\mathcal{D}_{out}^{train}$
4: Perform batch energy normalization based on Eq. 6
5: Perform gradient descent on model f with $\mathcal{L}_{total}$ based on Eq. 11
6: **end for**

Inference
Input: Outlier distribution $\mathcal{P}^{out}$; Fine-tuned model f
Data: Test dataset $\mathcal{D}_{in\cup out}^{test}$

1: **for** each sample x in dataset $\mathcal{D}_{in\cup out}^{test}$ **do**
2: Adapt the outlier distribution $\mathcal{P}^{out}$ with sample x and model f based on Eq. 4
3: Obtain calibrated global energy score $\mathbb{G}^{\mathcal{P}}(x)$ for sample x as OOD score using the outlier distribution $\mathcal{P}^{out}$ based on Eq. 5
4: **end for**

C Discussion of Stable Margin Hyperparameters in DNE

C.1 Margin Hyperparameters in DNE-C

To identify OOD samples, we expect ID samples to have high energy, while OOD samples to have low energy. Formally, let $\mathbf{x}^{in} = \{x_1^{in}, x_2^{in}, ..., x_{b^{in}}^{in}\}, \mathbf{x}^{in} \in X^{in}$ be one training batch of ID data, with b^{in} be its batch size, and a corresponding batch of outlier data $\mathbf{x}^{out} = \{x_1^{out}, x_2^{out}, ..., x_{b^{out}}^{out}\}$ whose batch size is b^{out}, the sum of class-normalized energy $F_j(x)$ for all training samples in one batch (ID samples and outlier samples in one batch) in each class j would be one:

$$\mathbf{C}_j(\mathbf{x}^{in}) + \mathbf{C}_j(\mathbf{x}^{out}) = \sum_{i=1}^{b^{in}} F_j(x_i^{in}) + \sum_{i=1}^{b^{out}} F_j(x_i^{out}) = 1. \tag{13}$$

DNE-C class-wisely constrains the energy that optimizes the class energy of ID samples to be large for each class, while the class energy of outlier samples is small for each class. Therefore, the expected class-wise normalized energy for the batch of outlier samples $\mathbf{C}_j(\mathbf{x^{in}}), j \in \{1, 2, ..., k\}$ would be 1 on all classes, while the expected class-wise normalized energy for the batch of ID samples $\mathbf{C}_j(\mathbf{x^{out}}), j \in \{1, 2, ..., k\}$ would be zero on all classes:

$$\begin{cases} \mathbf{C}_j(\mathbf{x}^{in}) \to 1, \\ \mathbf{C}_j(\mathbf{x}^{out}) \to 0. \end{cases} \tag{14}$$

To this end, we set $m_{in}^c = 1$ and $m_{out}^c = 0$ for each class margin in Eq. 7, which optimizes the class-wise normalized energy for the batch of outlier samples $\mathbf{C}_j(\mathbf{x^{in}})$ on each class j towards one, while at the same time optimizing the class-wise normalized energy for the batch of ID samples $\mathbf{C}_j(\mathbf{x^{out}})$ on each class j towards zero. In this way, these energy margin hyperparameters do not rely on the training dataset and/or the imbalance factor.

C.2 Margin Hyperparameters in DNE-S

Similarly, the sum of class-normalized energy $F_j(x)$ for all training samples in one batch (ID samples and outlier samples in one batch) in each class j would be one. Furthermore, the sum of class-normalized energy $F_j(x)$ for all training samples in one batch over all classes would be k since there

are k categories in the ID data:

$$\sum_{i=1}^{b^{in}} \mathbf{S}(x_i^{in}) + \sum_{i=1}^{b^{out}} \mathbf{S}(x_i^{out}) = \sum_{i=1}^{b^{in}} \sum_{j=1}^{k} F_j(x_i^{in}) + \sum_{i=1}^{b^{out}} \sum_{j=1}^{k} F_j(x_i^{out}) = k. \tag{15}$$

DNE-S sample-wisely constrains the energy that optimizes the global energy of each ID sample to be large, while being well-balanced between head samples and tail samples. Therefore, the expected sample-wise normalized energy $\mathbf{S}(x), x \in \mathbf{x}^{in}$ would be $\frac{k}{b^{in}}$ for each ID sample, which is evenly divided the same scale to each ID sample (for either head samples or tail samples). And the expected sample-wise normalized energy $\mathbf{S}(x), x \in \mathbf{x}^{out}$ would be zero for each outlier sample:

$$\begin{cases} \mathbf{S}(x) \to \frac{k}{b^{in}}, x \in \mathbf{x}^{in}, \\ \mathbf{S}(x) \to 0, x \in \mathbf{x}^{out}. \end{cases} \tag{16}$$

Therefore, we set $m_{in}^s = \frac{k}{b^{in}}$ and $m_{out}^s = 0$, which optimizes the sum of class-normalized energy on all classes $\mathbf{S}(x)$ for each ID sample to $\frac{k}{b^{in}}$ and optimizes the sum of class-normalized energy on all classes $\mathbf{S}(x)$ for each outlier sample to zero. After doing this, we can regularize the global energy of the ID data, particularly the low global energy for tail samples, reducing the over-confident prediction of head samples. The same as the DNE-C loss, these specified margin parameters for training in DNE-S also do not rely on the training dataset and/or the imbalance factor.

D More Experimental Results

D.1 More Results for DODA

Table 7 presents the results of our proposed component DODA in enabling two popular baselines OE [12] and EnergyOE [24] on CIFAR10/100-LT on the six OOD test datasets. It shows that DODA can consistently enhance the OOD detection for both baselines in all four metrics across all six datasets, demonstrating the strong capability of DODA in reducing the learned outlier distribution gap to the true OOD. Nevertheless, these DODA-enabled baselines underperform AdaptOD, indicating that the other component of AdaptOD (*i.e.*, DNE) helps to produce a largely enhanced vanilla outlier distribution for DODA.

D.2 Differentiating OOD Data from Head and Tail Samples.

To evaluate the effectiveness in distinguishing OOD data from head and tail samples, we perform two particular inference settings: one with only tail samples and OOD samples, and another one with only head samples and OOD samples. Table 8 shows the averaged results over the six OOD test datasets on CIFAR10/100-LT of the baseline EnergyOE [24], previous SOTA model COCL [30], and our AdaptOD. It can be observed that AdaptOD does a better job than the two methods in both scenarios, resulting in significantly enhanced overall detection performance.

D.3 More Ablation Study

D.3.1 Imbalance Ratio

In the Experiments section, we use the default imbalance ratio $\rho = 100$ on both CIFAR10-LT and CIFAR100-LT. In this section, we show that our method can work well under different imbalance ratios. Table 9 shows the comparison of AdaptOD with two SOTA long-tailed OOD detection methods EnergyOE [24] and COCL [30] on CIFAR10-LT with $\rho = 50$ and $\rho = 10$. Our approach can significantly outperform these baselines in not only OOD detection performance but also ID classification accuracy by a considerable margin with different imbalance ratios. Furthermore, our approach AdaptOD performs better in more imbalanced datasets, indicating the superiority of AdaptOD for OOD detection in long-tail recognition.

Table 10 shows the comparison of our approach AdaptOD with two SOTA TTA methods AUTO and AdaODD for OOD detection on CIFAR10-LT with $\rho = 50$. To have a straightforward and extensive comparison, we compare DODA (the component of AdaptOD) with the two TTA methods, all of which are added on top of the same training method. In the experiments, we use three training

https://doi.org/10.52202/079017-4199

Table 7: Results of original and DODA-enabled OE and EnergyOE, and AdaptOD.

OOD Dataset	Method	ID Dataset: CIFAR10-LT				ID Dataset: CIFAR100-LT			
		AUC↑	AP-in↑	AP-out↑	FPR↓	AUC↑	AP-in↑	AP-out↑	FPR↓
Texture [5]	OE [12]	92.30	96.01	82.57	48.65	76.01	85.28	57.47	87.45
	OE [12]+**DODA(Ours)**	95.02	96.97	84.40	46.99	77.93	86.51	62.48	82.75
	EnergyOE [24]	95.53	97.42	92.93	18.44	79.56	86.03	70.88	79.45
	EnergyOE [24]+**DODA(Ours)**	97.32	97.88	93.65	16.28	80.85	87.34	75.52	77.00
	AdaptOD(Ours)	**98.22**	**98.81**	**94.91**	**11.60**	**83.88**	**89.43**	**76.47**	**58.47**
SVHN [31]	OE [12]	94.86	91.59	97.00	29.11	81.82	73.25	89.10	80.98
	OE [12]+**DODA(Ours)**	95.95	92.01	98.16	25.75	84.20	75.78	91.68	74.86
	EnergyOE [24]	96.63	92.33	98.46	14.37	86.19	81.42	91.74	34.36
	EnergyOE [24]+**DODA(Ours)**	97.24	92.88	98.73	12.86	90.26	88.13	95.30	21.73
	AdaptOD(Ours)	**98.13**	**94.34**	**99.11**	**10.33**	**93.09**	**91.32**	**96.86**	**17.63**
CIFAR [19]	OE [12]	83.32	84.06	80.83	65.82	62.60	66.16	57.77	93.53
	OE [12]+**DODA(Ours)**	85.52	86.03	83.15	60.99	66.02	72.11	62.03	90.85
	EnergyOE [24]	84.44	85.74	84.63	61.73	61.15	67.12	56.66	91.42
	EnergyOE [24]+**DODA(Ours)**	86.71	87.86	87.01	54.33	70.42	76.10	68.66	89.87
	AdaptOD(Ours)	**89.05**	**89.93**	**88.22**	**45.51**	**72.77**	**76.37**	**70.58**	**86.04**
TIN [20]	OE [12]	86.35	89.88	79.30	64.50	68.22	79.36	51.82	88.54
	OE [12]+**DODA(Ours)**	88.39	90.88	82.70	61.40	70.36	79.72	53.44	88.38
	EnergyOE [24]	88.40	91.65	84.95	46.23	70.78	79.40	55.90	90.74
	EnergyOE [24]+**DODA(Ours)**	89.93	92.46	86.12	44.02	71.25	79.80	57.91	89.42
	AdaptOD(Ours)	**91.40**	**93.85**	**88.18**	**42.77**	**72.87**	**82.06**	**58.92**	**88.24**
LSUN [51]	OE	91.57	93.06	88.37	53.99	76.81	85.33	60.94	83.79
	OE [12]+**DODA(Ours)**	93.09	93.42	91.30	48.39	77.83	86.24	62.10	82.44
	EnergyOE [24]	94.00	94.78	93.70	28.42	81.61	86.57	69.16	80.57
	EnergyOE [24]+**DODA(Ours)**	94.92	95.77	94.56	26.17	82.54	88.12	70.88	77.68
	AdaptOD(Ours)	**96.16**	**96.84**	**95.86**	**24.12**	**85.70**	**90.55**	**72.70**	**70.20**
Place365 [58]	OE [12]	90.20	82.09	95.24	57.06	75.68	60.99	86.51	83.55
	OE [12]+**DODA(Ours)**	91.74	83.99	96.64	52.57	76.39	62.48	87.52	82.72
	EnergyOE [24]	92.51	84.26	97.14	33.63	79.12	63.38	89.09	81.43
	EnergyOE [24]+**DODA(Ours)**	94.03	86.15	97.75	31.23	81.08	65.85	90.94	80.09
	AdaptOD(Ours)	**95.19**	**89.56**	**98.44**	**29.22**	**83.27**	**68.82**	**91.44**	**71.63**

Table 8: Comparison results on separating head/tail samples from OOD samples.

ID dataset	method	Head Samples				Tail Samples			
		AUC↑	AP-in↑	AP-out↑	FPR↓	AUC↑	AP-in↑	AP-out↑	FPR↓
CIFAR10-LT	EnergyOE [24]	95.88	89.67	98.31	23.06	83.45	61.07	93.37	58.61
	COCL [30]	96.34	93.34	98.67	19.59	91.91	76.98	97.15	34.30
	AdaptOD(Ours)	**98.20**	**96.80**	**99.00**	**11.20**	**93.40**	**80.58**	**98.27**	**30.70**
CIFAR100-LT	EnergyOE [24]	84.22	69.70	92.81	69.42	67.63	35.85	85.96	81.77
	COCL [30]	87.73	73.84	93.94	66.01	74.85	47.76	87.59	77.01
	AdaptOD(Ours)	**91.81**	**80.42**	**96.43**	**58.49**	**78.34**	**56.67**	**91.15**	**70.82**

Table 9: Comparison results of imbalance ratio among EnergyOE [24], COCL [30], and our approach AdaptOD on CIFAR10-LT.

Imbalance Ratio	Method	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
$\rho = 100$	EnergyOE [24]	91.92	91.03	91.97	33.80	74.57
	COCL [30]	93.28	92.24	92.89	30.88	81.56
	AdaptOD(Ours)	**94.69**	**93.89**	**94.12**	**27.26**	**82.27**
$\rho = 50$	EnergyOE [24]	93.48	92.68	93.05	29.74	81.23
	COCL [30]	94.30	93.85	93.31	26.98	84.89
	AdaptOD(Ours)	**95.14**	**94.53**	**94.66**	**24.43**	**85.77**
$\rho = 10$	EnergyOE [24]	95.03	94.34	94.83	25.26	88.47
	COCL [30]	95.71	95.12	95.33	20.91	89.65
	AdaptOD(Ours)	**96.34**	**95.72**	**95.86**	**18.33**	**90.24**

methods, including two SOTA long-tailed OOD detection methods and our proposed DNE-based training method. It is impressive that our DODA component consistently remains the best performer when the TTA methods are combined with all three different training methods on the CIFAR10-LT datasets.

D.3.2 Network Architectures

In the Experiments section, we use the standard ResNet18 as the backbone model on both CIFAR10-LT and CIFAR100-LT. To show the generality of our method, we also perform a long-tailed OOD detection experiment using both ResNet34 and ResNet18. The results are shown in Table 11 and

Table 10: Comparison to different TTA-based OOD detection methods on CIFAR10-LT with $\rho = 50$.

Training	Test	AUC↑	AP-in↑	AP-out↑	FPR↓
EnergyOE [24]	w/o TTA	93.48±0.25	92.68±0.33	93.05±0.30	29.74±0.22
	AUTO [49]	93.85±0.29	92.84±0.25	93.34±0.9	29.10±0.35
	AdaODD [56]	94.14±0.33	92.92±0.32	93.60±0.33	29.01±0.20
	DODA(Ours)	94.60±0.28	93.46±0.36	93.91±0.26	28.42±0.24
COCL [30]	w/o TTA	94.30±0.25	93.85±0.25	93.31±0.44	26.98±0.28
	AUTO [49]	94.62±0.31	93.91±0.33	93.52±0.40	26.49±0.37
	AdaODD [56]	94.41±0.29	93.84±0.31	93.35±0.46	26.67±0.35
	DODA(Ours)	94.82±0.24	94.13±0.29	94.21±0.36	26.02±0.30
DNE (Ours)	w/o TTA	93.85±0.38	93.43±0.28	93.62±0.39	27.47±0.38
	AUTO [49]	94.59±0.43	93.68±0.33	93.98±0.38	26.50±0.35
	AdaODD [56]	94.75±0.42	93.78±0.32	94.22±0.44	25.64±0.37
	DODA(Ours)	**95.14±0.41**	**94.53±0.27**	**94.66±0.36**	**24.43±0.32**

Table 11: Comparison results of model structure among EnergyOE [24], COCL [30], and our approach AdaptOD on CIFAR10-LT.

Model	Method	AUC↑	AP-in↑	AP-out↑	FPR↓	ACC↑
ResNet18	EnergyOE [24]	91.92	91.03	91.97	33.80	74.57
	COCL [30]	93.28	92.24	92.89	30.88	81.56
	AdaptOD(Ours)	**94.69**	**93.89**	**94.12**	**27.26**	**82.27**
ResNet34	EnergyOE [24]	92.25	91.37	92.31	32.44	74.89
	COCL [30]	93.52	92.93	92.83	30.74	81.75
	AdaptOD(Ours)	**94.98**	**94.33**	**94.52**	**26.61**	**83.47**

Table 12: Comparison to different TTA-based OOD detection methods on CIFAR10-LT using ResNet34. The results are averaged over the six OOD test datasets in the SC-OOD benchmark.

Training	Test	AUC↑	AP-in↑	AP-out↑	FPR↓
EnergyOE [24]	w/o TTA	92.25±0.32	91.37±0.31	92.31±0.28	32.44±0.37
	AUTO [49]	92.98±0.40	91.93±0.26	92.60±0.38	32.03±0.38
	AdaODD [56]	93.16±0.36	92.10±0.50	92.90±0.46	31.87±0.47
	DODA(Ours)	93.81±0.32	92.68±0.28	93.27±0.39	29.65±0.36
COCL [30]	w/o TTA	93.52±0.36	92.93±0.48	92.83±0.27	30.74±0.38
	AUTO [49]	93.73±0.45	93.04±0.40	93.26±0.31	29.60±0.40
	AdaODD [56]	93.90±0.47	93.19±0.35	93.46±0.44	29.31±0.41
	DODA(Ours)	94.27±0.39	93.57±0.38	93.82±0.36	28.78±0.34
DNE (Ours)	w/o TTA	93.28±0.30	92.64±0.25	92.95±0.32	31.18±0.33
	AUTO [49]	93.77±0.33	92.83±0.50	93.36±0.45	29.99±0.47
	AdaODD [56]	93.84±0.42	93.04±0.39	93.61±0.37	29.53±0.42
	DODA(Ours)	**94.98±0.35**	**94.33±0.33**	**94.52±0.40**	**26.61±0.37**

Table 12. Table 11 shows the comparison of AdaptOD with two SOTA long-tailed OOD detection methods EnergyOE [24] and COCL [30] on CIFAR10-LT using different backbone models. Table 12 shows the comparison of AdaptOD with two SOTA TTA methods AUTO and AdaODD for OOD detection on CIFAR10-LT using ResNet34. AdaptOD maintains its superiority with different network architectures.

D.3.3 Sensitivity

In the Approach section, we utilize a Z-score-based method based on training ID data to implement the OOD filter, which predicts true OOD samples for adapting outlier distribution in DODA. The threshold R in the OOD filter is calculated with only training ID data and can be directly used during inference. α is a hyperparameter to adjust the threshold R. A too high value of R can misclassify a large number of ID samples as OOD samples. On the other hand, a too low value of R will filter out too many true OOD samples. In both cases, the outlier distribution adaptation becomes ineffective. Fig. 4 shows the sensitivity of AdaptOD with respect to α in Eq. 3, showing that the performance of AdaptOD is relatively stable with a relatively large range of α values, *e.g.*, $[2.5, 3.5]$.

D.3.4 Computational Overhead

AdaptOD performs normalization on the logit output of each class for each batch of training samples before energy training. Fig. 13 shows the computational overhead of AdaptOD compared to the baseline EnergyOE and the previous SOTA method BERL using the same backbone on CIFAR100-LT

https://doi.org/10.52202/079017-4199

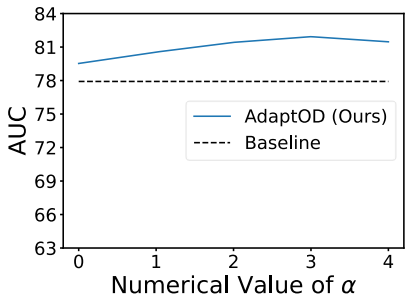


Figure 4: Average performance of AdaptOD w.r.t. α over six OOD datasets on CIFAR100-LT.

with a single NVIDIA RTX 3090, in which they are all fine-tuning-based and energy-based methods. It shows that the training speed of AdaptOD is similar with the previous methods on both ResNet18 backbone and ResNet34 backbones.

Table 13: Comparison results of training time (seconds) on CIFAR100-LT.

Model	Training Time (seconds)		
	EnergyOE [24]	BERL [4]	AdaptOD(Ours)
ResNet18	8.56	9.12	8.84
ResNet34	12.65	13.12	12.89

Table 14: Comparison results on synthetic OOD datasets with CIFAR10-LT.

Dataset	Method	AUC↑	AP-in↑	AP-out↑	FPR↓
Gaussian	EnergyOE [24]	99.74	99.76	99.33	1.96
	BERL [4]	99.76	99.34	99.16	0.49
	COCL [30]	99.68	99.79	99.39	**0.02**
	AdaptOD(Ours)	**99.83**	**99.87**	**99.58**	0.08
Rademacher	EnergyOE [24]	99.13	99.25	97.16	2.32
	BERL [4]	99.00	99.06	96.26	1.42
	COCL [30]	99.76	**99.84**	99.56	**0.01**
	AdaptOD(Ours)	**99.78**	99.65	**99.61**	0.04
Blobs	EnergyOE [24]	90.16	93.25	85.39	9.44
	BERL [4]	93.18	96.87	89.34	6.54
	COCL [30]	98.75	99.17	97.49	1.04
	AdaptOD(Ours)	**99.12**	**99.43**	**98.62**	**0.53**
Average	EnergyOE [24]	96.34	97.42	93.96	4.57
	BERL [4]	97.32	98.42	94.92	2.81
	COCL [30]	99.40	99.60	98.81	0.35
	AdaptOD(Ours)	**99.58**	**99.65**	**99.27**	**0.22**

D.4 Experiment Results on Synthetic OOD Datasets

To demonstrate the superiority of our approach AdaptOD on diverse OOD datasets, we also evaluate our approach AdaptOD with three synthetic OOD datasets on CIFAR10-LT, including Gaussian, Rademacher, and Blobs. Specifically, *Gaussian* noises have each dimension sampled from an isotropic Gaussian distribution. *Rademacher* noises are images where each dimension is -1 or 1 with equal probability, so each dimension is sampled from a symmetric Rademacher distribution. *Blobs* noises consist of algorithmically generated amorphous shapes with definite edges. We use three SOTA methods for comparison, including EnergyOE [24], BERL [4], and COCL [30]. As in Table 14, our approach AdaptOD achieves similarly significant improvement over these methods on these synthetic OOD datasets as on the other OOD datasets.

E Limitation and Broader Impacts

E.1 Limitation

While AdaptOD offers a straightforward and competitive solution for out-of-distribution detection in long-tailed recognition, it necessitates the incorporation of additional outlier data to learn an enhanced vanilla outlier distribution, increasing the difficulty of applying it to real-world scenarios. The DODA component in AdaptOD is an attempt that utilizes the detected OOD samples during the inference stage to improve OOD detection performance. This requires online updating of the learned outlier distribution. An alternative way is to optimize the outlier distribution and reduce its gap to the OOD distribution during training, eliminating the online updating step. The lack of true OOD data remains as a major challenge in such approaches. We leave it for future work.

E.2 Broader Impacts

OOD detection is a branch of anomaly detection that typically plays a positive role in enhancing model security in various safety-critical applications such as autonomous driving. When applying our methods, we need to ensure that the models are only used for the purpose of enhancing the safety of deep learning models in real-world environments and do not infringe on human privacy.

https://doi.org/10.52202/079017-4199

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed in Appendix E.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There is no theoretical result in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our paper fully discloses all the information needed to reproduce the main experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-4199

Answer: [Yes]

Justification: Our paper provides open access to the data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our paper specifies all the training and testing details

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our paper shows average results over six runs with different random seeds and report the variance for our metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our paper provides sufficient information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The societal impacts of the work are discussed in Appendix E.2.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-4199

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Our paper gives proper acknowledgement of the original papers that produced the code packages or datasets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-4199