ATOM: Robustifying Out-of-Distribution Detection Using Outlier Mining

Jiefeng Chen[1(✉)], Yixuan Li[1], Xi Wu[2], Yingyu Liang[1], and Somesh Jha[1]

[1] Department of Computer Sciences, University of Wisconsin-Madison,
1210 W. Dayton Street, Madison, WI, USA
{jiefeng,sharonli,yliang,jha}@cs.wisc.edu
[2] Google, Madison, WI, USA
wuxi@google.com

Abstract. Detecting out-of-distribution (OOD) inputs is critical for safely deploying deep learning models in an open-world setting. However, existing OOD detection solutions can be brittle in the open world, facing various types of adversarial OOD inputs. While methods leveraging auxiliary OOD data have emerged, our analysis on illuminative examples reveals a key insight that the majority of auxiliary OOD examples may not meaningfully improve or even hurt the decision boundary of the OOD detector, which is also observed in empirical results on real data. In this paper, we provide a theoretically motivated method, *Adversarial Training with informative Outlier Mining* (ATOM), which improves the robustness of OOD detection. We show that, by mining informative auxiliary OOD data, one can significantly improve OOD detection performance, and somewhat surprisingly, generalize to unseen adversarial attacks. ATOM achieves **state-of-the-art** performance under a broad family of classic and adversarial OOD evaluation tasks. For example, on the CIFAR-10 in-distribution dataset, ATOM reduces the FPR (at TPR 95%) by up to 57.99% under adversarial OOD inputs, surpassing the previous best baseline by a large margin.

Keywords: Out-of-distribution detection · Outlier mining · Robustness

1 Introduction

Out-of-distribution (OOD) detection has become an indispensable part of building reliable open-world machine learning models [2]. An OOD detector determines whether an input is from the same distribution as the training data, or different distribution. As of recently a plethora of exciting literature has emerged to combat the problem of OOD detection [16,20,21,24,26–29,33].

The full version of this paper with a detailed appendix can be found at https://arxiv.org/pdf/2006.15207.pdf.

N. Oliver et al. (Eds.): ECML PKDD 2021, LNAI 12977, pp. 430–445, 2021.
https://doi.org/10.1007/978-3-030-86523-8_26

Despite the promise, previous methods primarily focused on clean OOD data, while largely underlooking the robustness aspect of OOD detection. Concerningly, recent works have shown the brittleness of OOD detection methods under adversarial perturbations [5,16,37]. As illustrated in Fig. 1, an OOD image (*e.g.*, mailbox) can be perturbed to be misclassified by the OOD detector as in-distribution (traffic sign data). Failing to detect such an *adversarial OOD example*[1] can be consequential in safety-critical applications such as autonomous driving [12]. Empirically on CIFAR-10, our analysis reveals that the false positive rate (FPR) of a competitive method Outlier Exposure [19] can increase from 3.66% to 99.94% under adversarial attack.

Motivated by this, we make an important step towards the robust OOD detection problem, and propose a novel training framework, ***A****dversarial* ***T****raining with informative* ***O****utlier* ***M****ining* (ATOM). Our key idea is to *selectively* utilize auxiliary outlier data for estimating a tight decision boundary between ID and OOD data, which leads to robust OOD detection performance. While recent methods [16,19,32,33] have leveraged auxiliary OOD data, we show that *randomly* selecting outlier samples for training yields a large portion of uninformative samples, which do not meaningfully improve the decision boundary between ID and OOD data (see Fig. 2). Our work demonstrates that by mining low OOD score data for training, one can significantly improve the robustness of an OOD detector, and somewhat surprisingly, generalize to unseen adversarial attacks.

We extensively evaluate ATOM on common OOD detection benchmarks, as well as a suite of adversarial OOD tasks, as illustrated in Fig. 1. ATOM achieves state-of-the-art performance, significantly outperforming competitive methods using standard training on random outliers [19,32,33], or using adversarial training on random outlier data [16]. On the classic OOD evaluation task (clean OOD data), ATOM achieves comparable and often better performance than current state-of-the-art methods. On L_∞ OOD evaluation task, ATOM outperforms the best baseline ACET [16] by a large margin (e.g. **53.9%** false positive rate deduction on CIFAR-10). Moreover, our ablation study underlines the importance of having both adversarial training and outlier mining (ATOM) for achieving robust OOD detection.

Lastly, we provide theoretical analysis for ATOM, characterizing how outlier mining can better shape the decision boundary of the OOD detector. While hard negative mining has been explored in different domains of learning, e.g., object detection, deep metric learning [11,13,38], the vast literature of OOD detection has not explored this idea. Moreover, most uses of hard negative mining are on a heuristic basis, but in this paper, we derive precise formal guarantees with insights. Our **key contributions** are summarized as follows:

- We propose a novel training framework, adversarial training with outlier mining (ATOM), which facilitates efficient use of auxiliary outlier data to regularize the model for robust OOD detection.

[1] Adversarial OOD examples are constructed w.r.t the OOD detector, which is different from the standard notion of adversarial examples (constructed w.r.t the classification model).

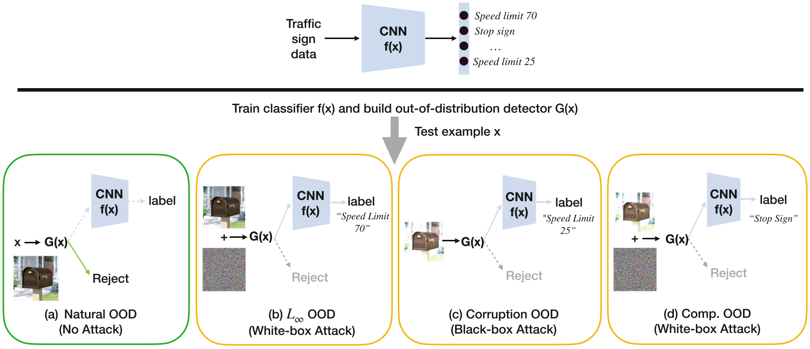
Traffic sign data
CNN f(x)
Speed limit 70
Stop sign
…
Speed limit 25
Train classifier f(x) and build out-of-distribution detector G(x)
Test example x
x → G(x)
label
Reject
(a) Natural OOD (No Attack)
+ → G(x)
"Speed Limit 70"
(b) L_∞ OOD (White-box Attack)
"Speed Limit 25"
(c) Corruption OOD (Black-box Attack)
"Stop Sign"
(d) Comp. OOD (White-box Attack)

Fig. 1. Robust out-of-distribution detection. When deploying an image classification system (OOD detector $G(\mathbf{x})$ + image classifier $f(\mathbf{x})$) in an open world, there can be multiple types of OOD examples. We consider a broad family of OOD inputs, including (a) Natural OOD, (b) L_∞ OOD, (c) corruption OOD, and (d) Compositional OOD. A detailed description of these OOD inputs can be found in Sect. 4.1. In (b-d), a perturbed OOD input (e.g., a perturbed mailbox image) can mislead the OOD detector to classify it as an in-distribution sample. This can trigger the downstream image classifier $f(\mathbf{x})$ to predict it as one of the in-distribution classes (e.g., speed limit 70). Through *adversarial training with informative outlier mining* (ATOM), our method can robustify the decision boundary of OOD detector $G(\mathbf{x})$, which leads to improved performance across all types of OOD inputs. Solid lines are actual computation flow.

- We perform extensive analysis and comparison with a diverse collection of OOD detection methods using: (1) pre-trained models, (2) models trained on randomly sampled outliers, (3) adversarial training. ATOM establishes **state-of-the-art** performance under a broad family of clean and adversarial OOD evaluation tasks.
- We contribute theoretical analysis formalizing the intuition of mining informative outliers for improving the robustness of OOD detection.
- Lastly, we provide a unified evaluation framework that allows future research examining the robustness of OOD detection algorithms under a broad family of OOD inputs. Our code and data are released to facilitate future research on robust OOD detection: https://github.com/jfc43/informative-outlier-mining.

2 Preliminaries

We consider the setting of multi-class classification. We consider a training dataset $\mathcal{D}_{\text{in}}^{\text{train}}$ drawn i.i.d. from a data distribution $P_{\boldsymbol{X},Y}$, where $\boldsymbol{X}$ is the sample space and $Y = \{1, 2, \cdots, K\}$ is the set of labels. In addition, we have an auxiliary outlier data $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$ from distribution $U_{\mathbf{X}}$. The use of auxiliary outliers helps regularize the model for OOD detection, as shown in several recent works [16,25,29,32,33].

Robust Out-of-Distribution Detection. The goal is to learn a detector $G : \mathbf{x} \to \{-1, 1\}$, which outputs 1 for an in-distribution example $\mathbf{x}$ and output -1

for a clean or perturbed OOD example $\mathbf{x}$. Formally, let $\Omega(\mathbf{x})$ be a set of small perturbations on an OOD example $\mathbf{x}$. The detector is evaluated on $\mathbf{x}$ from $P_{\mathbf{X}}$ and on the worst-case input inside $\Omega(\mathbf{x})$ for an OOD example $\mathbf{x}$ from $Q_{\mathbf{X}}$. The false negative rate (FNR) and false positive rate (FPR) are defined as:

$$\mathrm{FNR}(G) = \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{X}}}\mathbb{I}[G(\mathbf{x}) = -1], \quad \mathrm{FPR}(G; Q_{\mathbf{X}}, \Omega) = \mathbb{E}_{\mathbf{x}\sim Q_{\mathbf{X}}} \max_{\delta\in\Omega(\mathbf{x})} \mathbb{I}[G(\mathbf{x}+\delta) = 1].$$

Remark. Note that test-time OOD distribution $Q_{\mathbf{X}}$ is unknown, which can be different from $U_{\mathbf{X}}$. The difference between the auxiliary data $U_{\mathbf{X}}$ and test OOD data $Q_{\mathbf{X}}$ raises the fundamental question of how to effectively leverage $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$ for improving learning the decision boundary between in- vs. OOD data. For terminology clarity, we refer to training OOD examples as *outliers*, and exclusively use *OOD data* to refer to test-time anomalous inputs.

3 Method

In this section, we introduce *Adversarial Training with informative Outlier Mining* (ATOM). We first present our method overview, and then describe details of the training objective with informative outlier mining.

Method Overview: A Conceptual Example. We use the terminology *outlier mining* to denote the process of selecting informative outlier training samples from the pool of auxiliary outlier data. We illustrate our idea with a toy example in Fig. 2, where in-distribution data consists of class-conditional Gaussians. Outlier training data is sampled from a uniform distribution from outside the support of in-distribution. Without outlier mining (*left*), we will almost sample those "easy" outliers and the decision boundary of the OOD detector learned can be loose. In contrast, with outlier mining (*right*), selective outliers close to the decision boundary between ID and OOD data, which improves OOD detection. This is particularly important for robust OOD detection where the boundary needs to have a margin from the OOD data so that even adversarial perturbation (red color) cannot move the OOD data points across the boundary. We proceed with describing the training mechanism that achieves our novel conceptual idea and will provide formal theoretical guarantees in Sect. 5.

3.1 ATOM: Adversarial Training with Informative Outlier Mining

Training Objective. The classification involves using a mixture of ID data and outlier samples. Specifically, we consider a $(K+1)$-way classifier network f, where the $(K+1)$-th class label indicates out-of-distribution class. Denote by $F_\theta(\mathbf{x})$ the softmax output of f on $\mathbf{x}$. The robust training objective is given by

$$\underset{\theta}{\text{minimize}} \quad \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{in}}^{\text{train}}}[\ell(\mathbf{x}, y; F_\theta)] + \lambda \cdot \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}^{\text{train}}} \max_{\mathbf{x}'\in\Omega_{\infty,\epsilon}(\mathbf{x})} [\ell(\mathbf{x}', K+1; F_\theta)] \tag{1}$$

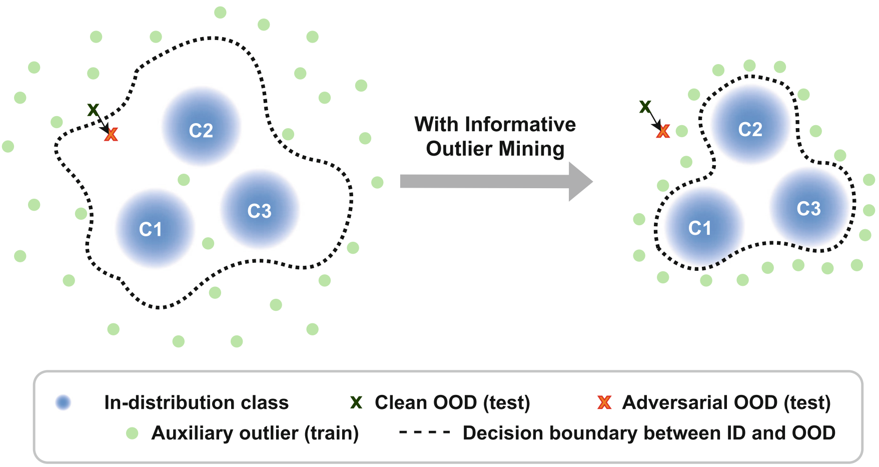
C1
C2
C3
With Informative Outlier Mining
C1
C2
C3
In-distribution class
Clean OOD (test)
Adversarial OOD (test)
Auxiliary outlier (train)
Decision boundary between ID and OOD

Fig. 2. A toy example in 2D space for illustration of informative outlier mining. With informative outlier mining, we can tighten the decision boundary and build a robust OOD detector.

where ℓ is the cross entropy loss, and $\mathcal{D}_{\text{out}}^{\text{train}}$ is the OOD training dataset. We use Projected Gradient Descent (PGD) [30] to solve the inner max of the objective, and apply it to half of a minibatch while keeping the other half clean to ensure performance on both clean and perturbed data.

Once trained, the OOD detector $G(\mathbf{x})$ can be constructed by:

$$G(\mathbf{x}) = \begin{cases} -1 & \text{if } F(\mathbf{x})_{K+1} \geq \gamma, \\ 1 & \text{if } F(\mathbf{x})_{K+1} < \gamma, \end{cases} \tag{2}$$

where γ is the threshold, and in practice can be chosen on the in-distribution data so that a high fraction of the test examples are correctly classified by G. We call $F(\mathbf{x})_{K+1}$ the *OOD score* of $\mathbf{x}$. For an input labeled as in-distribution by G, one can obtain its semantic label using $\hat{F}(\mathbf{x})$:

$$\hat{F}(\mathbf{x}) = \underset{y\in\{1,2,\cdots,K\}}{\arg\max} \ F(\mathbf{x})_y \tag{3}$$

Informative Outlier Mining. We propose to adaptively choose OOD training examples where the detector is uncertain about. Specifically, during each training epoch, we randomly sample N data points from the auxiliary OOD dataset $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$, and use the current model to infer the OOD scores[2]. Next, we sort the data points according to the OOD scores and select a subset of $n < N$ data points, starting with the qN^{th} data in the sorted list. We then use the selected samples as OOD training data $\mathcal{D}_{\text{out}}^{\text{train}}$ for the next epoch of training.

[2] Since the inference stage can be fully parallel, outlier mining can be applied with relatively low overhead.

Algorithm 1: ATOM: Adv. Training with informative Outlier Mining

```
Input: $\mathcal{D}_{\text{in}}^{\text{train}}$, $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$, $F_\theta$, $m$, $N$, $n$, $q$
Output: $\hat{F}$, $G$
for $t = 1, 2, \cdots, m$ do
    Randomly sample $N$ data points from $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$ to get a candidate set $\mathcal{S}$;
    Compute OOD scores on $\mathcal{S}$ using current model $F_\theta$ to get set
      $V = \{F(\mathbf{x})_{K+1} \mid \mathbf{x} \in \mathcal{S}\}$. Sort scores in $V$ from the lowest to the highest;
    $\mathcal{D}_{\text{out}}^{\text{train}} \leftarrow V[qN : qN + n]$ ;        /* $q \in [0, 1 - n/N]$ */
    Train $F_\theta$ for one epoch using the training objective of (1);
end
Build $G$ and $\hat{F}$ using (2) and (3) respectively;
```

Intuitively, q determines the *informativeness* of the sampled points w.r.t the OOD detector. The larger q is, the less informative those sampled examples become. Note that informative outlier mining is performed on (non-adversarial) auxiliary OOD data. Selected examples are then used in the robust training objective (1).

We provide the complete training algorithm using informative outlier mining in Algorithm 1. Importantly, the use of informative outlier mining highlights the key difference between ATOM and previous work using ***randomly*** sampled outliers [16,19,32,33].

4 Experiments

In this section, we describe our experimental setup and show that ATOM can substantially improve OOD detection performance on both clean OOD data and adversarially perturbed OOD inputs. We also conducted extensive ablation analysis to explore different aspects of our algorithm.

4.1 Setup

In-Distribution Datasets. We use CIFAR-10, and CIFAR-100 [22] datasets as in-distribution datasets. We also show results on SVHN in Appendix B.8.

Auxiliary OOD Datasets. By default, we use 80 Million Tiny Images (TinyImages) [45] as $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$, which is a common setting in prior works. We also use ImageNet-RC, a variant of ImageNet [7] as an alternative auxiliary OOD dataset.

Out-of-Distribution Datasets. For OOD test dataset, we follow common setup in literature and use six diverse datasets: `SVHN`, `Textures` [8], `Places365` [53], `LSUN (crop)`, `LSUN (resize)` [50], and `iSUN` [49].

Hyperparameters. The hyperparameter q is chosen on a separate validation set from TinyImages, which is different from test-time OOD data (see Appendix B.9). Based on the validation, we set $q = 0.125$ for CIFAR-10 and $q = 0.5$ for

CIFAR-100. For all experiments, we set $\lambda = 1$. For CIFAR-10 and CIFAR-100, we set $N = 400,000$, and $n = 100,000$. More details about experimental set up are in Appendix B.1.

Robust OOD Evaluation Tasks. We consider the following family of OOD inputs, for which we provide details and visualizations in Appendix B.5:

- **Natural OOD:** This is equivalent to the classic OOD evaluation with clean OOD input $\mathbf{x}$, and $\Omega = \emptyset$.
- **L_∞ attacked OOD (white-box):** We consider small L_∞-norm bounded perturbations on an OOD input $\mathbf{x}$ [1,30], which induce the model to produce a high confidence score (or a low OOD score) for $\mathbf{x}$. We denote the adversarial perturbations by $\Omega_{\infty,\epsilon}(\mathbf{x})$, where ϵ is the adversarial budget. We provide attack algorithms for all eight OOD detection methods in Appendix B.4.
- **Corruption attacked OOD (black-box):** We consider a more realistic type of attack based on common corruptions [17], which could appear naturally in the physical world. For each OOD image, we generate 75 corrupted images (15 corruption types × 5 severity levels), and then select the one with the lowest OOD score.
- **Compositionally attacked OOD (white-box):** Lastly, we consider applying L_∞-norm bounded attack and corruption attack jointly to an OOD input $\mathbf{x}$, as considered in [23].

Evaluation Metrics. We measure the following metrics: the false positive rate (FPR) at 5% false negative rate (FNR), and the area under the receiver operating characteristic curve (AUROC).

4.2 Results

ATOM vs. Existing Methods. We show in Table 1 that ATOM outperforms competitive OOD detection methods on both classic and adversarial OOD evaluation tasks. There are several salient observations. **First**, on classic OOD evaluation task (clean OOD data), ATOM achieves comparable or often even better performance than the current state-of-the-art methods. **Second**, on the existing adversarial OOD evaluation task, L_∞ OOD, ATOM outperforms current state-of-the-art method ACET [16] by a large margin (e.g. on CIFAR-10, our method outperforms ACET by **53.9%** measured by FPR). **Third**, while ACET is somewhat brittle under the new Corruption OOD evaluation task, our method can generalize surprisingly well to the unknown corruption attacked OOD inputs, outperforming the best baseline by a large margin (e.g. on CIFAR-10, by up to **30.99%** measured by FPR). **Finally**, while almost every method fails under the hardest compositional OOD evaluation task, our method still achieves impressive results (e.g. on CIFAR-10, reduces the FPR by **57.99%**). The performance is noteworthy since our method is not trained explicitly on corrupted OOD inputs. Our training method leads to improved OOD detection while preserving classification performance on in-distribution data (see Appendix B.14). Consistent performance improvement is observed on *alternative in-distribution*

Table 1. Comparison with competitive OOD detection methods. We use DenseNet as network architecture for all methods. We evaluate on four types of OOD inputs: (1) natural OOD, (2) corruption attacked OOD, (3) L_∞ attacked OOD, and (4) compositionally attacked OOD inputs. The description of these OOD inputs can be found in Sect. 4.1. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages and are averaged over six different OOD test datasets described in Sect. 4.1. **Bold** numbers are superior results. Results on additional in-distribution dataset SVHN are provided in Appendix B.8. Results on a different architecture, WideResNet, are provided in Appendix B.12.

$\mathcal{D}_{\text{in}}^{\text{test}}$	Method	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑
		Natural OOD		Corruption OOD		L_∞ OOD		Comp. OOD	
CIFAR-10	MSP [18]	50.54	91.79	100.00	58.35	100.00	13.82	100.00	13.67
	ODIN [27]	21.65	94.66	99.37	51.44	99.99	0.18	100.00	0.01
	Mahalanobis [26]	26.95	90.30	91.92	43.94	95.07	12.47	99.88	1.58
	SOFL [33]	2.78	99.04	62.07	88.65	99.98	1.01	100.00	0.76
	OE [19]	3.66	98.82	56.25	90.66	99.94	0.34	99.99	0.16
	ACET [16]	12.28	97.67	66.93	88.43	74.45	78.05	96.88	53.71
	CCU [32]	3.39	98.92	56.76	89.38	99.91	0.35	99.97	0.21
	ROWL [37]	25.03	86.96	94.34	52.31	99.98	49.49	100.00	49.48
	ATOM (ours)	**1.69**	**99.20**	**25.26**	**95.29**	**20.55**	**88.94**	**38.89**	**86.71**
CIFAR-100	MSP [18]	78.05	76.11	100.00	30.04	100.00	2.25	100.00	2.06
	ODIN [27]	56.77	83.62	100.00	36.95	100.00	0.14	100.00	0.00
	Mahalanobis [26]	42.63	87.86	95.92	42.96	95.44	15.87	99.86	2.08
	SOFL [33]	43.36	91.21	99.93	45.23	100.00	0.35	100.00	0.27
	OE [19]	49.21	88.05	99.96	45.01	100.00	0.94	100.00	0.59
	ACET [16]	50.93	89.29	99.53	54.19	76.27	59.45	99.71	38.63
	CCU [32]	43.04	90.95	99.90	48.34	100.00	0.75	100.00	0.48
	ROWL [37]	93.35	53.02	100.00	49.69	100.00	49.69	100.00	49.69
	ATOM (ours)	**32.30**	**93.06**	**93.15**	**71.96**	**38.72**	**88.03**	**93.44**	**69.15**

datasets (SVHN and CIFAR-100), *alternative network architecture* (WideResNet, Appendix B.12), and with *alternative auxiliary dataset* (ImageNet-RC, see Appendix B.11).

Adversarial Training Alone is not Able to Achieve Strong OOD Robustness. We perform an ablation study that isolates the effect of outlier mining. In particular, we use the same training objective as in Eq. (1), but with randomly sampled outliers. The results in Table 2 show AT (no outlier mining) is in general less robust. For example, under L_∞ OOD, AT displays 23.76% and 31.61% reduction in FPR on CIFAR-10 and CIFAR-100 respectively. This validates the importance of outlier mining for robust OOD detection, which provably improves the decision boundary as we will show in Sect. 5.

Effect of adversarial Training. We perform an ablation study that isolates the effect of adversarial training. In particular, we consider the following objective without adversarial training:

$$\underset{\theta}{\text{minimize}} \quad \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{in}}^{\text{train}}}[\ell(\mathbf{x},y;\hat{F}_\theta)] + \lambda \cdot \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}^{\text{train}}}[\ell(\mathbf{x},K+1;\hat{F}_\theta)], \qquad (4)$$

which we name *Natural Training with informative Outlier Mining* (NTOM). In Table 2, we show that NTOM achieves comparable performance as ATOM on

Table 2. Ablation on ATOM training objective. We use DenseNet as network architecture. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages and are averaged over six different OOD test datasets described in Sect. 4.1.

$\mathcal{D}_{\text{in}}^{\text{test}}$	Method	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑
		Natural OOD		Corruption OOD		L_∞ OOD		Comp. OOD	
CIFAR-10	AT (no outlier mining)	2.65	99.11	42.28	91.94	44.31	68.64	65.17	72.62
	NTOM (no adversarial training)	1.87	**99.28**	30.58	94.67	99.90	1.22	99.99	0.45
	ATOM (ours)	**1.69**	99.20	**25.26**	**95.29**	**20.55**	**88.94**	**38.89**	**86.71**
CIFAR-100	AT (no outlier mining)	51.50	89.62	99.70	58.61	70.33	58.84	99.80	34.98
	NTOM (no adversarial training)	36.94	92.61	98.17	65.70	99.97	0.76	100.00	0.16
	ATOM (ours)	**32.30**	**93.06**	**93.15**	**71.96**	**38.72**	**88.03**	**93.44**	**69.15**

natural OOD and corruption OOD. However, NTOM is less robust under L_∞ OOD (with 79.35% reduction in FPR on CIFAR-10) and compositional OOD inputs. This underlies the importance of having both adversarial training and outlier mining (ATOM) for overall good performance, particularly for robust OOD evaluation tasks.

Effect of Sampling Parameter q**.** Table 3 shows the performance with different sampling parameter q. For all three datasets, training on auxiliary outliers with large OOD scores (*i.e.*, too easy examples with $q = 0.75$) worsens the performance, which suggests the necessity to include examples on which the OOD detector is uncertain. Interestingly, in the setting where the in-distribution data and auxiliary OOD data are disjoint (*e.g.*, SVHN/TinyImages), $q = 0$ is optimal, which suggests that the hardest outliers are mostly useful for training. However, in a more realistic setting, the auxiliary OOD data can almost always contain data similar to in-distribution data (*e.g.*, CIFAR/TinyImages). Even without removing near-duplicates exhaustively, ATOM can adaptively avoid training on those near-duplicates of in-distribution data (e.g. using $q = 0.125$ for CIFAR-10 and $q = 0.5$ for CIFAR-100).

Ablation on a Different Auxiliary Dataset. To see the effect of the auxiliary dataset, we additionally experiment with ImageNet-RC as an alternative. We observe a consistent improvement of ATOM, and in many cases with performance better than using TinyImages. For example, on CIFAR-100, the FPR under natural OOD inputs is reduced from 32.30% (w/TinyImages) to 15.49% (w/ImageNet-RC). Interestingly, in all three datasets, using $q = 0$ (hardest outliers) yields the optimal performance since there are substantially fewer near-duplicates between ImageNet-RC and in-distribution data. This ablation suggests that ATOM's success does not depend on a particular auxiliary dataset. Full results are provided in Appendix B.11.

5 Theoretical Analysis

In this section, we provide theoretical insight on mining informative outliers for robust OOD detection. We proceed with a brief summary of our key results.

Table 3. Ablation study on q. We use DenseNet as network architecture. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages and are averaged over six natural OOD test datasets mentioned in Sect. 4.1. Note: the hyperparameter q is chosen on a separate validation set, which is different from test-time OOD data. See Appendix B.9 for details.

$\mathcal{D}_{\text{in}}^{\text{test}}$	Model	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑
		Natural OOD		Corruption OOD		L_∞ OOD		Comp. OOD	
SVHN	ATOM (q = 0.0)	0.07	99.97	5.47	98.52	7.02	98.00	96.33	49.52
	ATOM (q = 0.125)	1.30	99.63	34.97	94.97	39.61	82.92	99.92	6.30
	ATOM (q = 0.25)	1.36	99.60	41.98	94.30	52.39	71.34	99.97	1.35
	ATOM (q = 0.5)	2.11	99.46	44.85	93.84	59.72	65.59	99.97	3.15
	ATOM (q = 0.75)	2.91	99.26	51.33	93.07	66.20	57.16	99.96	2.04
CIFAR-10	ATOM (q = 0.0)	2.24	99.20	40.46	92.86	36.80	73.11	66.15	73.93
	ATOM (q = 0.125)	1.69	99.20	25.26	95.29	20.55	88.94	38.89	86.71
	ATOM (q = 0.25)	2.34	99.12	22.71	95.29	24.93	94.83	41.58	91.56
	ATOM (q = 0.5)	4.03	98.97	33.93	93.51	22.39	95.16	45.11	90.56
	ATOM (q = 0.75)	5.35	98.77	41.02	92.78	21.87	93.37	43.64	91.98
CIFAR-100	ATOM (q = 0.0)	44.38	91.92	99.76	60.12	68.32	65.75	99.80	49.85
	ATOM (q = 0.125)	26.91	94.97	98.35	71.53	34.66	87.54	98.42	68.52
	ATOM (q = 0.25)	32.43	93.93	97.71	72.61	40.37	82.68	97.87	65.19
	ATOM (q = 0.5)	32.30	93.06	93.15	71.96	38.72	88.03	93.44	69.15
	ATOM (q = 0.75)	38.56	91.20	97.59	58.53	62.66	78.70	97.97	54.89

Results Overview. At a high level, our analysis provides two important insights. **First**, we show that with informative auxiliary OOD data, *less* in-distribution data is needed to build a robust OOD detector. **Second**, we show using outlier mining achieves a robust OOD detector in a more *realistic* case when the auxiliary OOD data contains many outliers that are far from the decision boundary (and thus non-informative), and may contain some in-distribution data. The above two insights are important for building a robust OOD detector in practice, particularly because labeled in-distribution data is expensive to obtain while auxiliary outlier data is relatively cheap to collect. *By performing outlier mining, one can effectively reduce the sample complexity while achieving strong robustness.* We provide the main results and intuition here and refer readers to Appendix A for the details and the proofs.

5.1 Setup

Data Model. To establish formal guarantees, we use a Gaussian $\mathcal{N}(\mu, \sigma^2 I)$ to model the in-distribution $P_{\mathbf{X}}$ and the test OOD distribution can be any distribution largely supported outside a ball around μ. We consider robust OOD detection under adversarial perturbation with bounded ℓ_∞ norm, i.e., the perturbation $\|\delta\|_\infty \leq \epsilon$. Given $\mu \in \mathbb{R}^d, \sigma > 0, \gamma \in (0, \sqrt{d}), \epsilon_\tau > 0$, we consider the following data model:

- $P_{\mathbf{X}}$ **(in-distribution data)** is $\mathcal{N}(\mu, \sigma^2 I)$. The in-distribution data $\{\mathbf{x}_i\}_{i=1}^n$ is drawn from $P_{\mathbf{X}}$.
- $Q_{\mathbf{X}}$ **(out-of-distribution data)** can be any distribution from the family $\mathcal{Q} = \{Q_{\mathbf{X}} : \Pr_{\mathbf{x} \sim Q_{\mathbf{X}}}[\|\mathbf{x} - \mu\|_2 \leq \tau] \leq \epsilon_\tau\}$, where $\tau = \sigma\sqrt{d} + \sigma\gamma + \epsilon\sqrt{d}$.

- **Hypothesis class of OOD detector**: $\mathcal{G} = \{G_{u,r}(\mathbf{x}) : G_{u,r}(\mathbf{x}) = 2 \cdot \mathbb{I}[\|\mathbf{x} - u\|_2 \leq r] - 1, u \in \mathbb{R}^d, r \in \mathbb{R}_+\}$.

Here, γ is a parameter indicating the margin between the in-distribution and OOD data, and ϵ_τ is a small number bounding the probability mass the OOD distribution can have close to the in-distribution.

Metrics. For a detector G, we are interested in the False Negative Rate $\mathrm{FNR}(G)$ and the worst False Positive Rate $\sup_{Q_{\mathbf{X}} \in \boldsymbol{Q}} \mathrm{FPR}(G; Q_{\mathbf{X}}, \Omega_{\infty,\epsilon}(\mathbf{x}))$ over all the test OOD distributions $\boldsymbol{Q}$ under ℓ_∞ perturbations of magnitude ϵ. For simplicity, we denote them as $\mathrm{FNR}(G)$ and $\mathrm{FPR}(G; \boldsymbol{Q})$.

While the Gaussian data model may be simpler than the practical data, its simplicity is desirable for our purpose of demonstrating our insights. Finally, the analysis can be generalized to mixtures of Gaussians which better models real-world data.

5.2 Learning with Informative Auxiliary Data

We show that informative auxiliary outliers can reduce the sample complexity for in-distribution data. Note that learning a robust detector requires to estimate μ to distance $\gamma\sigma$, which needs $\tilde{\Theta}(d/\gamma^2)$ in-distribution data, for example, one can compute a robust detector by:

$$u = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \quad r = (1 + \gamma/4\sqrt{d})\hat{\sigma}, \tag{5}$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|_2^2$. Then we show that with informative auxiliary data, we need much less in-distribution data for learning. We model the auxiliary data $U_{\mathbf{X}}$ as a distribution over the sphere $\{\mathbf{x} : \|\mathbf{x} - \mu\|_2^2 = \sigma_o^2 d\}$ for $\sigma_o > \sigma$, and assume its density is at least η times that of the uniform distribution on the sphere for some constant $\eta > 0$, i.e., it's surrounding the boundary of $P_{\mathbf{X}}$. Given $\{\mathbf{x}_i\}_{i=1}^n$ from $P_{\mathbf{X}}$ and $\{\tilde{\mathbf{x}}_i\}_{i=1}^{n'}$ from $U_{\mathbf{X}}$, a natural idea is to compute $\bar{\mathbf{x}}$ and r as above as an intermediate solution, and refine it to have small errors on the auxiliary data under perturbation, i.e., find u by minimizing a natural "margin loss":

$$u = \mathop{\arg\min}_{p: \|p - \bar{\mathbf{x}}\|_2 \leq s} \frac{1}{n'} \sum_{i=1}^{n'} \max_{\|\delta\|_\infty \leq \epsilon} \mathbb{I}\left[\|\tilde{\mathbf{x}}_i + \delta - p\|_2 < t\right] \tag{6}$$

where s, t are hyper-parameters to be chosen. We show that with $\tilde{O}(d/\gamma^4)$ in-distribution data and sufficient auxiliary data can give a robust detector. See proof in Appendix A.2.

5.3 Learning with Informative Outlier Mining

In this subsection, we consider a more realistic data distribution where the auxiliary data can contain non-informative outliers (far away from the boundary), and in some cases mixed with in-distribution data. The non-informative outliers may not provide useful information to distinguish a good OOD detector statistically, which motivates the need for outlier mining.

Uninformative Outliers can Lead to Bad Detectors. To formalize, we model the non-informative ("easy" outlier) data as $Q_q = \mathcal{N}(0, \sigma_q^2 I)$, where σ_q is large to ensure they are obvious outliers. The auxiliary data distribution U_{mix} is then a mixture of $U_{\mathbf{X}}$, Q_q and $P_{\mathbf{X}}$, where Q_q has a large weight. Formally, $U_{\text{mix}} = \nu U_{\mathbf{X}} + (1-2\nu)Q_q + \nu P_{\mathbf{X}}$ for a small $\nu \in (0,1)$. Then we see that the previous learning rule cannot work: those robust detectors (with u of distance $O(\sigma\gamma)$ to μ) and those bad ones (with u far away from μ) cannot be distinguished. There is only a small fraction of auxiliary data from $U_{\mathbf{X}}$ for distinguishing the good and bad detectors, while the majority (those from Q_q) do not differentiate them and some (those from $P_{\mathbf{X}}$) can even penalize the good ones and favor the bad ones.

Informative Outlier Mining Improves the Detector with Reduced Sample Complexity. The above failure case suggests that a more sophisticated method is needed. Below we show that outlier mining can help to identify informative data and improve the learning performance. It can remove most data outside $U_{\mathbf{X}}$, and keep the data from $U_{\mathbf{X}}$, and the previous method can work after outlier mining. We first use in-distribution data to get an intermediate solution $\bar{\mathbf{x}}$ and r by Eqs. (5). Then, we use a simple thresholding mechanism to only pick points close to the decision boundary of the intermediate solution, which removes *non-informative outliers*. Specifically, we only select outliers with mild "confidence scores" w.r.t. the intermediate solution, i.e., the distances to $\bar{\mathbf{x}}$ fall in some interval $[a, b]$:

$$S := \{i : \|\tilde{\mathbf{x}}_i - \bar{\mathbf{x}}\|_2 \in [a, b], 1 \le i \le n'\} \tag{7}$$

The final solution u_{om} is obtained by solving Eq. (6) on only S instead of all auxiliary data. We can prove:

Proposition 1. (Error bound with outlier mining). *Suppose* $\sigma^2\gamma^2 \ge C\epsilon\sigma_o d$ *and* $\sigma\sqrt{d} + C\sigma\gamma^2 < \sigma_o\sqrt{d} < C\sigma\sqrt{d}$ *for a sufficiently large constant* C*, and* $\sigma_q\sqrt{d} > 2(\sigma_o\sqrt{d} + \|\mu\|_2)$*. For some absolute constant* c *and any* $\alpha \in (0,1)$*, if the number of in-distribution data* $n \ge \frac{Cd}{\gamma^4}\log\frac{1}{\alpha}$ *and the number of auxiliary data* $n' \ge \frac{\exp(C\gamma^4)}{\nu^2\eta^2}\log\frac{d\sigma}{\alpha}$*, then there exist parameter values* s, t, a, b *such that with probability* $\ge 1-\alpha$*, the detector* $G_{u_{\text{om}},r}$ *computed above satisfies:*

$$\text{FNR}(G_{u_{\text{om}},r}) \le \exp(-c\gamma^2), \quad \text{FPR}(G_{u_{\text{om}},r}; \mathcal{Q}) \le \epsilon_\tau.$$

This means that even in the presence of a large amount of uninformative or even harmful auxiliary data, we can successfully learn a good detector. Furthermore, this can reduce the sample size n by a factor of γ^2. For example, when $\gamma = \Theta(d^{1/8})$, we only need $n = \tilde{\Theta}(\sqrt{d})$, while in the case without auxiliary data, we need $n = \tilde{\Theta}(d^{3/4})$.

Remark. We note that when $U_{\mathbf{X}}$ is as ideal as the uniform distribution over the sphere (i.e., $\eta = 1$), then we can let u be the average of points in S after mining, which will require $n' = \tilde{\Theta}(d/(\nu^2\gamma^2))$ auxiliary data, much less than that for more general η. We also note that our analysis and the result also hold for many other auxiliary data distributions U_{mix}, and the particular U_{mix} used here is for the ease of explanation; see Appendix A for more discussions.

6 Related Work

OOD Detection. [18] introduced a baseline for OOD detection using the maximum softmax probability from a pre-trained network. Subsequent works improve the OOD uncertainty estimation by using deep ensembles [24], the calibrated softmax score [27], the Mahalanobis distance-based confidence score [26], as well as the energy score [29]. Some methods regularize the model with auxiliary anomalous data that were either realistic [19,33,35] or artificially generated by GANs [25]. Several other works [3,31,41] also explored regularizing the model to produce lower confidence for anomalous examples. Recent works have also studied the computational efficiency aspect of OOD detection [28] and large-scale OOD detection on ImageNet [21].

Robustness of OOD Detection. Worst-case aspects of OOD detection have been studied in [16,37]. However, these papers are primarily concerned with L_∞ norm bounded adversarial attacks, while our evaluation also includes common image corruption attacks. Besides, [16,32] only evaluate adversarial robustness of OOD detection on random noise images, while we also evaluate it on natural OOD images. [32] has shown the first provable guarantees for worst-case OOD detection on some balls around uniform noise, and [5] studied the provable guarantees for worst-case OOD detection not only for noise but also for images from related but different image classification tasks. Our paper proposes ATOM which achieves state-of-the-art performance on a broader family of clean and perturbed OOD inputs. The key difference compared to prior work is introducing the informative outlier mining technique, which can significantly improve the generalization and robustness of OOD detection.

Adversarial Robustness. Adversarial examples [4,14,36,44] have received considerable attention in recent years. Many defense methods have been proposed to mitigate this problem. One of the most effective methods is adversarial training [30], which uses robust optimization techniques to render deep learning models resistant to adversarial attacks. [6,34,46,52] showed that unlabeled data could improve adversarial robustness for classification.

Hard Example Mining. Hard example mining was introduced in [43] for training face detection models, where they gradually grew the set of background examples by selecting those examples for which the detector triggered a false alarm. The idea has been used extensively for object detection literature [11,13,38]. It also has been used extensively in deep metric learning [9,15,39,42,47] and deep embedding learning [10,40,48,51]. Although hard example mining has been used in various learning domains, to the best of our knowledge, we are the first to explore it to improve the robustness of out-of-distribution detection.

7 Conclusion

In this paper, we propose Adversarial Training with informative Outlier Mining (ATOM), a method that enhances the robustness of the OOD detector. We show the merit of adaptively selecting the OOD training examples which the OOD detector is uncertain about. Extensive experiments show ATOM can significantly improve the decision boundary of the OOD detector, achieving state-of-the-art performance under a broad family of *clean and perturbed* OOD evaluation tasks. We also provide a theoretical analysis that justifies the benefits of outlier mining. Further, our unified evaluation framework allows future research to examine the robustness of the OOD detector. We hope our research can raise more attention to a broader view of robustness in out-of-distribution detection.

Acknowledgments. The work is partially supported by Air Force Grant FA9550-18-1-0166, the National Science Foundation (NSF) Grants CCF-FMitF-1836978, IIS-2008559, SaTC-Frontiers-1804648 and CCF-1652140, and ARO grant number W911NF-17-1-0405. Jiefeng Chen and Somesh Jha are partially supported by the DARPA-GARD problem under agreement number 885000.

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: ICML, pp. 274–283. PMLR (2018)
2. Bendale, A., Boult, T.: Towards open world recognition. In: CVPR. pp. 1893–1902 (2015)
3. Bevandić, P., Krešo, I., Oršić, M., Šegvić, S.: Discriminative out-of-distribution detection for semantic segmentation. arXiv preprint arXiv:1808.07703 (2018)
4. Biggio, B., et al.: Evasion attacks against machine learning at test time. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013. LNCS (LNAI), vol. 8190, pp. 387–402. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40994-3_25
5. Bitterwolf, J., Meinke, A., Hein, M.: Certifiably adversarially robust detection of out-of-distribution data. In: NeurIPS 33 (2020)
6. Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J.C., Liang, P.S.: Unlabeled data improves adversarial robustness. In: NeurIPS, pp. 11190–11201 (2019)
7. Chrabaszcz, P., Loshchilov, I., Hutter, F.: A downsampled variant of imagenet as an alternative to the cifar datasets. arXiv preprint arXiv:1707.08819 (2017)
8. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR (2014)
9. Cui, Y., Zhou, F., Lin, Y., Belongie, S.: Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In: CVPR, pp. 1153–1162 (2016)

10. Duan, Y., Chen, L., Lu, J., Zhou, J.: Deep embedding learning with discriminative sampling policy. In: CVPR, pp. 4964–4973 (2019)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2009)
12. Filos, A., Tigkas, P., McAllister, R., Rhinehart, N., Levine, S., Gal, Y.: Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In: ICML, pp. 3145–3153. PMLR (2020)
13. Gidaris, S., Komodakis, N.: Object detection via a multi-region and semantic segmentation-aware CNN model. In: ICCV, pp. 1134–1142 (2015)
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. ICLR (2015)
15. Harwood, B., Kumar BG, V., Carneiro, G., Reid, I., Drummond, T.: Smart mining for deep metric learning. In: ICCV, pp. 2821–2829 (2017)
16. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: CVPR, pp. 41–50 (2019)
17. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)
18. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
19. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: ICLR (2019)
20. Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: detecting out-of-distribution image without learning from out-of-distribution data. In: CVPR (2020)
21. Huang, R., Li, Y.: Towards scaling out-of-distribution detection for large semantic space. In: CVPR (2021)
22. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
23. Laidlaw, C., Feizi, S.: Functional adversarial attacks. In: NeurIPS, pp. 10408–10418 (2019)
24. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS, pp. 6402–6413 (2017)
25. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: ICLR (2018)
26. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS, pp. 7167–7177 (2018)
27. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: ICLR (2018)
28. Lin, Z., Dutta, S., Li, Y.: Mood: Multi-level out-of-distribution detection. In: CVPR (2021)
29. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: NeurIPS (2020)
30. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
31. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. In: NeurIPS, pp. 7047–7058 (2018)
32. Meinke, A., Hein, M.: Towards neural networks that provably know when they don't know. In: ICLR (2020)

33. Mohseni, S., Pitale, M., Yadawa, J., Wang, Z.: Self-supervised learning for generalizable out-of-distribution detection. AAAI **34**, 5216–5223 (2020)
34. Najafi, A., Maeda, S.I., Koyama, M., Miyato, T.: Robustness to adversarial perturbations in learning from incomplete data. In: NeurIPS, pp. 5541–5551 (2019)
35. Papadopoulos, A., Rajati, M.R., Shaikh, N., Wang, J.: Outlier exposure with confidence control for out-of-distribution detection. Neurocomputing **441**, 138–150 (2021)
36. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387. IEEE (2016)
37. Sehwag, V., Bhagoji, A.N., Song, L., Sitawarin, C., Cullina, D., Chiang, M., 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 (2019)
38. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: CVPR, pp. 761–769 (2016)
39. Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F.: Discriminative learning of deep convolutional feature point descriptors. In: ICCV, pp. 118–126 (2015)
40. Smirnov, E., Melnikov, A., Oleinik, A., Ivanova, E., Kalinovskiy, I., Luckyanets, E.: Hard example mining with auxiliary embeddings. In: CVPR Workshops, pp. 37–46 (2018)
41. Subramanya, A., Srinivas, S., Babu, R.V.: Confidence estimation in deep neural networks via density modelling. arXiv preprint arXiv:1707.07013 (2017)
42. Suh, Y., Han, B., Kim, W., Lee, K.M.: Stochastic class-based hard example mining for deep metric learning. In: CVPR, pp. 7251–7259 (2019)
43. Sung, K.K.: Learning and example selection for object and pattern detection. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA (1995)
44. Szegedy, C., et al.: Intriguing properties of neural networks. In: ICLR (2014)
45. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1958–1970 (2008)
46. Uesato, J., Alayrac, J.B., Huang, P.S., Stanforth, R., Fawzi, A., Kohli, P.: Are labels required for improving adversarial robustness? NeurIPS (2019)
47. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV, pp. 2794–2802 (2015)
48. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: ICCV, pp. 2840–2848 (2017)
49. Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J.: Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755 (2015)
50. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
51. Yuan, Y., Yang, K., Zhang, C.: Hard-aware deeply cascaded embedding. In: ICCV, pp. 814–823 (2017)
52. Zhai, R., et al.: Adversarially robust generalization just requires more unlabeled data. arXiv preprint arXiv:1906.00555 (2019)
53. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2017)