

# Revisiting Energy-Based Model for Out-of-Distribution Detection

Yifan Wu, *Student Member, IEEE*, Xichen Ye, Songmin Dai, Dengye Pan, Xiaoqiang Li, *Member, IEEE*, Weizhong Zhang, and Yifan Chen, *Member, IEEE*

**Abstract**—Out-of-distribution (OOD) detection is an essential approach to robustifying deep learning models, enabling them to identify inputs that fall outside of their trained distribution. Existing OOD detection methods usually depend on crafted data, such as specific outlier datasets or elaborate data augmentations; this characteristic is reasonable while the frequent mismatch between crafted data and OOD data limits model robustness and generalizability. In response to this issue, we introduce Outlier Exposure by Simple Transformations (OEST), a framework that enhances OOD detection by leveraging “peripheral-distribution” (PD) data; specifically, PD data are samples generated through *simple data transformations*, thus an efficient alternative to manually curated outliers.

We further adopt the energy-based models (EBMs) to study PD data. We first recognize the “energy barrier” in OOD detection which characterizes the energy difference between in-distribution (ID) / OOD samples and eases the detection; the in-between PD data are introduced to establish the energy barrier in training. Furthermore, this energy barrier concept motivates a theoretically grounded energy-barrier loss to replace the classical energy-bounded loss, which leads to an improved paradigm, OEST\*, and brings a more effective and theoretically sound separation between ID and OOD samples. We perform empirical validation to provide sanity checks of our proposal, and extensive experiments across various benchmarks demonstrate that OEST\* achieves better or similar accuracy compared with state-of-the-art methods. The source code of our method is available at: <https://github.com/victor-yifanwu/Outlier-Exposure-by-Simple-Transformations>.

**Index Terms**—Out-of-distribution detection, Outlier exposure, Energy-based models, Data augmentation.

## I. INTRODUCTION

THE predominant assumption in model training is that test data are drawn independently and identically distributed (*i.i.d.*) from the same distribution as the training data. Such distribution alignment is generally known as in-distribution (ID). Although the ID assumption leads to simple formulation, it rarely holds in *open-world* scenarios as distribution shifts

Yifan Wu and Xichen Ye are with the School of Computer Engineering and Science, Shanghai University, Shanghai, China, and also with the School of Computer Science, Fudan University, Shanghai, China (e-mail: victorwu001219@gmail.com; yexichen0930@outlook.com).

Songmin Dai, Dengye Pan, and Xiaoqiang Li are with the School of Computer Engineering and Science, Shanghai University, Shanghai, China (e-mail: laodar@shu.edu.cn; pandy@shu.edu.cn; xqli@shu.edu.cn).

Weizhong Zhang is with the School of Data Science, Fudan University, Shanghai, China (e-mail: weizhongzhang@fudan.edu.cn).

Yifan Chen is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China (e-mail: yifanc@hkbu.edu.hk).

Yifan Wu and Xichen Ye contributed equally to this work.

Corresponding authors: Xiaoqiang Li and Yifan Chen.

inevitably exist between training and testing data. This discrepancy poses significant challenges to a few existing models [1]–[4]. It is essential to recognize these deviations as outliers, namely out-of-distribution (OOD) samples [5]–[13], instead of blindly categorizing unseen samples into known classes with high confidence [14], [15]. Due to its broad application scenarios (*e.g.*, autonomous vehicles [16] and medical tasks [17], [18]), a number of methodologies for out-of-distribution detection have been developed [19], [20].

An intuitive approach to alleviate the misclassification of unknown samples as known categories, is to incorporate a significant amount of auxiliary external data in training [21]–[30]; the utilization of real outliers generally brings about better performance. However, naturally the similarity between the crafted and the real OOD samples makes a significant difference, as revealed in recent studies [31], [32]. This turns into an issue particularly in specialized fields such as medical or industrial imaging, where data characteristics greatly differ from those found in public vision datasets and high-quality external datasets are scarce [33]. Even worse, another challenge arises in practice that the external auxiliary OOD dataset tends to contain quite a few ID samples, necessitating either elaborate algorithms or tremendous human labor to filter these samples. These complications deteriorate the effectiveness of OOD detection.

In response to these issues, we manage to eliminate the reliance on real outliers in OOD detection, through exploiting *simple transformations* over ID samples; we further recognize and term the transformed samples as *peripheral-distribution* (PD) data. We consider this PD samples as neither ID nor regularly OOD. This ambiguity arises from the versatility of the collection of simple transformations: samples augmented with certain transformations, *e.g.*, rotation [34], remain semantically ID, whereas those augmented with sobel filtering [35] are distinct from ID samples. Experimental observation in Figure 1a further confirms PD data is an interpolation between ID and OOD samples.

In addition, we revisit energy-based OOD detection [23, EBO] and connect the aforementioned concept of PD data to the energy-based model [36, EBM]. We aim to encourage higher energy for samples from peripheral-distribution while lowering that of ID samples; we therefore create an “energy barrier” between the ID and the PD data (and thus the OOD data). The energy barrier we propose is formulated in Assumption 1, and we verify its statistical benefits in Theorem 1.

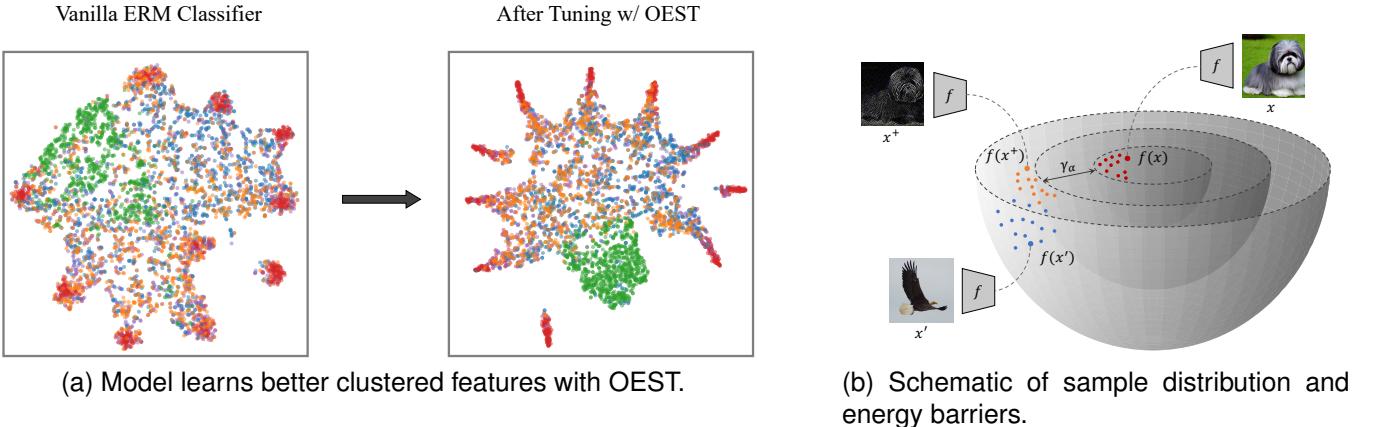


Fig. 1. (a) The t-SNE visualization of representations from CIFAR-10 (Red) test samples, rotated CIFAR-10 (Orange) test samples, CIFAR-100 (Blue), SVHN (Green) and ImageNet (Purple) before and after applying our training strategy OEST. Specifically, the embedding features are extracted from the penultimate layer of ResNet-18 classifier (trained on CIFAR-10).

(b) We illustrate the feature space as a series of concentric spherical shells, where each shell corresponds to a certain energy level. The innermost shell contains in-distribution samples with the lowest energy, represented by  $x$ . Moving outward, the orange points indicate augmented peripheral-distribution samples, denoted by  $x^+$ . OEST establishes an energy barrier (reading  $\gamma_\alpha$ ) between ID ( $x$ ) and PD ( $x^+$ ) data, thus separating ID and out-of-distribution samples ( $x'$ ).

Combining all the pieces above, we propose a new training paradigm, Outlier Exposure by Simple Transformations (**OEST**), which is illustrated in Figure 1b. This approach features a comprehensive use of numerous data augmentation techniques, including those previously deemed non-contributory in [37]. Remarkably, OEST achieves outstanding performance in both near-OOD and far-OOD detection tasks, with solely an extra 10-epoch tuning. The main contributions of this paper are summarized as follows: `leftmargin=*`

- We introduce peripheral-distribution (PD) data for OOD detection, which consists of samples augmented through various simple transformations.
- We revisit the energy-based model, and suggest the energy polarization of ID and OOD samples can benefit OOD detection.
- We propose to establish an energy barrier between ID and PD data, which consequently and provably enhance the distinction between ID and OOD samples.
- We devise a targeted tuning strategy for existing classifiers built upon the establishment of energy barrier, which achieves state-of-the-art results with a large margin under both near- and far-OOD scenarios.

The preliminary 4-page version of this manuscript was presented in ICIP 2023 [38], where we solely suggested applying simple transformations for OOD detection. In this extended paper, we provide a more comprehensive investigation into the proposed methodology; we newly recognize the energy barrier between ID and PD data (in Section IV-B) to provably explain the empirical success of OEST, develop a theoretically rigorous energy loss function (in Section IV-C), and broaden the experimental evaluation across public benchmarks (in Section V).

The rest of the paper is organized as follows. In Section II, we review related works on OOD detection. In Section III, we detail the preliminaries of this work, including its theoretical basis and a short introduction to the energy-based model. In

Section IV, we illustrate the proposed method in detail, along with its theoretical analysis. Experimental results and analyses are provided in Section V. Finally, we conclude the paper in Section VI.

## II. RELATED WORKS

In this section, we review prior works on out-of-distribution (OOD) detection, involving three primary approaches: ① OOD scoring methods (see Section II-A), ② training-based methods (see Section II-B), and ③ methods with outlier exposure (see Section II-C).

### A. OOD Scoring Methods

In general, OOD scoring methods assess the likelihood that a sample originates from the training distribution, *i.e.*, is in-distribution, based on sample features or model outputs.

From a feature perspective, early studies employed parametric density estimation, assuming the feature embedding space consists of a mixture of multivariate Gaussian distributions, to score samples based on the Mahalanobis distance [39] or the gram matrix [40]. A more recent approach [41] utilized the distance between the sample feature and its  $k$ -th nearest neighbor (KNN) as the score; likewise, SHE [42] leveraged the similarity between the sample feature and class centers for OOD detection.

For model outputs, one common OOD score is the maximum softmax prediction (MSP) [6]. Subsequently, ODIN [43] was proposed to utilize temperature scaling and input perturbation to maximize the MSP gap between ID and OOD data. Follow-up studies revealed that the key to ODIN's effectiveness is transforming the softmax score back to the logit space through temperature scaling; therefore, methods like the maximum logit scores [44] and the standardized max logits [16] were developed.

However, raw softmax or logit scores are found prone to overconfidence issues, which prompts the development of the

energy-based models (EBM) [23], [45]. EBM employs an energy-based function to transform logits into a more reliable scoring metric and is theoretically underpinned via a likelihood perspective [46], [47]. Recent studies [48]–[52] began to focus not only on the last or penultimate layers of the model but also on the hidden layers, as well as the activations among them.

All the aforementioned OOD scoring methods (including ours) can be classified as *post-hoc methods* because the OOD scores within, derived from feature or model outputs, can be implemented without modifying the training procedure or objective. These approaches avoid both the overhead cost of retraining and any detrimental impact on the ID accuracy of the original classifier.

### B. Training-based Methods

Another genre of OOD detection intervenes in the model training. We note, although methods with outlier exposure also fall under training-based approaches, we defer the related discussion to the next subsection and focus exclusively on the methods that do not utilize outlier samples in this subsection.

Training-based approaches involve the following methodologies. ① From the perspective of confidence estimation, [53] proposed to modify the model structure, while [54] designed a new Softmax layer. ② Regarding modifying training objectives, G-ODIN [55] introduced a specialized objective called DeConf-C based on ODIN [43], and [56] advocated for training with logit normalization (LogitNorm), a straightforward modification to the standard cross-entropy loss aimed at mitigating overconfidence. ③ From the perspective of enhancing model representation, some studies have employed adversarial training [15], [25], [57]–[59] or stronger data augmentation [60]–[64] to enrich ID samples. ④ Additionally, self-supervised methods have been utilized to improve classifier robustness in OOD detection. [65], [66] introduced an additional training objective, image transformation prediction, during model training. ⑤ Moreover, to enhance sensitivity to covariate shifts, [37] treats original and augmented samples as positive and negative examples, respectively. We remark this approach diverges from the traditional contrastive learning framework [67] by actively separating positive and negative samples in the feature space.

### C. Methods with Outlier Exposure

As a broadly studied technique, outlier exposure in OOD detection can be divided into two main categories based on the source of outliers. ① The first category utilizes a collected set of real-world OOD samples (*real outliers*) to aid models in learning the discrepancy between ID and OOD, while ② the second category focuses on *generating outlier data* to enhance model robustness against various unforeseen OOD samples.

① For the methodology based on real outliers, the initial approach was proposed by [21], which encourages high-entropic predictions on given outlier samples. Subsequently, MCD [22] employed a dual-branch network to distinguish between ID and OOD data, and a follow-up work [23] further tuned the classifier on both ID and OOD samples with energy-based loss (see more discussion on Section III-B).

TABLE I  
DETAILED DESCRIPTION OF THE MAIN NOTATIONS

Notation	Definition
$\mathcal{D}_{ID}$	joint distribution of ID data $(x, y)$
$\mathcal{D}_{OOD}$	joint distribution of OOD data $(x', y')$
$\mathcal{X}_{ID}$	support space of ID inputs
$\mathcal{X}_{OOD}$	support space of OOD inputs
$\mathcal{X}_{PD}$	support space of PD inputs
$\mathcal{Y}_{ID}$	support space of ID labels
$f_\theta$	the neural classifier with parameters $\theta$
$\mathcal{L}_{CE}$	cross-entropy loss function
$\mathcal{L}_{energy}$	energy-based loss function

Other straightforward methods [24]–[26] treat the given OOD samples as the  $(k+1)^{\text{th}}$ -class. Recent studies [25], [27]–[30] started to focus on a selected set of meaningful outliers among numerous OOD samples.

② For outlier generation, earlier studies tended to generate data based on low-dimensional feature spaces; specifically, they utilized KL divergence [68], low-density regions [69], high-confidence regions [70], or meta-learning [71]. Subsequent work instead proposes a new paradigm for generating outliers that can be implemented using both GANs and diffusion models [72], [73]. Additionally, considering the challenges of image generation in the high-dimensional pixel space, recent approaches, such as VOS [74] and NPOS [75], have proposed to generate outlier data by injecting perturbations into ID sample features.

In summary, OOD detection leveraging real outliers can achieve superior performance. However, the effectiveness of these methods can be significantly influenced by the correlations between the provided and the actual OOD samples [76].

## III. PRELIMINARIES

In this section, we provide a formulation of out-of-distribution (OOD) detection in Section III-A, followed by a revisit of the energy-based model (EBM) for OOD detection in Section III-B. For the reader's convenience, we list a collection of defined notations in Table I.

### A. Out-of-Distribution Detection

The emergence of out-of-distribution (OOD) detection was driven by the practical need for models to discern and reject inputs that are semantically different from the training distribution. To discuss this concept within a rigorous framework, we embrace the widely acknowledged definition of in-distribution data and out-of-distribution data, as outlined in the previous literature [6], [19], [20].

In this paper, we consider a typical  $C$ -class classification problem, where we have access to independently and identically distributed (*i.i.d.*) samples  $(x, y)$  drawn from the ID distribution,  $\mathcal{D}_{ID}$ . Specifically, we denote the input space as  $\mathcal{X}_{ID}$ , and the label space as  $\mathcal{Y}_{ID} = [C]$ . In contrast, out-of-distribution (OOD) samples are drawn from a different distribution,  $\mathcal{D}_{OOD}$ . For each OOD sample  $(x', y')$ , the input  $x'$  has semantics that differ from those of any ID samples, and notably the label  $y'$  does not belong to any of the  $C$  classes

present in the training dataset, *i.e.*,  $y' \notin \mathcal{Y}_{\text{ID}}$ . Typically, we denote the set of OOD inputs as  $\mathcal{X}_{\text{OOD}}$ .

The goal of OOD detection is to design a score function:

$$s_{\theta}(\mathbf{x}) \in \mathcal{R}, \quad (1)$$

where  $\theta$  is the learnable parameters. The desired outcome is that in-distribution samples receive higher scores than out-of-distribution samples, and consequently an OOD discriminator can be straightforwardly defined using this score function.

As a side note, one might consider modeling  $p(\mathbf{x})$  using a generative model, and intuitively take  $p(\cdot)$  as a score function, given the strong modeling capabilities in modern generative modeling. However, previous research has shown that the density functions estimated by deep generative models cannot be reliably used for OOD detection [77].

Moreover, in this work we consider the scenario in which we only have access to a trained classifier, and our objective is to repurpose it as an OOD discriminator. To address this practical constraint, energy-based models (EBMs) [36] offer an alternative approach to constructing a score function to distinguish OOD samples with a classifier. We will shortly detail EBM in the next subsection.

### B. Energy-Based Model

As an OOD scoring method (see Section II-A), the essence of the energy-based model is to construct an energy function  $E(\cdot, \cdot)$  that maps each sample  $(\mathbf{x}, y)$  to a scalar,  $E(\mathbf{x}, y)$ , known as the *energy*. Energy values can be converted into a probability density  $p(\mathbf{x}, y)$  in the form of *Gibbs distribution* ( $T$  is the temperature parameter):

$$p(\mathbf{x}, y) = \frac{\exp(-E(\mathbf{x}, y)/T)}{Z}, \quad (2)$$

where  $Z = \int_{\mathbf{x}} \sum_{i=1}^C \exp(-E(\mathbf{x}, i)/T)$  is the normalizing constant (also known as the partition function). The probability density  $p(\mathbf{x})$  can then be computed as:

$$p(\mathbf{x}) = \sum_y p(\mathbf{x}, y) = \frac{\sum_y \exp(-E(\mathbf{x}, y)/T)}{Z}. \quad (3)$$

The normalization constant  $Z$  is usually intractable to compute or reliably estimate over the input space. To address this, standard approaches in log-concave sampling is to take the negative logarithm of both sides in Eq. (3) [78], giving:

$$-\log p(\mathbf{x}) = -\log \sum_y \exp(-E(\mathbf{x}, y)/T) + \log Z. \quad (4)$$

The equation above indicates that omitting the term  $Z$  does not affect OOD detection, as  $\log Z$  is constant to each sample. Consequently, we can define the *Helmholtz free energy*  $E(\mathbf{x})$  (here we reload the notation  $E(\cdot)$  for convenience) as a surrogate of  $-\log p(\mathbf{x})$ :

$$E(\mathbf{x}) = -T \log \sum_y \exp(-E(\mathbf{x}, y)/T). \quad (5)$$

Now, let us consider a *fixed* classifier,  $f_{\theta} : \mathbb{R}^D \mapsto \mathbb{R}^C$ , with the model parameters  $\theta$  that have already been trained on  $\mathcal{D}_{\text{ID}}$ . A straightforward approach to construct  $E(\cdot, \cdot)$  is to define the

parameterized energy function as  $E(\mathbf{x}, y; f_{\theta}) = -f_{\theta}^{(y)}(\mathbf{x})/T$ , where  $f_{\theta}^{(i)}(\mathbf{x})$  represents the  $i$ -th output of  $f_{\theta}(\mathbf{x})$ . The parameterized Helmholtz free energy,  $E(\mathbf{x}; f_{\theta})$ , then becomes:

$$E(\mathbf{x}; f_{\theta}) = -T \log \sum_{i=1}^C \exp(f_{\theta}^{(i)}(\mathbf{x})/T). \quad (6)$$

Here, we simply set  $T = 1$  for computational convenience in the following sections. To this end, the score function is defined as  $s_{\theta}(\mathbf{x}) = -E(\mathbf{x}; f_{\theta})$ . The OOD discriminator  $D(\mathbf{x}; \tau, f_{\theta})$  is then given by:

$$D(\mathbf{x}; \tau, f_{\theta}) = \begin{cases} 0 & \text{if } -E(\mathbf{x}; f_{\theta}) \leq \tau, \\ 1 & \text{if } -E(\mathbf{x}; f_{\theta}) > \tau, \end{cases} \quad (7)$$

where  $\tau$  is a threshold. Samples with higher energy (lower score) values are considered as OOD inputs and vice versa.

As a closing remark, we note it is also feasible to further tune the classifier  $f_{\theta}(\cdot)$  for better OOD detection performance [23]. We recall Eq. (6) can serve as a surrogate for the negative log-likelihood of  $p(\mathbf{x})$  for *fixed*  $E(\mathbf{x}, y)$ , and following the spirit of MLE (maximum likelihood estimation) in score-based methods, users can intuitively turn to minimize the energy for better modeling the data. We will revisit this tuning strategy in Section IV-C (see the paragraph “Issues of  $\mathcal{L}_{\text{energy}}$ ”).

## IV. METHODOLOGY

In this section, we first introduce the novel concept of peripheral-distribution samples in Section IV-A. Following that, we present a fresh understanding of Energy-Based Models (EBMs) through a new concept “energy barrier” proposed in Section IV-B. This barrier effectively separates in-distribution and out-of-distribution samples. Finally, in Section IV-C, we propose a straightforward tuning strategy that leverages PD samples to improve the robustness of existing classifiers.

### A. Peripheral-Distribution Samples

To effectively dissect in- and out-of-distribution data, or even near-distribution data [37], this paper introduces a new concept, peripheral-distribution (PD) samples. This concept arises from the lack of real outliers (OOD data) in training; due to this lack, augmented samples from a specially defined distribution are usually taken as proxies for such outliers.

It remains an open problem how those augmented samples are related to ID data. [67] recognized augmented samples as positive, while [37] discovered some augmentations (*e.g.*, rotation) are beneficial when they are treated as negative. Here, as shown in Figure 1a, certain augmented samples are found to be peripheral to the ID data. Motivated by the observation, we conceptually refer those augmented samples to peripheral-distribution data, an interpolation in the feature space to connect in-distribution and out-of-distribution samples.

From a practical standpoint, PD data can be generated by applying specific data augmentation transformations to ID samples. This view is inspired by the principle of contrastive learning [67], which suggests that transformed data tends to

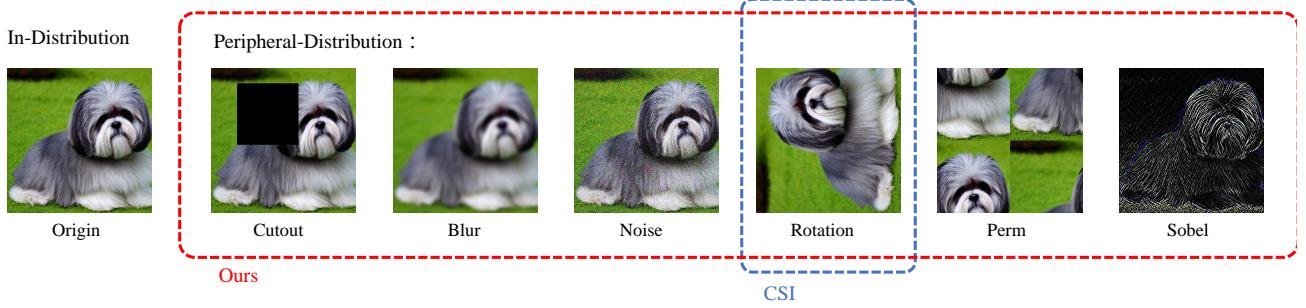


Fig. 2. Visualization of the original image and the considered simple transformations. The difference between our design and a baseline method CSI [37] is also exhibited. CSI must select different suitable transformations elaborately for each particular scenario, and specifically for CIFAR-10 CSI chooses rotation. However, our method utilizes all kinds of transformations.

remain close to the original data in the feature space while still exhibiting a shift in feature distribution. As demonstrated in Figure 1a, the representations of the augmented data indeed are located in the hypothesized interpolation zone between in- and out-of-distribution samples.

We give the formulation of PD data as follows. Consider a set  $\mathcal{S}$  comprising different transformations, which can be either random or deterministic. For a given batch of ID samples  $\mathcal{B} = \{\mathbf{x}_i\}_{i=1}^B, \forall \mathbf{x}_i \in \mathcal{X}_{ID}$ , we can generate peripheral-distribution samples  $\mathcal{B}^+$  by augmenting  $\mathcal{B}$  with transformations from the predefined collection  $\mathcal{S}$  (which is fixed and thus we omit the dependence on it in the notation of  $\mathcal{B}^+$ ). We thereby define the peripheral-distribution samples as

$$\mathcal{B}^+ = \bigcup_{S \in \mathcal{S}} \{\mathcal{B}_S\}, \text{ where } \mathcal{B}_S := \{S(\mathbf{x}_i)\}_{i=1}^B. \quad (8)$$

Later, for the given  $\mathcal{X}_{ID}$ , we denote the support of peripheral-distribution data as  $\mathcal{X}_{PD}$ , with  $\mathcal{B}^+ \subset \mathcal{X}_{PD}$ .

### B. Energy Barrier Assumption

In this section, we demonstrate that each peripheral-distribution sample can provide an ‘‘energy barrier’’ (defined in Assumption 1), which benefits the classifier tuning. Similar to [79], [80], we make the following assumption regarding the representations of peripheral-distribution data and show how the energy barrier can indirectly expose the differences between in- and out-of-distribution samples.

Specifically, we consider a linear classifier (adopted in the most common softmax classifier):

$$f(\mathbf{x}) := \mathbf{C}\mathbf{x},$$

where  $\mathbf{C} \in \mathbb{R}^D \times \mathbb{R}^C$  maps a sample  $\mathbf{x}$  from input space  $\mathbb{R}^D$  to  $C$  values, a.k.a. logits. We denote the  $i$ -th row in  $\mathbf{C}$  as  $\mathbf{c}_i$ , a length- $D$  vector indicating the corresponding class; we sometimes call  $\mathbf{c}_i$  the ‘‘class representation’’ of class  $i$ . Under the image classification setting, the linear classifier  $f(\mathbf{x}) : \mathbb{R}^D \mapsto \mathbb{R}^C$  is usually the last layer of a neural network.

**Assumption 1** (Energy Barrier Assumption on Peripheral-Distribution Samples). *With the classifier  $f(\cdot)$  and an out-of-distribution instance  $\mathbf{x}'$ , we assume all the representations, including the class representations  $\mathbf{c}_i$ 's, lie in a bounded domain with radius  $B > 0$ . Moreover, for a random ID*

*sample  $\mathbf{x}$  and a certain probability level  $\alpha \in (0, 1)$ , there exists a certain augmented sample  $\mathbf{x}^+$  such that*

$$E(\mathbf{x}^+; f) - E(\mathbf{x}; f) > B\|\mathbf{x}' - \mathbf{x}^+\| + \gamma_\alpha \quad (9)$$

*will hold with probability  $1 - \alpha$ , where  $\gamma_\alpha \geq 0$  is a constant.*

**Remark.** In addition to the usual compact domain assumption, we require there exists a large enough energy barrier ( $\gamma_\alpha$  in Eq. (9)) between the original samples and one peripheral augmented sample  $\mathbf{x}^+$ ; the high probability inequality also implies there is supposed to be one augmented sample  $\mathbf{x}^+$  closer to the out-of-distribution sample  $\mathbf{x}'$  than to most ID samples (otherwise the inequality will be invalid if  $\|\mathbf{x}' - \mathbf{x}^+\|$  is overly large), which is heavily utilized in contrastive learning theory [79], [80]<sup>1</sup>. In this regard, peripheral augmented samples can help differentiate confusing OOD images close to ID samples.

Through lifting the energy barrier  $E(\mathbf{x}^+; f) - E(\mathbf{x}; f)$  as in Assumption 1, we can construct a gap between OOD and ID samples; the finding is formulated as follows.

**Theorem 1.** *When Assumption 1 holds, we then have*

$$E(\mathbf{x}'; f) - E(\mathbf{x}; f) > \gamma_\alpha$$

*holds with probability  $1 - \alpha$ . The OOD sample  $\mathbf{x}'$  will be guaranteed to have higher energy than a random ID sample  $\mathbf{x}$  with high probability.*

The proof is deferred to Appendix A.

**Remark.** To close this subsection, we remark the validity of Theorem 1 heavily depends on the energy barrier assumption Assumption 1, which may not necessarily hold for the trained classifier  $f(\cdot)$ . However, the theoretical result motivates our following empirical design, which aims to establish an energy barrier between the original samples and the augmented ones.

### C. Establishing the Energy Barrier via PD Samples

There are two requirements implied by Assumption 1, that ① the augmented samples constitute a qualified semantics interpolation between ID and OOD samples and ② there is an energy barrier between the aforementioned augmented samples

<sup>1</sup>For example, in scenarios where vehicle images are concerned, the certain augmentation cropping, which transforms an image of a vehicle to merely a tire, obviously moves its semantic boundary toward the out-of-distribution category.

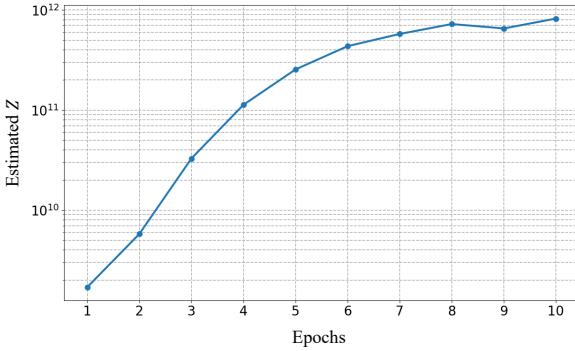


Fig. 3. Visualization of  $Z$  during the tuning process of OEST on CIFAR-10. Here,  $Z$  is computed as the empirical aggregation of all images used in [20].

and the ID data. Our training strategy is accordingly composed of two parts, ① choices of proper data augmentations, and ② a carefully designed tuning objective.

① For the choices of data augmentations, we consider a flurry of regular transformations illustrated in Figure 2: leftmargin=\*

- Geometric transformation: cutout [62], permutation [37], and rotation [34].
- Appearance transformations: Gaussian noise, Gaussian blur, and Sobel filtering [35].

We note the transformations are beneficial considering they all contribute to a qualified interpolation when the classifier has already learned discriminative features from ID data, allowing augmented samples' representation to reside near the in-distribution samples' in the feature space, without completely overlapping with them. The effects of each transformation are verified through the experiments in Section V-C1.

② Considering the practical goal of OOD detection, which involves both classification and distinguishing OOD samples, the tuning objective consists of two components: the standard cross-entropy loss and an energy-based loss. Thus, the overall tuning objective is roughly:

$$\min_{\theta} \mathbb{E}_{(\mathbf{x}, y), (\mathbf{x}', y') \sim \mathcal{D}_{\text{ID}}} \mathcal{L}_{\text{CE}}(\mathbf{x}, y) + \alpha \cdot \mathcal{L}_{\text{energy}}(\mathbf{x}, \mathbf{x}'), \quad (10)$$

where  $\alpha$  is a loss scaling factor and we note  $\mathcal{L}_{\text{energy}}(\mathbf{x}, \mathbf{x}')$  depends on the ID sample  $\mathbf{x}$  and another i.i.d. copy  $\mathbf{x}'$ . Previously, inspired by the energy-bounded learning objective proposed in [23], which was originally designed for tuning with real outliers, we introduced a similar energy term for OEST [38]. Particularly, the *energy-bounded loss*  $\mathcal{L}_{\text{energy}}$  in Eq. (10) for an ID input pair  $(\mathbf{x}_{\text{in}}, \mathbf{x}'_{\text{in}})$  is:

$$\begin{aligned} \mathcal{L}_{\text{energy}}(\mathbf{x}_{\text{in}}, \mathbf{x}'_{\text{in}}) = & (\max(0, E(\mathbf{x}_{\text{in}}) - m_{\text{in}}))^2 \\ & + (\max(0, m_{\text{per}} - E(\mathbf{x}_{\text{per}})))^2, \end{aligned} \quad (11)$$

where  $m_{\text{in}}$  and  $m_{\text{per}}$  are the margin hyper-parameters for the energy gap, and  $\mathbf{x}_{\text{per}}$  is the random PD sample augmented from  $\mathbf{x}'_{\text{in}}$ . Therefore, the energy loss  $\mathcal{L}_{\text{energy}}$  penalizes the in-distribution samples whose energy values are higher than  $m_{\text{in}}$  and the peripheral-distribution samples whose energy values are respectively lower than  $m_{\text{per}}$ .

**Issues of  $\mathcal{L}_{\text{energy}}$ .** Revisiting Eq. (4), we recall the energy  $E(\mathbf{x})$  represents  $-\log p(\mathbf{x})$ , where we generally omit  $\log Z$

---

**Algorithm 1** OEST\* Tuning Algorithm for OOD Detection

---

**Require:** Training data; original classifier  $f_{\theta}(\cdot)$ ; transformation set  $\mathcal{S}$ ; learning rate  $\eta$ ; transformation ratio  $\tau$ ; number of epochs  $T$ .

**Ensure:** Tuned model  $f_{\theta^*}$

- 1: Initialize the model parameter  $\theta$
- 2: **for** each epoch  $t = 1, 2, \dots, T$  **do**
- 3:   **for** each mini-batch  $(\mathbf{x}, y)$  and  $(\mathbf{x}', y')$  **do**
- 4:     Compute the total loss:  

$$\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}}(\mathbf{x}, y) + \alpha \cdot \mathcal{L}_{\text{energy}^*}(\mathbf{x}, \mathbf{x}').$$

(In  $\mathcal{L}_{\text{energy}^*}$ , a transformation from  $\mathcal{S}$  is applied to  $\mathbf{x}'$  with ratio  $\tau$ .)
- 5:     Perform backpropagation and update  $\theta$  as:  

$$\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}$$
- 6:   **end for**
- 7: **end for**
- 8: **return** Tuned neural network parameters  $\theta^*$ .

---

in practice due to the intractability of computing  $\log Z$ . As discussed in Section III-B, it is theoretically justifiable to ignore the normalization constant during inference using the fixed classifier  $f(\cdot)$ , but its omission becomes questionable in tuning  $f(\cdot)$  with the energy-bounded loss  $\mathcal{L}_{\text{energy}}$ .

As a score-based method, the implicit goal of tuning is to maximize  $p(\mathbf{x}_{\text{in}})$  for ID samples; although we once commented in Eq. (6) that the energy function  $E(\mathbf{x}_{\text{in}})$  can serve as a surrogate for the negative log-likelihood of  $p(\mathbf{x}_{\text{in}})$  for *fixed*  $E(\cdot)$ , we note in tuning, minimizing  $E(\mathbf{x})$  **does not equal** maximizing  $p(\mathbf{x})$  considering  $\log Z$  is changing as well. As shown in Figure 3,  $Z$  undergoes significant fluctuations during the tuning process, which echoes that ignoring  $Z$  in such scenarios could mislead the MLE objective.

Therefore, we propose a new energy loss called *energy-barrier loss*, which is given as:

$$\mathcal{L}_{\text{energy}^*}(\mathbf{x}_{\text{in}}, \mathbf{x}'_{\text{in}}) = \left[ \log \sigma((E(\mathbf{x}_{\text{per}}) - E(\mathbf{x}_{\text{in}})) / \beta) \right], \quad (12)$$

where again  $\mathbf{x}_{\text{per}}$  is the random PD sample augmented from  $\mathbf{x}'_{\text{in}}$ ,  $\sigma(\cdot)$  is the sigmoid function, and  $\beta$  is the hyper-parameter. In Eq. (12), the energy difference term  $E(\mathbf{x}_{\text{per}}) - E(\mathbf{x}_{\text{in}})$  successfully removes the dependence on  $\log Z$ , making the formulation theoretically rigorous for the MLE spirit beneath the energy-based model [36].

Overall,  $\mathcal{L}_{\text{energy}^*}$  emphasizes the relative energy differences between in-distribution and peripheral-distribution samples, in line with the principles discussed in Section IV-B, *i.e.*, establishing an energy barrier around the peripheral-distribution samples. In the follow-up experiments, we will denote *OEST* as the method with the energy loss  $\mathcal{L}_{\text{energy}}$ , and *OEST\** as the method with the energy loss  $\mathcal{L}_{\text{energy}^*}$ .

## V. EXPERIMENTS

In this section, we conduct extensive experiments to validate the effectiveness of our method and compare its performance against existing approaches. Additionally, we perform comprehensive ablation studies to assess the impact of different components of the framework. It is important to note that

TABLE II

OOD DETECTION PERFORMANCE (%) ON CIFAR-10. ALL THE RESULTS ARE AVERAGE VALUES OBTAINED FROM 3 RANDOM RUNS. THE TOP-1 RESULTS ARE IN **BOLD**, WHILE THE SECOND- AND THIRD-BEST RESULTS ARE UNDERLINED.

Method	CIFAR-100				Tin		MNIST			SVHN			Textures			Places365			Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	ID ACC ↑			
<b>Post-Hoc Inference Methods</b>																				
ASH [50]	74.11	87.31	76.44	86.25	83.16	70.00	73.46	83.64	77.45	84.59	79.89	77.89	77.42	81.61	95.06					
SHE [42]	80.31	81.00	82.76	78.30	90.43	42.22	86.38	62.74	81.57	84.60	82.89	76.36	84.06	70.87	95.06					
ODIN [43]	82.18	77.00	83.55	75.38	95.24	23.83	84.58	68.61	86.94	67.70	85.07	70.36	86.26	63.81	95.06					
MSP [6]	87.19	53.08	88.87	43.27	92.63	23.64	91.46	25.82	89.89	34.96	88.92	42.47	89.83	37.21	95.06					
MLS [44]	86.31	66.59	88.72	56.06	94.15	25.06	91.69	35.09	89.41	51.73	89.14	54.84	89.90	48.23	95.06					
EBO [23]	86.36	66.60	88.80	56.08	94.32	24.99	91.79	35.12	89.47	51.82	89.25	54.85	90.00	48.24	95.06					
TempScale [81]	87.17	55.81	89.00	46.11	93.11	23.53	91.66	26.97	90.01	38.16	89.11	45.27	90.01	39.31	95.06					
GEN [82]	87.21	58.75	89.20	48.59	93.83	23.00	91.97	28.14	90.14	40.74	89.46	47.03	90.30	41.04	95.06					
KNN [41]	89.73	37.64	91.56	30.37	94.26	20.05	92.67	22.60	93.16	24.06	91.77	30.38	92.19	27.52	95.06					
<b>Training methods from scratch</b>																				
MOS [83]	70.57	79.38	72.34	78.05	74.81	65.95	73.66	57.79	70.35	76.78	86.81	51.09	74.76	68.17	94.83					
ARPL [84]	86.76	43.38	88.12	37.28	92.62	21.49	87.69	35.68	88.57	35.19	88.57	37.21	88.72	35.04	93.66					
VOS [74]	86.57	61.57	88.84	52.49	91.56	35.92	92.18	31.50	89.68	46.53	89.90	47.78	89.79	45.97	94.31					
CSI [37]	88.16	37.57	90.87	29.74	92.55	24.41	95.18	17.56	90.71	28.95	89.56	34.76	91.17	28.83	91.16					
ConfBranch [85]	88.91	34.44	90.77	28.11	94.49	15.79	95.42	14.06	91.10	27.24	90.39	28.85	91.85	24.75	94.88					
NPOS [75]	88.57	35.71	90.99	29.57	92.64	22.96	98.88	6.41	94.44	20.80	90.32	32.19	92.64	24.61	/					
CIDER [86]	89.47	35.60	91.94	28.61	93.30	24.76	98.06	8.04	93.71	25.05	93.77	25.03	93.38	24.52	/					
G-ODIN [55]	88.14	48.86	90.09	42.21	98.95	4.53	97.76	10.72	95.02	27.27	90.31	43.30	93.38	29.48	94.70					
LogitNorm [56]	90.95	34.37	93.70	24.30	99.14	3.93	98.25	8.33	94.77	21.94	94.79	21.04	95.27	18.99	94.30					
RotPred [66]	91.19	34.24	94.17	22.04	97.52	9.24	98.89	3.20	97.30	9.87	92.76	26.61	95.31	17.53	95.35					
<b>Tuning method</b>																				
OEST (Ours)	91.27	33.46	94.62	21.57	<b>99.65</b>	<b>1.83</b>	99.10	5.07	97.87	11.03	94.80	20.70	96.22	15.61	95.00					
OEST* (Ours)	<b>91.47</b>	<b>32.60</b>	<b>94.81</b>	<b>21.02</b>	98.98	4.97	<b>99.28</b>	3.75	<b>98.18</b>	<b>9.52</b>	<b>95.10</b>	<b>20.13</b>	<b>96.30</b>	<b>15.33</b>	94.97					

in practical applications, the out-of-distribution (OOD) detection task is indeed twofold: a) accurately categorizing in-distribution samples, as in conventional classification tasks, and b) enabling a well-trained classifier to distinguish out-of-distribution samples during the inference phase correctly.

In Section V-A, we first introduce our experimental setup, including the datasets, evaluation metrics, and training details. Then, in Section V-B, we present the main experimental results, showcasing the efficacy of our method across various datasets. We provide a thorough analysis of its performance, comparing it with other methods to demonstrate the robustness and reliability of our approach. Finally, in Section V-C, we carry out exhaustive ablation studies to systematically examine the contribution of each component, offering deeper insights into their individual roles. These experiments are designed to comprehensively validate the effectiveness of our method, ensuring it performs well on both in-distribution classification and OOD detection.

### A. Setup

1) *Datasets*: We follow a common setup in the out-of-distribution detection field and mainly report results using two widely used datasets, CIFAR-10 and CIFAR-100 [87]. In terms of OOD datasets, we primarily adhere to the practices outlined in OpenOOD [19]. When CIFAR-10 [87] is used as the in-distribution dataset, CIFAR-100 [87] and Tiny ImageNet (Tin) [88] are used as near OOD datasets, while MNIST [89], SVHN [90], Textures [91], and Place365 [92] are employed as far OOD datasets. Similarly, for CIFAR-100 [87] as the in-distribution dataset, we adopt CIFAR-10 [87] and Tiny ImageNet (Tin) [88] as near OOD datasets, and MNIST [89],

SVHN [90], Textures [91], and Place365 [92] as far OOD datasets. For a detailed description of all datasets used, please refer to Appendix C-A.

Additionally, for the ablation studies, we use KMNIST [93] as the in-distribution dataset, with CIFAR-10 [87] and EMNIST [94] serving as real outliers for tuning, and MNIST [89] for OOD evaluation. Moreover, we conduct further experiments with MNIST and SVHN as in-distribution datasets, the details of which can be found in Appendix C-B.

2) *Evaluation Metrics*: As mentioned earlier, OOD detection in practical applications has two key objectives: accurate classification of in-distribution samples and reliable detection of out-of-distribution (OOD) samples. To rigorously evaluate the effectiveness of our methods, we assess the results using three widely recognized metrics. The first two metrics focus on OOD detection performance. The first is the Area Under the Receiver Operating Characteristic Curve (AUROC) [95], which provides a probabilistic measure of the likelihood that a positive sample receives a higher discriminative score than a negative one [96]. AUROC serves as a comprehensive indicator of the model's ability to differentiate between in-distribution and OOD samples. The second metric is the False Positive Rate at 95% True Positive Rate (FPR95) [43], which measures how often negative samples are mistakenly classified as positive. FPR95 is particularly useful for assessing the model's reliability in scenarios where precise classification of positive samples is critical. The third metric, in-distribution testing accuracy (ID-ACC), reflects the model's performance on the original classification task. This metric ensures that the model maintains strong classification capabilities, which is crucial along with OOD detection. By incorporating these

TABLE III

OOD DETECTION PERFORMANCE (%) ON CIFAR-100. ALL THE RESULTS ARE AVERAGE VALUES OBTAINED FROM 3 RANDOM RUNS. THE TOP-1 RESULTS ARE IN **BOLD**, WHILE THE SECOND- AND THIRD-BEST RESULTS ARE UNDERLINED.

Method	CIFAR-10				Tin		MNIST			SVHN		Textures		Places365		Average	
	AUROC↑	FPR95↓	ID ACC ↑														
<b>Post-Hoc Inference Methods</b>																	
OpenMax [97]	74.38	60.17	78.44	52.99	76.01	53.82	82.07	53.20	80.56	56.12	79.29	<u>54.85</u>	78.46	55.19	<u>77.25</u>		
MSP [6]	78.47	58.91	82.07	50.70	76.08	57.23	78.42	59.07	77.32	61.88	79.22	<u>56.62</u>	78.60	57.40	<u>77.25</u>		
TempScale [81]	79.02	<b>58.72</b>	82.79	50.26	77.27	56.05	79.79	57.71	78.11	61.56	79.80	56.46	79.46	56.79	<u>77.25</u>		
ODIN [43]	78.18	60.64	81.63	55.19	83.79	45.94	74.54	67.41	79.33	62.37	79.45	59.71	79.49	58.54	<u>77.25</u>		
MLS [44]	<u>79.21</u>	<u>59.11</u>	82.90	51.83	78.91	52.95	81.65	53.90	78.39	62.39	79.75	57.68	80.14	56.31	<u>77.25</u>		
EBO [23]	79.05	59.21	82.76	52.03	79.18	52.62	82.03	53.62	78.35	62.35	79.52	57.75	80.15	56.26	<u>77.25</u>		
GEN [82]	<b>79.38</b>	<u>58.87</u>	<u>83.25</u>	49.98	78.29	53.92	81.41	55.45	78.74	61.23	80.28	56.25	80.23	55.95	<u>77.25</u>		
ReAct [48]	78.65	61.30	82.88	51.47	78.37	56.04	83.01	50.41	80.15	55.04	80.03	<u>55.30</u>	80.52	54.93	<u>77.25</u>		
KNN [41]	77.02	72.80	<u>83.34</u>	<u>49.65</u>	82.36	48.58	84.15	51.75	83.66	53.56	79.43	60.70	81.66	56.17	<u>77.25</u>		
RMDS [98]	77.75	61.37	82.55	<u>49.56</u>	79.74	52.05	84.89	51.65	83.65	53.99	<b>83.40</b>	<u>53.57</u>	82.00	53.70	<u>77.25</u>		
<b>Training methods from scratch</b>																	
CSI [37]	69.50	72.62	73.40	67.90	51.79	80.54	80.24	67.21	62.22	90.51	70.99	69.41	68.02	74.70	61.60		
ConfBranch [85]	68.80	74.56	74.41	<u>65.86</u>	74.29	55.95	65.51	76.01	65.39	85.43	70.42	69.90	69.80	71.29	76.59		
ARPL [84]	73.38	64.84	76.50	58.27	73.77	59.12	76.45	<u>59.76</u>	69.93	71.66	74.62	62.01	74.11	62.61	<u>70.70</u>		
CIDER [86]	67.55	82.71	78.65	61.33	68.14	75.32	<u>97.17</u>	<u>17.82</u>	82.21	54.43	74.43	69.30	78.03	60.15	/		
MOS [83]	78.54	60.60	82.26	51.49	80.68	52.70	81.59	56.33	79.92	61.24	78.50	58.86	80.25	56.87	<u>76.98</u>		
LogitNorm [56]	74.57	73.88	82.37	51.89	90.69	34.12	82.80	47.52	72.37	77.38	80.25	55.44	80.51	56.71	<u>76.34</u>		
NPOS [75]	75.37	72.50	81.32	54.21	73.26	66.98	92.43	30.67	<u>85.55</u>	47.39	77.92	59.47	80.98	55.20	/		
VOS [74]	<u>79.14</u>	59.23	82.73	51.89	82.29	48.56	84.23	47.23	78.41	62.55	<u>80.34</u>	56.44	81.19	54.32	<u>77.20</u>		
G-ODIN [55]	73.04	78.82	81.26	56.34	<u>91.15</u>	<u>27.19</u>	83.74	42.68	<u>89.62</u>	<b>35.83</b>	78.17	65.03	82.83	50.98	74.46		
RotPred [66]	71.11	72.00	81.75	53.17	<u>93.10</u>	22.77	95.39	<u>15.64</u>	88.16	40.03	76.95	59.56	84.41	<u>43.86</u>	76.03		
<b>Tuning method</b>																	
OEST (Ours)	77.43	63.35	82.71	52.20	90.75	33.24	88.54	36.18	81.44	55.12	79.03	58.58	<u>83.32</u>	<u>49.78</u>	76.87		
OEST* (Ours)	75.22	69.94	<b>85.75</b>	<u>47.29</u>	<b>95.68</b>	<u>19.39</u>	<b>98.56</b>	<u>8.04</u>	<b>90.32</b>	<u>36.30</u>	<u>82.63</u>	<u>54.51</u>	<b>88.03</b>	<u>39.25</u>	<b>77.63</b>		

three metrics, we emphasize the primary goal of OOD detection: achieving a balanced performance across both the classification task and the detection of OOD samples.

3) *Training Details*: We primarily report results obtained by using ResNet [99]. For the CIFAR-10 and CIFAR-100 datasets, we use the trained ResNet-18 model provided by [20], which is trained with SGD optimizer using a learning rate of 0.1, momentum of 0.9, and weight decay of  $5 \times 10^{-4}$  for 100 epochs. We further tune the trained model for an additional 10 epochs, again using the SGD optimizer. For CIFAR-10, we apply six augmentations—cutout, blur, noise, rotation, permutation, and sobel—comprising a mix of geometric and appearance transformations to generate peripheral-distribution data. The ratio of in-distribution to peripheral-distribution data is set to 1:1, with a batch size of 128 for the in-distribution data. For CIFAR-100, in addition to these six augmentations, we also apply a seventh augmentation, *RandAugment*, a more diverse transformation technique from the PyTorch library. This inclusion provides a wider range of transformations, further enriching the peripheral-distribution data for CIFAR-100. In this case, the ratio of in-distribution to peripheral-distribution data is set to 1:2, with a batch size of 128 for the in-distribution data. For OEST, we apply a weight of  $\alpha = 0.01$  for  $L_{\text{energy}}$ , with the margin values  $m_{\text{in}}$  and  $m_{\text{pre}}$  set to -25 and -7, respectively, following the setup in [23]. The learning rate follows a cosine annealing schedule, starting at  $1 \times 10^{-4}$  and gradually decaying to  $1 \times 10^{-8}$  throughout tuning. For OEST\*, we use weights of  $\alpha = 0.2$  and  $\beta = 10$  for  $L_{\text{energy}*}$ . The learning rate also follows a cosine annealing schedule, starting at  $1 \times 10^{-3}$  and gradually decaying to  $1 \times 10^{-7}$  during tuning. Specifically, for our methods, we further tune the three trained

models provided by [20] using a random seed of 1 to ensure fairness and reproducibility in our comparisons. Additional ablation studies on other hyper-parameters are provided in the subsequent subsection.

## B. Main Results

To thoroughly assess the performance of the proposed method across various scenarios, we conduct a rigorous comparison not only against baseline models but also against several algorithms that have gained recognition in recent years. This comprehensive evaluation allows us to gain a deeper understanding of the proposed method. We select methods from the best average detection performance under a unified evaluation benchmark [19]. The results are presented in Table II and Table III. It should be noted that certain approaches, such as PixMix [64], are not directly comparable to ours, as they leverage a manually curated set of auxiliary images for mixing.

1) *CIFAR-10 as ID*: Table II presents the OOD detection performance for CIFAR-10 as the in-distribution dataset, evaluated across six out-of-distribution test datasets. Our methods consistently achieve the top-3 AUROC on all six OOD datasets. Notably, the average AUROC of OEST\* reaches 96.30%, surpassing RotPred by 0.99%. Additionally, our methods also achieve the lowest FPR95 on almost OOD datasets so that OEST\* achieves the lowest average FPR95 (15.33%). The outstanding average performance on both AUROC and FPR95 demonstrates our methods' superior ability to maintain detection accuracy while minimizing false positives. Moreover, with a more theoretically solid loss function, OEST\* demonstrates a more balanced and superior performance compared to OEST.

TABLE IV

THE COMPARISON WITH DIFFERENT SINGLE DATA AUGMENTATION BY EACH SIMPLE TRANSFORMATIONS ABOUT AUROC (%) WHEN CIFAR-10 IS THE GIVEN IN-DISTRIBUTION. THE FIRST MODEL IS THE TRAINED CLASSIFIER, THE LAST ONE IS THE MODEL FURTHER TUNED WITH THE COMPOSITION OF ALL CONSIDERED DATA AUGMENTATIONS. THE TOP-1 RESULTS ARE IN **BOLD**, WHILE THE SECOND-BEST RESULTS ARE UNDERLINED.

Simple Transform	CIFAR-100		Tin		MNIST		SVHN		Textures		Places365		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
EBO [23]	86.36	66.60	88.80	56.08	94.32	24.99	91.79	35.12	89.47	51.82	89.25	54.85	90.00	48.24
Sobel	90.68	43.83	90.23	41.06	<u>99.01</u>	<u>4.38</u>	93.73	38.43	91.34	39.34	92.02	36.04	93.00	33.85
Blur	90.04	40.88	92.20	30.71	<u>97.17</u>	12.49	<b>99.77</b>	<b>0.61</b>	92.17	34.00	91.97	31.84	93.89	25.09
Noise	83.68	61.84	89.97	44.06	85.50	50.39	87.84	38.17	93.06	33.16	88.13	50.68	88.03	46.38
Cutout	89.49	43.21	92.77	30.06	<b>99.94</b>	<b>0.14</b>	93.29	19.33	93.76	25.92	94.77	21.09	94.00	23.29
Perm	91.10	34.13	<u>94.66</u>	<u>22.89</u>	96.93	12.71	96.87	12.31	97.60	11.42	<b>96.45</b>	<b>16.04</b>	94.00	23.29
Rotation	<b>91.49</b>	<u>32.79</u>	94.41	22.97	97.45	10.41	99.17	4.36	<u>97.70</u>	<u>10.34</u>	93.76	26.00	<u>95.83</u>	<u>16.18</u>
OEST*	91.47	<b>32.60</b>	<b>94.81</b>	<b>21.02</b>	98.98	4.97	99.28	<u>3.75</u>	<b>98.18</b>	<b>9.52</b>	<u>95.10</u>	<u>20.13</u>	<b>96.30</b>	<b>15.33</b>

TABLE V

AUROC (%) PERFORMANCE COMPARISON BETWEEN PERIPHERAL-DISTRIBUTION SAMPLES AND REAL OUTLIERS. THE TOP-1 RESULTS ARE IN **BOLD**, WHILE THE SECOND-BEST RESULTS ARE UNDERLINED.

$\mathcal{D}_{ID}$	External dataset?	$\mathcal{X}_{OOD}$ or $\mathcal{X}_{PD}$	MNIST
KMNIST	✗	EBO [23]	89.6
	✓	CIFAR-10	91.4
	✓	EMNIST	<b>99.8</b>
	✗	OEST* (Ours)	<u>98.6</u>

Since post-hoc inference methods do not alter the model’s structure and only adjust the computation of the final scoring function, the accuracy on in-distribution samples remains unchanged. In contrast, methods that involve training from scratch or continual tuning typically result in a decrease in ID-ACC. This reduction arises from the inherent trade-off between optimizing for out-of-distribution (OOD) detection and maintaining in-distribution accuracy. During training, Empirical Risk Minimization (ERM) focuses on minimizing classification error for in-distribution samples, while OOD detection often requires a different optimization direction. This divergence can create a mismatch between the objectives of enhancing OOD detection performance and preserving the accuracy of the original classification task, leading to a potential drop in ID-ACC. However, our tuning strategy resulted in only a 0.06% decrease in classification accuracy for OEST and a 0.09% decrease for OEST\*—differences so small they can be considered negligible, particularly in the context of CIFAR-10 tasks.

2) *CIFAR-100 as ID*: As shown in Table III, our approach significantly surpasses other baselines, achieving the highest average AUROC of 88.03% (+3.62%) and the lowest average FPR95 of 39.25% (-3.61%). Notably, OEST\* exhibits exceptional AUROC performance on Tiny ImageNet (85.75%), MNIST (95.68%), SVHN (98.56%), and Texture (90.32%), underscoring its effectiveness across both near-OOD and far-OOD datasets. Furthermore, our tuning process not only enhances out-of-distribution detection but also improves the original model’s classification accuracy, reflected in a higher ID-ACC score of 77.63%. This improvement suggests that our

method promotes more robust feature representations, benefiting both in-distribution classification and OOD detection. In this context, OEST\* outperforms OEST on five out of six out-of-distribution datasets, highlighting the advantages of the energy-barrier loss  $\mathcal{L}_{energy^*}$ . With a more rigorous theoretical foundation than the energy-bounded loss, the energy-barrier loss enables OEST\* to achieve superior results, further emphasizing the necessity of accounting for  $\log Z$  during training, as it cannot be simply disregarded. However, we acknowledge that our methods are less effective on CIFAR-10. This phenomenon is discussed in detail in Section V-C4.

### C. Experimental Analysis

Building on the experimental results discussed above, we conducted a series of comprehensive experiments to explore various factors of our method. More discussion can be found in Appendix C-C.

1) *The Ablation Study for Simple Transformations*: As illustrated in Table IV, we have observed that almost all transformations individually improve the classifier’s performance. Consequently, we decided to combine all the transformations together to evaluate their collective impact. The results confirm that the mixed version effectively enhances the performance of the trained classifier, demonstrating better comprehensive OOD detection performance in terms of both AUROC and FPR95 metrics. Although the mixed version may not consistently outperform individual transformations in certain benchmarks, it still yields the highest overall improvement for OOD detection.

2) *Limitation of Real Outliers*: We evaluate classifiers on KMNIST as the in-distribution dataset, with MNIST as the OOD dataset, as shown in Table V. All models are based on the LeNet backbone; the first model is a trained classifier, and the other three are further tuned based on the first one, utilizing either external datasets or samples augmented with simple transformations (cutout, blur, noise, and permutation). The results indicate that using CIFAR-10 as auxiliary data provides only a marginal improvement in OOD detection performance (AUROC 91.4%). This is likely because KMNIST [93], EMNIST [94], and MNIST [89] are all grayscale datasets containing handwritten images, whereas CIFAR-10 [87] comprises colorful images of various natural scenes,

TABLE VI  
THE COMPARISON OF OUR METHOD WITH DIFFERENT BACKBONES [100] ABOUT AUROC (%) WHEN CIFAR-10 IS THE GIVEN IN-DISTRIBUTION. **BOLD** DENOTES THE BEST RESULTS.

Backbone	Further Tuned?	CIFAR-100		Tin		MNIST		SVHN		Textures		Places365		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Res18	✗	86.7	49.3	87.0	45.2	94.9	24.6	95.0	24.9	90.8	39.3	89.4	40.3	89.8	37.3
	✓	<b>91.5</b>	<b>32.6</b>	<b>94.8</b>	<b>21.0</b>	<b>99.0</b>	<b>5.0</b>	<b>99.3</b>	<b>3.8</b>	<b>98.2</b>	<b>9.5</b>	<b>95.1</b>	<b>20.1</b>	<b>96.3</b>	<b>15.3</b>
Res50	✗	88.7	51.0	89.0	47.0	98.0	10.7	85.6	55.2	86.9	57.5	90.6	42.2	88.2	43.9
	✓	<b>92.6</b>	<b>35.1</b>	<b>93.8</b>	<b>25.4</b>	<b>99.9</b>	<b>0.2</b>	<b>99.5</b>	<b>2.6</b>	<b>97.3</b>	<b>12.5</b>	<b>94.9</b>	<b>22.1</b>	<b>96.3</b>	<b>16.3</b>
WRN34	✗	87.8	44.2	87.5	41.1	93.6	28.1	91.9	39.9	85.7	42.6	88.1	39.9	89.1	39.3
	✓	<b>93.2</b>	<b>38.4</b>	<b>93.6</b>	<b>30.5</b>	<b>99.8</b>	<b>0.2</b>	<b>99.8</b>	<b>1.0</b>	<b>97.7</b>	<b>14.6</b>	<b>95.4</b>	<b>23.8</b>	<b>96.4</b>	<b>18.1</b>

making it stylistically distinct from MNIST. By contrast, using EMNIST as auxiliary data, which has a closer resemblance in style and content to MNIST, significantly boosts performance, achieving a top AUROC of 99.8%. This finding underscores the importance of choosing an appropriate auxiliary dataset in outlier-based methods and reveals how dataset selection can limit the applicability of these approaches. Thus, for OOD detection, it is generally assumed that auxiliary OOD training data is not accessible. Additionally, our proposed peripheral-distribution samples demonstrate performance comparable to real outliers (AUROC 98.6%), further validating the effectiveness of our approach in the absence of specific auxiliary datasets.

3) *Ablation Study on Different Backbones:* We perform ablation studies using various backbone architectures, specifically ResNet18, ResNet50, and WideResNet34. As shown in Table VI, we present the AUROC values for each backbone. The results demonstrate that our method consistently enhances the performance of the trained classifier, regardless of the backbone architecture used. Additionally, we observed that using WideResNet as the backbone yielded the best AUROC performance, which aligns with the general understanding of the neural network's capacity and expressiveness.

4) *Assumption Validity and the Influence of Backbone Strength:* In the analysis presented in Table III, we observe a marginal decline in our model's performance on CIFAR-10 test metrics, particularly in comparison to results achieved solely using the trained model. We attribute this decline to a deviation from our initial Energy Barrier Assumption on Peripheral-Distribution (Assumption 1). Notably, due to the stylistic similarity between CIFAR-10 and CIFAR-100, a result of their similar data collection methods, certain augmented samples in our peripheral distribution may inadvertently overlap with CIFAR-10 samples. This overlap challenges our core assumption that a sufficiently large energy barrier exists to clearly differentiate ID data from peripheral-distribution samples and OOD samples.

To investigate this hypothesis, we conducted a series of experiments using ResNet architectures with varying depths. The results reveal that a stronger backbone mitigates the observed performance drop, producing notable improvements in both AUROC and FPR95 metrics. For example, switching from ResNet18 to ResNet34 yields a 10.1% increase in AUROC and a 24.87% reduction in FPR95. When using ResNet50, these improvements become even more substantial: a 12.29%

TABLE VII  
OOD DETECTION PERFORMANCE (%) OF CIFAR-100 CLASSIFIER WITH DIFFERENT BACKBONES ON CIFAR-10

Backbone	Method	AUROC ↑	FPR95 ↓
ResNet18	EBO [23]	79.05	59.21
	OEST* (Ours)	75.22(-3.83)	69.94(+10.73)
ResNet34	EBO [23]	79.43	82.41
	OEST* (Ours)	89.53(+10.10)	57.54(-24.87)
ResNet50	EBO [23]	82.27	91.35
	OEST* (Ours)	94.56(+12.29)	33.38(-57.95)

increase in AUROC and a 57.95% decrease in FPR95. These outcomes suggest that more powerful feature extraction allows the classifier to create a more compact clustering of ID representations, thereby enhancing the reliability of peripheral samples generated from ID data. This observation underscores the necessity of a strong feature extraction backbone to maintain the energy barrier, reinforcing the criticality of our initial assumption.

## VI. CONCLUSION

In this work, we introduce an out-of-distribution (OOD) detection framework, OEST, which leverages the principles beneath energy-based models (EBMs) to enhance classifier robustness without substantial reliance on expensive real outlier data. Specifically, we generate peripheral-distribution data to offer a practical and theoretically sound solution for OOD detection; by employing peripheral-distribution data, OEST builds an energy barrier around in-distribution samples, consequently distinguishing them from OOD samples through a spectrum of data transformations. In contrast to training-based methods (*cf.* Section II-B), OEST solely further tunes trained models and allows efficient deployment without substantial computational demands. Furthermore, we devise the energy-barrier loss to displace the energy-bounded loss in [38] (inducing the advanced version, OEST\*), provide statistical guarantee under the EBM framework, and successfully improve OOD detection performance across various benchmarks. Our experiments show that OEST\* consistently outperforms baseline models across various tasks. We are confident that OEST will pave the way to new out-of-distribution detection and open-world object detection.

## ACKNOWLEDGMENTS

This work is supported by the Shanghai Engineering Research Center of Intelligent Computing System (No. 19DZ2252600) and the Research Grants Council (RGC) under grant ECS-22303424. The authors also thank Prof. Cheng Jin for providing the computational resources that significantly contributed to the success of this research.

APPENDIX A  
PROOF OF THEOREM 1

**Theorem 1.** *When Assumption 1 holds, we then have*

$$E(\mathbf{x}'; f) - E(\mathbf{x}; f) > \gamma_\alpha$$

holds with probability  $1 - \alpha$ . The out-of-distribution sample  $\mathbf{x}'$  will be guaranteed to have higher energy than a random ID sample  $\mathbf{x}$  with high probability.

*Proof.* Inserting the augmented sample  $\mathbf{x}^+$ , we first reformulate the target gap  $E(\mathbf{x}'; f) - E(\mathbf{x}; f)$  as

$$(E(\mathbf{x}'; f) - E(\mathbf{x}^+; f)) + (E(\mathbf{x}^+; f) - E(\mathbf{x}; f)).$$

It then suffices to bound  $E(\mathbf{x}'; f) - E(\mathbf{x}^+; f)$  from below, considering in Assumption 1 we already have

$$E(\mathbf{x}^+; f) - E(\mathbf{x}; f) > B \cdot \|\mathbf{x}' - \mathbf{x}^+\| + \gamma_\alpha.$$

We expand  $E(\mathbf{x}'; f) - E(\mathbf{x}^+; f)$  as

$$T \cdot \log \left[ \frac{\sum_{i=1}^C \exp(\langle \mathbf{x}^+, \mathbf{c}_i \rangle / T)}{\sum_{i=1}^C \exp(\langle \mathbf{x}', \mathbf{c}_i \rangle / T)} \right]. \quad (13)$$

For the fraction of the form  $(\sum a_i) / (\sum b_i)$ , we notice

$$\frac{\sum_{i=1}^C a_i}{\sum_{i=1}^C b_i} = \sum_i \left( \frac{a_i}{b_i} \cdot \frac{b_i}{\sum_{j=1}^C b_j} \right),$$

which indicates the internal fraction in Eq. (13) is as well a weighted sum of the following positive terms

$$\exp(\langle \mathbf{x}^+, \mathbf{c}_i \rangle / T) / \exp(\langle \mathbf{x}', \mathbf{c}_i \rangle / T),$$

implying the fraction is no lower than the term above for a certain  $i \in [C]$ . We thus have

$$\begin{aligned} E(\mathbf{x}'; f) - E(\mathbf{x}^+; f) &\geq T \cdot \log \frac{\exp(\langle \mathbf{x}^+, \mathbf{c}_i \rangle / T)}{\exp(\langle \mathbf{x}', \mathbf{c}_i \rangle / T)} \\ &= \langle \mathbf{x}^+ - \mathbf{x}', \mathbf{c}_i \rangle \\ &\geq -\|\mathbf{x}^+ - \mathbf{x}'\| \|\mathbf{c}_i\| \\ &\geq -B \cdot \|\mathbf{x}^+ - \mathbf{x}'\|. \end{aligned}$$

Combining Assumption 1, we can attain the claim in Theorem 1 and the proof is complete.  $\square$

## REFERENCES

- [1] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *Proc. ACM SIGMOD*, 2001.
- [2] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, 2004.
- [3] I. Ben-Gal, "Outlier detection," in *Data Min. Knowl. Discov. Handb.* Springer, 2005.
- [4] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, 2019.
- [5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.
- [6] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," in *Int. Conf. Learn. Represent.*, 2017.
- [7] T. G. Dietterich, "Steps toward robust artificial intelligence," *AI Mag.*, 2017.
- [8] J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg, "AI safety gridworlds," *arXiv preprint arXiv:1711.09883*, 2017.
- [9] N. A. Smuha, "The EU approach to ethics guidelines for trustworthy artificial intelligence," *Comput. Law Rev. Int.*, 2019.
- [10] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems," *ACM Trans. Interact. Intell. Syst.*, 2020.
- [11] S. Mohseni, H. Wang, Z. Yu, C. Xiao, Z. Wang, and J. Yadawa, "Practical Machine Learning Safety: A Survey and Primer," *arXiv preprint arXiv:2106.04823*, 2021.
- [12] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, "Unsolved problems in ml safety," *arXiv preprint arXiv:2109.13916*, 2021.
- [13] D. Hendrycks and M. Mazeika, "X-risk analysis for AI research," *arXiv preprint arXiv:2206.05862*, 2022.
- [14] A. Nguyen, J. Yosinski, and J. Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 427–436.
- [15] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why ReLU Networks Yield High-Confidence Predictions Far Away from the Training Data and How to Mitigate the Problem," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 41–50.
- [16] S. Jung, J. Lee, D. Gwak, S. Choi, and J. Choo, "Standardized Max Logits: A Simple Yet Effective Approach for Identifying Unexpected Road Obstacles in Urban-Scene Segmentation," in *IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15 425–15 434.
- [17] J. Gaspar, E. Catumbela, B. Marques *et al.*, "A systematic review of outliers detection techniques in medical data-preliminary study," in *Proc. Int. Conf. Health Informatics*. SCITEPRESS, 2011, pp. 575–582.
- [18] M. Hauskrecht, I. Batal, M. Valko *et al.*, "Outlier detection for patient monitoring and alerting," *J. Biomed. Inform.*, vol. 46, no. 1, pp. 47–55, 2013.
- [19] J. Yang, P. Wang, D. Zou, Z. Zhou, K. Ding, W. Peng, H. Wang, G. Chen, B. Li, Y. Sun, and et al., "OpenOOD: Benchmarking Generalized Out-of-Distribution Detection," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 32 598–32 611, 2022.
- [20] J. Zhang, J. Yang, P. Wang, H. Wang, Y. Lin, H. Zhang, Y. Sun, X. Du, K. Zhou, W. Zhang *et al.*, "OpenOOD v1. 5: Enhanced Benchmark for Out-of-Distribution Detection," *arXiv preprint arXiv:2306.09301*, 2023.
- [21] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep Anomaly Detection with Outlier Exposure," in *Int. Conf. Learn. Represent.*, 2019.
- [22] Q. Yu and K. Aizawa, "Unsupervised Out-of-Distribution Detection by Maximum Classifier Discrepancy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9518–9526.
- [23] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-Based Out-of-Distribution Detection," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 21 464–21 475, 2020.
- [24] S. Mohseni, M. Pitale, J. B. S. Yadawa, and Z. Wang, "Self-Supervised Learning for Generalizable Out-of-Distribution Detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 04, 2020, pp. 5216–5223.
- [25] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "Atom: Robustifying out-of-distribution detection using outlier mining," *ECML&PKDD*, 2021.
- [26] S. Thulasidasan, S. Thapa, S. Dhaubadel, G. Chennupati, T. Bhattacharya, and J. Bilmes, "An effective baseline for robustness to distributional shift," *arXiv preprint arXiv:2105.07107*, 2021.

- [27] Y. Li and N. Vasconcelos, "Background data resampling for outlier-aware classification," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.
- [28] A.-A. Papadopoulos, M. R. Rajati, N. Shaikh, and J. Wang, "Outlier Exposure with Confidence Control for Out-of-Distribution Detection," *Neurocomputing*, vol. 441, pp. 138–150, 2021.
- [29] Y. Ming, Y. Fan, and Y. Li, "Poem: Out-of-distribution detection with posterior sampling," in *ICML*, 2022.
- [30] J. Zhang, N. Inkawich, R. Linderman, Y. Chen, and H. Li, "Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 5531–5540.
- [31] P. Liznerski, L. Ruff, R. A. Vandermeulen *et al.*, "Exposing Outlier Exposure: What Can Be Learned From Few, One, and Zero Outlier Images," *Trans. Mach. Learn. Res.*, 2022.
- [32] Q. Wang, Z. Fang, Y. Zhang, F. Liu, Y. Li, and B. Han, "Learning to augment distributions for out-of-distribution detection," in *Adv. Neural Inform. Process. Syst.*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023, pp. 73 274–73 286.
- [33] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Muller, "A Unifying Review of Deep and Shallow Anomaly Detection," *Proc. IEEE*, vol. 109, pp. 756–795, 2020.
- [34] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised Representation Learning by Predicting Image Rotations," in *Int. Conf. Learn. Represent.*, 2018.
- [35] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an Image Edge Detection Filter Using the Sobel Operator," *IEEE J. Solid-State Circuits*, vol. 23, no. 2, pp. 358–367, 1988.
- [36] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A Tutorial on Energy-Based Learning," *Predicting Structured Data*, vol. 1, no. 0, 2006.
- [37] J. Tack, S. Mo, J. Jeong, and J. Shin, "CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 11 839–11 852, 2020.
- [38] Y. Wu, S. Dai, D. Pan, and X. Li, "OEST: Outlier Exposure by Simple Transformations for Out-of-Distribution Detection," in *2023 IEEE Int. Conf. Image Process. (ICIP)*. IEEE, 2023, pp. 2170–2174.
- [39] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
- [40] C. S. Sastry and S. Oore, "Detecting Out-of-Distribution Examples with Gram Matrices," in *Int. Conf. Mach. Learn.*, 2020, pp. 8491–8501.
- [41] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-Distribution Detection with Deep Nearest Neighbors," in *Int. Conf. Mach. Learn.*, 2022, pp. 20 827–20 840.
- [42] J. Zhang, Q. Fu, X. Chen, L. Du, Z. Li, G. Wang, xiaoguang Liu, S. Han, and D. Zhang, "Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy," in *The Eleventh International Conference on Learning Representations*, 2023.
- [43] S. Liang, Y. Li, and R. Srikant, "Enhancing The Reliability of Out-of-Distribution Image Detection in Neural Networks," in *Int. Conf. Learn. Represent.*, 2018.
- [44] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," in *Int. Conf. Mach. Learn.*, 2022, pp. 8759–8773.
- [45] Z. Lin, S. D. Roy, and Y. Li, "Mood: Multi-level out-of-distribution detection," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 313–15 323.
- [46] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your Classifier is Secretly an Energy Based Model and You Should Treat It Like One," in *Int. Conf. Learn. Represent.*, 2020.
- [47] P. Morteza and Y. Li, "Provable guarantees for understanding out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, 2022, pp. 7831–7840.
- [48] Y. Sun, C. Guo, and Y. Li, "REAct: Out-of-Distribution Detection with Rectified Activations," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 144–157, 2021.
- [49] X. Dong, J. Guo, A. Li, W.-T. Ting, C. Liu, and H. Kung, "Neural mean discrepancy for efficient out-of-distribution detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.
- [50] A. Djuricic, N. Bozanic, A. Ashok, and R. Liu, "Extremely simple activation shaping for out-of-distribution detection," in *The Eleventh International Conference on Learning Representations*, 2023.
- [51] K. Xu, R. Chen, G. Franchi, and A. Yao, "Scaling for training time and post-hoc out-of-distribution detection enhancement," in *The Twelfth International Conference on Learning Representations*, 2024.
- [52] W. Wan, W. Zhang, and C. Jin, "Out-of-distribution detection using neural activation prior," 2024.
- [53] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865*, 2018.
- [54] Y. Wang, B. Li, T. Che, K. Zhou, Z. Liu, and D. Li, "Energy-based open-world uncertainty modeling for confidence calibration," in *Int. Conf. Comput. Vis.*, 2021.
- [55] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized ODIN: Detecting Out-of-Distribution Image Without Learning from Out-of-Distribution Data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10 951–10 960.
- [56] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating Neural Network Overconfidence with Logit Normalization," in *Int. Conf. Mach. Learn.*, 2022, pp. 23 631–23 644.
- [57] J. Bitterwolf, A. Meinke, and M. Hein, "Certifiably adversarially robust detection of out-of-distribution data," in *Adv. Neural Inform. Process. Syst.*, 2020.
- [58] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "Robust out-of-distribution detection for neural networks," *arXiv preprint arXiv:2003.09711*, 2020.
- [59] S. Choi and S.-Y. Chung, "Novelty detection via blurring," in *Int. Conf. Learn. Represent.*, 2020.
- [60] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *Adv. Neural Inform. Process. Syst.*, 2019.
- [61] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
- [62] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.
- [63] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv preprint arXiv:1912.02781*, 2019.
- [64] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt, "PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 16 783–16 792.
- [65] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *Adv. Neural Inform. Process. Syst.*, 2018.
- [66] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
- [67] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
- [68] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in *Int. Conf. Learn. Represent.*, 2018.
- [69] S. Vernekar, A. Gaurav, V. Abdelzad, T. Denoudin, R. Salay, and K. Czarnecki, "Out-of-distribution detection in classifiers via generation," in *Adv. Neural Inform. Process. Syst. Worksh.*, 2019.
- [70] K. Sricharan and A. Srivastava, "Building robust classifiers through generation of confident out of distribution examples," in *Adv. Neural Inform. Process. Syst. Worksh.*, 2018.
- [71] T. Jeong and H. Kim, "Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification," in *Adv. Neural Inform. Process. Syst.*, 2020.
- [72] S. Dai, J. Li, L. Wang, C. Zhu, Y. Wu, and X. Li, "Unsupervised Learning of Multi-level Structures for Anomaly Detection," *arXiv preprint arXiv:2104.12102*, 2021.
- [73] S. Dai, Y. Wu, X. Li, and X. Xue, "Generating and reweighting dense contrastive patterns for unsupervised anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1454–1462.
- [74] X. Du, Z. Wang, M. Cai, and Y. Li, "VOS: Learning What You Don't Know by Virtual Outlier Synthesis," in *Int. Conf. Learn. Represent.*, 2022.
- [75] L. Tao, X. Du, J. Zhu, and Y. Li, "Non-Parametric Outlier Synthesis," in *Int. Conf. Learn. Represent.*, 2022.
- [76] A. Shafaei, M. Schmidt, and J. J. Little, "A less biased evaluation of out-of-distribution sample detectors," in *Brit. Mach. Vis. Conf.*, 2019.

- [77] E. T. Nalisnick, A. Matsukawa, Y. W. Teh, D. Görür, and B. Lakshminarayanan, “Do deep generative models know what they don’t know?” in *Int. Conf. Learn. Represent.*, 2019.
- [78] S. Chewi, “Log-concave sampling,” *Book draft available at <https://chewisinho.github.io>*, 2023.
- [79] Y. Wang, Q. Zhang, Y. Wang, J. Yang, and Z. Lin, “Chaos is a Ladder: A New Theoretical Understanding of Contrastive Learning via Augmentation Overlap,” in *Int. Conf. Learn. Represent.*, 2022.
- [80] K. Shen, R. M. Jones, A. Kumar, S. M. Xie, J. Z. HaoChen, T. Ma, and P. Liang, “Connect, Not Collapse: Explaining Contrastive Learning for Unsupervised Domain Adaptation,” in *Int. Conf. Mach. Learn.*, 2022, pp. 19 847–19 878.
- [81] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in *Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.
- [82] X. Liu, Y. Lochman, and C. Zach, “GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23 946–23 955.
- [83] R. Huang and Y. Li, “MOS: Towards Scaling Out-of-Distribution Detection for Large Semantic Space,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8710–8719.
- [84] G. Chen, P. Peng, X. Wang, and Y. Tian, “Adversarial Reciprocal Points Learning for Open Set Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8065–8081, 2021.
- [85] T. DeVries and G. W. Taylor, “Learning Confidence for Out-of-Distribution Detection in Neural Networks,” *arXiv preprint arXiv:1802.04865*, 2018.
- [86] Y. Ming, Y. Sun, O. Dia *et al.*, “How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection?” in *Int. Conf. Learn. Represent.*, 2022.
- [87] A. Krizhevsky, G. Hinton *et al.*, “Learning Multiple Layers of Features from Tiny Images,” 2009.
- [88] Y. Le and X. Yang, “Tiny ImageNet Visual Recognition Challenge,” *CS 231N*, vol. 7, no. 7, p. 3, 2015.
- [89] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-Based Learning Applied to Document Recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [90] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, “Multi-Digit Number Recognition from Street View Imagery Using Deep Convolutional Neural Networks,” *arXiv preprint arXiv:1312.6082*, 2013.
- [91] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, “Describing Textures in the Wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3606–3613.
- [92] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million Image Database for Scene Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, 2017.
- [93] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, “Deep Learning for Classical Japanese Literature,” *arXiv preprint arXiv:1812.01718*, 2018.
- [94] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, “EMNIST: Extending MNIST to Handwritten Letters,” in *Int. Joint Conf. Neural Netw. (IJCNN)*. IEEE, 2017, pp. 2921–2926.
- [95] J. Davis and M. Goadrich, “The Relationship Between Precision-Recall and ROC Curves,” in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 233–240.
- [96] T. Fawcett, “An Introduction to ROC Analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.
- [97] A. Bendale and T. E. Boult, “Towards Open Set Deep Networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1563–1572.
- [98] J. Ren, S. Fort, J. Liu, A. G. Roy, S. Padhy, and B. Lakshminarayanan, “A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection,” *arXiv preprint arXiv:2106.09022*, 2021.
- [99] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [100] S. Zagoruyko and N. Komodakis, “Wide Residual Networks,” in *Brit. Mach. Vis. Conf.*, 2016.
- [101] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.
- [102] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, “LSUN: Construction of a Large-Scale Image Dataset Using Deep Learning with Humans in the Loop,” *arXiv preprint arXiv:1506.03365*, 2015.
- [103] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [104] W. J. Moon, J.-H. Kim, and J.-P. Heo, “Tailoring Self-Supervision for Supervised Learning,” in *Eur. Conf. Comput. Vis.*, 2022.

**Yifan Wu** (Student Member, IEEE) received the B.E. degree in intelligent science and technology from the School of Computer Engineering and Science, Shanghai University, Shanghai, China, in 2024. He is currently a research assistant at the School of Computer Science, Fudan University, Shanghai, China. His research interests include anomaly detection, out-of-distribution detection, and noisy label learning.

**Xichen Ye** received the B.E. degree in computer science and technology with the School of Information Science and Technology, Hangzhou Normal University, Zhejiang, China, in 2021, and the M.E. degree in computer science and technology with the School of Computer Engineering and Science, Shanghai University, Shanghai, China, in 2024. He is currently a research assistant at the School of Computer Science, Fudan University, Shanghai, China. His research interests include robust machine learning within the field of computer vision.

**Songmin Dai** received the B.E. degree in applied physics from the College of Sciences, Shanghai University, Shanghai, China, in 2016, and the Ph.D. degree in computer science and technology from the School of Computer Engineering and Science, Shanghai University, Shanghai, China, in 2021. He is currently a researcher with the Hithink RoyalFlush AI Research Institute, Hangzhou, China. His research interests include unsupervised anomaly detection and generation models within the field of computer vision.

**Dengye Pan** received the B.E. degree from Qian Weichang College, Shanghai University, Shanghai, China, in 2022, and is currently pursuing the M.S. degree in computer science and technology with the School of Computer Engineering and Science, Shanghai University, Shanghai, China. Her research interests include out-of-distribution detection within the field of computer vision.

**Xiaoqiang Li** (Member, IEEE) received the Ph.D. degree in computer science from Fudan University, Shanghai, China, in 2004. He is currently an Associate Professor of computer science with Shanghai University, China. He is currently an ACM Member and a Senior Member of the Chinese Computer Society. He is the Deputy Director of the Multimedia Special Committee of the Shanghai Computer Society. His current research interests include image processing, pattern recognition, computer vision, and machine learning. He has published over 100 conference and journal papers in these areas, including CVPR, ICCV, Neurips, IEEE TCSV, IEEE TMM, IEEE TII, IEEE TCYB, IEEE TIP, etc.

**Weizhong Zhang** received the B.S. and Ph.D. degrees from Zhejiang University in 2012 and 2017, respectively. He is currently a tenure-track Professor with the School of Data Science, Fudan University. His research interests include sparse neural network training, robustness, and out-of-distribution generalization. He has published over 40 conference and journal papers in these areas, including ICML, Neurips, ICLR, JMLR, IEEE TPAMI, IEEE TKDE, IEEE TIP, IEEE TIT, etc.

**Yifan Chen** received the B.S. degree from Fudan University, Shanghai, China, in 2018, and the PhD degree in Statistics from University of Illinois Urbana-Champaign in 2023. He is currently an assistant professor in computer science and math at Hong Kong Baptist University. He is broadly interested in developing efficient machine learning algorithms, encompassing both statistical and deep learning models. He has published several papers in these areas, including ICML, Neurips, KDD, etc.

# Supplementary Material for “Revisiting Energy-Based Model for Out-of-Distribution Detection”

## APPENDIX B SUPPLEMENTARY ALGORITHM FOR OOD DETECTION

In this section, we provide a supplementary description of the OOD detection process using the parameterized energy function as detailed in the main paper. Algorithm 2 outlines the steps to compute the energy-based score for a given test image sample and to determine its in- or out-of-distribution status based on a predefined threshold.

---

**Algorithm 2** OOD Detection Using the Parameterized Energy Function

---

**Require:** Classifier  $f_\theta$  with parameters  $\theta$ , temperature  $T$ , and threshold  $\tau$

- 1: **Input:** Test image sample  $x$
- 2: Compute the logits from the fully connected layer of  $f_\theta$ , denoted as  $f_\theta^{(i)}(x)$  for each class  $i = 1, \dots, C$
- 3: Compute the energy function  $E(x; f_\theta)$  using the logits:

$$E(x; f_\theta) = -\log \left( \sum_{i=1}^C \exp f_\theta^{(i)}(x) \right)$$

- 4: Define the score function  $s_\theta(x)$  as the negative energy:

$$s_\theta(x) = -E(x; f_\theta)$$

- 5: Set the OOD discriminator  $D(x; \tau, f_\theta)$  as:

$$D(x; \tau, f_\theta) = \begin{cases} 1, & \text{if } s_\theta(x) > \tau \\ 0, & \text{if } s_\theta(x) \leq \tau \end{cases}$$

- 6: **Output:**  $D(x; \tau, f_\theta)$ , where  $D = 1$  indicates OOD, and  $D = 0$  indicates ID
- 

## APPENDIX C EXPERIMENTS

### A. Datasets

In this section, we provide a detailed description of all the datasets used in our experiments:

**CIFAR-10** [87]: A dataset of 60,000 color images in 10 classes, with 50,000 images used for training and 10,000 for testing. Each image is 32x32 pixels.

**CIFAR-100** [87]: A dataset of 60,000 color images, categorized into 100 classes. It consists of 50,000 training images and 10,000 test images, with 600 images per class.

**MNIST** [89]: A dataset consisting of 70,000 grayscale images of handwritten digits, where 60,000 images are used for training and 10,000 for testing. Each image is 28x28 pixels.

**KMNIST** [93]: The Kuzushiji-MNIST (KMNIST) dataset consists of 70,000 grayscale images of handwritten Japanese characters, spanning 10 classes. It includes 60,000 images for training and 10,000 images for testing, with each image being 28x28 pixels in size.

**SVHN** [90]: A real-world dataset containing over 600,000 images of street view house numbers. It is split into 73,257 training images and 26,032 testing images, with an additional 531,131 extra training images. The dataset contains 10 digit classes.

**Tin** [88]: Tiny ImageNet is a popular dataset derived from the larger ImageNet dataset. It consists of 110,000 color images across 200 different classes. Each image has a resolution of 64x64 pixels, which is smaller than the original ImageNet dataset.

**Textures** [91]: A dataset of texture images with various surface patterns, used for evaluating models under non-object-like OOD settings.

**Place365** [92]: A scene recognition dataset containing 1.8 million images across 365 scene categories.

**EMNIST** [94]: The Extended MNIST (EMNIST) dataset contains 814,255 grayscale images of handwritten characters. In our experiments, we specifically denote EMNIST as the subset of the EMNIST dataset containing only handwritten English letters.

**FMNIST** [101]: The Fashion MNIST dataset contains 70,000 grayscale images of 10 different fashion items, including t-shirts, trousers, and shoes. Each image is 28x28 pixels.

**LSUN** [102]: The LSUN dataset is a large-scale dataset designed for scene understanding tasks, containing millions of labeled images across multiple scene and object categories. For our OOD experiments, we specifically use the LSUN-Crop subset, which consists of 10 scene categories. Each LSUN image has a larger resolution (typically 256x256 pixels), but for consistency in our experiments, the images are cropped to match the format of our in-distribution datasets.

**ImageNet** [103]: ImageNet is a large-scale dataset containing over 1.2 million color images categorized into 1,000 classes. In our experiments, we use the ImageNet-Resize subset, where the images have been resized to 32x32 pixels for consistency with other datasets. This subset is often used for OOD detection, as it provides a broad range of natural images.

TABLE VIII  
AUROC (%) FOR OOD DETECTION PERFORMANCE ON MNIST AND SVHN. THE TOP-1 RESULTS ARE IN **BOLD**.

Methods	MNIST→			SVHN→			
	FMNIST	EMNIST	CIFAR-10	CIFAR-10	CIFAR-100	LSUN-Crop	ImageNet-Resize
baseline [6]	97.2	88.3	99.6	93.8	93.5	94.5	93.9
CEDA [15]	99.4	89.5	<b>100.0</b>	96.0	95.9	98.4	95.5
ACET [15]	99.8	91.2	<b>100.0</b>	97.3	97.1	<b>99.7</b>	97.7
OEST (Ours)	<b>100.0</b>	95.7	<b>100.0</b>	99.4	<b>99.1</b>	99.5	<b>99.9</b>
OEST* (Ours)	<b>100.0</b>	<b>96.1</b>	<b>100.0</b>	<b>99.5</b>	<b>99.1</b>	<b>99.7</b>	<b>99.9</b>

### B. Experiments on MNIST and SVHN

To validate the broad applicability of our proposed method, we also conduct experiments utilizing MNIST [89] and SVHN [90] as in-distribution datasets. When MNIST [89] is used as the in-distribution dataset, FMNIST [101], EMNIST [94], and CIFAR-10 [87] are adopted for OOD testing. For SVHN [90] as the in-distribution dataset, we evaluate using CIFAR-10 [87], CIFAR-100 [87], LSUN-Crop [102], and ImageNet-Resize [103] as OOD datasets.

For MNIST, we first train a LeNet model [89] as the pretrained model using the SGD optimizer with a learning rate of 0.01, momentum of 0.9, and a weight decay of  $5 \times 10^{-4}$  for 60 epochs. We then fine-tune this pretrained model for an additional 10 epochs, still using the SGD optimizer. During fine-tuning, the learning rate follows a cosine annealing schedule, starting at  $1 \times 10^{-4}$  and decaying gradually to  $1 \times 10^{-8}$ . Four simple transformations—noise, blur, perm, and sobel—are applied to generate the peripheral-distribution data.

For SVHN, we train a ResNet-18 model [99] as the pretrained model using the SGD optimizer with a learning rate of 0.1, momentum of 0.9, and a weight decay of  $5 \times 10^{-4}$  for 100 epochs. We then fine-tune this pretrained model for an additional 10 epochs, once again using the SGD optimizer. During fine-tuning, the learning rate follows a cosine annealing schedule, starting at  $1 \times 10^{-4}$  and gradually decaying to  $1 \times 10^{-8}$ . To generate peripheral-distribution data, we apply six simple transformations: noise, blur, perm, rotation, and sobel.

As shown in Table VIII, our method continues to demonstrate superior performance compared to other baselines.

### C. Experimental Analysis

1) *Comparison with the Contrastive Training Scheme.*: We also observed that CSI performs poorly in this case, largely due to certain image categories having insignificant appearance differences even after transformations like rotation, which leads to model confusion. CSI relies on a single transformation type and treats the transformed images as negative class samples, which may not sufficiently capture the complexities of the data. In contrast, our method introduces several innovative improvements. By integrating multiple data transformation techniques and incorporating the concept of peripheral-distribution, our approach addresses these limitations. This is reflected in the performance gains, where we significantly outperform CSI in both AUROC and FPR95.

For the scheme of [37] and [104], both of them aim to figure out better data augmentation operation to generate negative samples for contrastive learning. Specifically, [37] proposes rotation, and [104] proposes two novel transformations, named LoRot-I and LoRot-W. However, ours will not be troubled by this, because every transformation can be useful in our training scheme just as shown in Table IV. Furthermore, considering a simple effective transformation, rotation, as CSI shown in Table II and rotation in Table IV, our training scheme is superior to CSI in all six benchmarks, which means ours is a better scheme to make full use of it.

2) *Hyper-parameters Analysis*: We conducted a systematic analysis of the hyper-parameters  $\alpha$  and  $\beta$  to evaluate their impact on the OOD detection performance of the ResNet-18 classifier. The results are shown in Figure 4, with AUROC and Accuracy representing the performance metrics to be maximized, and FPR95 representing the robustness metric to be minimized.

First, we analyzed the effect of  $\alpha$  by fixing  $\beta$  at 10, as shown in Figure 4a. As  $\alpha$  increases from 0.1 to 1.0, AUROC consistently increases and FPR95 steadily decreases, indicating that larger  $\alpha$  values improve the model's ability to distinguish in-distribution (ID) and out-of-distribution (OOD) samples. However, this improvement comes at the cost of a decrease in

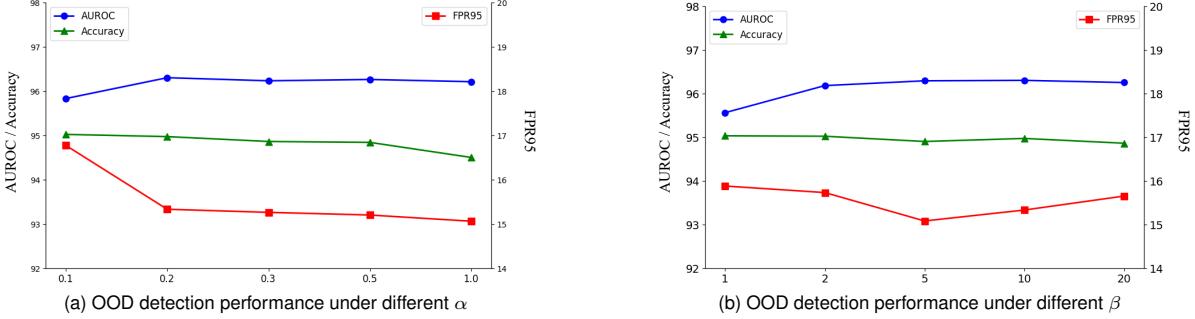


Fig. 4. OOD detection performance of a ResNet-18 classifier trained on CIFAR-10 as the in-distribution dataset, evaluated under varying values of hyperparameters  $\alpha$  and  $\beta$ . In (a),  $\beta$  is fixed at 10, and the effect of changing  $\alpha$  is shown. In (b),  $\alpha$  is fixed at 0.2, and the impact of changing  $\beta$  is shown. Higher AUROC and Accuracy for both experiments indicate better performance, while lower FPR95 reflects better performance.

classification accuracy for in-distribution samples, which is an undesirable side effect. To balance these competing objectives, we chose  $\alpha = 0.2$  as the final value, which provides a compromise between maximizing OOD detection performance and maintaining satisfactory in-distribution accuracy.

Next, we investigated the effect of  $\beta$  by fixing  $\alpha$  at 0.2, as shown in Figure 4b. As  $\beta$  increases from 0.05 to 10, AUROC shows an overall increasing trend, suggesting improved OOD detection capability. However, this improvement comes at the cost of a slight decrease in classification accuracy for in-distribution samples. Meanwhile, FPR95 initially decreases, reflecting enhanced robustness, but starts to increase again beyond certain values of  $\beta$ . To balance these effects, we chose  $\beta = 10$  as the final value, which provides a reasonable trade-off between maximizing OOD detection performance and maintaining satisfactory in-distribution accuracy.

Based on these observations, we selected  $\alpha = 0.2$  and  $\beta = 10$  as the optimal hyper-parameter settings for our final model.

TABLE IX

TO VERIFY THE IMPORTANCE OF PRE-TRAIN + FINE-TUNE PROCESS, WE USE CIFAR-10 AS IN-DISTRIBUTION DATA AND ROTATION TO GENERATE PERIPHERAL-DISTRIBUTION DATA FOR FINE-TUNE.

$D_{in}^{train}$	pre-train + fine-tune	AUROC $\uparrow$	FPR95 $\downarrow$
CIFAR-10	training from scratch ✓	92.1 <b>96.3</b>	31.1 <b>15.3</b>

3) *The Importance of Fine-Tune*: To further validate the necessity of both the pre-train and fine-tune steps, we conducted experiments as illustrated in Table IX. Compared to training from scratch, the pre-train plus fine-tune scheme yields better results, effectively enhancing the model's performance and reducing the false positive rate. A possible reason is that training from scratch with augmented samples together with the original samples can lead the model to prematurely learn the pattern differences between samples. However, these differences are merely low-level semantic information at the texture and color levels. In contrast, when training is divided into pre-train plus fine-tune stages, the model has already learned higher-level semantic information during the pre-train stage. This can help the model better understand the differences between the data-augmented samples and the original samples during the fine-tune stage, leading to superior out-of-distribution detection results. This experimental outcome further validates the necessity of pre-train and fine-tune in enhancing model performance. Also, this scheme can reduce the consumption of computational resources, as we can use well-trained classifier and only need to fine-tune 10 epochs.

4) *Visualization of t-SNE*: We performed t-SNE visualization of the features before and after fine-tune to provide a clear illustration. As shown in Figure 1a, red represents the test samples of CIFAR-10, orange represents the rotated CIFAR-10 samples, blue, purple and green represent three out-of-distribution test datasets of CIFAR-100, SVHN, and ImageNet, respectively. Feature embeddings are obtained from the ultimate convolutional layer. We observed that rotated CIFAR-10 are located in the peripheral area of in-distribution, which supports the notion of the peripheral augmented samples lying between the in-distribution and out-of-distribution samples. Moreover, by assigning different energy scores to the rotated CIFAR-10 samples and the original samples during finetune, the decision boundaries of the classifier become much clearer and can effectively discriminate the CIFAR-100 with the highest similarity to the original distribution samples.