DECIPHERING THE DEFINITION OF ADVERSARIAL ROBUSTNESS FOR POST-HOC OOD DETECTORS

Peter Lorenz [1 2] **Mario Fernandez** [1 3] **Jens Müller** [2] **Ullrich Köthe** [2]

Abstract

Detecting out-of-distribution (OOD) inputs is critical for safely deploying deep learning models in real-world scenarios. In recent years, many OOD detectors have been developed, and even the benchmarking has been standardized, i.e. OpenOOD. The number of post-hoc detectors is growing fast and showing an option to protect a pre-trained classifier against natural distribution shifts, claiming to be ready for real-world scenarios. However, its efficacy in handling adversarial examples has been neglected in the majority of studies. This paper investigates the adversarial robustness of the 16 post-hoc detectors on several evasion attacks and discuss a roadmap towards adversarial defense in OOD detectors.

1. INTRODUCTION

Adversarial robustness in the context of out-of-distribution (OOD) detection refers to the ability of a detector to correctly identify OOD samples even when they have been adversarial perturbed to evade a deep neural network (DNN). Evasion attacks, e.g. PGD (Madry et al., 2017), which are designed to fool deep learning classifiers, are difficult to spot as an outlier for OOD detectors. To prevent errors in real-world applications, it is crucial to detect OOD cases, not only for natural distribution shifts (Taori et al., 2020) but also for adversarial examples (Sehwag et al., 2019a) without degrading the generalizability of the underlying pre-trained classifier (Yang et al., 2021).

Current standardized benchmarks such as OpenOOD (Zhang et al., 2023) and RoboDepth (Kong et al., 2024) merely focus on natural distribution shifts and corruptions (Hendrycks & Dietterich, 2019). Especially OpenOOD aims to make a fair comparison across methods initially developed for anomaly detection, model uncertainty, open set recognition, and OOD detection methods.

Table 1. Post-hoc OOD detectors architecture comparison: I) Features: output of layers before the last layer. II) Logits: raw output of the last layer. III) Probabilities: normalized output of the last layer. IV) Adversarial robust against evasive attacks (see Section 2.3 and Table 3). The '~' means that the detector only partly fulfills a certain property. All methods are in the OpenOOD benchmark suite (Zhang et al., 2023).

Detector	Venue	Detector Architecture			Adversarial Robust
		Features	Logits	Probs	
SCALE	ICLR'24	✓	✓		
NNGUIDE	NeurIPS'23	✓		✓	
GEN	CVPR'23			✓	
ASH	ICLR'23	✓			
DICE	ECCV'22	✓			
KNN	ICML'22	✓			
VIM	CVPR'22	✓	✓		
KLM	ICML22			✓	
MLS	ICML22		✓		
REACT	NeurIPS'21	✓			
RMDS	ARXIV'21	✓		✓	~
GRAM	ICML'20		✓		
EBO	NeurIPS'20	✓			
ODIN	ICLR'18			✓	
MDS	NeurIPS'18	✓			~
MSP	ICLR'17			✓	

OpenOOD benchmark suite (Yang et al., 2021) evaluates methods on semantic shift (e.g., samples that are semantically different from the training data, representing truly novel or unseen concepts.) (Hendrycks & Gimpel, 2016) and covariate shift (e.g. samples that come from a different distribution than the training data, but still belong to the same semantic categories). Current OOD detection methods, as some listed in Table 1, achieve outstanding results on prominent OOD benchmarks, such as the OpenImage-O (Wang et al., 2022), ImageNet-O (Hendrycks et al., 2021), Texture (Cimpoi et al., 2014), and iNaturalist (Huang & Li, 2021; Van Horn et al., 2018). OOD detection is a very quickly growing field due to the number of methods added to OpenOOD. More specific post-hoc methods with their plug-and-play capabilities on pre-trained classifiers are more flexible and scalable compared to methods that require full retraining on new OOD data (Yang et al., 2022a; Cong & Prakash, 2022). Simple post-hoc methods like KNN (Sun

[1]Fraunhofer ITWM, Germany [2]Heidelberg University, Germany [3]École Normale Supérieure, PSL University, France. Correspondence to: Peter Lorenz <kg183@uni-heidelberg.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

et al., 2022) are highlighted maintaining good performance on toy-datasets (e.g. MNIST (Deng, 2012), CIFAR-10 or CIAR-100 (Krizhevsky et al., 2009), and also show outstanding performance on the more realistic dataset like ImageNet (Deng et al., 2009) according to (Yang et al., 2022a). These experiments do neglect the adversarial robustness of "state-of-the-art" detectors and the real-world capabilities are questionable as past studies had shown (Sehwag et al., 2019a; Song et al., 2020; Chen et al., 2020; Salehi et al., 2021). Adversarial examples remain challenging because they share the same semantics as the training data, but aim to modify the classifier's output.

In this study, we investigate the adversarial robustness of post-hoc ODD detectors. Our contributions can be summarized as follows:

- We revise the definition of adversarial OOD regarding post-hoc OOD methods to finally have a common understanding of robust adversarial OOD detection.
- We examine 16 post-hoc OOD detectors by delving into their current ability to detect adversarial examples — an aspect that has been disregarded.
- We expand the OpenOOD framework with evasive attacks and provide adversarial OOD datasets: github.com/adverML/AdvOpenOOD.

2. RELATED WORK

2.1. Evasion Attacks Crafting Inliers

The objective of evasion attacks is to generate adversarial examples that will result in the misclassification of inputs at deep learning models (Biggio et al., 2013). It is possible to distinguish between two types of attack: black-box attacks (Zheng et al., 2023), where the classifier is queried, and white-box attacks (Carlini & Wagner, 2017b), where the network is under the attacker's complete control. A white-box threat model is strictly stronger. They try to find the smallest possible perturbation, often imperceptible to humans, to manipulate the model's decision boundaries.

More formally, for an input $\boldsymbol{x}$ with the ground-truth label y, an adversarial example $\boldsymbol{x}'$ is crafted by adding small noise $\boldsymbol{\delta}$ to $\boldsymbol{x}$ such that the predictor model loss $\boldsymbol{J}(\boldsymbol{x}', y)$ is maximized. The L^p norm of the adversarial noise should be less than a specified value ϵ, i.e., $\|\boldsymbol{x} - \boldsymbol{x}'\| \leq \epsilon$, e.g., $\epsilon = 8/255$ (Croce et al., 2020), to ensure that the image does not change semantically. The attack method Fast Gradient Sign Method (FGSM) by (Goodfellow et al., 2014) maximizes the loss function in a single step by taking a step towards the sign of the gradient of $\boldsymbol{J}(\boldsymbol{x}, y)$ w.r.t. to $\boldsymbol{x}$: $\boldsymbol{x}' = \boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} \boldsymbol{J}(\boldsymbol{x}', y))$, where the noise meets the L^∞ norm bound ϵ. Furthermore, this approach can be applied iteratively, as shown by (Kurakin et al., 2018), using a reduced step size α: $\boldsymbol{x}'_0 = \boldsymbol{x}$, $\boldsymbol{x}'_{t+1} = \boldsymbol{x}'_t + \alpha\epsilon \text{sign}(\nabla_{\boldsymbol{x}} \boldsymbol{J}(\boldsymbol{x}'_t, y))$, where in each step, the perturbation should be constrained within the L^∞ ball of radius ϵ. This constraint is a characteristic of the Projected Gradient Descent (PGD) attack proposed by (Madry et al., 2017), which is commonly considered a standard attack for the evaluation of model robustness (Liu et al., 2023a). Masked PGD (mPGD) (Xu et al., 2023b) is a variant of the PGD attack that restricts perturbations to a specific region within an image. The PGD attack is extended towards a patch: $\boldsymbol{x}'_{t+1} = \text{Clip}_{\boldsymbol{x}'}\left(\boldsymbol{x}'_t + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{x}} \boldsymbol{J}(\boldsymbol{x}'_t, y, \boldsymbol{\theta})[patch]), \epsilon\right)$. In this context, the term "patch" refers to the region specified as $[x : x + h, y : y + w]$, where $[x, y, h, w]$ are the provided patch's coordinates and dimensions.

Lastly, DeepFool (DF) (Moosavi-Dezfooli et al., 2016) assumes that the network's decision boundary is linear, even though in reality, it may not be. It aims to find the minimal perturbation, corresponding to the orthogonal projection onto the hyperplane.

2.2. Advantages of Post-Hoc OOD Detectors

Post-hoc OOD detection methods, which make use of specific layers of the pre-trained classifier, have been demonstrated to outperform retraining-based approaches, thereby underscoring their empirical efficacy (Zhang et al., 2023). Their plug-and-play nature allows seamless integration with pre-trained models, without necessitating alterations to the training procedure or access to original training data (Zhang et al., 2023). In Table 1 are 16 post-hoc detectors listed, which also shows if a method uses features, logits, or the probabilities of the pre-trained model. Post-hoc methods underline their simplicity to outperform others on natural distribution shifts datasets by being as lightweight as possible. The latest post-hoc OOD detector in Table 1 is SCALE (Xu et al., 2023a). In contrast to the previous activation shaping method, ASH (Djurisic et al., 2022), that involves pruning and scaling of activations, SCALE demonstrates state-of-the-art results are solely derived from scaling.

OOD detection has a more expansive scope when compared to anomaly detection or open-set recognition (OSR) (Scheirer et al., 2012). Anomaly detection is concerned with the identification of rare deviations within a single distribution. OSR addresses the issue of unknown classes during inference. OOD detection methods aim to identify any test sample that deviates from the training data distribution (Zhang et al., 2023).

Moreover, post-hoc OOD methods can be augmented with other techniques, such as those employed in OSR (Gillert & von Lukas, 2021) or uncertainty estimation (Schwaiger et al., 2020). Combining different techniques means also the post-hoc methods become more complex and thus the attack surface might become larger, i.e. attacks against uncertainty estimation (Ledda et al., 2023) differ from the

evasion attacks. It is not necessarily the case that a post-hoc method must be combined with other techniques. This demonstrates SAFE (Wilson et al., 2023).

2.3. OOD Adversarial Detection

Ensuring the protection of deployed DL models is the aim of OOD detectors, but the task of providing comprehensive defense (Carlini et al., 2019) against unknown threats is challenging. Every defense mechanism can be circumvented at some point (Carlini & Wagner, 2017a). Many OOD detectors can be easily evaded by slightly perturbing benign OOD inputs, creating OOD adversarial examples that reveal a severe limitation of current open-world learning frameworks (Sehwag et al., 2019b; Azizmalayeri et al., 2022). Even adversarial training-based defense methods, effective against ID adversarial attacks, struggle against OOD adversarial examples (Azizmalayeri et al., 2022). In the past years, various defensive techniques (Wu et al., 2023) and combinations of them have surfaced to combat threats such as adversarial training (Madry et al., 2017; Wang et al., 2023; Bai et al., 2024), gradient masking for obfuscation (Papernot et al., 2017), input transformations such as input purification (Nie et al., 2022; Lin et al., 2024). However, attackers consistently adapt their adversarial attacks to the specific defense mechanisms (Tramer et al., 2020). According to (Croce et al., 2022), optimization-based defenses could be a promising future, because they can adapt during test-time towards the input.

OOD detectors have benefited from insights in the adversarial machine learning (AML) field, but still lack comprehensive defense against unknown threats. There are adversarial training based methods, e.g. ALOE (Chen et al., 2020), OSAD (Shao et al., 2020), or ATOM (Chen et al., 2021). A discriminator-based method, ADT (Azizmalayeri et al., 2022), significantly outperforms previous methods by addressing their vulnerabilities to strong adversarial attacks. More recently, the post-hoc method SAFE (Wilson et al., 2023) leverages the most sensitive layers in a pre-trained classifier through targeted input-level adversarial perturbations. To this end, "adversarial robust" OOD detection methods lag, being a comprehensive defense against unknown and adaptive threats remains an intricate challenge.

3. ROBUSTNESS DEFINITION

The robustness definition in the field of OOD detectors has been ambiguous when it comes to attack methods. There are two categories of adversarial examples. The first one merely attacks the underlying pre-trained classifier and the second one aims to fool the OOD detector itself. Adversarial robustness can be considered in a classifier (Unified Robustness) or the OOD detector (Robust OOD Detection) according to (Karunanayake et al., 2024). In this work, we focus on unified robustness, which belong to the covariate shift. For image classification, a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i); \boldsymbol{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$ sample from a training distribution $\hat{P}_{\text{data}}(\boldsymbol{x}, y)$ is used to train some classifier $C : \mathcal{X} \to \mathcal{Y}$. In real-world deployments, distribution shift occurs when classifier C receives data from test distribution $\hat{P}_{\text{test}}(\boldsymbol{x}, y)$ where $\hat{P}_{\text{data}}(\boldsymbol{x}, y) \neq \hat{P}_{\text{test}}(\boldsymbol{x}, y)$ (Moreno-Torres et al., 2012). An OOD detector is a scoring function s that maps an image $\boldsymbol{x}$ to a real number $\mathbb{R}$ such that some threshold τ arrives at the detection rule $f(\boldsymbol{x})$: ID if $s(\boldsymbol{x}) \geq \tau$, OOD otherwise. Table 4 in Appendix A gives an overview of several detectors as well as considered ID and OOD datasets and model architectures. The ImageNet-1K dataset together with ResNet-50 architecture has become standard for ID. Popular OD datasets are iNaturalist, SUN, Places, and Textures. Some OOD detectors, such as ALOE, OSAD, ADT, ATOM (see Section 2.3), aim to be adversarial robust. They usually take evading attacks (see Section 2.1) such as FGSM or PGD as OOD. These computationally expensive methods only show empirical results on the small-scale CIFAR-10. A robust OOD detector is built to distinguish whether a perturbed input is OOD. Standardized OOD benchmark frameworks, i.e. OpenOOD (Zhang et al., 2023) or RoboDepth (Kong et al., 2024) do not include unified robustness in their benchmarks at the moment. Consequently, both frameworks, give a false sense of encompassing open-world capabilities. They focus on natural distribution (Hendrycks et al., 2021), where OOD detection in large-scale semantic space has attracted increasing attention (Hendrycks et al., 2019), see Appendix A. Some OOD datasets have issues, where ID classes are part of the OD dataset (Bitterwolf et al., 2023). Recently, (Yang et al., 2023) found a clean semantic shift dataset that minimizes the interference of covariate shift. The experiments show that state-of-the-art OOD detectors are more sensitive to covariate shift and the advances in semantic shift detection are minimal. The investigation of adversarial examples could gain insights into understanding the covariate shift and towards a generalized OOD detection (Yang et al., 2021). The difference between benign $\boldsymbol{x}$ and the attacked counterpart $\boldsymbol{x}'$ is the different attention of the pre-trained classifier C per sample. Let's define an attention map $\boldsymbol{A}$ (Guo et al., 2022): The input image is passed through a classifier C to obtain a feature map $\boldsymbol{F}$. A possible tool could be Grad-CAM (Selvaraju et al., 2017) to investigate the attention change between benign $\boldsymbol{x}$ and adversarial example $\boldsymbol{x}'$ on $\boldsymbol{F}$ (Rieger & Hansen, 2020).

4. EXPERIMENTS

Experiment Setup We extend the OpenOOD framework (Zhang et al., 2023) to consider adversarial attacks. We attack the pre-trained classifiers on the corresponding test sets and evaluate 16 post-hoc OOD detectors. As **at-**

tack methods, we choose FGSM(-L^∞), PGD(-L^∞), DF(-L^2) from FoolBox (Rauber et al., 2017); and the mPGD-(L^∞). The attacked **models** are ResNet-18 (He et al., 2016a), ResNet-50 (He et al., 2016a), and Swin-T (Liu et al., 2021). The **datasets** are CIFAR-10 & CIFAR-100 (Krizhevsky et al., 2009), ImageNet-1K (Deng et al., 2009) and its variant ImageNet-200 with just 200 classes. The efficacy of the attacks is not absolute and depends on a multitude of factors, including the hyperparameters, model architecture, and the dataset. The attack success rate (ASR) is presented in Table 2. The attacks PGD and FGSM do have an epsilon size, we use an epsilon size of $8/255$ for CIFAR-10/100 and $4/255$ for the ImageNet. The mPGD randomly attacks an area of the image (8×8 px for CIFAR-10/100 and 60×60 px for the ImageNet) without an epsilon constraint, leading to perceptible perturbations.

Table 2. Setup. The attack success rate (ASR) from various attacks on different models and datasets. $\mathcal{A}_{\text{std}}$ refers to the standard accuracy of the pre-trained classifier.

Dataset	Arch	$\mathcal{A}_{\text{std}}$ (%)	Attack	ASR (%)
CIFAR-10	ResNet-18	95.32	PGD	99.88
			FGSM	59.21
			DF	100
			mPGD	68.06
CIFAR-100	ResNet-18	77.19	PGD	100
			FGSM	91.99
			DF	100
			mPGD	88.14
ImageNet-200	ResNet-18	86.27	PGD	99.9
			FGSM	95.46
			DF	100
			mPGD	96.53
ImageNet-1K	ResNet-50	76.19	PGD	99.97
			FGSM	93.33
			DF	100
			mPGD	98.48
	Swin-T	95.99	PGD	99.99
			FGSM	75.09
			DF	100
			mPGD	98.84

We utilize two **metrics** to assess the OOD detection performance, elaborated as follows: 1) FPR95↓ stands for false positive rate measured when true positive rate (TPR) sits at 95%. Intuitively, FPR95 measures the portion of samples that are falsely recognized as ID data. 2) AUROC↑ refers to the area under the receiver operating characteristic curve, for binary classification problems like OOD detection.

4.1. DISCUSSION

All methods do not show sufficient results, when evaluating 16 post-hoc methods as unveiled in the Table 3. Only two of them - Mahalanobis distance-based methods - (Lee et al., 2018; Ren et al., 2021) show partly detection capabilities against the FGSM, PGD, and mPGD attacks on ResNet-50 with ImageNet-1K. The Mahalanobis distance's robustness has been thoroughly studied (Kamoi & Kobayashi, 2020; Eustratiadis et al., 2021; Yang et al., 2022b; Anthony & Kamnitsas, 2023), and its adversarial robustness is attributed to its covariance structure consideration (Eustratiadis et al., 2021). This highlights the conflict between AML and OOD detection, where detectors often excel in either adversarial or natural distributions, but not both. High detection rates on OOD samples are the fundament for further defense mechanisms.

Level of Adversarial Robustness - From Detectors towards Defenses A defense is more sophisticated than detectors to mitigate the attacker's efforts to fool a classifier. A step towards adversarial defense could be to improve adversarial robustness in OOD detectors. We suggest a possible roadmap to evaluate detectors and lift them toward an adversarial defense:

1. Evaluate on strong attacks (Carlini & Wagner, 2017a) and avoid hyperparameters that weaken the strength of the attack's effect. The FGSM is not recommended because it performs a single step to find the adversarial perturbation, making it less effective than PGD (Li et al., 2020). Furthermore, the attack hyperparameter space is huge (Cinà et al., 2024) and could mitigate the attacks' strength.
2. Use different models and other datasets than the simple ResNet-18 trained on CIFAR-10. We suggest using ImageNet-1K because its complexity in terms of resolution and objects is more closely to real-world scenarios.
3. Elaborate your strategy to countermeasure the attack as demonstrated in (Sehwag et al., 2019b). New defense mechanisms have often been broken quickly again (Carlini & Wagner, 2017a). For example, a differentiable OOD detector can be easily fooled if the attacker approximates the gradients of the network during the backward pass in a differentiable manner, known as BPDA (Athalye et al., 2018).
4. Let your method fail against sophisticated attacks, such as adaptive attacks (Athalye et al., 2018; Croce et al., 2022) or design OOD adversarial examples that convert OD to ID samples (Sehwag et al., 2019a). Adversarial robustness is an iterative process, where defenses are proposed, evaluated, and then improved upon in response to new attacks or discovered vulnerabilities.

5. CONCLUSION

In this study, we assess the performance of the 16 post-hoc OOD detectors in their ability to detect various evasive attacks. We conducted prominent white-box adversarial attacks, such as PGD and DeepFool, on the CIFAR-10 and ImageNet-1K datasets. Our discovery indicates that current post-hoc methods are not ready for real-world applications as long as they are vulnerable against well-known threat — adversarial examples. We hope that our experiments give a baseline for further research by improving post-hoc methods towards robustness and will find a place as a standardized benchmark, such as OpenOOD.

Future Work We propose to extend the experiments towards transferability because adversarial examples transfer effectively across different datasets (Alhamoud et al., 2022) and models (Gu et al., 2023). Finally, we would suggest using black-box attacks for a realistic open-world scenario.

Table 3. Results. We evaluate the post-hoc OOD detectors using the metrics FPR95↓ (%) and AUROC↑ (%). The norm-bounded attacks PGD and FGSM do have an epsilon size of $8/255$ for CIFAR-10/100 and $4/255$ for the ImageNet.

Detector	Attacks	CIFAR-10		CIFAR-100		ImageNet-200		ImageNet-1K			
		ResNet-18						ResNet-50		Swin-T	
		FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
SCALE	PGD	99.67	34.53	99.97	16.18	95.49	35.14	100.00	0.20	95.49	35.14
	FGSM	85.74	77.50	49.69	85.64	79.75	76.88	89.75	66.28	79.75	76.88
	DF	67.07	81.73	69.22	68.69	79.75	76.88	87.82	57.77	79.75	76.88
	mPGD	88.50	70.69	85.58	59.67	93.24	42.17	100.00	6.90	93.24	42.17
NNGUIDE	PGD	99.39	30.29	98.85	17.07	96.44	33.60	100.00	0.12	96.44	33.60
	FGSM	93.10	53.01	68.14	77.62	83.21	75.13	85.27	73.53	83.21	75.13
	DF	92.08	63.25	85.36	64.14	83.21	75.13	82.19	62.81	83.21	75.13
	mPGD	92.94	58.90	90.98	57.26	94.30	42.87	99.99	9.52	94.30	42.87
GEN	PGD	99.51	41.75	99.90	26.03	89.17	40.03	100.00	0.21	89.17	40.03
	FGSM	70.14	81.29	45.66	87.10	72.06	79.00	83.63	73.28	72.06	79.00
	DF	44.32	85.98	71.38	65.82	72.06	79.00	80.01	62.96	72.06	79.00
	mPGD	83.65	74.35	75.59	66.56	88.10	47.17	99.96	12.40	88.10	47.17
ASH	PGD	99.67	31.23	99.96	24.49	97.06	32.97	100.00	0.19	97.06	32.97
	FGSM	86.94	70.60	42.86	88.14	83.81	74.57	85.61	69.99	83.81	74.57
	DF	77.10	74.62	74.98	65.44	83.81	74.57	83.29	60.88	83.81	74.57
	mPGD	90.75	64.35	78.87	66.61	95.20	40.17	100.00	8.49	95.20	40.17
DICE	PGD	96.65	36.09	99.93	23.45	95.03	34.24	100.00	0.11	95.03	34.24
	FGSM	75.86	72.65	46.27	86.51	75.62	78.73	86.99	71.20	75.62	78.73
	DF	68.84	73.44	76.73	65.46	75.62	78.73	84.72	62.92	75.62	78.73
	mPGD	89.46	66.25	80.48	66.68	91.60	42.99	99.99	8.63	91.60	42.99
KNN	PGD	64.91	69.18	90.06	43.07	85.23	55.53	78.63	55.74	85.23	55.53
	FGSM	61.23	82.08	47.81	84.69	76.88	73.23	75.89	68.43	76.88	73.23
	DF	38.78	85.84	78.54	63.06	76.88	73.23	86.65	58.09	76.88	73.23
	mPGD	76.02	75.02	78.93	64.03	88.63	54.44	87.09	48.40	88.63	54.44
VIM	PGD	92.45	56.83	98.16	42.79	89.17	46.44	100.00	5.16	89.17	46.44
	FGSM	54.89	84.43	54.72	74.75	71.60	69.44	75.28	71.31	71.60	69.44
	DF	43.92	84.92	78.57	61.75	71.60	69.44	82.95	59.88	71.60	69.44
	mPGD	80.70	74.35	80.56	60.01	90.16	50.71	99.86	24.51	90.16	50.71
KLM	PGD	91.43	60.91	91.94	45.75	90.87	54.58	95.49	40.77	90.87	54.58
	FGSM	96.90	66.02	72.31	80.83	80.71	74.44	80.54	71.56	80.71	74.44
	DF	80.84	71.64	91.38	59.58	80.71	74.44	85.48	59.30	80.71	74.44
	mPGD	97.52	62.26	89.69	61.30	91.96	55.14	94.81	41.19	91.96	55.14
MLS	PGD	99.58	39.65	99.96	24.43	94.43	35.64	100.00	0.12	94.43	35.64
	FGSM	75.61	80.97	43.04	87.63	74.47	79.06	85.30	74.16	74.47	79.06
	DF	51.12	84.89	74.91	65.41	74.47	79.06	81.44	63.37	74.47	79.06
	mPGD	85.11	73.78	78.81	66.41	90.90	44.19	99.97	10.65	90.90	44.19
REACT	PGD	98.84	45.19	99.89	25.13	94.49	35.90	100.00	4.15	94.49	35.90
	FGSM	79.55	79.84	42.83	88.27	76.12	78.15	80.31	74.14	76.12	78.15
	DF	54.80	84.16	74.99	65.49	76.12	78.15	79.97	62.88	76.12	78.15
	mPGD	85.32	73.19	78.70	66.52	91.26	44.59	99.71	20.13	91.26	44.59
GRAM	PGD	99.82	22.50	99.94	17.12	98.75	25.92	100.00	0.07	98.75	25.92
	FGSM	94.77	56.54	79.42	79.23	91.60	69.07	96.06	59.67	91.60	69.07
	DF	88.14	60.87	90.78	58.04	91.60	69.07	93.66	55.04	91.60	69.07
	mPGD	93.21	55.56	91.49	58.72	97.83	34.92	100.00	4.36	97.83	34.92
RMDS	PGD	49.03	82.70	66.28	77.08	53.69	76.40	**37.47**	**95.23**	53.69	76.40
	FGSM	68.03	80.66	76.00	79.46	67.10	76.26	71.65	73.33	67.10	76.26
	DF	43.26	85.75	64.52	80.96	67.10	76.26	73.74	65.11	67.10	76.26
	mPGD	77.05	75.84	84.46	74.74	79.37	63.99	90.78	50.99	79.37	63.99
EBO	PGD	99.58	39.61	99.96	24.49	94.47	35.41	100.00	0.12	94.47	35.41
	FGSM	75.62	81.04	42.86	88.14	74.58	79.17	85.35	74.42	74.58	79.17
	DF	51.23	84.71	74.99	65.44	74.58	79.17	81.63	63.79	74.58	79.17
	mPGD	85.11	73.80	78.87	66.61	91.00	44.10	99.97	10.64	91.00	44.10
MDS	PGD	47.81	**84.48**	50.82	**84.18**	57.34	76.11	**0.05**	**99.95**	57.34	76.11
	FGSM	64.24	79.22	86.06	51.31	91.41	49.61	90.61	52.93	91.41	49.61
	DF	54.04	79.60	89.76	52.53	91.41	49.61	93.19	49.18	91.41	49.61
	mPGD	78.25	69.86	91.87	50.63	83.81	65.37	51.74	**89.98**	83.81	65.37
ODIN	PGD	99.73	33.25	99.97	17.58	97.20	31.19	100.00	0.91	97.20	31.19
	FGSM	73.56	83.16	39.74	90.19	77.39	76.38	85.44	72.56	77.39	76.38
	DF	60.16	83.99	72.94	68.00	77.39	76.38	82.00	65.46	77.39	76.38
	mPGD	86.49	74.11	85.66	62.09	91.72	45.12	99.90	15.89	91.72	45.12
MSP	PGD	99.34	43.62	100.00	27.04	87.66	42.60	100.00	4.08	87.66	42.60
	FGSM	67.16	81.14	47.87	85.08	72.12	77.77	80.39	70.45	72.12	77.77
	DF	39.78	86.85	70.63	65.92	72.12	77.77	72.65	60.76	72.12	77.77
	mPGD	82.76	74.68	75.03	66.23	87.48	49.07	100.00	22.21	87.48	49.07

References

Alhamoud, K., Hammoud, H. A. A. K., Alfarra, M., and Ghanem, B. Generalizability of adversarial robustness under distribution shifts. *arXiv preprint arXiv:2209.15042*, 2022.

Anthony, H. and Kamnitsas, K. On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pp. 136–146. Springer, 2023.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Azizmalayeri, M., Soltani Moakhar, A., Zarei, A., Zohrabi, R., Manzuri, M., and Rohban, M. H. Your out-of-distribution detection method is not robust! *NeurIPS*, 2022.

Bai, Y., Zhou, M., Patel, V. M., and Sojoudi, S. Mixednuts: Training-free accuracy-robustness balance via nonlinearly mixed classifiers. *arXiv preprint arXiv:2402.02263*, 2024.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pp. 387–402. Springer, 2013.

Bitterwolf, J., Mueller, M., and Hein, M. In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826*, 2023.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 3–14, 2017a.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017b.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. Robust out-of-distribution detection for neural networks. *arXiv preprint arXiv:2003.09711*, 2020.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. 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*, pp. 430–445. Springer, 2021.

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

Cinà, A. E., Rony, J., Pintor, M., Demetrio, L., Demontis, A., Biggio, B., Ayed, I. B., and Roli, F. Attackbench: Evaluating gradient-based attacks for adversarial examples. *arXiv preprint arXiv:2404.19460*, 2024.

Cong, T. and Prakash, A. Sneakoscope: Revisiting unsupervised out-of-distribution detection, 2022. URL https://openreview.net/forum?id=xdNcdoHdBER.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Croce, F., Gowal, S., Brunner, T., Shelhamer, E., Hein, M., and Cemgil, T. Evaluating the adversarial robustness of adaptive test-time defenses. In *International Conference on Machine Learning*, pp. 4421–4435. PMLR, 2022.

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

Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., and Sun, J. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13733–13742, 2021.

Djurisic, A., Bozanic, N., Ashok, A., and Liu, R. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Eustratiadis, P., Gouk, H., Li, D., and Hospedales, T. Weight-covariance alignment for adversarially robust neural networks. In *International Conference on Machine Learning*, pp. 3047–3056. PMLR, 2021.

Gillert, A. and von Lukas, U. F. Towards combined open set recognition and out-of-distribution detection for fine-grained classification. In *VISIGRAPP (5: VISAPP)*, pp. 225–233, 2021.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gu, J., Jia, X., de Jorge, P., Yu, W., Liu, X., Ma, A., Xun, Y., Hu, A., Khakzar, A., Li, Z., et al. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*, 2023.

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M., and Hu, S.-M. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645. Springer, 2016b.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

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

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., and Song, D. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *CVPR*, pp. 15262–15271, 2021.

Hoiem, D., Divvala, S. K., and Hays, J. H. Pascal voc 2008 challenge. *World Literature Today*, 24(1):1–4, 2009.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Huang, R. and Li, Y. Mos: Towards scaling out-of-distribution detection for large semantic space. In *CVPR*, pp. 8710–8719, 2021.

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34: 677–689, 2021.

Kamoi, R. and Kobayashi, K. Why is the mahalanobis distance effective for anomaly detection? *arXiv preprint arXiv:2003.00402*, 2020.

Karunanayake, N., Gunawardena, R., Seneviratne, S., and Chawla, S. Out-of-distribution data: An acquaintance of adversarial examples–a survey. *arXiv preprint arXiv:2404.05219*, 2024.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

Kong, L., Xie, S., Hu, H., Ng, L. X., Cottereau, B., and Ooi, W. T. Robodepth: Robust out-of-distribution depth estimation under corruptions. *Advances in Neural Information Processing Systems*, 36, 2024.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6(1):1, 2009.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. 2018.

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

Ledda, E., Angioni, D., Piras, G., Fumera, G., Biggio, B., and Roli, F. Adversarial attacks against uncertainty quantification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4599–4608, 2023.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS*, 31, 2018.

Li, B., Wang, S., Jana, S., and Carin, L. Towards understanding fast adversarial training. *arXiv preprint arXiv:2006.03089*, 2020.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Lin, G., Li, C., Zhang, J., Tanaka, T., and Zhao, Q. Adversarial training on purification (atop): Advancing both robustness and generalization. *arXiv preprint arXiv:2401.16352*, 2024.

Liu, C., Dong, Y., Xiang, W., Yang, X., Su, H., Zhu, J., Chen, Y., He, Y., Xue, H., and Zheng, S. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *arXiv preprint arXiv:2302.14301*, 2023a.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *NeurIPS*, 2020.

Liu, X., Lochman, Y., and Zach, C. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23946–23955, 2023b.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, pp. 2574–2582, 2016.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A. Y., et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, Spain, 2011.

Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., and Anandkumar, A. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Park, J., Jung, Y. G., and Teoh, A. B. J. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1686–1695, 2023.

Quattoni, A. and Torralba, A. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 413–420. IEEE, 2009.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428–10436, 2020.

Rauber, J., Brendel, W., and Bethge, M. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

Ren, J., Fort, S., Liu, J., Roy, A. G., Padhy, S., and Lakshminarayanan, B. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021a.

Ridnik, T., Lawen, H., Noy, A., Ben Baruch, E., Sharir, G., and Friedman, I. Tresnet: High performance gpu-dedicated architecture. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1400–1409, 2021b.

Rieger, L. and Hansen, L. K. A simple defense against adversarial attacks on heatmap explanations. *arXiv preprint arXiv:2007.06381*, 2020.

Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M. H., and Sabokrou, M. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., and Boult, T. E. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2012.

Schwaiger, A., Sinhamahapatra, P., Gansloser, J., and Roscher, K. Is uncertainty quantification in deep learning sufficient for out-of-distribution detection? *Aisafety@ ijcai*, 54, 2020.

Sehwag, V., Bhagoji, A. N., Song, L., Sitawarin, C., Cullina, D., Chiang, M., and Mittal, P. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 105–116, 2019a.

Sehwag, V., Bhagoji, A. N., Song, L., Sitawarin, C., Cullina, D., Chiang, M., and Mittal, P. Better the devil you know: An analysis of evasion attacks using out-of-distribution adversarial examples. *arXiv preprint arXiv:1905.01726*, 2019b.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

Shao, R., Perera, P., Yuen, P. C., and Patel, V. M. Open-set adversarial defense. In *ECCV*. Springer, 2020.

Song, L., Sehwag, V., Bhagoji, A. N., and Mittal, P. A critical evaluation of open-world machine learning. *arXiv preprint arXiv:2007.04391*, 2020.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32: 323–332, 2012.

Sun, Y. and Li, Y. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pp. 691–708. Springer, 2022.

Sun, Y., Guo, C., and Li, Y. React: Out-of-distribution detection with rectified activations. *NeurIPS*, 34:144–157, 2021.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *ICML*. PMLR, 2022.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *NeurIPSs*, 2020.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

Touvron, H., Vedaldi, A., Douze, M., and Jégou, H. Fixing the train-test resolution discrepancy. *Advances in neural information processing systems*, 32, 2019.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Tramer, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., and Belongie, S. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 595–604, 2015.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. Open-set recognition: A good closed-set classifier is all you need? *arXiv preprint arXiv:2110.06207*, 2021.

Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, pp. 4921–4930, 2022.

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pp. 36246–36263. PMLR, 2023.

Wilson, S., Fischer, T., Dayoub, F., Miller, D., and Sünderhauf, N. Safe: Sensitivity-aware features for out-of-distribution object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23565–23576, 2023.

Wu, B., Wei, S., Zhu, M., Zheng, M., Zhu, Z., Zhang, M., Chen, H., Yuan, D., Liu, L., and Liu, Q. Defenses in adversarial machine learning: A survey. *arXiv preprint arXiv:2312.08890*, 2023.

Wu, F., Wang, D., Hwang, M., Hao, C., Lu, J., Zhang, J., Chou, C., Darrell, T., and Bayen, A. Decentralized vehicle coordination: The berkeley deepdrive drone dataset. *arXiv preprint arXiv:2209.08763*, 2022.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Xu, K., Chen, R., Franchi, G., and Yao, A. Scaling for training time and post-hoc out-of-distribution detection enhancement. *arXiv preprint arXiv:2310.00227*, 2023a.

Xu, K., Xiao, Y., Zheng, Z., Cai, K., and Nevatia, R. Patchzero: Defending against adversarial patch attacks by detecting and zeroing the patch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4632–4641, 2023b.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

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

Yang, W., Zhang, B., and Russakovsky, O. Imagenet-ood: Deciphering modern out-of-distribution detection algorithms. *arXiv preprint arXiv:2310.01755*, 2023.

Yang, X., Guo, Y., Dong, M., and Xue, J.-H. Toward certified robustness of distance metric learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022b.

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

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., Li, Y., Liu, Z., Chen, Y., and Li, H. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

Zheng, M., Yan, X., Zhu, Z., Chen, H., and Wu, B. Blackboxbench: A comprehensive benchmark of black-box adversarial attacks. *arXiv preprint arXiv:2312.16979*, 2023.

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

A. OOD Definition per Method

This section extends the Section 3. In Table 4, we compare the chosen ID and OOD datasets of each OOD detector. In this comparison, we pick those experiments with largest dimensioned available datasets. Furthermore, we extend this comparison by appending the adversarial robust OOD detectors from the related work in Section 2.3. It can be observed that the more recent post-hoc detectors tend to have ImageNet-1K as standard ID datasets. In contrast, the adversarial robust OOD detectors are benchmarked more likely on the smaller and less complex CIFAR-10 dataset.

Following attacks, models, and datasets are used to heuristically evaluate OOD samples (more details in Table 4):

Attacks: FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2017)

Model Architectures:

- ResNet-18/ResNet-50 (He et al., 2016a), ResNetv2-101 (He et al., 2016b), ResNet-50-D (Touvron et al., 2019), TResNet-M (Ridnik et al., 2021b), WideResNet (Zagoruyko & Komodakis, 2016)
- BiT (Kolesnikov et al., 2020), VIT-B-16 (Dosovitskiy et al., 2020), ViT (Touvron et al., 2021), DeiT (Touvron et al., 2021), Swin-T (Liu et al., 2021)
- DenseNet-121 & DenseNet-101 (Huang et al., 2017)
- MobileNet (Howard et al., 2017), MobileNetV2 (Sandler et al., 2018)
- RegNet & RegNetX4.0 (Radosavovic et al., 2020)
- RepVGG (Ding et al., 2021)
- Mixer-B-16 (Tolstikhin et al., 2021)
- CLIP (Radford et al., 2021)

Datasets: CIFAR-10/CIFAR-100 (Krizhevsky et al., 2009), BDD-Anomaly (Hendrycks et al., 2019), DeepDrive (Wu et al., 2022) ImageNet-1K (Deng et al., 2009), ImageNet-21K (Ridnik et al., 2021a), ImageNet-O (Hendrycks & Dietterich, 2019), iNaturalist (Van Horn et al., 2018), ISUN (Quattoni & Torralba, 2009), GSTRB (Stallkamp et al., 2012), LSUN (Yu et al., 2015), NINCO (Bitterwolf et al., 2023), OpenImage-O (Wang et al., 2022), Pascal-VOC (Hoiem et al., 2009), Places (Zhou et al., 2017), Textures (Cimpoi et al., 2014), TinyImageNet (Le & Yang, 2015), Species (Van Horn et al., 2015), StreetHazards (Hendrycks et al., 2019), SSB-hard (Vaze et al., 2021), SVHN (Netzer et al., 2011), SUN (Xiao et al., 2010).

Table 4. Overview of ID and OOD definition from several OOD detectors. The detectors are divided into the following categories (Zhang et al., 2023): Classifiction-based , Density-based , Distance-based . We also mark: Supervised and Adversarial Robust .

Methods	ID	OOD	Model Architectures
		PostHoc Methods	
SCALE (Xu et al., 2023a)	ImageNet-1K	Near-OOD: NINCO, SSB-hard; Far-OOD: iNaturalist, OpenImage-O, Textures	ResNet-50
NNGuide (Park et al., 2023)	ImageNet-1K	Near-OOD: iNaturalist, OpenImage-O; Far-OOD: Textures; Overlapping: SUN and Places	MobileNet, RegNet, ResNet-50, ViT
GEN (Liu et al., 2023b)	ImageNet-1K	ImageNet-O, iNaturalist, OpenImage-O, Texture,	BiT, DeiT, RepVGG, ResNet-50, ResNet-50-D, Swin-T, ViT
ASH (Djurisic et al., 2022)	ImageNet-1K	iNaturalist, Places, SUN, Textures	MobileNetV2, ResNet-50
DICE (Sun & Li, 2022)	ImageNet-1K	iNaturalist, Places, SUN, Textures	DenseNet-101
KNN (Sun et al., 2022)	ImageNet-1K	iNaturalist, Places, SUN, Textures	ResNet-50
VIM (Wang et al., 2022)	ImageNet-1K	ImageNet-O, iNaturalist, OpenImage-O, Texture	BiT-S, DeiT, RepVGG, ResNet-50, ResNet-50-D, Swin-T, VIT-B-16
KLM; MLS (Hendrycks et al., 2019)	ImageNet-21K; ImageNet-1K, Places	Species (categories); BDD-Anomaly, StreetHazards (segmentation)	Mixer-B-16; ResNet-50, TResNet-M, ViTB-16
REACT (Sun et al., 2021)	ImageNet-1K	iNaturalist, Places, SUN, Textures	MobileNet, ResNet
GRAM (Huang et al., 2021)	ImageNet-1K	iNaturalist, SUN, Places, Textures	DenseNet-121, ResNetv2-101
RMDS (Ren et al., 2021)	CIFAR-10, CIFAR-100	CIFAR-10, CIFAR-100	BiT, CLIP, VIT-B-16
EBO (Liu et al., 2020)	CIFAR-10	ISUN, Places, Texture, SVHN, LSUN	WideResNet
MDS (Lee et al., 2018)	CIFAR-10	SVHN, TinyImageNet, LSUN, Adversarial Examples	DenseNet, ResNet
ODIN (Liang et al., 2017)	CIFAR-10	LSUN, SVHN, TinyImageNet	DenseNet, ResNet
MSP (Hendrycks & Gimpel, 2016)	CIFAR-10	SUN (Gaussian)	WideResNet 40-4
		OOD Detectors for Adversarial Robustness	
ALOE (Chen et al., 2020)	CIFAR-10, CIFAR-100, GSTRB	PGD attack	DenseNet
OSAD (Shao et al., 2020)	CIFAR-10, SVHN, Tiny-ImageNet	FGSM, PGD attack	ResNet-18
ADT (Azizmalayeri et al., 2022)	CIFAR-10, CIFAR-100	FGSM, PGD attack	ViT
ATOM (Chen et al., 2021)	CIFAR-10, CIFAR-100, SVHN	PGD attack	WideResNet
SAFE (Wilson et al., 2023)	PASCAL-VOC, DeepDrive	FGSM attack	RegNetX4.0, ResNet-50