Robust Out-of-distribution Detection via Informative Outlier Mining

Jiefeng Chen [1] **Yixuan Li** [2] **Xi Wu** [3] **Yingyu Liang** [1] **Somesh Jha** [1]

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 under small adversarial perturbations. In this paper, we propose a simple and effective method, *Adversarial Training with informative Outlier Mining* (ATOM), to robustify OOD detection. Our key observation is that while unlabeled data can be used as auxiliary OOD training data, the majority of these data points are not informative to improve the decision boundary of the OOD detector. We show that, by carefully choosing which outliers to train on, one can significantly improve the robustness of the OOD detector, and somewhat surprisingly, generalize to some adversarial attacks not seen during training. We provide additionally a unified evaluation framework that allows future research examining the robustness of OOD detection algorithms. ATOM achieves state-of-the-art performance under a broad family of natural and perturbed OOD evaluation tasks, surpassing previous methods by a large margin. Finally, we provide theoretical insights for the benefit of outlier mining.

1. Introduction

Out-of-distribution (OOD) uncertainty estimation has become an indispensable part of building reliable open-world machine learning models (Amodei et al., 2016). An OOD detector determines whether an input is from the same distribution as the training data (in-distribution), or a different distribution (out-of-distribution). The performance of the OOD detector is central for safety-critical applications such as autonomous driving (Eykholt et al., 2018) or rare disease identification (Blauwkamp et al., 2019).

Despite exciting progress made in OOD detection, previous methods mostly focused on natural OOD data (Hendrycks & Gimpel, 2016; Liang et al., 2017; Lee et al., 2018; Lakshminarayanan et al., 2017; Hendrycks et al., 2018; Mohseni et al., 2020). Scant attention has been paid to the robustness aspect of OOD detection. Recently, Sehwag et al. demonstrated that OOD detection methods can be evaded by worst-case adversarial perturbations (Papernot et al., 2016; Goodfellow et al., 2014; Biggio et al., 2013; Szegedy et al., 2013). For example, an OOD image (e.g., mailbox) can be perturbed to be misclassified by the OOD detector as in-distribution (traffic sign data). Such an adversarial OOD example[1] is then passed to the image classifier and trigger undesirable prediction and action (e.g., speed limit 70). Therefore, the failure mode leads to the following question: how can we make out-of-distribution detection algorithm robust in the presence of small perturbations to OOD inputs?

Motivated by this, we propose a method called *Adversarial Training with informative Outlier Mining* (ATOM), which achieves state-of-the-art performance on a broad family of natural and perturbed OOD inputs. Our key observation is that while unlabeled data (Hendrycks et al., 2018) provides the abundance of OOD data to train on, the majority OOD examples can be too easy to provide useful information and meaningfully improve the decision boundary of OOD detector. We show that, by carefully choosing which OOD data to train on, one can significantly improve the robustness of an OOD detector, and somewhat surprisingly, generalize to unseen adversarial attacks.

Contributions. We provide a unified framework that allows examining the robustness of OOD detection algorithms under a broad family of OOD inputs, as illustrated in Figure 1. Our evaluation goes beyond previous approaches that primarily focus on natural OOD inputs. Under this taxonomy, we extensively examine the robustness of eight common OOD detection methods. Our experiments reveal that existing methods have heterogeneous performance across various types of perturbations. In particular, we show that methods relying on pre-trained neural networks (Hendrycks & Gimpel, 2016; Liang et al., 2017; Lee et al., 2018) are fragile across all types of perturbations. While (Hein et al., 2019) provides robustness against adversarial OOD examples gen-

[1]University of Wisconsin-Madison [2]Stanford University [3]Google. Correspondence to: Jiefeng Chen <jchen662@wisc.edu>.

ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

[1]We note here that the adversarial OOD examples are constructed w.r.t the OOD detectors $G(x)$, rather than the image classification model $f(x)$.

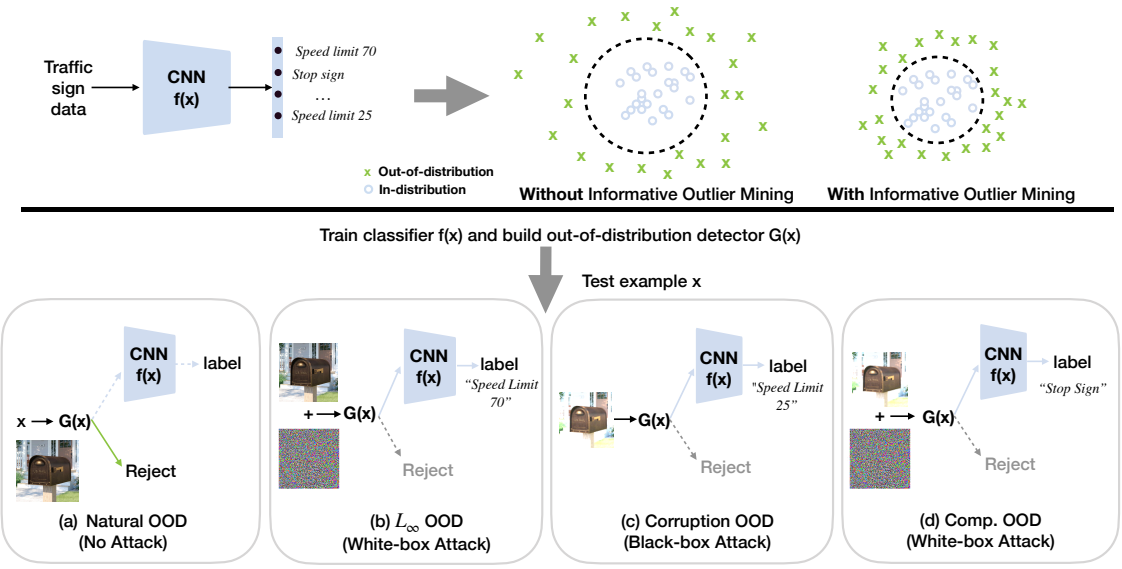


Figure 1: When deploying an image classification system (OOD detector $G(x)$ + image classifier $f(x)$) in an open world, there can be multiple types of out-of-distribution examples. We consider a broad family of OOD inputs, including (a) natural OOD, (b) adversarially perturbed OOD under L_∞ norm perturbation, and (c) corruption attacked OOD. 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 $f(x)$ 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 $G(x)$, which leads to improved performance across all types of OOD inputs. Solid lines are actual computation flow.

erated by L_∞-norm bounded attacks, such defense can be somewhat brittle under attacks not seen during training.

To this end, we devise a simple and effective method, ATOM, which improves the OOD detection performance on both clean and perturbed inputs. The key idea of our method is to adaptively choose informative OOD training examples that the OOD detector is mildly uncertain about. When evaluating on natural OOD data, our method outperforms state-of-the-art method SOFL (Mohseni et al., 2020) on all datasets. On CIFAR-10, our method outperforms the best baseline (Hein et al., 2019) by **60.55%** (FPR) under L_∞ attacked OOD inputs. Our method can generalize surprisingly well to unknown corrupted OOD inputs, outperforming the best baseline by **29.55%** measured by FPR. Finally, while almost every method fails under the strongest compositional attack, our method reduces the FPR by **61.78%**. Our method leads to improved OOD detection while maintaining similar classification accuracy on in-distribution data as a pre-trained model. We conduct ablation analysis to explore the effect of informative outlier sampling.

Lastly, we provide theoretical analysis formalizing the intuition behind our method. Under a Gaussian model of the data, we show that using outlier mining helps learn a correct detector in the presence of non-informative examples. Our theoretical results justify using auxiliary unlabeled data and outlier mining for robust OOD detection. Our code is available at: https://github.com/jfc43/informative-outlier-mining.

2. Problem Statement

Preliminaries. 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. An classifier $f(\mathbf{x})$ is trained on the in-distribution $P_{\boldsymbol{X}}$, the marginal distribution of $P_{\boldsymbol{X},Y}$. The OOD examples are revealed during test time, which are from a different distribution $Q_{\boldsymbol{X}}$, potentially with perturbations added. Formally, let $\Omega(\mathbf{x})$ be a set of small perturbations on an OOD example $\mathbf{x}$. The task of *robust out-of-distribution detection* is to learn a detector $G : \mathbf{x} \to \{-1, 1\}$, which outputs 1 for $\mathbf{x}$ from $P_{\mathbf{X}}$, and output -1 for the worst-case input inside $\Omega(\mathbf{x})$ given an OOD example from $Q_{\mathbf{X}}$. The false negative rate (FNR) and false positive rate (FPR) are defined as:

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

Robust OOD Evaluation Tasks. We consider the following family of natural and adversarial OOD inputs:

- **Natural OOD**: This is equivalent to the classic OOD evaluation where clean input $\mathbf{x}$ is used and $\Omega = \emptyset$.
- L_∞ **attacked OOD (white-box)**: We consider small L_∞-norm bounded perturbations on $\mathbf{x}$ (Madry et al.,

2017; Athalye et al., 2018), which induce the model to produce high confidence scores (or low OOD scores) for OOD inputs. The set of adversarial perturbations is $\Omega_{\infty,\epsilon}(\mathbf{x})$, where ϵ is the adversarial budget. We provide attack algorithms for all eight OOD detection methods in Appendix E.4.

- **Corruption attacked OOD (black-box)**: We consider a more realistic type of attack based on common corruptions (Hendrycks & Dietterich, 2019), which could appear naturally in the physical world. Some corruptions include noise, blur, and weather, etc. We provide details in Appendix E.4.
- **Compositionally attacked OOD (white-box)**: Lastly, we consider combining L_∞ attack and corruption attack, as considered in (Laidlaw & Feizi, 2019).

We show visualizations for four types of OOD samples in Appendix E.5. To our knowledge, we are the first to study the performance of OOD methods under all these tasks.

3. ATOM: Adversarial Training with Informative Outlier Mining

In this section, we introduce our method, *Adversarial Training with informative Outlier Mining* (ATOM), for robust OOD detection.

Training Objective. We consider a $(K+1)$-way classifier network $\hat{f}$, where the $(K+1)$-th class label indicates out-of-distribution class. Denote by $\hat{F}_\theta(\mathbf{x})$ the softmax output of $\hat{f}$ on $\mathbf{x}$. The robust training objective is given by

$$\begin{aligned} &\underset{\theta}{\text{minimize}} \quad \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{in}}^{\text{train}}}[\ell(\mathbf{x},y;\hat{F}_\theta)] \\ &+\lambda\cdot\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}^{\text{train}}}\max_{\mathbf{x}'\in\Omega_{\infty,\epsilon}(\mathbf{x})}[\ell(\mathbf{x}',K+1;\hat{F}_\theta)], \end{aligned} \tag{2}$$

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) (Madry et al., 2017) to solve the inner max of the objective, and apply it to half of a minibatch while keeping the other half clean to ensure proper performance on both clean and perturbed data. Once trained , the OOD detector $G(\mathbf{x})$ can be constructed by:

$$G(\mathbf{x}) = \begin{cases} -1 & \text{if } \hat{F}(\mathbf{x})_{K+1} \geq \gamma, \\ 1 & \text{if } \hat{F}(\mathbf{x})_{K+1} < \gamma, \end{cases} \tag{3}$$

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

$$F(\mathbf{x}) = \underset{y\in\{1,2,\cdots,K\}}{\arg\max}\ \hat{F}(\mathbf{x})_y \tag{4}$$

Informative Outlier Mining. When training a neural network, we may not have data from the test out-of-distribution $Q_{\mathbf{X}}$, but instead, have an unlabeled auxiliary dataset $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$ from distribution $U_{\mathbf{X}}$. While unlabeled data gives rise to the abundance of OOD data to train, we observe that the above training objective quickly converges to the solution where OOD training data yield high OOD scores (see Figure 2). Continuing the training process on these OOD data points can no longer provide meaningful information that improves the decision boundary of the OOD detector.

Motivated by this, we propose to adaptively choose OOD training examples where the detector is mildly uncertain about. We provide the complete training algorithm using informative outlier mining in Appendix B. Our method is different from random sampling as used in previous works (Hendrycks et al., 2018; Hein et al., 2019; Meinke & Hein, 2019; Mohseni et al., 2020). Specifically, during each training epoch, we randomly sample N data points from the unlabeled OOD dataset $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$, and use the current model to infer the OOD scores . After that, 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. 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. Our empirical and theoretical studies reveal the importance of outlier mining for the robust OOD detection task. We report ablation analysis in Appendix A.

4. Experiments

4.1. Setup

In-distribution Datasets. we use CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) as in-distribution datasets.

Out-of-distribution Datasets. For auxiliary outlier dataset $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$, we use 80 Million Tiny Images (Torralba et al., 2008). For OOD test dataset, we follow the procedure in (Liang et al., 2017; Hendrycks et al., 2018) and use six natural image datasets: `SVHN`, `Textures`, `Places365`, `LSUN (crop)`, `LSUN (resize)`, and `iSUN`.

Hyperparameters. We set $\lambda = 1$, $N = 4\times10^5$, and $n = 10^5$. The hyperparameter q is chosen on a validation set from Tiny Images (Torralba et al., 2008), which does not depend on test-time OOD data (see Appendix E.8). We set $q = 0.125$ for CIFAR-10 and $q = 0.25$ for CIFAR-100.

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

More details about experimental set up are in Appendix E.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	MSP	50.54	91.79	100.00	58.34	100.00	13.83	100.00	13.67
	ODIN	21.72	94.72	99.30	52.32	99.99	0.17	100.00	0.01
	Mahalanobis	28.50	89.60	94.58	37.76	97.67	3.90	99.93	0.32
	SOFL	2.78	99.04	61.82	88.72	99.98	1.08	100.00	0.77
	OE	3.66	98.82	56.44	90.66	99.95	0.35	99.99	0.16
	ACET	13.13	97.61	68.54	88.00	75.86	77.66	97.86	52.99
	CCU	3.39	98.92	56.50	89.34	99.90	0.36	99.98	0.21
	ROWL	43.14	77.78	94.19	52.26	93.40	52.65	97.86	50.42
	ATOM (ours)	**2.07**	**99.11**	**26.95**	**94.96**	**15.31**	**97.33**	**36.08**	**93.78**
CIFAR-100	MSP	78.05	76.11	100.00	30.08	100.00	2.35	100.00	2.13
	ODIN	53.03	84.45	100.00	36.36	100.00	0.40	100.00	0.01
	Mahalanobis	43.25	85.65	**96.62**	33.47	95.13	26.71	99.91	10.32
	SOFL	43.36	91.21	99.92	45.20	100.00	0.42	100.00	0.30
	OE	49.21	88.05	99.96	45.10	100.00	0.97	100.00	0.59
	ACET	47.69	88.47	99.86	43.38	79.33	50.59	98.60	24.96
	CCU	43.04	90.95	99.90	48.32	100.00	0.78	100.00	0.47
	ROWL	95.82	51.90	100.00	49.80	99.99	49.81	100.00	49.80
	ATOM (ours)	**34.06**	**93.79**	99.08	**72.27**	**52.89**	**82.61**	**96.83**	**68.93**

Table 1: Comparison with competitive OOD detection 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. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages and are averaged over six OOD test datasets described in section 4.1. **Bold** numbers are superior results.

4.2. Results

We show in Table 1 that ATOM outperforms existing methods on both natural and perturbed OOD evaluation tasks. First, when evaluating on natural OOD data, ATOM outperforms current state-of-the-art method SOFL (Mohseni et al., 2020). On CIFAR-10, our method outperforms the best baseline ACET (Hein et al., 2019) by **60.55%** (FPR) under the L_∞ attacked OOD task. While ACET is somewhat brittle under unknown attacks, our method can generalize surprisingly well to unknown corruption attacked OOD inputs, outperforming the best baseline by **29.55%** measured by FPR. Finally, while almost every method fails under the strongest compositional attack, our method reduces the FPR by **61.78%**. The performance is noteworthy since our method is not trained explicitly on corrupted OOD inputs.

We also conduct ablation study on the effect of outlier mining, and provide details in Appendix A. The ablation reveals that (1) using outlier mining considerably outperforms random sampling, and (2) sampling from a mild range of OOD scores is important.

5. Theoretical Insights

We provide analysis in a Gaussian model which formalizes the intuition behind outlier mining. Detailed proofs and additional results are deferred to Appendix D.

Gaussian data model. Given $\mu \in \mathbb{R}^d, \sigma > 0, \nu > 0$, in our model: (1) $P_{\mathbf{X}}$ is $\mathcal{N}(\mu, \sigma^2 I)$; (2) $Q_{\mathbf{X}}$ can be any distribution from the family $\mathcal{Q} = \{\mathcal{N}(-\mu + v, \sigma^2 I) : v \in \mathbb{R}^d, \|v\|_2 \leq \nu\}$; (3) the hypothesis class for detectors is $\mathcal{G} = \{G_\theta(\mathbf{x}) = \text{sign}(\theta^\top \mathbf{x}) : \theta \in \mathbb{R}^d\}$. Assume the following parameter values: first choose an integer $n_0 > 0$, then let

$$\epsilon \in (0, 1/2), \|\mu\|_2^2 = d \gg n_0/\epsilon^4, \sigma^2 = \sqrt{dn_0}, \nu \leq \|\mu\|_2/4.$$

We consider the FNR and the FPR under ℓ_∞ perturbations of magnitude ϵ. Since $Q_{\mathbf{X}}$ is not accessible during training time, we bound $\sup_{Q_{\mathbf{X}} \in \mathcal{Q}} \text{FPR}(G; Q_{\mathbf{X}})$.

Now assume we have n in-distribution data, and n' auxiliary outliers from the distribution U_{mix}, a uniform mixture of $\mathcal{N}(-\mu, \sigma^2 I)$ and $\mathcal{N}(-10\mu, \sigma^2 I)$. Consider the following algorithm: first average in-distribution data to get an intermediate solution $\hat{\theta}_{\text{int}}$, then select outliers $\tilde{\mathbf{x}}$ with confidence scores $f(\tilde{\mathbf{x}}) = 1/(1 + e^{-\tilde{\mathbf{x}}^\top \hat{\theta}_{\text{int}}/d}) \in [a, b]$, and the final solution $\hat{\theta}_{\text{om}}$ is -1 times the average of the selected outliers. By only picking points with *mild* confidence scores, we remove easy outliers far from in-distribution (e.g., most points in $\mathcal{N}(-\mu_o, \sigma^2 I)$), and also difficult outliers (e.g., tail of $\mathcal{N}(-\mu, \sigma^2 I)$ in the support of in-distribution). This yields a good detector:

Proposition 1. *There exist thresholds a and b for $\hat{\theta}_{\text{om}}$, and a universal constant c such that for the parameter setting (5) with $\sqrt{d/n_0} \geq c(\log d + 1/\epsilon^2)$, we have that if $n \geq cn_0 \log d$ and $n' \geq (d + n_0 \cdot 4\epsilon^2)\sqrt{d/n_0}$, then*

$$\mathbb{E}_{\hat{\theta}_{\text{om}}} \text{FNR}(G_{\hat{\theta}_{\text{om}}}) \leq 10^{-3}$$
$$\mathbb{E}_{\hat{\theta}_{\text{om}}} \sup_{Q_{\mathbf{X}} \in \mathcal{Q}} \text{FPR}(G_{\hat{\theta}_{\text{om}}}; Q_{\mathbf{X}}) \leq 10^{-3}.$$

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2): 151–175, 2010.

Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Timothy A Blauwkamp, Simone Thair, Michael J Rosen, Lily Blair, Martin S Lindner, Igor D Vilfan, Trupti Kawli, Fred C Christians, Shivkumar Venkatasubrahmanyam, Gregory D Wall, et al. Analytical and clinical validation of a microbial cell-free dna sequencing test for infectious disease. *Nature microbiology*, 4(4):663–674, 2019.

Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Hard negative mining for metric learning based zero-shot classification. In *European Conference on Computer Vision*, pp. 524–531. Springer, 2016.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pp. 11190–11201, 2019.

Hyunsun Choi and Eric Jang. Generative ensembles for robust anomaly detection. 2018.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159, 2011.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634, 2018.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE international conference on computer vision*, pp. 1134–1142, 2015.

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

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

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

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

Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 10408–10418, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

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

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

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018.

Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. *arXiv preprint arXiv:1909.12180*, 2019.

Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. 2020.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387. IEEE, 2016.

Phillip Pope, Yogesh Balaji, and Soheil Feizi. Adversarial robustness of flow-based generative models. *arXiv preprint arXiv:1911.08654*, 2019.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pp. 14680–14691, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pp. 5014–5026, 2018.

Vikash Sehwag, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 105–116, 2019.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016.

Akshayvarun Subramanya, Suraj Srinivas, and R Venkatesh Babu. Confidence estimation in deep neural networks via density modelling. *arXiv preprint arXiv:1707.07013*, 2017.

Kah-Kay Sung. Learning and example selection for object and pattern detection. 1996.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pp. 4790–4798, 2016.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

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

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

Supplementary Material

A. Ablation Study: Informative Outlier Mining

How does informative outlier mining compare to random sampling? We perform an ablation study on how different informative outlier mining strategies affect the performance of ATOM. Table 2 shows the performance where we vary the sampling interval with different q. First, we observe that using informative outlier mining leads to a more robust decision boundary than random sampling. For example, on CIFAR-10, our method ($q = 0.125$) achieves FPR 26.95% under unseen corrupted OOD inputs, compared to random sampling (45.70%). On natural OOD inputs, our method ($q = 0.5$) reduces the FPR by 23.15% on more complex dataset CIFAR-100. We provide theoretical reasoning for the benefit of outlier mining in Section 5.

How does the sampling parameter affect performance? The ablation also reveals the importance of sampling from a mild range of OOD scores. On the one hand, training on OOD inputs primarily with large OOD scores (i.e., too easy examples with $q = 0.75$) worsens the performance on CIFAR-10, which suggests the necessity to include examples on which the OOD detector is uncertain. On the other hand, training on the hardest examples (i.e., $q = 0$) with lowest OOD scores might be harmful since some of the data points resemble in-distribution data, as visually evidenced in Figure 2(c).

$\mathcal{D}_{\text{in}}^{\text{test}}$	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**	
CIFAR-10	ATOM (rand. sample)	2.93	99.09	45.70	90.98	52.19	71.13	69.19	73.37
	ATOM (q=0.0)	2.54	99.06	40.16	92.87	52.27	65.40	68.30	68.18
	ATOM (q=0.125)	2.07	99.11	26.95	94.96	15.31	97.33	36.08	93.78
	ATOM (q=0.25)	2.71	99.13	32.21	93.89	25.25	95.47	44.02	92.01
	ATOM (q=0.5)	5.03	98.79	40.22	92.64	36.49	92.48	59.38	88.21
	ATOM (q=0.75)	7.18	98.41	61.26	87.45	20.80	95.79	62.32	87.18
CIFAR-100	ATOM (rand. sample)	54.16	88.96	99.88	54.58	62.29	69.94	95.95	44.99
	ATOM (q=0.0)	57.46	88.28	99.94	49.84	60.33	76.27	96.56	45.43
	ATOM (q=0.125)	47.38	91.62	99.57	63.98	44.75	89.44	90.10	65.54
	ATOM (q=0.25)	34.06	93.79	99.08	72.27	52.89	82.61	96.83	68.93
	ATOM (q=0.5)	31.01	93.27	97.01	64.05	63.77	76.53	95.87	59.25
	ATOM (q=0.75)	33.18	92.46	96.90	64.85	53.16	86.13	96.78	64.63

Table 2: Ablation study on informative outlier mining. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages and are averaged over six OOD test datasets mentioned in section 4.1.

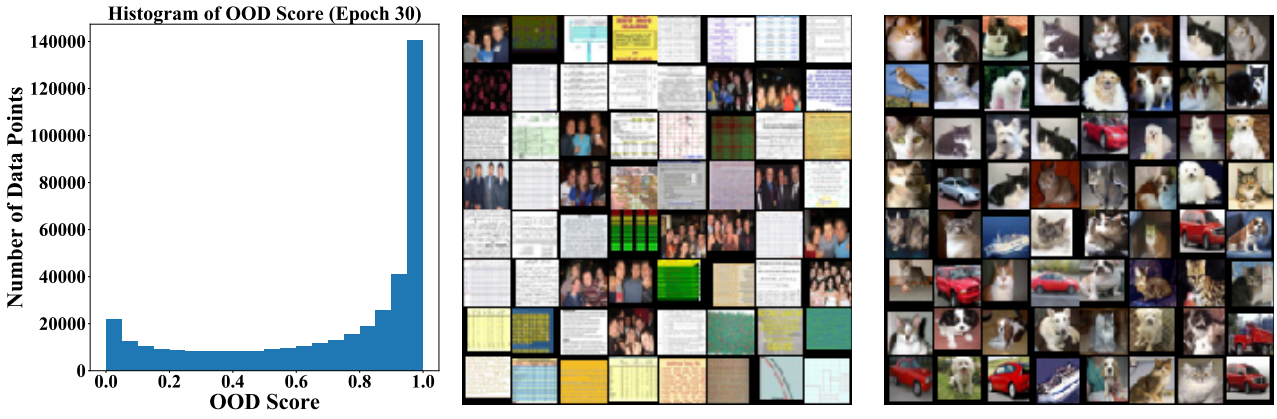
B. ATOM: Extended Description

Algorithm 1 ATOM: Adversarial Training with informative Outlier Mining.

input $\mathcal{D}_{\text{in}}^{\text{train}}$, $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$, $\hat{F}_\theta$, m, N, n, q
output 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 $\hat{F}_\theta$ to get set $V = \{\hat{F}(\mathbf{x})_{K+1} \mid \mathbf{x} \in \mathcal{S}\}$.
    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 $\hat{F}_\theta$ for one epoch using the training objective of (2).
  end for
  Build G and F using objective (3) and (4) respectively.

C. Related Work

Discriminative Based Out-of-Distribution Detection. Hendrycks et al. (Hendrycks & Gimpel, 2016) introduced a baseline approach for OOD detection using the maximum softmax probability from a pre-trained network. Several works attempt to



(a) OOD score distribution (b) highest OOD scores (c) lowest OOD scores

Figure 2: On CIFAR-10, we train a DenseNet with objective (2) for 100 epochs **without** informative outlier mining. At epoch 30, we randomly sample 400,000 data points from $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$, and plot the OOD score frequency distribution (a). We observe that the model quickly converges to solution where OOD score distribution becomes dominated by *easy* examples with score closer to 1, as shown in (b). Therefore, training on these easy OOD data points can no longer help improve the decision boundary of OOD detector. We also observe that training on *too hard* examples with score closer to 0 might be harmful since those examples resemble in-distribution data (c).

improve the OOD uncertainty estimation by using deep ensembles (Lakshminarayanan et al., 2017), the calibrated softmax score (Liang et al., 2017), and the Mahalanobis distance-based confidence score (Lee et al., 2018). Some methods also modify the neural networks by re-training or fine-tuning on some auxiliary anomalous data that are or realistic (Hendrycks et al., 2018; Mohseni et al., 2020) or artificially generated by GANs (Lee et al., 2017). Many other works (Subramanya et al., 2017; Malinin & Gales, 2018; Bevandić et al., 2018) also regularize the model to have lower confidence for anomalous examples. Worst-case aspects of OOD detection have previously been studied in (Hein et al., 2019; Meinke & Hein, 2019; Sehwag et al., 2019). However, these papers are primarily concerned with L_∞ norm bounded adversarial attacks. In this paper, we consider a broader family of clean and perturbed OOD inputs to examine the robustness of OOD detection algorithms.

Generative Modeling Based Out-of-distribution Detection. Generative models (Dinh et al., 2016; Kingma & Welling, 2013; Rezende et al., 2014; Van den Oord et al., 2016; Tabak & Turner, 2013) can be alternative approaches for detecting OOD examples, as they directly estimate the in-distribution density and can declare a test sample to be out-of-distribution if it lies in the low-density regions. However, as shown by (Nalisnick et al., 2018), deep generative models can assign a high likelihood to out-of-distribution data. Deep generative models can be more effective for out-of-distribution detection using alternative metrics (Choi & Jang, 2018), likelihood ratio (Ren et al., 2019; Serrà et al., 2019), and modified training technique (Hendrycks et al., 2018). Recently, (Pope et al., 2019) shows that flow-based generative models are sensitive under adversarial attacks. Note that we mainly considered discriminative-based approaches, which can be more competitive due to the availability of label information (and, in some cases, auxiliary outlier data (Hein et al., 2019; Hendrycks et al., 2018; Meinke & Hein, 2019; Mohseni et al., 2020)).

Adversarial Robustness. Adversarial examples (Goodfellow et al., 2014; Papernot et al., 2016; Biggio et al., 2013; Szegedy et al., 2013) 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 (Madry et al., 2017), which uses robust optimization techniques to render deep learning models resistant to adversarial attacks. (Carmon et al., 2019) shows that unlabeled data can improve adversarial robustness on in-distribution via self-training. Our method is also related to self-training, but our focus is to improve the generalization and robustness of OOD detection.

Hard Example Mining. Hard example mining was introduced in the work (Sung, 1996) for training face detection models, where they gradually grow the set of background examples by selecting those examples for which the detector triggers a false alarm. The idea has been used extensively for object detection literature (Felzenszwalb et al., 2009; Gidaris & Komodakis, 2015; Shrivastava et al., 2016). (Bucher et al., 2016) uses hard negative mining for zero-shot classification. To the best of our knowledge, we are the first to explore hard example mining for out-of-distribution detection.

D. Theoretical Analysis

D.1. A General Error Bound

The interesting questions related to our method are: **(1)** why auxiliary data from $U_{\mathbf{X}}$ helps? **(2)** why the detector G trained on $U_{\mathbf{X}}$ generalizes to different OOD distributions $Q_{\mathbf{X}}$ in test time?

To see what $Q_{\mathbf{X}}$ can benefit from the auxiliary data, we adopt the domain adaption framework (Ben-David et al., 2010). Recall that in domain adaptation there are two domains $\boldsymbol{s}, \boldsymbol{t}$, each being a distribution over the input space $\mathcal{X}$ and label space $\{-1, 1\}$. A classifier is trained on $\boldsymbol{s}$ then applied on $\boldsymbol{t}$. At a high level, we view our OOD detection problem as classification, where the source domain $\boldsymbol{s}$ is $P_{\mathbf{X}}$ with labels 1 and $U_{\mathbf{X}}$ with labels -1, and the target domain $\boldsymbol{t}$ is $P_{\mathbf{X}}$ with labels 1 and $Q_{\mathbf{X}}$ with label -1.

We focus on the FPR metric below; the argument for FNR is similar. Suppose we learn the OOD detector from a hypothesis class $\mathcal{G}$. Following (Ben-David et al., 2010), we define (a variant) of the divergence of $Q_{\mathbf{X}}$ and $U_{\mathbf{X}}$ w.r.t. the hypothesis class $\mathcal{G}$ as

$$d_{\mathcal{G}}(Q_{\mathbf{X}}, U_{\mathbf{X}}) = \sup_{G, G' \in \mathcal{G}} v(G, G'; Q_{\mathbf{X}}) - v(G, G'; U_{\mathbf{X}})$$

where

$$v(G, G'; D) = \mathrm{FPR}(G; D, \Omega) - \mathrm{FPR}(G'; D, \Omega)$$

is the error difference of G and G' on the distribution D.

The divergence upper bounds the change of the hypothesis error difference between $Q_{\mathbf{X}}$ and $U_{\mathbf{X}}$. If it is small, then for any $G, G' \in \mathcal{G}$ where G has a smaller error than G' in $U_{\mathbf{X}}$, we know that G will also have a smaller (or not too larger) error than G' in $Q_{\mathbf{X}}$. That is, if the divergence is small, then the ranking of the hypotheses w.r.t. the error is roughly the same in both distributions. This *rank-preserving* property thus makes sure that a good hypothesis learned in $U_{\mathbf{X}}$ will also be good for $Q_{\mathbf{X}}$.

Now we show that, if $d_{\mathcal{G}}(Q_{\mathbf{X}}, U_{\mathbf{X}})$ is small (i.e., $Q_{\mathbf{X}}$ and $U_{\mathbf{X}}$ are aligned w.r.t. the class $\mathcal{G}$), then a detector G with small FPR on $U_{\mathbf{X}}$ will also have small FPR on $Q_{\mathbf{X}}$.

Proposition 2. *For any $G \in \mathcal{G}$,*

$$\mathrm{FPR}(G; Q_{\mathbf{X}}, \Omega) \leq \inf_{G^* \in \mathcal{G}} \mathrm{FPR}(G^*; Q_{\mathbf{X}}, \Omega) + \mathrm{FPR}(G; U_{\mathbf{X}}, \Omega) + d_{\mathcal{G}}(Q_{\mathbf{X}}, U_{\mathbf{X}}).$$

Proof. For simplicity, we omit Ω from $\mathrm{FPR}(G; Q_{\mathbf{X}}, \Omega)$. For any $G^* \in \mathcal{G}$, we have

$$\begin{aligned}
\mathrm{FPR}(G; Q_{\mathbf{X}}) &= \mathrm{FPR}(G^*; Q_{\mathbf{X}}) + \mathrm{FPR}(G; Q_{\mathbf{X}}) - \mathrm{FPR}(G^*; Q_{\mathbf{X}}) && (5)\\
&= \mathrm{FPR}(G^*; Q_{\mathbf{X}}) + \mathrm{FPR}(G; U_{\mathbf{X}}) - \mathrm{FPR}(G^*; U_{\mathbf{X}}) && (6)\\
&\quad + [(\mathrm{FPR}(G; Q_{\mathbf{X}}) - \mathrm{FPR}(G^*; Q_{\mathbf{X}})) - (\mathrm{FPR}(G; U_{\mathbf{X}}) - \mathrm{FPR}(G^*; U_{\mathbf{X}}))]. && (7)
\end{aligned}$$

The last term is

$$\begin{aligned}
&(\mathrm{FPR}(G; Q_{\mathbf{X}}) - \mathrm{FPR}(G^*; Q_{\mathbf{X}})) - (\mathrm{FPR}(G; U_{\mathbf{X}}) - \mathrm{FPR}(G^*; U_{\mathbf{X}})) && (8)\\
&= v(G, G^*; Q_{\mathbf{X}}) - v(G, G^*; U_{\mathbf{X}}) && (9)\\
&\leq d_{\mathcal{G}}(Q_{\mathbf{X}}, U_{\mathbf{X}}). && (10)
\end{aligned}$$

Therefore,

$$\mathrm{FPR}(G; Q_{\mathbf{X}}) \leq \mathrm{FPR}(G^*; Q_{\mathbf{X}}) + \mathrm{FPR}(G; U_{\mathbf{X}}) + d_{\mathcal{G}}(Q_{\mathbf{X}}, U_{\mathbf{X}}). \quad (11)$$

Taking $\inf$ over $G^* \in \mathcal{G}$ completes the proof. □

The error of the detector is bounded by three terms: the best error, the error on the training distributions, and the divergence between $Q_{\mathbf{X}}$ and $U_{\mathbf{X}}$. Assuming that there exists a ground-truth detector with a small test error, and that the optimization can lead to a small training error, the test error is then characterized by the divergence. So in this case, as long as the rankings of the hypotheses (according to the error) on $Q_{\mathbf{X}}$ and $U_{\mathbf{X}}$ are similar, detectors learned on $U_{\mathbf{X}}$ can generalize to $Q_{\mathbf{X}}$.

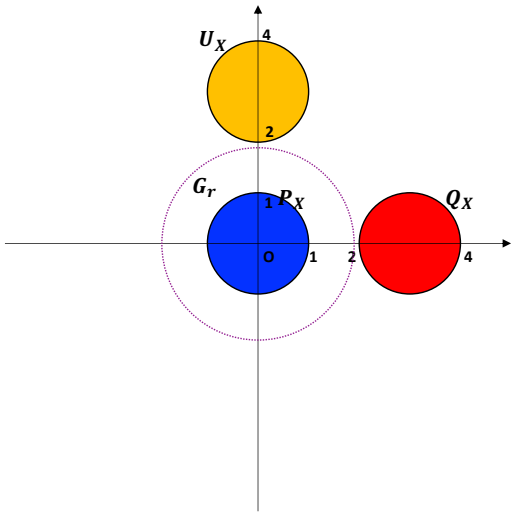


Figure 3: An illustration example to explain why $U_{\mathbf{X}}$ helps to get a good detector G_r. With $U_{\mathbf{X}}$, we can prune away hypotheses G_r for any $r \geq 1.9$. Thus, the resulting detector G_r can detect OOD samples from $Q_{\mathbf{X}}$ successfully and robustly.

An illustration example. In this example, the in-distribution $P_{\mathbf{X}}$ is uniform over the disk around the origin in $\mathbb{R}^2$ with radius 1, $U_{\mathbf{X}}$ is uniform over the disk around $(0, 3)$ with radius 1, and $Q_{\mathbf{X}}$ is uniform over the disk around $(3, 0)$ with radius 1. Assume the adversary budget is $\epsilon = 0.1$, i.e., $\Omega_{\infty,\epsilon} = \{\|\delta\|_\infty \leq 0.1\}$. The hypothesis class for the detector contains all functions of the form $G_r(x) = 2\mathbb{I}[\|x\|_2 \leq r] - 1$ with parameter r. See Figure 3.

The example first shows the effect of the auxiliary outlier data: $U_{\mathbf{X}}$ helps prune away hypotheses G_r for any $r \geq 1.9$. Furthermore, it also shows how learning over $U_{\mathbf{X}}$ can generalize to $Q_{\mathbf{X}}$. Although $Q_{\mathbf{X}}$ and $U_{\mathbf{X}}$ have non-overlapping supports, $U_{\mathbf{X}}$ helps to calibrate the error of the hypotheses, so any good detector trained on $P_{\mathbf{X}}$ and $U_{\mathbf{X}}$ can be used for distinguishing $P_{\mathbf{X}}$ and $Q_{\mathbf{X}}$. Formally, the $d_{\mathcal{G}}$ is small in Proposition 2.

The analysis also shows the importance of training on *perturbed* instances from the unlabeled data $U_{\mathbf{X}}$. Not using perturbation is equivalent to using $\Omega = \{\mathbf{0}\}$. In this case, the analysis shows that it only guarantees the error on *unperturbed* instances from $Q_{\mathbf{X}}$, even if $Q_{\mathbf{X}}$ and $U_{\mathbf{X}}$ has small divergence and the learned detector can have small training error on $U_{\mathbf{X}}$.

D.2. Analysis in a Gaussian Model

To understand how the outlier training data affect the generalization, we study a concrete distributional model, which is inspired by the models in (Schmidt et al., 2018; Carmon et al., 2019). In this model, we establish a separation of the in-distribution sample sizes needed in the two cases: with and without auxiliary outlier data for training. We also demonstrate the benefit of outlier mining when the auxiliary data consists of uninformative outliers.

While the theoretical model is simple (in fact, much simpler than the practical data distributions), its simplicity is actually desired for our analytical purpose. More precisely, the separation of the sample sizes under this simple model suggests the same phenomenon can happen in more complicated models. This then means the auxiliary outlier data not only help training but are *necessary* for obtaining detectors with reasonable performance when in-distribution data is limited.

Gaussian Model. To specify a distributional model for our robust OOD formulation, we need in-distribution $P_{\mathbf{X}}$, family of OOD distributions $\mathcal{Q}$, and the hypothesis class $\mathcal{H}$ for the OOD detector G. When auxiliary data is available, we also need to specify their distribution $U_{\mathbf{X}}$. Let $\mu \in \mathbb{R}^d$ be the mean vector, $\sigma > 0$ be the variance parameter, and $\nu > 0$ be a parameter. In our (μ, σ, ν)-Gaussian model:

- $P_{\mathbf{X}}$ is $\mathcal{N}(\mu, \sigma^2 I)$.
- $\mathcal{Q} = \{\mathcal{N}(-\mu + v, \sigma^2 I) : v \in \mathbb{R}^d, \|v\|_2 \leq \nu\}$.
- $\mathcal{H} = \{G_\theta(x) = \text{sign}(\theta^\top \mathbf{x}) : \theta \in \mathbb{R}^d\}$. Here $G_\theta(\mathbf{x}) = 1$ means it predicts x to be an in-distribution example, and

$G_\theta(\mathbf{x}) = -1$ means it predicts an OOD example.

We are interested in the False Negative Rate $\mathrm{FNR}(G)$ and worst False Positive Rate $\sup_{Q_\mathbf{X} \in \mathcal{Q}} \mathrm{FPR}(G; Q_\mathbf{X}, \Omega_{\infty,\epsilon}(\mathbf{x}))$ over $Q_\mathbf{X} \in \mathcal{Q}$ under ℓ_∞ perturbations of magnitude ϵ. For simplicity, we denote them as $\mathrm{FNR}(G)$ and $\mathrm{FPR}(G; Q_\mathbf{X})$ in our proofs.

Parameter Setting. The model parameters are set such that:

1. There exists a classifier that achieves very low errors FPR and FNR.
2. We need n_1 in-distribution data from $P_\mathbf{X}$ to learn a classifier with non-trivial robust errors.
3. Using n_0 in-distribution examples from $P_\mathbf{X}$ and n_{out} auxiliary outliers from $U_\mathbf{X}$ where n_0 is much smaller than n_1, we can learn a classifier with non-trivial robust errors.

Here n_0, n_1, n_{out} are sample sizes whose values are specified later in our analysis.

To achieve the three goals, the following parameter values are used (repeating (5)):

$$\epsilon \in (0, 1/2), \quad \|\mu\|_2^2 = d \gg n_0/\epsilon^4, \quad \sigma^2 = \sqrt{dn_0}, \quad \nu \leq \|\mu\|_2/4. \tag{12}$$

To interpret the parameter setting, one can view ϵ as fixed and d/n_0 as a large number. In the following subsections, we show how these three goals are achieved.

D.2.1. Existence of Robust Classifier

We give closed forms of the errors, and show that using $\theta = \mu$ gives small errors under the chosen parameter setting in (5).

Closed Forms of the Errors. By definition, the FNR of a detector G_θ (on $P_\mathbf{X}$) is:

$$\mathrm{FNR}(G_\theta) = \mathbb{P}_{\mathbf{x}\sim P_\mathbf{X}}[\theta^\top \mathbf{x} \leq 0] = \mathbb{P}_{\mathbf{x}\sim P_\mathbf{X}}\left[\mathcal{N}\left(\frac{\mu^\top\theta}{\sigma\|\theta\|_2}, 1\right) \leq 0\right] =: \Phi\left(\frac{\mu^\top\theta}{\sigma\|\theta\|_2}\right) \tag{13}$$

where

$$\Phi(\mathbf{x}) := \frac{1}{\sqrt{2\pi}} \int_\mathbf{x}^\infty e^{-t^2/2} dt \tag{14}$$

is the Gaussian error function.

Given a test OOD distribution $Q_v = \mathcal{N}(-\mu + v, \sigma^2 I)$, the robust FPR of G_θ on Q_v is:

$$\begin{aligned}
\mathrm{FPR}(G_\theta; Q_v) &= \mathbb{P}_{\mathbf{x}\sim Q_v}\left[\inf_{\|\delta\|_\infty \leq \epsilon} \theta^\top(\mathbf{x}+\delta) \geq 0\right] && (15)\\
&= \mathbb{P}_{\mathbf{x}\sim Q_v}\left[\theta^\top\mathbf{x} + \epsilon\|\theta\|_1 \geq 0\right] && (16)\\
&= \mathbb{P}_{\mathbf{x}\sim Q_v}\left[\mathcal{N}((\mu+v)^\top\theta, (\sigma\|\theta\|_2)^2) \geq -\epsilon\|\theta\|_1\right] && (17)\\
&= \Phi\left(\frac{(\mu+v)^\top\theta}{\sigma\|\theta\|_2} - \frac{\epsilon\|\theta\|_1}{\sigma\|\theta\|_2}\right). && (18)
\end{aligned}$$

Then the worst robust FPR of G_θ on $\mathcal{Q}$ is:

$$\begin{aligned}
\sup_{Q_v\in\mathcal{Q}} \mathrm{FPR}(G_\theta; Q_v) &= \sup_{\|v\|_2\leq\nu} \Phi\left(\frac{(\mu+v)^\top\theta}{\sigma\|\theta\|_2} - \frac{\epsilon\|\theta\|_1}{\sigma\|\theta\|_2}\right) && (19)\\
&= \Phi\left(\frac{\mu^\top\theta}{\sigma\|\theta\|_2} - \frac{\nu}{\sigma} - \frac{\epsilon\|\theta\|_1}{\sigma\|\theta\|_2}\right) && (20)\\
&\leq \Phi\left(\frac{\mu^\top\theta}{\sigma\|\theta\|_2} - \frac{\nu}{\sigma} - \frac{\epsilon\sqrt{d}}{\sigma}\right). && (21)
\end{aligned}$$

Small Errors of G_μ. Given the closed forms, we can now show that G_μ achieves small FNR and FPR in our parameter setting.

$$\mathrm{FNR}(G_\mu) = \Phi\left(\frac{\|\mu\|_2}{\sigma}\right) = \Phi\left(\left(\frac{d}{n_0}\right)^{1/4}\right) \leq e^{-\frac{1}{2}\sqrt{d/n_0}}. \tag{22}$$

$$\sup_{Q_v \in \mathcal{Q}} \mathrm{FPR}(G_\mu; Q_v) \leq \Phi\left(\frac{\|\mu\|_2}{\sigma} - \frac{\nu}{\sigma} - \frac{\epsilon\sqrt{d}}{\sigma}\right) \tag{23}$$

$$\leq \Phi\left(\left(1 - \frac{1}{4} - \epsilon\right)\left(\frac{d}{n_0}\right)^{1/4}\right) \leq e^{-\frac{1}{32}\sqrt{d/n_0}}. \tag{24}$$

Therefore, in the regime $d/n_0 \gg 1$, the detector G_μ achieves both small FNR on $P_{\mathbf{X}}$ and robust FPR on any test OOD distribution in $\mathcal{Q}$.

D.2.2. Learning Without Auxiliary Outlier Data

Given in-distribution data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, we consider the detector $G_{\hat{\theta}_n}$ given by

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i. \tag{25}$$

As shown in the closed form solutions, the key factor determining the errors is $\frac{\mu^\top \hat{\theta}_n}{\sigma\|\hat{\theta}_n\|_2}$. To bound this term, we cite the following lemma from existing work:

Lemma 1 (Lemma 1 in (Carmon et al., 2019))**.** *There exist numerical constants c_0, c_1, c_2 such that under parameter setting (5) and $d/n_0 > c_0$,*

$$\frac{\mu^\top \hat{\theta}_n}{\sigma\|\hat{\theta}_n\|_2} \geq \left(\sqrt{\frac{n_0}{d}} + \frac{n_0}{n}\left(1 + c_1\left(\frac{n_0}{d}\right)^{1/8}\right)\right)^{-1/2} \tag{26}$$

with probability $\geq 1 - e^{-c_2(d/n_0)^{1/4}\min\{n,(d/n_0)^{1/4}\}}$.

This lemma leads to the following guarantee about learning the OOD detector from in-distribution data only.

Proposition 3. *There exists a universal constant c such that for the parameter setting (5) with $\sqrt{d/n_0} \geq c/\epsilon^2$, we have that if $n \geq n_0 \cdot 4\epsilon^2\sqrt{d/n_0}$, then*

$$\mathbb{E}_{\hat{\theta}_n}\mathrm{FNR}(G_{\hat{\theta}_n}) \leq 10^{-3}, \quad \mathbb{E}_{\hat{\theta}_n}\sup_{Q_{\mathbf{X}} \in \mathcal{Q}} \mathrm{FPR}(G_{\hat{\theta}_n}; Q_{\mathbf{X}}, \Omega_{\infty,\epsilon}(\mathbf{x})) \leq 10^{-3}. \tag{27}$$

Proof. By Lemma 1, we have

$$\frac{\mu^\top \hat{\theta}_n}{\sigma\|\hat{\theta}_n\|_2} \geq \left(2\left(\sqrt{\frac{n_0}{n}} + \sqrt{\frac{n_0}{d}}\right)\right)^{-1/2} \tag{28}$$

with probability $\geq 1 - e^{-c_2(d/n_0)^{1/4}\min\{n,(d/n_0)^{1/4}\}}$. The proposition then comes from the parameter setting (5) and the closed form expressions (13) and (19) of the errors. □

Next we show that the above sample size is nearly optimal (up to a logarithmic factor). That is, a sample size of order $n_0 \cdot \frac{\epsilon^2\sqrt{d/n_0}}{\log d}$ is necessary for *all algorithms* to obtain both non-trivial robust FPR and FNR. We emphasize that this lower bound is information theoretic, i.e., it holds without restriction on the computational power of the learning algorithm and the hypothesis class used for the OOD detector. In particular, it applies not only to the linear classifier considered in Proposition 3 but also to any other learning algorithms.

*

Proof. The key for the proof is the observation that robust classification is a special case of our robust OOD problem. More precisely, consider the following robust classification problem. The data $(\mathbf{x}, y)$ with $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{-1, +1\}$ is generated as follows: first draw y uniformly at random, and then draw $\mathbf{x}$ from $\mathcal{N}(y \cdot \mu, \sigma^2 I)$. Given training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the goal is to find classifier $f_\theta(\mathbf{x}) = \text{sign}(\theta^\top \mathbf{x})$ with small robust classification error

$$\text{err}_{\infty,\epsilon}(f_\theta) = \mathbb{E}_{(\mathbf{x},y)} \max_{\|\delta\|_\infty \le \epsilon} \mathbb{I}[f_\theta(\mathbf{x} + \delta) \neq y]$$

under ℓ_∞ perturbation of magnitude ϵ. It has been shown that (Theorem 6 in (Schmidt et al., 2018) or Theorem 1 in (Carmon et al., 2019)) that when $\mu \sim \mathcal{N}(0, I)$ and $n \le n_0 \cdot \frac{\epsilon^2 \sqrt{d/n_0}}{8 \log d}$ and with the parameter setting (5), for any learning algorithm $\mathbb{A}_n$

$$\mathbb{E}\text{err}_{\infty,\epsilon}(\mathbb{A}_n(S)) \ge \frac{1}{2}(1 - d^{-1}). \tag{29}$$

Now consider the following variant of the robust OOD problem in the proposition. Suppose besides the data from $P_{\mathbf{X}}$, we also have n i.i.d. samples from a test OOD distribution $Q_0 = \mathcal{N}(-\mu, \sigma^2 I)$. Then the above robust classification problem can be reduced to this variant of robust OOD, by viewing the in-distribution data as with label $+1$ and viewing outliers as with label -1. Furthermore, it is clear that the sum of the FNR and FPR is larger than the robust classification error. Then

$$\mathbb{E}\left\{\text{FNR}(\mathbb{A}_n(S)) + \text{FPR}(\mathbb{A}_n(S); Q_0)\right\} \ge \frac{1}{2}(1 - d^{-1}). \tag{30}$$

Since this variant can be reduced to the original robust OOD problem in the proposition and furthermore $Q_0 \in \mathcal{Q}$, the statement then follows. □

D.2.3. Learning With Auxiliary Outlier Data

Assuming we have access to auxiliary outliers from a distribution $U_{\mathbf{X}}$ where:

- $U_{\mathbf{X}}$ is defined by the following distribution: first draw v uniformly at random from the ball $\{v : v \in \mathbb{R}^d, \|v\|_2 \le \nu\}$, then draw $\tilde{\mathbf{x}}$ from $\mathcal{N}(-\mu + v, \sigma^2 I)$.

Roughly speaking, $U_{\mathbf{X}}$ is a uniform mixture of distributions in $\mathcal{Q}$.

Given in-distribution data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from $P_{\mathbf{X}}$ and auxiliary outliers $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{n'}$ from $U_{\mathbf{X}}$, we consider the detector $G_{\hat{\theta}_{n,n'}}$ given by

$$\hat{\theta}_{n,n'} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i - \frac{1}{n'} \sum_{i=1}^{n'} \tilde{\mathbf{x}}_i. \tag{31}$$

We will show that with $n = n_0$ and sufficiently large n', the detector has small errors.

Again, as shown in the closed form solutions, the key factor determining the errors is $\frac{\mu^\top \hat{\theta}_{n,n'}}{\sigma \|\hat{\theta}_{n,n'}\|_2}$. The following lemma bounds this term.

Lemma 2. *There exist numerical constants c_0, c_1, c_2 such that under parameter setting (5) and $d/n_0 > c_0$,*

$$\frac{\mu^\top \hat{\theta}_{n,n'}}{\sigma \|\hat{\theta}_{n,n'}\|_2} \ge \left(\sqrt{\frac{n_0}{d}} + \frac{n_0}{n + n'} \left(1 + c_1 \left(\frac{n_0}{d} \right)^{1/8} \right) \right)^{-1/2} \tag{32}$$

with probability $\ge 1 - e^{-c_2 (d/n_0)^{1/4} \min\{n+n', (d/n_0)^{1/4}\}} - e^{-c_2 n'}$.

Proof. The proof follows the argument of Lemma 1 in (Carmon et al., 2019) but needs some modifications accommodating the difference in learning θ. Recall the generation of $\mathbf{x}_i'$: first draw v_i uniformly at random from the ball $\mathbb{B}(\nu) := \{v : v \in$

$\mathbb{R}^d, \|v\|_2 \leq \nu\}$, then draw $\bar{\mathbf{x}'}_i$ from $\mathcal{N}(\mu, \sigma^2 I)$, and finally let $\mathbf{x}'_i = v_i - \bar{\mathbf{x}'}_i$. So we have

$$\hat{\theta}_{n,n'} = \frac{1}{n+n'}\left(\sum_{i=1}^{n}\mathbf{x}_i + \sum_{i=1}^{n'}\bar{\mathbf{x}'}_i\right) - \frac{1}{n+n'}\left(\sum_{i=1}^{n'} v_i\right) \tag{33}$$

$$= \mu + \delta + \delta_v \tag{34}$$

where

$$\delta = \frac{1}{n+n'}\left(\sum_{i=1}^{n}\mathbf{x}_i + \sum_{i=1}^{n'}\bar{\mathbf{x}'}_i\right) - \mu \sim \mathcal{N}(0, \frac{\sigma^2}{n+n'}I), \tag{35}$$

$$\delta_v = -\frac{1}{n+n'}\left(\sum_{i=1}^{n'} v_i\right). \tag{36}$$

To lower bound the term $\frac{\mu^\top \hat{\theta}_{n,n'}}{\|\hat{\theta}_{n,n'}\|_2}$, we upper bound its squared inverse:

$$\frac{\|\hat{\theta}_{n,n'}\|_2^2}{(\mu^\top \hat{\theta}_{n,n'})^2} = \frac{\|\mu + \delta + \delta_v\|_2^2}{(\|\mu\|_2^2 + \mu^\top\delta + \mu^\top\delta_v)^2} \tag{37}$$

$$= \frac{1}{\|\mu\|_2^2} + \frac{\|\delta + \delta_v\|_2^2 - \frac{1}{\|\mu\|_2^2}(\mu^\top\delta + \mu^\top\delta_v)^2}{(\|\mu\|_2^2 + \mu^\top\delta + \mu^\top\delta_v)^2} \tag{38}$$

$$\leq \frac{1}{\|\mu\|_2^2} + \frac{2\|\delta\|_2^2 + 2\|\delta_v\|_2^2}{(\|\mu\|_2^2 + \mu^\top\delta + \mu^\top\delta_v)^2}. \tag{39}$$

For δ, we have

$$\|\delta\|_2^2 \sim \frac{\sigma^2}{n+n'}\chi_d^2 \text{ and } \frac{\mu^\top\delta}{\|\mu\|_2} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n+n'}\right). \tag{40}$$

So standard concentration bounds give

$$\mathbb{P}\left(\|\delta\|_2^2 \geq \frac{\sigma^2}{n+n'}\left(d + \frac{1}{\sigma}\right)\right) \leq e^{-d/8\sigma^2} \text{ and } \mathbb{P}\left(\frac{\mu^\top\delta}{\|\mu\|_2} \geq (\sigma\|\mu\|)^{1/2}\right) \leq 2e^{-(n+n')\|\mu\|_2/2\sigma}. \tag{41}$$

For δ_v, by subguassian concentration bounds, we have

$$\mathbb{P}\left(\|\delta_v\|_2 \geq \frac{C\nu}{\sqrt{n'}}\right) \leq e^{-cn'} \tag{42}$$

for some numeric constants c and C. Suppose the event $\|\delta_v\|_2 < \frac{C\nu}{\sqrt{n'}}$ is true. Then

$$|\mu^\top\delta_v| \leq \|\mu\|_2\|\delta_v\|_2 \leq \frac{C\nu\|\mu\|_2}{\sqrt{n'}}. \tag{43}$$

Plugging the concentration bounds in (37) and doing the same manipulation leads to the bound. To finish the proof, we also need to show $\mu^\top\hat{\theta}_{n,n'} > 0$, which can be shown by the same argument as in (Carmon et al., 2019). □

*

Proof. The proposition comes from Lemma 2, the parameter setting (5), and the closed form expressions (13) and (19) of the errors. □

D.3. Benefit of Outlier Mining

The above Gaussian example shows the benefit of having auxiliary outlier data for training. All the auxiliary data given in the example are implicitly related to the ideal parameter for the detector $\theta^* = \mu$ and thus are informative for learning the detector. However, this may not be the case in practice: typically only part of the auxiliary outlier data are informative, while the remaining are not very useful or even can be harmful for the learning. In this section, we study such an example, and shows that how outlier mining can help to identify informative data and improve the learning performance.

Suppose the algorithm gets n in-distribution data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ i.i.d. from $P_{\mathbf{X}}$ and n' auxiliary outliers $\{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_{n'}\}$ for training. Instead of from $U_{\mathbf{X}}$ specified above, the auxiliary outliers are i.i.d. from the distribution U_{mix}.

- U_{mix} is a uniform mixture of $\mathcal{N}(-\mu, \sigma^2 I)$ and $\mathcal{N}(-\mu_o, \sigma^2 I)$ for $\mu_o = 10\mu$.

That is, the distribution is defined by the following process: with probability $1/2$ sample the outlier from the informative part $\mathcal{N}(-\mu, \sigma^2 I)$, and with probability $1/2$ sample the outlier from the uninformative part $\mathcal{N}(-\mu_o, \sigma^2 I)$. We also note that $\mu_0 = 10\mu$ is chosen for simplicity of analysis. μ_0 can also be $c\mu$ for some sufficiently large $c > 1$, or even $\mu_o = c\mu + c'\mu_\perp$ for a sufficiently large $c > 1$, a small c' and a unit vector $\mu_\perp$ perpendicular to μ.

Naïve Method Without Outlier Mining. It is clear that naïvely applying the method in the previous section can lead to high errors: with n in-distribution examples from $P_{\mathbf{X}}$ and $n' = n$ auxiliary outliers from U_{mix}, when $n \to \infty$, we have $\hat{\theta}_{n,n'} \to -7\mu/4$ which has the worst errors among all detectors.

With Outlier Mining. Here we analyze the following algorithm using the outlier mining approach. The algorithm is simpler than what we used in Section 3 but shares the same intuition.

First, we use the n in-distribution data points to get an intermediate solution:

$$\hat{\theta}_{\text{int}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i. \tag{44}$$

We define the confidence score of a point $\tilde{\mathbf{x}}$ being in-distribution as:

$$f(\tilde{\mathbf{x}}) = \sigma(t) = \frac{1}{1+e^{-t}}, \text{ where } t(\tilde{\mathbf{x}}) = \frac{\tilde{\mathbf{x}}^\top \hat{\theta}_{\text{int}}}{d}. \tag{45}$$

Here $\sigma(t) = \frac{1}{1+e^{-t}}$ is the sigmoid function. We then select outlier training data whose confidence fall into an interval $[a, b]$ and use them to learn the final solution:

$$\hat{\theta}_{\text{om}} = \frac{\sum_{i=1}^{n'} (-\tilde{\mathbf{x}}_i)\mathbb{I}\{f(\tilde{\mathbf{x}}_i) \in [a,b]\}}{\sum_{i=1}^{n'} \mathbb{I}\{f(\tilde{\mathbf{x}}_i) \in [a,b]\}} \tag{46}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function.

Proposition 1. *There exist thresholds a and b for $\hat{\theta}_{\text{om}}$, and a universal constant c such that for the parameter setting (5) with $\sqrt{d/n_0} \geq c(\log d + 1/\epsilon^2)$, we have that if $n \geq cn_0 \log d$ and $n' \geq (d + n_0 \cdot 4\epsilon^2)\sqrt{d/n_0}$, then*

$$\mathbb{E}_{\hat{\theta}_{\text{om}}} \text{FNR}(G_{\hat{\theta}_{\text{om}}}) \leq 10^{-3}$$
$$\mathbb{E}_{\hat{\theta}_{\text{om}}} \sup_{Q_{\mathbf{X}} \in \mathcal{Q}} \text{FPR}(G_{\hat{\theta}_{\text{om}}}; Q_{\mathbf{X}}) \leq 10^{-3}.$$

Proof. Let $a = \sigma(-3/2), b = \sigma(-1/2)$. By definition we have

$$\delta_{\text{om}} := \hat{\theta}_{\text{om}} - \mu = \frac{\sum_{i=1}^{n'} (-\mu - \tilde{\mathbf{x}}_i)\mathbb{I}\{f(\tilde{\mathbf{x}}_i) \in [a,b]\}}{\sum_{i=1}^{n'} \mathbb{I}\{f(\tilde{\mathbf{x}}_i) \in [a,b]\}}. \tag{47}$$

By the closed form expressions (13) and (19) of the errors, it is sufficient to lower bound the key term $\frac{\mu^\top \hat{\theta}_{\text{om}}}{\|\hat{\theta}_{\text{om}}\|_2}$, which comes down to show that δ_{om} is small.

First, let's consider $\hat{\theta}_{\text{int}}$. Let $\delta_{\text{int}} := \hat{\theta}_{\text{int}} - \mu$. Then

$$\|\delta_{\text{int}}\|_2^2 \sim \frac{\sigma^2}{n}\chi_d^2 \text{ and } \frac{\mu^\top \delta_{\text{int}}}{\|\mu\|_2} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right). \tag{48}$$

So standard concentration bounds give

$$\mathbb{P}\left(\|\delta_{\text{int}}\|_2^2 \geq \frac{\sigma^2}{n}\left(d + \frac{1}{\sigma}\right)\right) \leq e^{-d/8\sigma^2} \text{ and } \mathbb{P}\left(\frac{|\mu^\top \delta_{\text{int}}|}{\|\mu\|_2} \geq \sqrt{\frac{d}{n}}\right) \leq 2e^{-d/2\sigma^2}. \tag{49}$$

So with probability $\geq 1 - 3e^{-d/8\sigma^2}$ over the randomness of the n in-distribution points, we have the good event $\mathcal{G}_{\text{int}}$: $\|\delta_{\text{int}}\|_2^2 \leq \frac{\sigma^2}{n}\left(d + \frac{1}{\sigma}\right)$ and $\frac{|\mu^\top \delta_{\text{int}}|}{\|\mu\|_2} \leq \sqrt{\frac{d}{n}}$.

Now, condition on a fix $\hat{\theta}_{\text{int}}$ satisfying $\mathcal{G}_{\text{int}}$, and consider $\hat{\theta}_{\text{om}}$. Define

$$z_i := -\mu - \tilde{\mathbf{x}}_i, \tag{50}$$
$$\mathbb{I}_{0i} := \mathbb{I}\{f(\tilde{\mathbf{x}}_i) \in [a, b]\}, \tag{51}$$
$$\mathbb{I}_{1i} := \mathbb{I}\{\tilde{\mathbf{x}}_i \text{ is from } \mathcal{N}(-\mu, \sigma^2 I)\}, \tag{52}$$
$$\mathbb{I}_{2i} := \mathbb{I}\{\tilde{\mathbf{x}}_i \text{ is from } \mathcal{N}(-\mu_o, \sigma^2 I)\}. \tag{53}$$

For simplicity, let's omit the subscript i and consider a sample $\tilde{\mathbf{x}}$ from U_{mix}, and the corresponding variables $z, \mathbb{I}_0, \mathbb{I}_1$, and $\mathbb{I}_2$. Since $\mathbb{I}_1 + \mathbb{I}_2 = 1$,

$$(-\mu - \tilde{\mathbf{x}})\mathbb{I}\{f(\tilde{\mathbf{x}}) \in [a, b]\} = z\mathbb{I}_0\mathbb{I}_1 + z\mathbb{I}_0\mathbb{I}_2. \tag{54}$$

Case 1. Let's first consider the case when $\tilde{\mathbf{x}}$ is from $\mathcal{N}(-\mu, \sigma^2 I)$. More precisely, we condition on a fixed $\hat{\theta}_{\text{int}}$ and condition on $\mathbb{I}_1 = 1$. Then $z \sim \mathcal{N}(0, \sigma^2 I)$ and it can be decomposed along the direction $\bar{\theta}_{\text{int}} := \hat{\theta}_{\text{int}}/\|\hat{\theta}_{\text{int}}\|_2$ as follows:

$$z = s \cdot \bar{\theta}_{\text{int}} + z_2 \tag{55}$$

where $s \sim \mathcal{N}(0, \sigma^2)$ and z_2 is a Gaussian distribution in the subspace orthogonal to $\bar{\theta}_{\text{int}}$. Then

$$t(\tilde{\mathbf{x}}) = \frac{\tilde{\mathbf{x}}^\top \hat{\theta}_{\text{int}}}{d} = -\frac{\mu^\top \hat{\theta}_{\text{int}}}{d} - \frac{s\|\hat{\theta}_{\text{int}}\|_2}{d}. \tag{56}$$

Therefore, we have

$$\mathbb{E}[z\mathbb{I}_0\mathbb{I}_1 | \mathbb{I}_1 = 1, \hat{\theta}_{\text{int}}] = \mathbb{E}[s \cdot \bar{\theta}_{\text{int}}\mathbb{I}_0 | \mathbb{I}_1 = 1, \hat{\theta}_{\text{int}}] + \mathbb{E}[z_2\mathbb{I}_0 | \mathbb{I}_1 = 1, \hat{\theta}_{\text{int}}] \tag{57}$$

Clearly the second term is 0 since $z_2\mathbb{I}_0$ is symmetric. So

$$\mathbb{E}[z\mathbb{I}_0\mathbb{I}_1 | \mathbb{I}_1 = 1, \hat{\theta}_{\text{int}}] = \mathbb{E}[s \cdot \bar{\theta}_{\text{int}}\mathbb{I}\{f(\tilde{\mathbf{x}}) \in [a, b]\} | \mathbb{I}_1 = 1, \hat{\theta}_{\text{int}}] \tag{58}$$
$$= \mathbb{E}\left[s \cdot \mathbb{I}\{s \in [a', b']\} | \mathbb{I}_1 = 1, \hat{\theta}_{\text{int}}\right] \cdot \bar{\theta}_{\text{int}} \tag{59}$$
$$= \mathbb{E}\left[s \cdot \mathbb{I}\{s \in [a', b']\}\right] \cdot \bar{\theta}_{\text{int}} \tag{60}$$

where

$$a' = -\frac{\mu^\top \hat{\theta}_{\text{int}}}{\|\hat{\theta}_{\text{int}}\|_2} - \frac{\sigma^{-1}(b)d}{\|\hat{\theta}_{\text{int}}\|_2} \tag{61}$$

$$= \frac{-2\mu^\top \hat{\theta}_{\text{int}} + d}{2\|\hat{\theta}_{\text{int}}\|_2} \tag{62}$$

$$= \frac{-2\mu^\top \delta_{\text{int}} - d}{2\|\hat{\theta}_{\text{int}}\|_2}, \tag{63}$$

$$b' = -\frac{\mu^\top \hat{\theta}_{\text{int}}}{\|\hat{\theta}_{\text{int}}\|_2} - \frac{\sigma^{-1}(a)d}{\|\hat{\theta}_{\text{int}}\|_2} \tag{64}$$

$$= \frac{-2\mu^\top \hat{\theta}_{\text{int}} + 3d}{2\|\hat{\theta}_{\text{int}}\|_2} \tag{65}$$

$$= \frac{-2\mu^\top \delta_{\text{int}} + d}{2\|\hat{\theta}_{\text{int}}\|_2}. \tag{66}$$

By the bound on $|\mu^\top \delta_{\text{int}}|$, we have

$$|\mathbb{E}\left[s \cdot \mathbb{I}\{s \in [a', b']\}\right]| \leq \int_{\frac{d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}(1-2/\sqrt{n})}^{\frac{d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}(1+2/\sqrt{n})} \sigma t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{67}$$

$$\leq \sigma \frac{d}{\sigma\|\hat{\theta}_{\text{int}}\|_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}(1-2/\sqrt{n})\right)^2} \tag{68}$$

$$\leq \frac{d}{\|\hat{\theta}_{\text{int}}\|_2} e^{-\frac{d^2}{32\sigma^2\|\hat{\theta}_{\text{int}}\|_2^2}} \tag{69}$$

Given the bound on $\|\delta_{\text{int}}\|_2^2$, we have

$$\|\hat{\theta}_{\text{int}}\|_2 \leq \|\mu\|_2 + \|\delta_{\text{int}}\|_2 \leq \sqrt{d} + \sqrt{\frac{\sigma^2}{n}\left(d + \frac{1}{\sigma}\right)} \leq \sqrt{d} + \sqrt{\frac{2\sigma^2 d}{n}}. \tag{70}$$

Since $n \geq Cn_0 \log d$ and $d \geq C^2 n_0 \log^2 d$ for a sufficiently large C, we have

$$\frac{\sigma^2\|\hat{\theta}_{\text{int}}\|_2^2}{d^2} \leq \frac{\sigma^2 d(1 + \sqrt{2\sigma^2/n})^2}{d^2} \leq 2\sqrt{\frac{n_0}{d}} + \frac{4n_0}{n} \leq \frac{6}{C \log d} \tag{71}$$

and thus

$$|\mathbb{E}\left[s \cdot \mathbb{I}\{s \in [a', b']\}\right]| \leq \frac{1}{d^2}. \tag{72}$$

Combining with $\mathbb{E}[z\mathbb{I}_0\mathbb{I}_1|\mathbb{I}_1 = 0, \hat{\theta}_{\text{int}}] = 0$ we get

$$\mathbb{E}[z\mathbb{I}_0\mathbb{I}_1|\hat{\theta}_{\text{int}}] = c_1 \cdot \bar{\theta}_{\text{int}} \tag{73}$$

for some c_1 satisfying $|c_1| \leq 1/d^2$. Furthermore, $z\mathbb{I}_0\mathbb{I}_1 \mid \hat{\theta}_{\text{int}}$ is truncated Gaussian and thus is sub-Gaussian with sub-Gaussian norm bounded by σ. Then by sub-Gaussian concentration bounds, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n'} \mu^\top z_i \mathbb{I}_{0i}\mathbb{I}_{1i} - \sum_{i=1}^{n'} \mu^\top \mathbb{E}[z_i \mathbb{I}_{0i}\mathbb{I}_{1i}|\hat{\theta}_{\text{int}}]\right| \geq \sqrt{n'd} \mid \hat{\theta}_{\text{int}}\right) \leq e^{-cd/\sigma^2}, \tag{74}$$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n'} z_i \mathbb{I}_{0i}\mathbb{I}_{1i}\right\|_2 \geq 4\sigma\sqrt{n'd} + 2\sqrt{n'd} \mid \hat{\theta}_{\text{int}}\right) \leq e^{-d/\sigma^2}. \tag{75}$$

for some constant $c > 0$. In other words, with probability $\geq 1 - 2e^{-cd/\sigma^2}$, we have

$$\left|\sum_{i=1}^{n'} \mu^\top z_i \mathbb{I}_{0i}\mathbb{I}_{1i}\right| \leq \sqrt{n'd} + \frac{n'}{d^{3/2}}, \tag{76}$$

$$\left\|\sum_{i=1}^{n'} z_i \mathbb{I}_{0i}\mathbb{I}_{1i}\right\|_2 \leq 6\sigma\sqrt{n'd}. \tag{77}$$

Conditioned on $\mathbb{I}_1 = 1$, we also have

$$\mathbb{E}[\mathbb{I}_0\mathbb{I}_1|\mathbb{I}_1 = 1, \hat{\theta}_{\text{int}}] = \mathbb{P}(s \in [a', b']) \tag{78}$$

$$\geq \int_{\frac{d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}(-1+2/\sqrt{n})}^{\frac{d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}(1-2/\sqrt{n})} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{79}$$

$$\geq 1 - 2\int_{\frac{d}{4\sigma\|\hat{\theta}_{\text{int}}\|_2}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{80}$$

$$\geq 1 - 2\int_{\frac{\sqrt{C\log d}}{12}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{81}$$

$$\geq 1 - \frac{1}{d}. \tag{82}$$

Let $m = \frac{n'}{2}\left(1 - \frac{1}{d}\right)$. Then by Chernoff's bound, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n'} \mathbb{I}_{0i}\mathbb{I}_{1i} - m\right| \geq \frac{1}{2}m\right) \leq e^{-c'm} \tag{83}$$

for an absolute constant $c' > 0$. That is, with probality $\geq 1 - e^{-cn'}$, we have $\sum_{i=1}^{n'} \mathbb{I}_{0i}\mathbb{I}_{1i} \geq n'/5$.

Case 2. Next, let's consider the case when $\tilde{\mathbf{x}}$ is from $\mathcal{N}(-\mu_o, \sigma^2 I)$. More precisely, we condition on a fixed $\hat{\theta}_{\text{int}}$ and condition on $\mathbb{I}_2 = 1$. Similar to case 1, we have

$$z = 9\mu + s \cdot \bar{\theta}_{\text{int}} + z_2 \tag{84}$$

where $s \sim \mathcal{N}(0, \sigma^2)$ and z_2 is a Gaussian distribution in the subspace orthogonal to $\bar{\theta}_{\text{int}}$. So

$$\mathbb{E}[(z - 9\mu)\mathbb{I}_0\mathbb{I}_2|\mathbb{I}_2 = 1, \hat{\theta}_{\text{int}}] = \mathbb{E}[s\mathbb{I}_0|\mathbb{I}_2 = 1, \hat{\theta}_{\text{int}}] \cdot \bar{\theta}_{\text{int}}. \tag{85}$$

For this,

$$\mathbb{E}[s\mathbb{I}\{f(\tilde{\mathbf{x}}) \in [a, b]\}|\mathbb{I}_2 = 1, \hat{\theta}_{\text{int}}] \cdot \bar{\theta}_{\text{int}} = \mathbb{E}\left[s \cdot \mathbb{I}\{s \in [a'', b'']\}\right] \cdot \bar{\theta}_{\text{int}} \tag{86}$$

where

$$a'' = \frac{-20\mu^\top \delta_{\text{int}} - 19d}{2\|\hat{\theta}_{\text{int}}\|_2}, \tag{87}$$

$$b'' = \frac{-20\mu^\top \delta_{\text{int}} - 17d}{2\|\hat{\theta}_{\text{int}}\|_2}. \tag{88}$$

By the bound on $|\mu^\top \delta_{\text{int}}|$ and $\|\delta_{\text{int}}\|_2$, we have

$$|\mathbb{E}\left[s \cdot \mathbb{I}\{s \in [a', b']\}\right]| \leq \left| \int_{\frac{-d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}(19+20/\sqrt{n})}^{\frac{-d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}(17-20/\sqrt{n})} \sigma t \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \right| \tag{89}$$

$$\leq \sigma \frac{20d}{\sigma\|\hat{\theta}_{\text{int}}\|_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{16d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}\right)^2} \tag{90}$$

$$\leq \frac{20d}{\|\hat{\theta}_{\text{int}}\|_2} e^{-\frac{32d^2}{\sigma^2\|\hat{\theta}_{\text{int}}\|_2^2}} \tag{91}$$

$$\leq \frac{1}{d^2}. \tag{92}$$

We also have

$$\left|\mathbb{E}\left[\mathbb{I}_0\mathbb{I}_2 | \mathbb{I}_2 = 1, \hat{\theta}_{\text{int}}\right]\right| = |\mathbb{E}\left[\mathbb{I}\{s \in [a', b']\}\right]| \tag{93}$$

$$\leq \int_{\frac{-d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}(19+20/\sqrt{n})}^{\frac{-d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}(17-20/\sqrt{n})} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{94}$$

$$\leq \frac{d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}(2 + 40/\sqrt{n}) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{16d}{2\sigma\|\hat{\theta}_{\text{int}}\|_2}\right)^2} \tag{95}$$

$$\leq \frac{1}{d^3}. \tag{96}$$

Combining the above, we have

$$\mathbb{E}[(z - 9\mu)\mathbb{I}_0\mathbb{I}_2 | \hat{\theta}_{\text{int}}] = c_1 \cdot \bar{\theta}_{\text{int}} \tag{97}$$

for a constant c_1 satisfying $|c_1| \leq 1/d^2$. Furthermore, $(z - 9\mu)\mathbb{I}_0\mathbb{I}_2 \mid \hat{\theta}_{\text{int}}$ is truncated Gaussian and thus is sub-Gaussian with sub-Gaussian norm bounded by σ. Then by sub-Gaussian concentration bounds, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n'} \mu^\top (z_i - 9\mu)\mathbb{I}_{0i}\mathbb{I}_{2i} - \sum_{i=1}^{n'} \mu^\top \mathbb{E}[(z_i - 9\mu)\mathbb{I}_{0i}\mathbb{I}_{2i} | \hat{\theta}_{\text{int}}]\right| \geq \sqrt{n'd} \mid \hat{\theta}_{\text{int}}\right) \leq e^{-cd/\sigma^2}, \tag{98}$$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n'} (z_i - 9\mu)\mathbb{I}_{0i}\mathbb{I}_{2i}\right\|_2 \geq 4\sigma\sqrt{n'd} + 2\sqrt{n'd} \mid \hat{\theta}_{\text{int}}\right) \leq e^{-d/\sigma^2}, \tag{99}$$

for some constant $c > 0$. Also by Hoeffding's bound, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n'} \mathbb{I}_{0i}\mathbb{I}_{2i} - \sum_{i=1}^{n'} \mathbb{E}[\mathbb{I}_{0i}\mathbb{I}_{2i} | \hat{\theta}_{\text{int}}]\right| \geq \sqrt{n'd/\sigma^2} \mid \hat{\theta}_{\text{int}}\right) \leq 2e^{-2d/\sigma^2}. \tag{100}$$

In other words, with probability $\geq 1 - 4e^{-cd/\sigma^2}$, we have

$$\left|\sum_{i=1}^{n'} \mu^\top z_i \mathbb{I}_{0i}\mathbb{I}_{2i}\right| \leq \sqrt{n'd} + 9d\sqrt{n'd}\sigma^2 + \frac{n'}{d^2}\sqrt{d} + 9d \cdot \frac{n'}{d^3} \leq \sqrt{n'd}\left(1 + \frac{9d}{\sigma}\right) + \frac{n'}{d^{3/2}}, \tag{101}$$

$$\left\|\sum_{i=1}^{n'} z_i \mathbb{I}_{0i}\mathbb{I}_{2i}\right\|_2 \leq 6\sigma\sqrt{n'd} + 9\sqrt{d}\left(\sqrt{\frac{n'd}{\sigma^2}} + \frac{n'}{d^3}\right) \leq \sqrt{n'd}\left(6\sigma + 9\sqrt{\frac{d}{\sigma^2}}\right) + \frac{9n'}{d^{5/2}}. \tag{102}$$

Combining (74)(75)(83) and (101)(102) together, we get with probability $\geq 1 - Ce^{-cd/\sigma^2}$,

$$|\mu^\top \delta_{\text{om}}| \leq C\sqrt{\frac{d}{n'}}\left(1 + \frac{9d}{\sigma}\right) + \frac{C}{d^{3/2}}, \tag{103}$$

$$\|\delta_{\text{om}}\|_2 \leq C\sqrt{\frac{d}{n'}}\left(6\sigma + 9\sqrt{\frac{d}{\sigma^2}}\right) + \frac{C}{d^{5/2}}. \tag{104}$$

Then $\frac{\mu^\top \hat{\theta}_{\text{om}}}{\|\hat{\theta}_{\text{om}}\|_2}$ can be lower bounded by

$$\frac{\mu^\top \hat{\theta}_{\text{om}}}{\|\hat{\theta}_{\text{om}}\|_2} = \frac{\mu^\top \mu + \mu^\top \delta_{\text{om}}}{\|\mu + \delta_{\text{om}}\|_2} \tag{105}$$

$$\geq \frac{\mu^\top \mu + \mu^\top \delta_{\text{om}}}{\|\mu\|_2 + \|\delta_{\text{om}}\|_2} \tag{106}$$

$$\geq \frac{d(1 - 1/\sqrt{d})}{\sqrt{d}(1 + 1/\sqrt{d})} \tag{107}$$

$$\geq \sqrt{d}\left(1 - \frac{2}{\sqrt{d}}\right). \tag{108}$$

The proof is completed by plugging the above into the closed form expressions (13) and (19) of the errors. □

E. Details of Experiments

E.1. Experimental Settings

Software and Hardware. We run all experiments with PyTorch and NVDIA GeForce RTX 2080Ti GPUs.

Number of Evaluation Runs. We run all experiments once with fixed random seeds.

In-distribution Dataset. We use CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) as in-distribution datasets that have 10 and 100 classes, respectively. Both datasets consist of 50,000 training images and 10,000 test images.

OOD Test Dataset. We provide the details of OOD test datasets below. All images are of size 32×32.

1. **SVHN.** The SVHN dataset (Netzer et al., 2011) contains color images of house numbers. There are ten classes of digits `0-9`. The original test set has 26,032 images. We randomly select 1,000 test images for each class and form a new test dataset of 10,000 images for evaluation.
2. **Textures.** The Describable Textures Dataset (DTD) (Cimpoi et al., 2014) contains textural images in the wild. We include the entire collection of 5640 images for evaluation.
3. **Places365.** The Places365 dataset (Zhou et al., 2017) contains large-scale photographs of scenes with 365 scene categories. There are 900 images per category in the test set. We randomly sample 10,000 images from the test set for evaluation.
4. **LSUN (crop) and LSUN (resize).** The Large-scale Scene UNderstanding dataset (LSUN) has a testing set of 10,000 images of 10 different scenes (Yu et al., 2015). We construct two datasets, `LSUN-C` and `LSUN-R`, by randomly cropping image patches of size 32×32 and downsampling each image to size 32×32, respectively.
5. **iSUN.** The iSUN (Xu et al., 2015) consists of a subset of SUN images. We include the entire collection of 8925 images in iSUN.

Architectures and Training Configurations. We use the state-of-the-art neural network architecture DenseNet (Huang et al., 2017). We follow the same setup as in (Huang et al., 2017), with depth $L = 100$, growth rate $k = 12$ (Dense-BC) and dropout rate 0. All neural networks are trained with stochastic gradient descent with Nesterov momentum (Duchi et al., 2011; Kingma & Ba, 2014). Specifically, we train Dense-BC for 100 epochs with momentum 0.9 and ℓ_2 weight decay with a coefficient of 10^{-4}. The initial learning rate of 0.1 decays by 0.1 at 50, 75, 90 epoch. We use batch size 64 for in-distribution data and 128 for out-of-distribution data. To solve the inner max of the robust training objective, we use PGD with $\epsilon = 8/255$, the number of iterations of 5, the step size of $2/255$, and random start.

Method	Training	Evaluation
MSP	2.5 h	4 h
ODIN	2.5 h	4 h
Mahalanobis	2.5 h	20 h
SOFL	14 h	4 h
OE	5 h	4 h
ACET	17 h	4 h
CCU	6.7 h	4 h
ROWL	24 h	4 h
ATOM (ours)	21 h	4 h

Table 3: The estimated average runtime for each result. h means hour. For MSP, ODIN, and Mahalanobis, we use standard training. The evaluation includes four OOD detection tasks listed in Section 2.

E.2. Average Runtime

We run our experiments using a single GPU on a machine with 4 GPUs and 32 cores. The estimated average runtime for each method is summarized in Table 3.

E.3. OOD Detection Methods

Maximum Softmax Probability (MSP). Hendrycks et al. (Hendrycks & Gimpel, 2016) propose to use $\max_i F_i(x)$ as confidence scores to detect OOD examples, where $F(x)$ is the softmax output of the neural network.

ODIN. Liang et al. (Liang et al., 2017) computes calibrated confidence scores using temperature scaling and input perturbation techniques. We choose temperature scaling parameter T and perturbation magnitude η by validating on a random noise data, which does not depend on prior knowledge of test OOD datasets. In all of our experiments, we set $T = 1000$. We set $\eta = 0.0014$ for CIFAR-10, and $\eta = 0.0028$ for CIFAR-100.

Mahalanobis. Lee et al. (Lee et al., 2018) propose to use Mahalanobis distance-based confidence scores to detect OOD samples. We use 500 examples randomly selected from $\mathcal{D}_{\text{in}}^{\text{train}}$ and adversarial examples generated by FGSM (Goodfellow et al., 2014) with perturbation size of 0.05 to train the Logistic Regression model and tune the noise perturbation magnitude η. η is chosen from 21 evenly spaced numbers between 0 and 0.004, and the optimal parameters are chosen to minimize the FPR at FNR 5%.

Outlier Exposure (OE). Outlier Exposure (Hendrycks et al., 2018) makes use of a large, unlabeled dataset $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$ to enhance the performance of existing OOD detection. We train from scratch for 100 epochs with $\lambda = 0.5$, and use in-distribution batch size of 64 and out-distribution batch size of 128 in our experiments.

Self-Supervised OOD Feature Learning (SOFL). Mohseni et al. (Mohseni et al., 2020) add an auxiliary head to the network and train in for the OOD detection task. They first use a full-supervised training to learn in-distribution training data for the main classification head and then a self-supervised training with OOD training set for the auxiliary head. Following the original setting, we set $\lambda = 5$ and use an in-distribution batch size of 64 and an out-distribution batch size of 320 in all of our experiments. In CIFAR-10, we use 5 reject classes, while in CIFAR-100, we use 10 reject classes. In CIFAR-10 and CIFAR-100, we train the model for 100 epochs with the full-supervised training and then continue to train for 100 epochs with the self-supervised OOD feature learning. We use the large, unlabeled dataset $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$ as out-of-distribution training dataset.

Adversarial Confidence Enhancing Training (ACET). Hein et al. (Hein et al., 2019) propose Adversarial Confidence Enhancing Training to enforce low model confidence for the OOD data point, as well as worst-case adversarial example in the neighborhood of an OOD example. We use the large, unlabeled dataset $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$ as an OOD training dataset instead of using random noise data for a fair comparison. In all of our experiments, we train for 100 epochs and set $\lambda = 1.0$. For both in-distribution and out-distribution, we use a batch size of 128. To solve the inner max of the training objective, we also apply PGD with $\epsilon = 8/255$, the number of iterations of 5, the step size of $2/255$, and random start to a half of a minibatch while keeping the other half clean to ensure proper performance on both perturbed and clean OOD examples for a fair comparison.

Certified Certain Uncertainty (CCU). Certified Certain Uncertainty (Meinke & Hein, 2019) gives guarantees on the confidence of the classifier decision far away from the training data. We use the same training set up as in the paper and code, except for an architectural difference (DenseNet).

Robust Open-World Deep Learning (ROWL). Sehwag et al. (Sehwag et al., 2019) propose to introduce additional background classes for OOD datasets and perform adversarial training on both the in- and out-of- distribution datasets to achieve robust open-world classification. When an input is classified as the background classes, it is considered as an OOD example. Thus, ROWL gives binary OOD scores (either 0 or 1) to the inputs. In our experiments, we only have one background class and randomly sample data points from the large, unlabeled dataset $\mathcal{D}_{\text{out}}^{\text{auxiliary}}$ to form the OOD dataset. To ensure data balance across classes, we include 5,000 OOD data points for CIFAR-10; while for CIFAR-100, we include 500 OOD data points. During training, we mix the in-distribution data and OOD data, use a batch size of 128, and train for 100 epochs. To solve the inner max of the training objective, we use PGD with $\epsilon = 8/255$, the number of iterations of 5, the step size of $2/255$, and random start.

E.4. Adversarial Attacks for OOD Detection Methods

We propose adversarial attack objectives for different OOD detection methods. We consider a family of adversarial perturbations for the OOD inputs: (1) L_∞-norm bounded attack (white-box attack); (2) common image corruptions attack (black-box attack); (3) compositional attack which combines common image corruptions attack and L_∞ norm bounded attack (white-box attack).

L_∞ norm bounded attack. For data point $\mathbf{x} \in \mathbb{R}^d$, the L_∞ norm bounded perturbation is defined as

$$\Omega_{\infty,\epsilon}(\mathbf{x}) = \{\delta \in \mathbb{R}^d \mid \|\delta\|_\infty \leq \epsilon \wedge \mathbf{x} + \delta \text{ is valid}\}, \tag{109}$$

where ϵ is the adversarial budget. $\mathbf{x} + \delta$ is considered valid if the values of $\mathbf{x} + \delta$ are in the image pixel value range.

For MSP, ODIN, OE, ACET, and CCU methods, we propose the following attack objective to generate adversarial OOD example on a clean OOD input $\mathbf{x}$:

$$\mathbf{x}' = \underset{\mathbf{x}' \in \Omega_{\infty,\epsilon}(\mathbf{x})}{\arg\max} -\frac{1}{K}\sum_{i=1}^{K} \log F(\mathbf{x}')_i \tag{110}$$

where $F(\mathbf{x})$ is the softmax output of the classifier network.

For Mahalanobis method, we propose the following attack objective to generate adverasrial OOD example on OOD input $\mathbf{x}$:

$$\mathbf{x}' = \underset{\mathbf{x}' \in \Omega_{\infty,\epsilon}(\mathbf{x})}{\arg\max} -\log \frac{1}{1 + e^{-(\sum_\ell \alpha_\ell M_\ell(\mathbf{x}') + b)}}, \tag{111}$$

where $M_\ell(\mathbf{x}')$ is the Mahalanobis distance-based confidence score of $\mathbf{x}'$ from the ℓ-th feature layer, $\{\alpha_\ell\}$ and b are the parameters of the logistic regression model.

For SOFL method, we propose the following attack objective to generate adversarial OOD example for an input $\mathbf{x}$:

$$\mathbf{x}' = \underset{\mathbf{x}' \in \Omega_{\infty,\epsilon}(\mathbf{x})}{\arg\max} -\log \sum_{i=K+1}^{K+R} \bar{F}(\mathbf{x}')_i \tag{112}$$

where $\bar{F}(\mathbf{x})$ is the softmax output of the whole neural network (including auxiliary head) and R is the number of reject classes.

For ROWL and ATOM method, we propose the following attack objective to generate adverasrial OOD example on OOD input $\mathbf{x}$:

$$\mathbf{x}' = \underset{\mathbf{x}' \in \Omega_{\infty,\epsilon}(\mathbf{x})}{\arg\max} -\log \hat{F}(\mathbf{x}')_{K+1} \tag{113}$$

where $\hat{F}(\mathbf{x})$ is the softmax output of the (K+1)-way neural network.

$\mathcal{D}_{\text{in}}^{\text{test}}$	Method	FNR	Pred. Acc.	End-to-end. Pred. Acc.
CIFAR-10	MSP	5.01	94.39	91.76
	ODIN	5.01	94.39	91.02
	Mahalanobis	5.01	94.39	89.71
	SOFL	5.01	95.11	91.60
	OE	5.01	94.79	91.86
	ACET	5.01	91.70	88.64
	CCU	5.01	94.89	91.88
	ROWL	1.30	89.45	89.45
	ATOM (ours)	5.01	94.98	91.14
CIFAR-100	MSP	5.01	75.05	73.87
	ODIN	5.01	75.05	73.50
	Mahalanobis	5.01	75.05	71.20
	SOFL	5.01	74.37	72.62
	OE	5.01	75.28	73.74
	ACET	5.01	74.99	73.43
	CCU	5.01	76.04	74.60
	ROWL	0.40	67.53	67.53
	ATOM (ours)	5.01	75.49	73.57

Table 4: The performance of OOD detector and classifier on in-distribution test data. We use three metrics: FNR, Prediction Accuracy and End-to-end Prediction Accuracy. We pick the threshold for the OOD detectors such that 95% of in-distribution test data points are classified as in-distribution. Prediction Accuracy measures the accuracy of the classifier on in-distribution test data. End-to-end Prediction Accuracy measures the accuracy of the open world classification system (detector+classifier), where an example is classified correctly if and only if the detector treats it as in-distribution and the classifier predicts its label correctly.

We use PGD with $\epsilon = 8/255$, the number of iterations of 40, the step size of 1/255 and random start to solve these attack objectives.

Common Image Corruptions attack. We use common image corruptions introduced in (Hendrycks & Dietterich, 2019). We apply 15 types of algorithmically generated corruptions from noise, blur, weather, and digital categories to each OOD image. Each type of corruption has five levels of severity, resulting in 75 distinct corruptions. Thus, for each OOD image, we generate 75 corrupted images and then select the one with the lowest OOD score (or highest confidence score to be in-distribution). Note that we only need the outputs of the OOD detectors to construct such adversarial OOD examples; thus it is a black-box attack.

Compositional Attack. For each OOD image, we first apply common image corruptions attack, and then apply the L_∞-norm bounded attack to generate adversarial OOD examples.

E.5. Visualizations of Four Types of OOD Samples

We show visualizations of four types of OOD samples in Figure 4.

E.6. Histogram of OOD Scores

In Figure 5, we show histogram of OOD scores for model snapshots trained on CIFAR-10 (in-distribution) using objective (2) **without** informative outlier mining. We plot every ten epochs for a model trained for a total of 100 epochs. We observe that the model quickly converges to a solution where OOD score distribution becomes dominated by *easy* examples with scores closer to 1. This is exacerbated as the model is trained for longer.

E.7. Performance of OOD Detector and Classifier on In-distribution Data

We summarize the performance of OOD detector $G(x)$ and image classifier $f(x)$ on in-distribution test data. See Table 4.

E.8. Choose Best q Using Validation Dataset

We create a validation OOD dataset by sampling 10,000 images from the 80 Million Tiny Images (Torralba et al., 2008), which is disjoint from our training data. We choose q from $\{0, 0.125, 0.25, 0.5, 0.75\}$. The results on the validation dataset are shown in Table 5. We select the best model based on the average AUROC across four types of OOD inputs. Based on

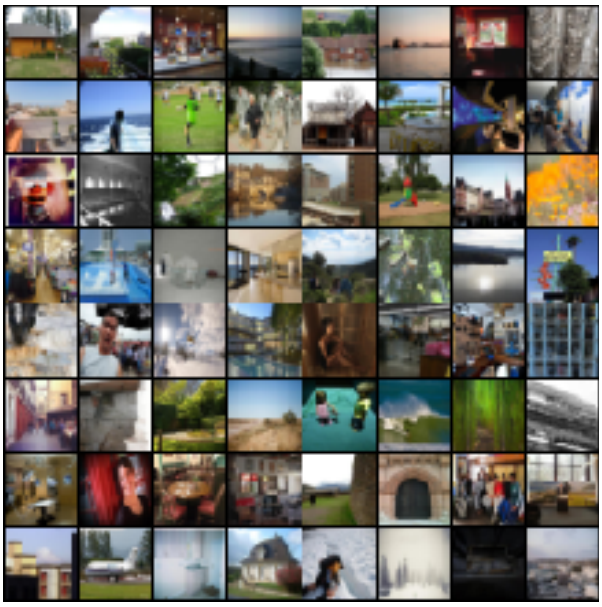
$\mathcal{D}_{\text{in}}^{\text{test}}$	Method	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑	FPR (5% FNR) ↓	AUROC ↑
		Natural OOD		Corruption OOD		L_∞ OOD		Comp. OOD	
CIFAR-10	ATOM (q=0.0)	5.39	**98.35**	39.65	92.47	35.24	91.06	60.44	80.65
	ATOM (q=0.125)	5.15	98.30	30.32	**93.85**	5.19	**98.26**	31.38	**93.81**
	ATOM (q=0.25)	6.02	98.06	33.79	92.55	22.56	95.12	43.66	91.04
	ATOM (q=0.5)	9.55	97.48	39.54	91.58	18.95	95.73	51.01	89.88
	ATOM (q=0.75)	13.98	96.61	56.88	87.00	14.10	96.61	57.02	87.09
CIFAR-100	ATOM (q=0.0)	45.25	91.53	98.84	58.54	43.14	90.22	94.68	55.53
	ATOM (q=0.125)	40.06	92.59	98.01	67.20	36.90	**92.79**	89.09	68.94
	ATOM (q=0.25)	35.84	**92.61**	96.31	**73.40**	35.03	92.70	94.63	**71.67**
	ATOM (q=0.5)	35.48	91.29	91.13	69.07	64.43	77.86	91.39	62.93
	ATOM (q=0.75)	43.13	88.42	89.83	63.89	43.17	88.45	90.05	63.84

Table 5: Evaluate models on validation dataset. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages and are averaged over six OOD test datasets mentioned in section 4.1.

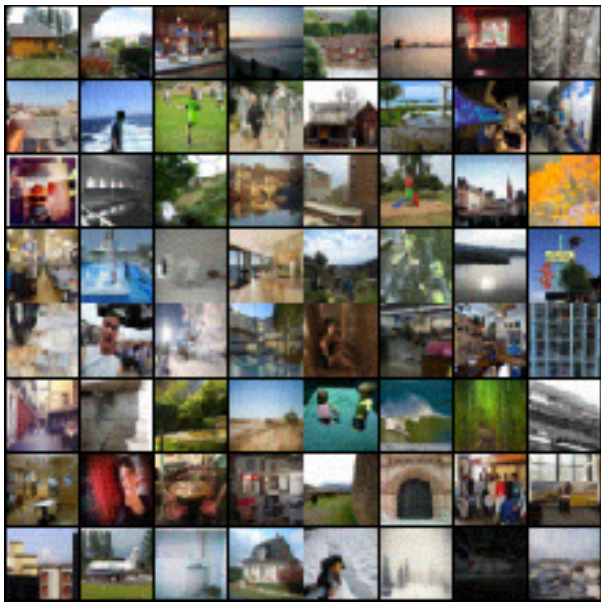
the results, the optimal q is 0.125 for CIFAR-10 and 0.25 for CIFAR-100.

E.9. Complete Experimental Results

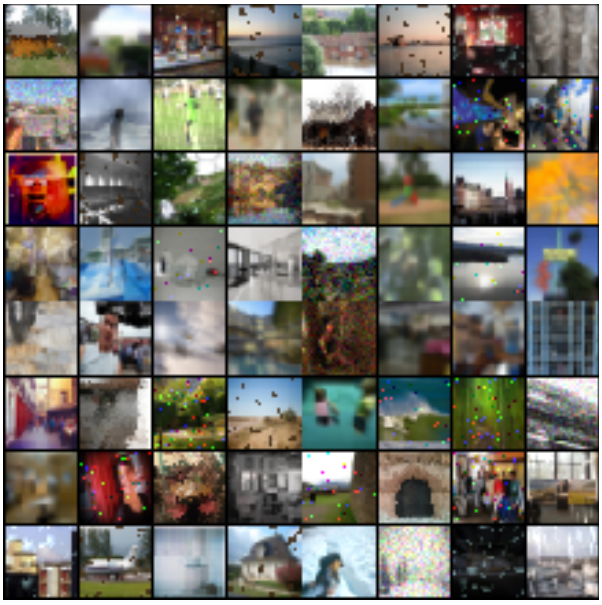
We report the performance of OOD detectors on each of the six OOD test datasets in Table 6 (CIFAR-10) and Table 7 (CIFAR-100).



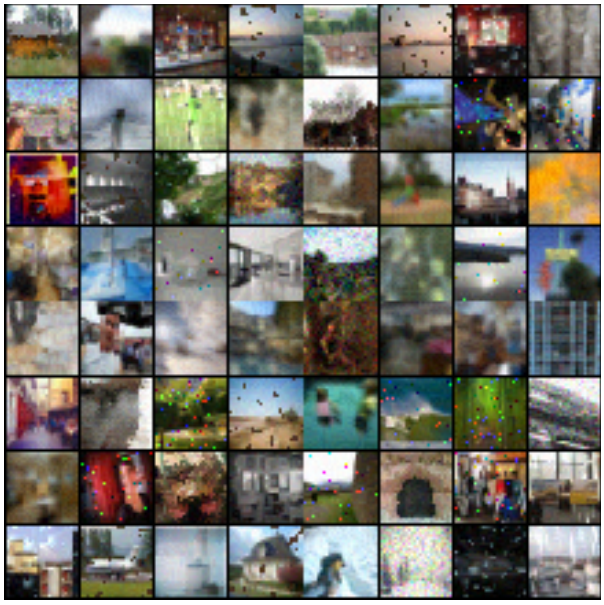
(a) Natural OOD



(b) L_∞ OOD



(c) Corruption OOD



(d) Comp. OOD

Figure 4: Examples of four types of OOD samples.

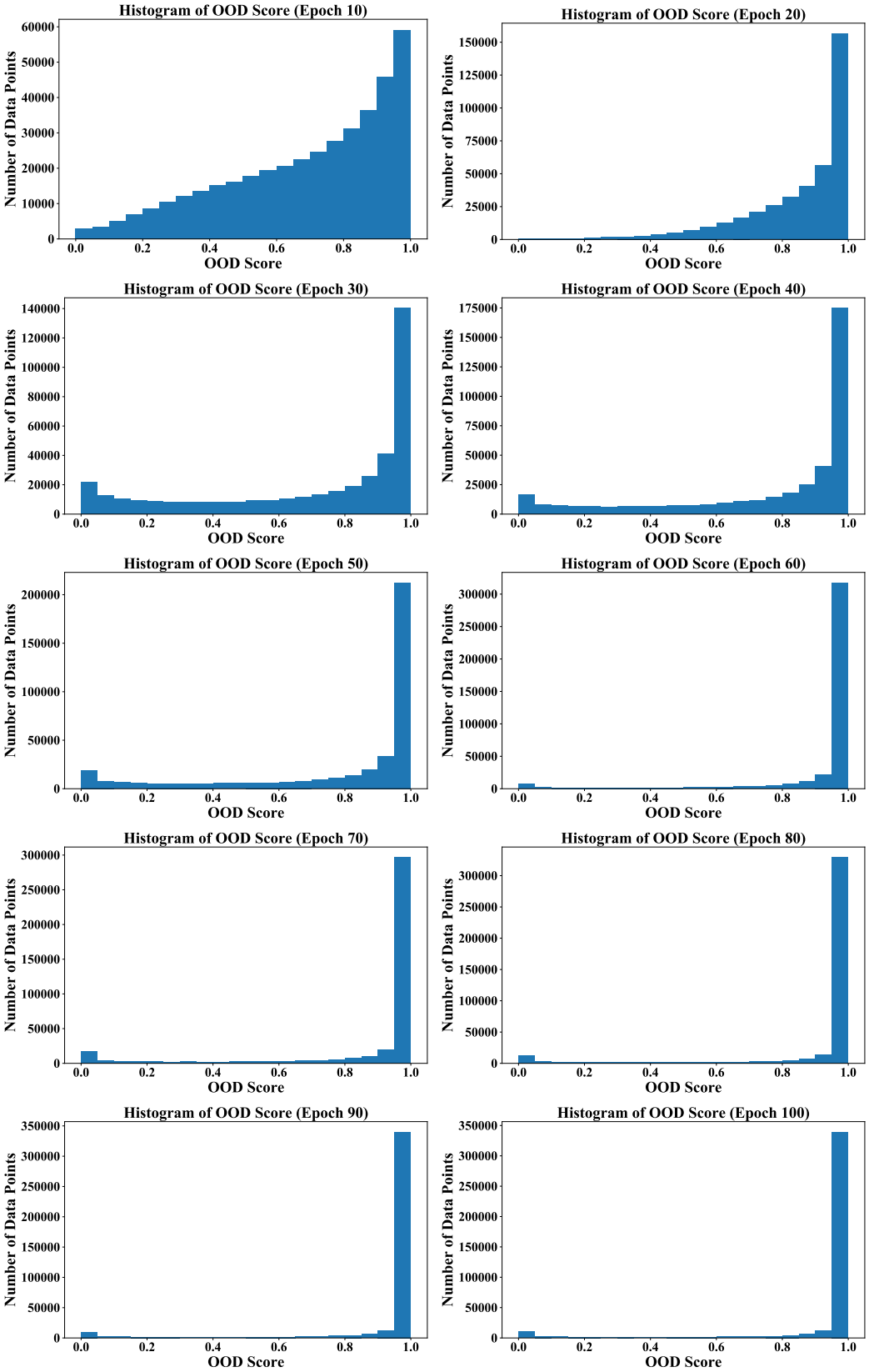


Figure 5: On CIFAR-10, we train the model with objective (2) for 100 epochs **without** informative outlier mining. For every 10 epochs, we randomly sample 400,000 data points from the large unlabeled dataset and use the current model snapshot to calculate the OOD scores.

$\mathcal{D}_{\text{out}}^{\text{test}}$	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	
LSUN-C	MSP	27.34	96.30	100.00	71.64	100.00	13.76	100.00	13.68
	ODIN	1.86	99.51	98.57	72.44	100.00	0.05	100.00	0.00
	Mahalanobis	17.79	93.30	96.83	36.41	98.97	1.61	99.91	0.49
	SOFL	0.39	99.40	55.61	93.09	100.00	2.52	100.00	1.96
	OE	0.97	99.52	41.19	93.96	99.98	0.42	100.00	0.27
	ACET	2.10	99.37	36.04	94.85	48.20	90.03	91.77	68.88
	CCU	0.62	**99.65**	33.57	94.59	99.93	0.39	100.00	0.05
	ROWL	22.65	88.02	95.46	51.62	80.54	59.08	97.40	50.65
	ATOM (ours)	**0.34**	99.57	**11.35**	**97.31**	**0.36**	**99.64**	**11.08**	**97.51**
LSUN-R	MSP	43.89	93.93	100.00	64.35	100.00	13.74	100.00	13.66
	ODIN	3.33	99.17	98.94	64.72	100.00	0.11	100.00	0.00
	Mahalanobis	6.68	98.04	98.77	37.01	97.98	4.87	100.00	0.06
	SOFL	1.67	99.29	55.99	90.62	100.00	0.55	100.00	0.43
	OE	0.99	**99.43**	51.61	92.22	99.98	0.14	100.00	0.04
	ACET	4.35	99.03	78.49	86.79	72.93	82.95	99.87	47.92
	CCU	1.53	99.28	57.05	90.50	100.00	0.03	100.00	0.10
	ROWL	55.82	71.44	98.74	49.98	99.16	49.77	99.77	49.46
	ATOM (ours)	**0.79**	99.10	**26.39**	**95.60**	**37.87**	**94.48**	**50.63**	**92.45**
iSUN	MSP	46.18	93.58	100.00	62.57	100.00	13.94	100.00	13.66
	ODIN	4.64	98.96	98.85	62.93	100.00	0.29	100.00	0.00
	Mahalanobis	8.28	97.83	98.01	40.70	94.85	9.22	100.00	0.09
	SOFL	2.24	99.22	53.64	90.99	100.00	0.54	100.00	0.50
	OE	1.14	**99.40**	47.94	92.52	99.98	0.19	100.00	0.04
	ACET	7.09	98.51	75.71	86.55	80.94	78.98	99.82	46.59
	CCU	1.74	99.27	52.32	91.01	100.00	0.06	100.00	0.14
	ROWL	58.38	70.16	98.35	50.17	99.51	49.60	99.79	49.46
	ATOM (ours)	**1.10**	99.17	**23.06**	**95.96**	**41.97**	**93.35**	**49.77**	**92.08**
Textures	MSP	64.66	87.64	100.00	51.85	100.00	14.20	100.00	13.72
	ODIN	51.68	85.26	99.56	39.44	99.96	0.57	100.00	0.07
	Mahalanobis	29.50	90.49	77.75	51.80	94.43	7.04	99.66	0.83
	SOFL	3.78	99.04	57.16	89.41	99.89	2.22	99.98	1.41
	OE	6.24	98.43	53.90	88.84	99.79	1.34	99.96	0.61
	ACET	12.66	97.86	56.29	89.91	66.67	76.56	95.94	54.61
	CCU	5.83	98.45	54.54	86.23	99.47	1.63	99.88	0.94
	ROWL	24.59	87.05	82.30	58.20	85.39	56.65	92.55	53.07
	ATOM (ours)	**1.95**	**99.43**	**22.94**	**94.80**	**3.44**	**99.06**	**26.47**	**94.29**
Places365	MSP	62.03	88.29	100.00	57.74	100.00	13.67	100.00	13.66
	ODIN	42.67	90.63	99.90	54.15	100.00	0.01	100.00	0.00
	Mahalanobis	86.40	65.89	99.54	20.39	99.91	0.34	100.00	0.13
	SOFL	7.73	97.81	60.46	88.28	100.00	0.50	100.00	0.26
	OE	11.08	97.00	68.24	87.47	100.00	0.02	100.00	0.02
	ACET	17.59	96.12	79.85	84.84	92.44	67.11	99.85	45.70
	CCU	8.49	97.63	66.43	85.79	99.99	0.04	100.00	0.01
	ROWL	61.01	68.84	98.49	50.10	99.31	49.69	99.77	49.46
	ATOM (ours)	**6.95**	**97.82**	**35.37**	**92.95**	**6.95**	**97.88**	**36.11**	**93.03**
SVHN	MSP	59.15	90.99	100.00	41.88	100.00	13.66	100.00	13.66
	ODIN	26.12	94.78	100.00	20.26	100.00	0.00	100.00	0.00
	Mahalanobis	22.36	92.08	96.57	40.23	99.91	0.32	99.99	0.31
	SOFL	**0.85**	99.47	88.09	79.91	100.00	0.13	100.00	0.09
	OE	1.55	99.16	75.77	88.96	100.00	0.01	100.00	0.01
	ACET	35.00	94.80	84.86	85.07	93.96	70.31	99.89	54.24
	CCU	2.14	99.25	75.10	87.94	100.00	0.00	100.00	0.00
	ROWL	36.42	81.14	91.78	53.46	96.49	51.10	97.86	50.42
	ATOM (ours)	1.27	**99.59**	**42.61**	**93.18**	**1.26**	**99.60**	**42.43**	**93.35**

Table 6: Comparison with competitive OOD detection methods. We use CIFAR-10 as in-distribution dataset. We evaluate the performance on all four types of OOD inputs: (1) natural OOD, (2) corruption attacked OOD, (3) L_∞ attacked OOD, and (4) compositionally attacked OOD inputs. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages. **Bold** numbers are superior results.

$\mathcal{D}_{\text{out}}^{\text{test}}$	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	
LSUN-C	MSP	62.03	84.78	100.00	32.56	100.00	2.73	100.00	2.49
	ODIN	**10.54**	**98.13**	99.99	50.42	100.00	0.76	100.00	0.03
	Mahalanobis	38.69	91.18	98.55	42.44	99.08	24.72	99.95	11.86
	SOFL	17.38	96.66	100.00	51.63	100.00	1.42	100.00	0.64
	OE	14.75	97.33	99.98	54.39	100.00	1.67	100.00	0.69
	ACET	13.69	97.55	99.78	59.52	42.55	88.00	97.34	39.64
	CCU	12.03	97.84	99.66	61.38	100.00	1.01	100.00	0.52
	ROWL	95.61	52.00	100.00	49.80	99.98	49.81	100.00	49.80
	ATOM (ours)	32.72	95.13	**98.36**	**81.30**	**32.43**	**95.20**	**95.84**	**79.86**
LSUN-R	MSP	77.48	76.40	100.00	32.23	100.00	1.98	100.00	1.81
	ODIN	31.96	94.04	100.00	41.10	100.00	0.55	100.00	0.00
	Mahalanobis	**16.58**	95.92	99.87	27.36	93.77	36.07	100.00	9.02
	SOFL	50.27	90.28	99.85	50.26	100.00	0.12	100.00	0.20
	OE	56.25	84.35	99.97	41.30	100.00	0.70	100.00	0.52
	ACET	51.59	86.09	99.87	37.13	99.41	13.29	99.65	9.48
	CCU	38.44	91.83	99.94	50.62	100.00	0.61	100.00	0.47
	ROWL	92.38	53.61	100.00	49.80	100.00	49.80	100.00	49.80
	ATOM (ours)	24.28	**96.15**	**99.00**	**74.58**	**82.34**	**64.04**	**95.90**	**66.73**
iSUN	MSP	78.87	75.69	100.00	31.77	100.00	2.14	100.00	1.83
	ODIN	34.89	93.08	100.00	39.49	100.00	0.82	100.00	0.00
	Mahalanobis	**18.66**	95.22	99.74	29.69	88.37	39.44	100.00	9.25
	SOFL	53.51	89.27	99.92	48.59	100.00	0.20	100.00	0.22
	OE	61.59	81.51	99.96	40.04	100.00	0.82	100.00	0.55
	ACET	54.34	84.75	99.92	36.81	99.37	14.89	99.64	11.97
	CCU	40.97	90.89	99.97	49.04	100.00	0.78	100.00	0.44
	ROWL	94.55	52.52	100.00	49.80	100.00	49.80	100.00	49.80
	ATOM (ours)	27.05	**95.66**	**99.25**	**71.68**	**81.39**	**62.63**	**96.97**	**64.48**
Textures	MSP	85.57	70.08	100.00	26.02	100.00	2.74	100.00	2.30
	ODIN	81.24	71.69	100.00	27.26	99.98	0.23	100.00	0.01
	Mahalanobis	**41.91**	84.82	**82.85**	45.78	89.75	27.95	99.50	12.49
	SOFL	57.00	87.35	99.75	43.98	99.98	0.62	100.00	0.39
	OE	59.86	86.17	99.91	43.10	100.00	1.55	100.00	0.70
	ACET	61.90	85.13	99.77	41.71	83.42	54.10	98.32	31.48
	CCU	60.80	86.34	99.88	44.85	100.00	1.36	100.00	0.58
	ROWL	97.11	51.25	100.00	49.80	99.98	49.81	100.00	49.80
	ATOM (ours)	45.25	**90.68**	98.74	**66.76**	**49.11**	**88.26**	**97.30**	**65.24**
Places365	MSP	83.65	73.71	100.00	32.23	100.00	1.87	100.00	1.94
	ODIN	80.25	76.20	100.00	36.22	100.00	0.01	100.00	0.00
	Mahalanobis	94.52	59.41	99.82	16.47	99.94	11.19	100.00	8.02
	SOFL	60.49	87.57	99.99	40.21	100.00	0.09	100.00	0.16
	OE	58.37	86.39	99.97	50.91	100.00	0.56	100.00	0.54
	ACET	56.81	86.75	99.82	48.27	92.20	51.07	98.41	27.41
	CCU	55.23	87.21	99.98	44.11	100.00	0.47	100.00	0.42
	ROWL	96.52	51.54	100.00	49.80	100.00	49.80	100.00	49.80
	ATOM (ours)	**52.63**	**88.67**	**99.24**	**69.22**	**49.92**	**89.02**	**97.95**	**66.64**
SVHN	MSP	80.71	76.00	100.00	25.65	100.00	2.62	100.00	2.39
	ODIN	79.27	73.55	100.00	23.68	100.00	0.00	100.00	0.00
	Mahalanobis	49.15	87.33	**98.91**	39.10	99.84	20.90	100.00	11.25
	SOFL	**21.50**	96.15	100.00	36.52	100.00	0.09	100.00	0.19
	OE	44.47	92.58	100.00	40.87	100.00	0.54	100.00	0.55
	ACET	47.80	90.55	100.00	36.84	59.05	82.17	98.25	29.76
	CCU	50.79	91.59	100.00	39.94	100.00	0.45	100.00	0.41
	ROWL	99.05	50.28	100.00	49.80	100.00	49.80	100.00	49.80
	ATOM (ours)	22.42	**96.43**	99.91	**70.09**	**22.15**	**96.50**	**97.01**	**70.64**

Table 7: Comparison with competitive OOD detection methods. We use CIFAR-100 as in-distribution dataset. We evaluate the performance on all four types of OOD inputs: (1) natural OOD, (2) corruption attacked OOD, (3) L_∞ attacked OOD, and (4) compositionally attacked OOD inputs. ↑ indicates larger value is better, and ↓ indicates lower value is better. All values are percentages. **Bold** numbers are superior results.