Out-of-Distribution Detection with Adversarial Outlier Exposure

Thomas Botschen[†] Konstantin Kirchheim[†] Frank Ortmeier

Department of Computer Science
University of Magdeburg, Germany
{firstname}.{lastname}@ovgu.de

Abstract

Machine learning models typically perform reliably only on inputs drawn from the distribution they were trained on, making Out-of-Distribution (OOD) detection essential for safety-critical applications. While exposing models to example outliers during training is one of the most effective ways to enhance OOD detection, recent studies suggest that synthetically generated outliers can also act as regularizers for deep neural networks. In this paper, we propose an augmentation scheme for synthetic outliers that regularizes a classifier's energy function by adversarially lowering the outliers' energy during training. We demonstrate that our method improves OOD detection performance and adversarial robustness on OOD data on several image classification benchmarks. Additionally, we show that our approach preserves in-distribution generalization. Our code is publicly available.[1]

1. Introduction

Out-of-Distribution (OOD) detection, which aims to identify inputs with low probability under the training distribution of a machine learning model, has garnered significant attention in recent years [34]. State-of-the-art methods are largely based on outlier exposure (OE), which improves OOD detection by training models against an auxiliary dataset of outliers [18]. Additionally, baseline classifiers can be interpreted as energy-based models [14, 25], where inputs are assigned a scalar energy value – low for in-distribution (ID) data and high for OOD data – making energy-based methods an efficient approach for OOD detection.

Beyond naturally occurring OOD samples, deep neural networks (DNNs) are also susceptible to adversarial examples – carefully manipulated inputs that resemble ID data yet cause high-confidence misclassifications. A common strategy to increase adversarial robustness is adversarial training, where models are trained against examples subjected to specific adversarial perturbations [13]. However, this approach has been shown to degrade generalization on ID data [32].

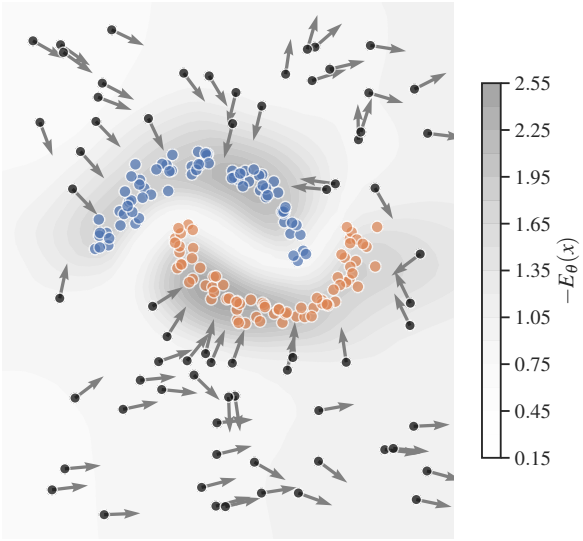


Figure 1. Illustration of Adversarial Outlier Exposure (AOE) on a two-dimensional dataset. AOE moves the training outliers (black) along the gradient of the model's energy function. Throughout the training, this process adjusts the energy function to assign higher outlier scores to augmented outliers, ultimately tightening the decision boundary.

While adversarial robustness and OOD detection are often studied independently, Song *et al.* [30] explored the relationship between them, identifying a tradeoff: improving robustness against adversarial attacks tends to weaken OOD detection performance.

In this work, we propose Adversarial Outlier Exposure (AOE), a novel approach that integrates adversarial training with outlier exposure by augmenting synthetic outliers with adversarial perturbations. Specifically, we modify outliers sampled from generative models by perturbing them along the direction of the classifier's energy gradient. This process

[†]These authors contributed equally to this work.
[1]https://github.com/2mey10/Adversarial-OE

regularizes the classifier's energy function in low-density regions of the input space, thereby enhancing its robustness. For an intuitive example, see Fig. 1.

Our contributions are as follows:

1. We introduce a novel adversarial training framework for outlier exposure, in which synthetic outliers are augmented with adversarial perturbations that decrease their energy-based outlier scores.
2. We demonstrate that our method not only improves adversarial robustness against OOD data but also preserves classification accuracy on ID data.
3. We provide empirical evidence that our approach enhances OOD detection performance, even for non-adversarial OOD data.

The remainder of this paper is structured as follows. First, we provide an overview of recent developments in OOD detection and adversarial training in Sec. 2. Next, we introduce our proposed Adversarial Outlier Exposure method in Sec. 3. In the following Sec. 4, we review related work. In Sec. 5, we present extensive experiments and ablation studies, demonstrating the effectiveness of AOE across multiple datasets. Finally, we conclude our work in Sec. 6.

2. Background

In the following, we will consider a classifier $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^K$ that maps inputs $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^M$ to logits to model the conditional distribution $p(y \mid \mathbf{x})$ by learning from a dataset sampled from a training distribution $p(\mathbf{x})$.

2.1. Out-of-Distribution Detection

Classifiers, as discussed above, can produce highly confident yet incorrect predictions for inputs that are unlikely under the data distribution $p(\mathbf{x})$. Any input $\mathbf{x} \in \mathcal{X}$ satisfying $p(\mathbf{x}) < \alpha$, for some suitably small threshold α, can be considered OOD – that is, unlikely to have been drawn from p. While we will use this definition of OOD in the following, it should be noted that other definitions exist. [34]

To identify such OOD points, a detector $D(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$ maps inputs to outlier scores. Classification as ID or OOD is determined by applying a threshold τ:

$$\text{outlier}(\mathbf{x}) = \begin{cases} \text{ID} & \text{if} \quad D(\mathbf{x}) < \tau, \\ \text{OOD} & \text{else.} \end{cases} \tag{1}$$

In recent years, OOD detection has developed into a broad field, with several publications providing comprehensive surveys [26, 34]. The following provides a brief overview.

Posteriors From the logits produced by the classifier, a posterior probability distribution over classes can be obtained via the softmax function. A baseline OOD detection method is Maximum Softmax Probability (MSP) [17], which uses $-\max_y p(y \mid x)$ as an outlier score, where p is the posterior predicted by the model. Other methods based on statistics of the posterior include KL-matching [16] and Monte Carlo Dropout [11].

Logits Logit-based methods omit the softmax normalization and directly compute outlier scores from the classifier's output. Examples include MaxLogit [16], which uses the negative maximum of the logits, $-\max_y f_{\boldsymbol{\theta}}(\mathbf{x})_y$, where $f_{\boldsymbol{\theta}}(\mathbf{x})_y$ denotes the y^{th} logit, and energy-based OOD detection (see Sec. 2.2), which can be viewed as a smooth approximation of the MaxLogit method.

Latent Representations Feature-based methods operate on latent representations from intermediate layers of a DNN classifier. Examples include the Mahalanobis method [23], which models the latent features of each class using a Gaussian distribution and computes the Mahalanobis distance as an outlier score. Simplified Hopfield Energy (SHE) [38] learns a center $\boldsymbol{\mu}_y$ for each ID class and uses $-\boldsymbol{\mu}_y^\top \phi(\mathbf{x})$ as the outlier score, where y is the maximum a posteriori class for $\mathbf{x}$ and $\phi(\mathbf{x})$ denotes its latent representation.

2.2. Energy-based OOD Detection

Grathwohl *et al.* [14] demonstrated that any classifier – as described above – can be reinterpreted as an energy-based model for the data distribution $p(\mathbf{x})$. Specifically, the density function takes the form

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{\exp(-E_{\boldsymbol{\theta}}(\mathbf{x}))}{Z(\boldsymbol{\theta})} \propto \exp(-E_{\boldsymbol{\theta}}(\mathbf{x})), \tag{2}$$

where $E_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}$ is the *energy function*, and $Z(\boldsymbol{\theta})$ is the *partition function*, which ensures proper normalization. For a classifier $f_{\boldsymbol{\theta}}$, the energy function is defined as

$$E_{\boldsymbol{\theta}}(\mathbf{x}) \triangleq -\log \sum_{i=1}^{K} \exp(f_{\boldsymbol{\theta}}(\mathbf{x})_i), \tag{3}$$

where $f_{\boldsymbol{\theta}}(\mathbf{x})_i$ denotes the i^{th} logit. As can be seen from Eq. (2), the energy function is theoretically aligned with the data density $p_{\boldsymbol{\theta}}(\mathbf{x})$, in the sense that higher energy corresponds to lower density. This makes the energy a natural criterion for OOD detection. Building on this interpretation, Liu *et al.* [25] introduced energy-based OOD detection (EBO), which utilizes the energy function to distinguish ID and OOD samples.

2.3. Outlier Exposure

Outlier exposure (OE), introduced by Hendrycks *et al.* [18], significantly improves outlier detection in neural networks by using a dataset of auxiliary outliers. OE achieves this by incorporating an additional penalty term, $\mathcal{L}_{OE}$, into the

training objective, promoting lower confidence for training outliers. The complete objective function is defined as:

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{ID}}}\Big[\mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}),y)\Big] + \lambda\mathbb{E}_{\mathbf{x}'\sim\mathcal{D}_{\text{OOD}}^{\text{train}}}\Big[\mathcal{L}_{\text{OE}}(f_{\boldsymbol{\theta}}(\mathbf{x}'),f_{\boldsymbol{\theta}}(\mathbf{x}),y)\Big], \quad (4)$$

where $\mathcal{D}_{\text{ID}}$ and $\mathcal{D}_{\text{OOD}}^{\text{train}}$ represent the ID and OOD training distributions, respectively, $\mathcal{L}$ is the loss used for ID data – typically cross-entropy – and $\mathcal{L}_{\text{OE}}$ is cross-entropy between the predicted posterior for $\mathbf{x}'$ and the uniform distribution:

$$\mathcal{L}_{OE} = \frac{1}{K}\sum_{i=1}^{K} f_{\boldsymbol{\theta}}(\mathbf{x}')_i - \log\sum_{i=1}^{K}\exp(f_{\boldsymbol{\theta}}(\mathbf{x}')_i)\,. \quad (5)$$

Variants of OE that directly regularize the energy function rather than the posteriors have also been proposed [25].

Synthetic Outliers While outlier exposure reliably increases the OOD detection performance of DNNs, it is based on the assumption that a set of representative example outliers is available during model optimization. However, it is possible to instead sample outliers from low likelihood regions of generative models of ID data, which can be done in the input space $\mathcal{X}$ [9, 21], as well as some latent space of the classifier $f_{\boldsymbol{\theta}}$ [10]. This strategy eliminates the need for a curated set of outliers and allows to flexibly parameterize $\mathcal{D}_{\text{OOD}}^{\text{train}}$ with various types of generators, such as Generative Adversarial Networks (GAN) [12] or denoising diffusion models [7]. Furthermore, the method can be used with pre-trained generative models, which makes it computationally efficient.

2.4. Adversarial Training

Adversarial attacks seek a small perturbation $\delta \in \mathbb{R}^M$ for an input $\mathbf{x}$ such that the classifier $f_{\boldsymbol{\theta}}$ produces an incorrect prediction for $\mathbf{x}+\delta$. Here, δ is usually norm-bounded by enforcing $\|\delta\|_p < \eta$, where $\|\cdot\|_p$ is some p-norm. [1, 4]

Various attack and defense strategies have been proposed. A computationally efficient, non-iterative attack is the Fast Gradient Sign Method (FGSM), defined as

$$\tilde{\mathbf{x}} \triangleq \mathbf{x} + \epsilon\ \text{sign}(\nabla_{\mathbf{x}}\mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}),y)), \quad (6)$$

where $\tilde{\mathbf{x}}$ is the adversarially perturbed input, and $\epsilon \in \mathbb{R}_+$ is a scalar step size. [13] Iterative methods such as Projected Gradient Descent (PGD) [27] offer stronger attacks at the cost of increased computational complexity.

A common method to increase the robustness of classifiers against such attacks is adversarial training, which involves training on adversarially perturbed images. However, training on ID data with adversarial perturbations has been shown to degrade generalization to clean in-distribution samples [32] and weaken the model's ability to detect OOD data [30].

3. Adversarial Outlier Exposure

While adversarial attacks are typically used to improve adversarial robustness on in-distribution training data, we propose applying adversarial perturbations to out-of-distribution data used during training instead. Such an attack can be performed by reducing the energy of an OOD input $\mathbf{x}$ – thus lowering its energy-based outlier score – by descending along the gradient of the model's energy function $\nabla_{\mathbf{x}}E_{\boldsymbol{\theta}}(\mathbf{x})$.

Inspired by FGSM (Eq. (6)), we propose to perturb the auxiliary training outliers with the sign of the gradient:

$$\hat{\mathbf{x}} \triangleq \mathbf{x} - \epsilon\ \text{sign}(\nabla_{\mathbf{x}}E_{\boldsymbol{\theta}}(\mathbf{x})). \quad (7)$$

Since FGSM is non-iterative, the magnitude of this perturbation δ in terms of the ℓ_∞-norm is simply

$$\|\epsilon\ \text{sign}(\nabla_{\mathbf{x}}E_{\boldsymbol{\theta}}(\mathbf{x}))\|_\infty = \|\delta\|_\infty = \epsilon. \quad (8)$$

This adversarial perturbation can shift $\mathbf{x}$ into a lower-energy region, which corresponds to a higher likelihood under the energy-based interpretation of Eq. (2). The second term in the outlier exposure objective (Eq. (4)) then becomes:

$$\lambda\mathbb{E}_{\mathbf{x}'\sim\mathcal{D}_{\text{OOD}}^{\text{train}}}\Big[\mathcal{L}_{\text{OE}}(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}'),f_{\boldsymbol{\theta}}(\mathbf{x}),y)\Big]. \quad (9)$$

During optimization, the model adapts its parameters to map perturbed OOD samples to high-energy regions, reinforcing their separation from ID data.

Intuition An intuitive visualization of this approach for a two-dimensional setting is shown in Fig. 1. Taking a step along the negative energy gradient of the classifier $f_{\boldsymbol{\theta}}$ can move OOD samples into lower-energy regions, possibly closer to ID samples. By training the model to reject such adversarially perturbed outliers, we expect the resulting decision boundary to be tighter.

Some examples of synthetic OOD images with increasing perturbation strength are provided in Fig. 2. Perturbed OOD images do not resemble ID data but tend to have lower energy-based outlier scores.

A histogram of energy values for ID and OOD samples from CIFAR-100, both pre- and post-perturbation, is shown in Fig. 3. We observe that applying small perturbations lowers the energy of OOD samples, aligning them more closely with the energy distribution of ID samples. This observation suggests that adversarially perturbed OOD samples could be misclassified as ID with high confidence. Training on such samples encourages the model to increase the sample's energy, thereby improving OOD detection.

Perturbing ID Data In principle, ID inputs could also be perturbed to increase their energy, incentivizing the model to maintain low energy scores for slightly modified inliers.

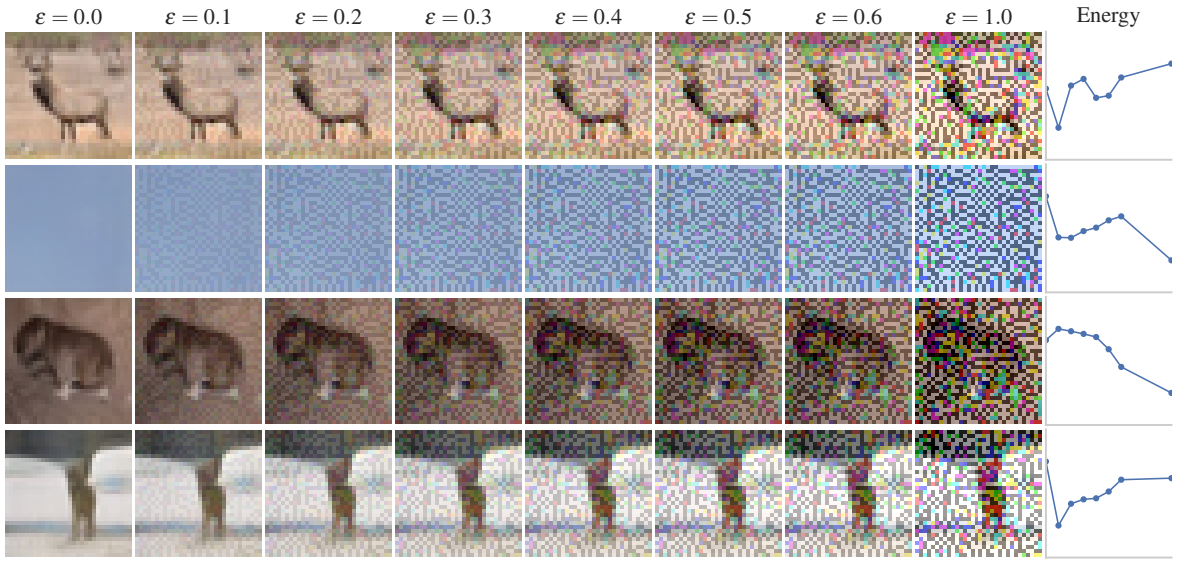


Figure 2. Synthetic outliers drawn from a BigGAN trained on CIFAR-100 (sampled at $\sigma^2 = 2$) with increasingly strong adversarial perturbation targeting the classifier's energy function. Perturbed OOD images do not resemble ID data yet tend to receive low energy-based outlier scores. Results are based on the pre-trained WideResNet provided by [17].

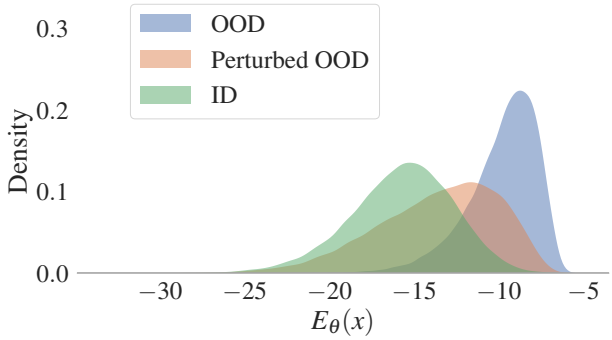


Figure 3. Distribution of energy scores for CIFAR-100 ID test data, as well as synthetic outliers before and after applying the adversarial perturbation with $\epsilon = 0.1$. The distribution of energy scores of the perturbed OOD samples more closely resembles the distribution of ID samples.

However, in line with previous findings, we observed that this significantly degrades performance. As a result, we restrict our augmentation scheme to OOD samples only.

4. Related Work

While several related approaches have been explored in the literature, to the best of our knowledge, we are the first to propose the specific training scheme presented in this work. Below, we provide a brief overview of closely related methods:

Augmenting Outliers The idea of applying perturbations to inputs to improve OOD discrimination is well-established. ODIN [24] perturbs test-time inputs via gradient ascent and uses temperature scaling to distinguish between in- and out-of-distribution samples. PixMix [19], on the other hand, enhances robustness by mixing in-distribution images with outliers during training to improve safety metrics.

In contrast to these methods, our approach exclusively augments outliers during training with adversarial perturbations and leaves ID inputs unchanged.

Robust Optimization Hein *et al.* [15] propose an adversarial training scheme that maximizes the model's log-confidence, using Gaussian noise as auxiliary outliers and a modified Projected Gradient Descent (PGD) strategy. Similarly, Chen *et al.* [5] introduces an adversarial training method using real auxiliary outlier images and PGD. Bitterwolf *et al.* [2] proposes a loss that enforces low softmax confidence in an ℓ_∞-ball around OOD inputs by penalizing the maximum logit difference.

In contrast, our method employs the more efficient FGSM and directly targets the model's energy function, which has a theoretical connection to the data density (see Sec. 2.2). Furthermore, while our approach is agnostic to the choice of $\mathcal{D}_{\text{OOD}}^{\text{train}}$, in our experiments, we use synthetic outliers sampled from a GAN.

Regularizing the Energy Function Several works attempt to regularize a model's energy landscape using auxiliary outliers. For instance, the energy-bounded learning loss [25]

encourages a separation between ID and OOD samples in terms of energy.

Unlike these methods, our approach actively perturbs outliers during training to shape the energy function more directly.

Synthetic Outliers Several prior works explore different strategies for generating outliers. Du *et al.* [9] use a diffusion-based model to synthesize outliers, while in Virtual Outlier Synthesis, Du *et al.* [10] sample from low-likelihood regions of a Gaussian fitted to the latent representations of ID data. Hein *et al.* [15] utilize simple noise-based distributions to sample outliers.

Our method does not impose specific assumptions on the outlier generator and could be readily applied to augment arbitrary outliers.

5. Experiments

For our experiments, we use a WideResNet [37] pre-trained on the respective in-distribution dataset, as this model is widely used in OOD detection papers [17, 25]. We apply standard data augmentation techniques, including random cropping and flipping, and fine-tune the pre-trained models using stochastic gradient descent (SGD) with a Nesterov momentum of 0.9. All models use ℓ_2 regularization with a coefficient of 5×10^{-4} and a cosine annealing schedule with an initial learning rate of 5×10^{-4} that decays gradually over 10 epochs. For AOE and OE, we set $\lambda = 0.5$ and for AOE $\epsilon = 0.5$ if not specified otherwise. Our implementation is based on PyTorch [29] and PyTorch-OOD [20].

The confidence intervals provided in the following figures correspond to the standard error of the mean of the respective metrics as estimated over eight different OOD datasets.

5.1. Datasets

As in-distribution data, we use CIFAR-10/100 [22].

Training Outliers As auxiliary training outliers $\mathcal{D}_{\text{OOD}}^{\text{train}}$, we use $\approx 10,000$ artificially generated images sampled from a BigGAN trained exclusively on ID data [3, 21]. To generate outliers, we sample from this generative model with $\sigma^2 = 2$. Some examples are provided in Fig. 2.

Test Outliers We test all models against eight different OOD datasets: resized images from Fooling Images [28] and Textures [6], cropped and resized TinyImageNet [8], and randomly cropped and resized images from the Large-Scale Scene Understanding dataset (LSUN) [36]. Additionally, we included 1000 samples of Gaussian and uniform noise each.

5.2. OOD Detection Metrics

We evaluate each method's OOD detection performance using two different metrics, averaging results across all outlier datasets.

AUROC The area under the receiver operating characteristic curve (ROC) measures the tradeoff between the false positive and true positive rates, providing a threshold-independent metric for binary classification. It ranges from 0 to 1, with higher values indicating better OOD detection performance. A value of 0.5 corresponds to random guessing.

FPR95 The false positive rate at 95% true positive rate represents the false positive rate at the threshold τ (see Eq. (1)) where the true positive rate is 95%. It corresponds to a single point on the ROC curve.

5.3. In-Distribution Classification

While adversarial training on in-distribution data typically increases robustness, it often reduces performance on clean ID samples. We compare the ID classification accuracy in Tab. 1. The results suggest that our training scheme does not significantly impact ID classification accuracy compared to adversarial training on ID data.

Table 1. In-distribution classification performance. All values in percent. Best values underlined. Even though we use an adversarial training scheme, our approach mostly maintains ID classification performance.

	Accuracy ↑	
Training Scheme	CIFAR-10	CIFAR-100
Cross-Entropy	<u>94.6</u>	<u>75.5</u>
Outlier Exposure [18]	94.5	<u>75.5</u>
Outlier Exposure + Noise	94.5	75.0
AOE (Ours)	94.4	74.8

Figure 4 shows the ID classification accuracy of models trained with different values of ϵ. We observe a steady but slight accuracy drop over the entire ϵ range, with a maximum reduction of 0.35% and 1.0% for CIFAR-10 and 100, respectively. These results indicate that AOE does not significantly degrade ID classification accuracy in the tested ϵ range.

5.4. Out-of-Distribution Detection

Results for OOD detection are provided in Tab. 2, showing that AOE outperforms outlier exposure for both the MSE and the EBO detector, as well as other methods that do not utilize auxiliary outliers during training.

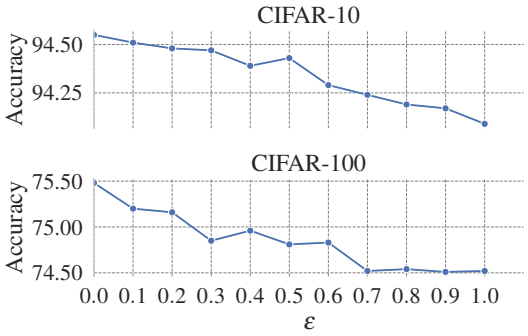


Figure 4. Classification accuracy for different levels of ϵ. Small perturbations only slightly impact ID classification performance.

The average AUROC for different values of ϵ is shown in Fig. 5. Note that $\epsilon = 0$ corresponds to vanilla outlier exposure without any perturbation. We observe that AOE consistently improves OOD detection performance. Performance increases as attack strength grows until saturation and then slightly decreases again. These trends hold for both CIFAR-10 and CIFAR-100.

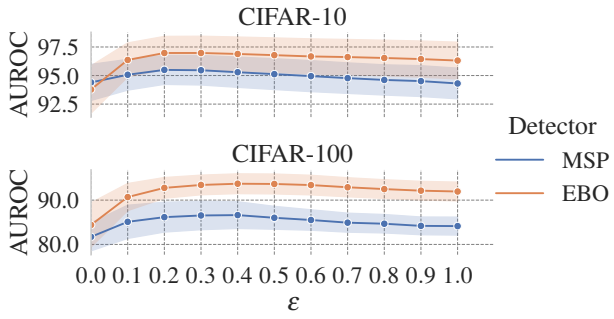


Figure 5. OOD detection performance vs. perturbation strength ϵ used during training. Note that $\epsilon = 0$ corresponds to standard outlier exposure.

5.5. Adversarial Robustness on OOD Data

Given that AOE trains models to assign high outlier scores to adversarially manipulated OOD data, we expect the training to increase the model's robustness to adversarially perturbed OOD data, similar to adversarial training on ID data.

To perform adversarial attacks on an OOD data point $\mathbf{x}$, we modify the point to maximize its outlier exposure loss $\mathcal{L}_{OE}$ as defined by Eq. (5), using FGSM. This attack incentivizes OOD samples to receive high maximum softmax values. We evaluate the adversarial robustness of models against an FGSM adversary that maximizes $\mathcal{L}_{OE}$ with varying attack budgets $\|\delta\|_\infty$. The results are shown in Fig. 6. When comparing the OOD detection performance of models trained with AOE to those trained with vanilla OE, we observe that AOE-trained models are significantly more robust across all perturbation levels and never degrade to random guessing. In contrast, for models trained with vanilla outlier exposure, perturbations with $\epsilon > 0.75$ reduce AUROC values below random guessing for some OOD datasets. Additionally, performance degradation occurs more gradually in models trained with AOE, as indicated by the lower slope in AUROC decline.

5.6. Near and Far OOD Data

OOD data can be broadly categorized into near OOD, which shares visual similarity with the ID data, and far OOD, which is more perceptually distinct [35]. In Fig. 7, we report CIFAR-100 results for OE and AOE for individual OOD datasets, ordered from far to near OOD. While all approaches show degraded performance on datasets that are more similar to the IN data, AOE overall maintains more stable performance across the near-to-far OOD continuum.

5.7. Ablation Studies

This section provides additional ablation studies to examine the contribution of individual components and the influence of different design aspects of our approach.

Comparison to Random Perturbations To better understand the contribution of our targeted perturbations, we contrast AOE with a model trained using random noise during outlier exposure. In this setting, auxiliary outliers are perturbed according to

$$\mathring{\mathbf{x}} = \mathbf{x}' + \delta \quad \text{with} \quad \delta \sim \mathcal{N}(\mathbf{0}, \epsilon \boldsymbol{I}) \tag{10}$$

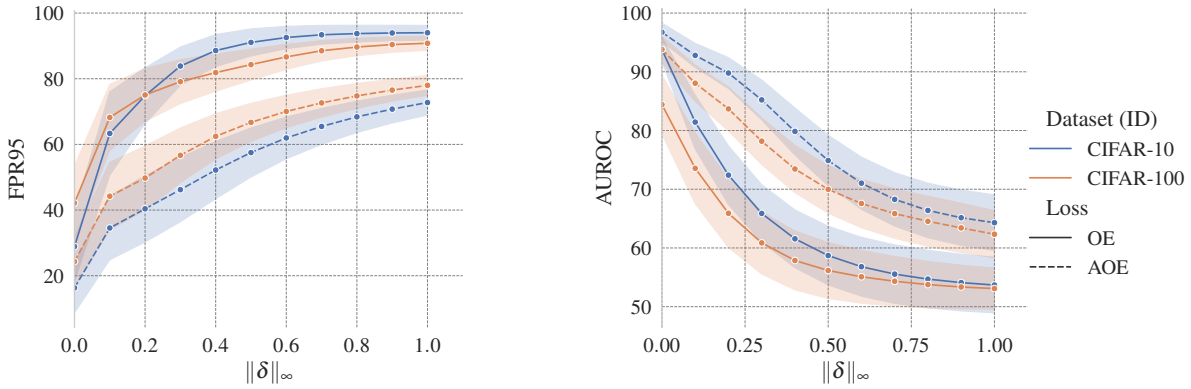
where $\mathbf{x}'$ represents the original auxiliary outlier. We evaluate performance across a range of ϵ values, as shown in Fig. 8. While injecting Gaussian noise can offer some gains, our gradient-guided perturbations consistently achieve higher OOD detection performance – particularly on the more challenging CIFAR-100 benchmark. These findings indicate that gradient-driven augmentations can enhance OOD detection more effectively than random noise.

Selection of Sampling Parameters Following [21], different sampling parameters σ^2 can be used to sample synthetic outliers from a noise-conditioned generative model. The impact of varying this hyperparameter is shown in Fig. 9. As expected, performance slightly improves for $\sigma^2 > 1$, where sampled outliers become increasingly dissimilar to ID data. However, for larger σ^2 values, performance remains relatively stable over a wide range of values, demonstrating the robustness of our approach to this hyperparameter.

Convergence Figure 10 depicts the performance over fine-tuning epochs. Most performance gains are obtained during

Table 2. OOD detection performance on CIFAR-10 and CIFAR-100. Higher AUROC values and lower FPR95 are preferable. Best values underlined. All values in percent.

Loss	Detector	ϵ	CIFAR-10		CIFAR-100	
			AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Cross-Entropy	ODIN [24]		84.0	43.5	80.2	53.0
Cross-Entropy	SHE [38]		84.6	48.3	56.1	100.0
Cross-Entropy	DICE [31]		87.0	43.6	82.0	47.1
Cross-Entropy	MaxLogit [16]		87.1	42.9	82.1	47.6
Cross-Entropy	KL-Matching [16]		87.2	56.3	81.0	53.3
Cross-Entropy	ViM [33]		95.1	20.5	90.1	31.1
Cross-Entropy	Mahalanobis [23]		95.6	17.0	86.8	36.3
Cross-Entropy	MSP [17]		89.4	33.1	79.2	54.2
Outlier Exposure [18]	MSP [17]		94.4	21.2	81.7	48.8
Outlier Exposure + Noise	MSP [17]	0.5	94.7	17.4	84.5	43.5
AOE (ours)	MSP [17]	0.5	95.1	16.9	86.1	42.0
Cross-Entropy	EBO [25]		86.9	43.7	81.8	47.5
Outlier Exposure [18]	EBO [25]		93.8	28.9	84.4	42.4
Outlier Exposure + Noise	EBO [25]	0.5	96.6	16.3	91.1	31.2
AOE (ours)	EBO [25]	0.5	96.8	16.0	93.7	24.7



(a) FPR95 (b) AUROC

Figure 6. We investigate the robustness of the EBO detector [25] against adversarial attacks on OOD data during test time. The plot illustrates FPR95 and AUROC for various attack budgets $\|\delta\|_\infty$, where $\|\delta\|_\infty = 0$ represents baseline performance. As we can see, the performance for models trained with AOE deteriorates less rapidly and always significantly exceeds random guessing, while the detection performance for models trained with vanilla OE rapidly deteriorates as the attack budget increases.

the first three epochs, after which the performance tends to converge.

Number of Auxiliary Outliers Fig. 11 depicts the model's performance based on the number of different outliers used during fine-tuning. As we can see, performance tends to increase as we increase the number of outliers, but even a small number of outliers leads to reasonable performance.

This is in line with Kirchheim *et al.* [21], who observed that a larger number of auxiliary outliers is associated with increased performance gains. However, in our experiments, performance saturates earlier, which we attribute to additional data augmentation.

Different Adversaries As described above, AOE improves OOD detection metrics on CIFAR-10 and CIFAR-100 when applied with FGSM-style training, as defined in Eq. (7). However, the method can also be used with more computationally intensive adversaries, which may offer stronger perturbations. Interestingly, we found that training against stronger adversaries, such as Projected Gradient Descent [27], is consistently outperformed by FGSM.

6. Conclusion

In this paper, we proposed a novel method for regularizing a classifier's energy function by incorporating synthetic outliers enhanced with adversarial perturbations during training.

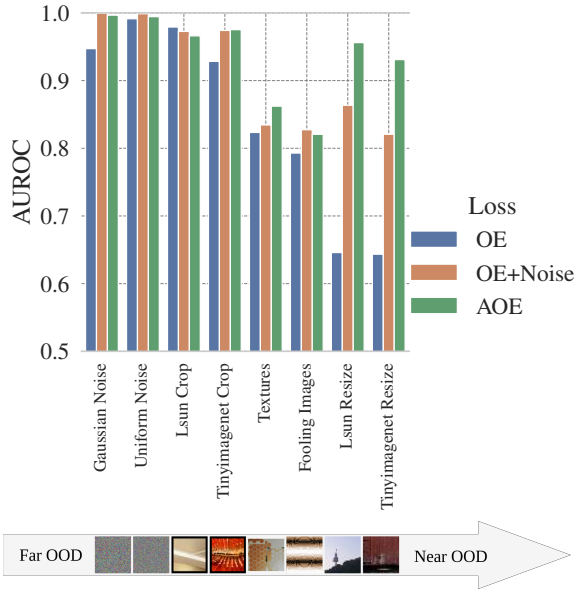


Figure 7. OOD detection performance of EBO on CIFAR-100 over individual OOD datasets for different training schemes.

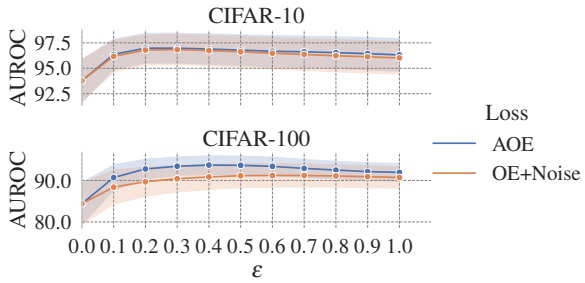


Figure 8. OOD detection performance of EBO detector for models trained with AOE vs. models trained with outliers augmented with Gaussian noise. $\epsilon = 0$ corresponds to vanilla outlier exposure. We observe that adversarial perturbations increase performance more than noise.

Our approach led to substantial improvements in OOD detection performance and adversarial robustness across multiple datasets while mostly preserving accuracy on in-distribution data.

Future research may explore the applicability of this method to higher-resolution datasets and alternative problem domains, such as image segmentation and object detection. Furthermore, evaluating its performance with real, non-synthetic outliers could provide valuable insights.

Acknowledgment

This research received funding from the *Federal Ministry for Economic Affairs and Climate Action (BMWK)* and the *European Union* under grant agreements 19I21039A.

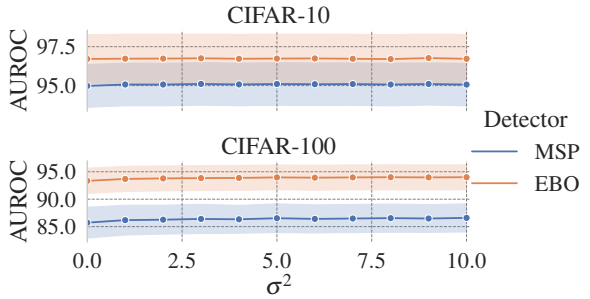


Figure 9. OOD detection performance for different values of the parameters σ^2 used to sample outliers from the BigGAN. Higher values produce more unrealistic auxiliary outliers on average. We observe that the results are stable, indicating the robustness of our method to this hyperparameter.

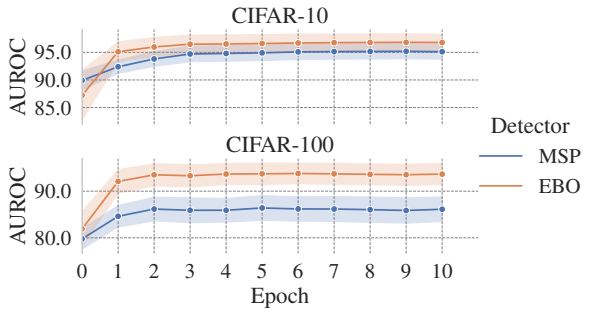


Figure 10. OOD detection performance over training epochs. Epoch 0 corresponds to pre-trained models without any AOE fine-tuning.

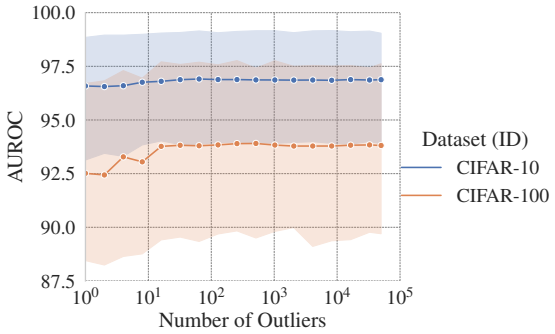


Figure 11. OOD detection performance for the EBO detector for models trained with AOE with varying numbers of synthetic outliers. While more outliers initially provide better performance, this effect saturates quickly.

References

[1] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9:155161–155196, 2021. 3

[2] Julian Bitterwolf, Alexander Meinke, and Matthias Hein. Certifiably adversarially robust detection of out-of-distribution data. *Advances in Neural Information Processing Systems*, 33:16085–16095, 2020. 4

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 5

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 3

[5] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 430–445. Springer, 2021. 4

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 5

[7] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[9] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 5

[10] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2021. 3, 5

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680, 2014. 3

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 3

[14] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2020. 1, 2

[15] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 41–50, 2019. 4, 5

[16] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *International Conference on Machine Learning*, 2022. 2, 7

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*, 2017. 2, 4, 5, 7

[18] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018. 1, 2, 5, 7

[19] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16783–16792, 2022. 4

[20] Konstantin Kirchheim, Marco Filax, and Frank Ortmeier. PyTorch-OOD: A library for out-of-distribution detection based on pytorch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4351–4360, June 2022. 5

[21] Konstantin Kirchheim and Frank Ortmeier. On outlier exposure with generative models. In *NeurIPS Machine Learning Safety Workshop*, December 2022. 3, 5, 6, 7

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[23] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31, 2018. 2, 7

[24] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018. 4, 7

[25] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 3, 4, 5, 7

[26] Shuo Lu, Yingsheng Wang, Lijun Sheng, Aihua Zheng,

Lingxiao He, and Jian Liang. Recent advances in ood detection: Problems and approaches. *arXiv preprint arXiv:2409.11884*, 2024. 2

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3, 7

[28] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 5

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[30] Liwei Song, Vikash Sehwag, Arjun Nitin Bhagoji, and Prateek Mittal. A critical evaluation of open-world machine learning. *arXiv preprint arXiv:2007.04391*, 2020. 1, 3

[31] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022. 7

[32] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018. 1, 3

[33] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022. 7

[34] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *ArXiv*, abs/2110.11334, 2021. 1, 2

[35] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131(10):2607–2622, 2023. 6

[36] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5

[37] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 5

[38] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 7