

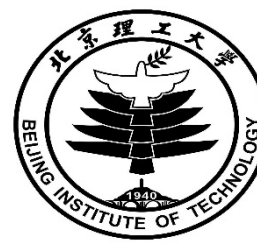
Detecting Adversarial Data by Probing Multiple Perturbations Using Expected Perturbation Score

Shuhai Zhang*, Feng Liu*, Jiahao Yang, Yifan Yang,
Changsheng Li, Bo Han and Mingkui Tan

South China University of Technology, The University of Melbourne,
Beijing Institute of Technology, Hong Kong Baptist University



ICML
International Conference
On Machine Learning



Contents

- 01** Background
- 02** **Expected Perturbation Score for Adversarial Detection**
 - Expected Perturbation Score (EPS)
 - Exploring EPS for Adversarial Detection
- 03** Experimental Results
- 04** Conclusion



Contents

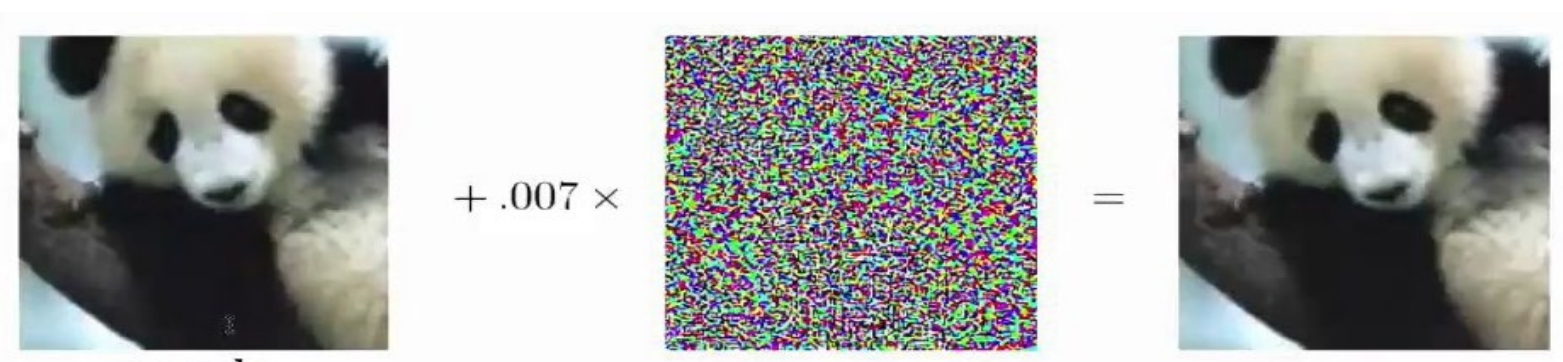
- 01 **Background**
- 02 Expected Perturbation Score for Adversarial Detection
 - Expected Perturbation Score (EPS)
 - Exploring EPS for Adversarial Detection
- 03 Experimental Results
- 04 Conclusion



Background: Adversarial Example

- Adversarial examples are generated by adding imperceptible perturbations to the input but may mislead the model to make unexpected predictions

Panda or Gibbon?



The diagram illustrates the concept of an adversarial example. It shows a sequence of three images: a panda, a small square of random noise, and a gibbon. The panda image is labeled 'panda' with '58% confidence'. This is followed by the text '+ .007 ×' and the noise image. This is followed by an equals sign and the gibbon image, which is labeled 'gibbon' with '99% confidence'.

panda
58% confidence

+ .007 ×

=

gibbon
99% confidence

How to solve this problem of so-called adversarial examples?

Background: Adversarial Robustness

□ Three types of advanced techniques to improve the robustness of models

- **Adversarial training** introduces adversarial data into training to improve the robustness of models but suffers from significant performance degradation and high computational complexity
- **Adversarial purification** relies on generative models to purify adversarial data before classification, which still has to compromise on unsatisfactory natural and adversarial accuracy
- **Adversarial detection** aims to tell whether a test sample is an adversarial one, for which the key is to find the discrepancy between the adversarial and natural distributions



*Existing adversarial detection approaches primarily **train a tailored detector for specific attacks** (Ma et al., 2018; Lee et al., 2018) or **for a specific classifier** (Deng et al., 2021), which largely overlook modeling the adversarial and natural distributions, **limiting their performance against unseen attacks or transferable attacks***

Ma et al., Characterizing adversarial subspaces using local intrinsic dimensionality. ICLR 2018.

Lee et al., A simple unified framework for detecting out-of-distribution samples and adversarial attacks. NeurIPS 2018.

Deng et al., A practical bayesian approach to adversarial detection. CVPR 2021.

Motivation

Intuition from score function $\nabla_{\mathbf{x}} \log p(\mathbf{x})$

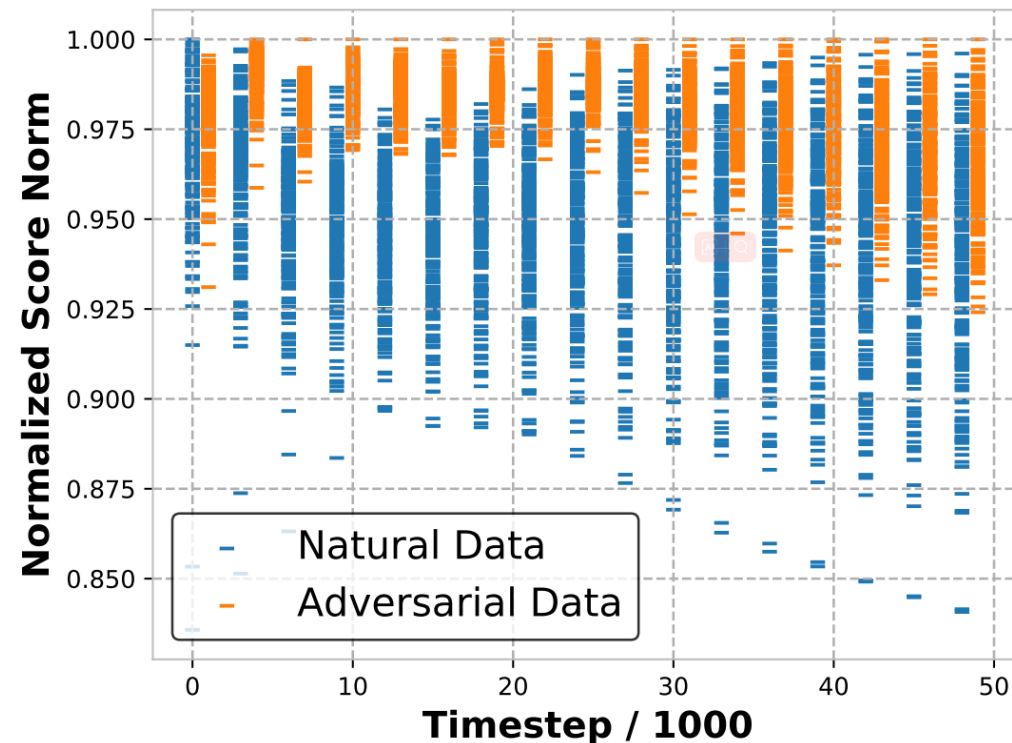
- Score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ represents the momentum of the sample towards **high density areas** of natural data (Song et al., 2019)



- A **lower** score norm indicates the sample is **closer** to the high-density areas of natural data

One score is useful but **not effective enough**

- Following a diffusion process, most natural samples have lower score norms than adversarial samples at the same timestep, but they are **very sensitive to the timesteps** due to the significant overlap across all timesteps

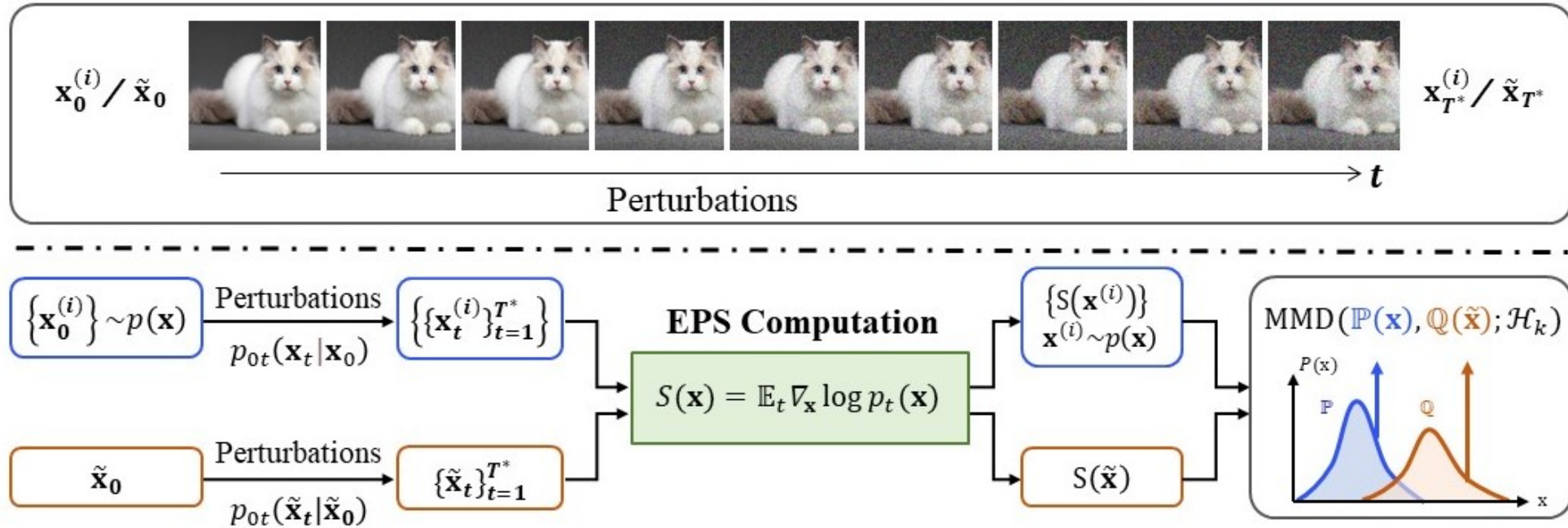


Contents

- 01 Background
- 02 Expected Perturbation Score for Adversarial Detection**
 - Expected Perturbation Score (EPS)
 - Exploring EPS for Adversarial Detection
- 03 Experimental Results
- 04 Conclusion



Adversarial Detection with Expected Perturbation Score



□ Computing expected perturbation score (EPS) using a pre-trained score model

- Add perturbations to a set of **natural images** and a **test image** following a diffusion process with time step T^* and obtain their EPSs via the score model

□ Adversarial detection with EPS (EPS-AD)

- Compute the *maximum mean discrepancy* (MMD) between the test sample and natural samples with EPS

Expected Perturbation Score (EPS)

Definition 1 (Expected Perturbation Score)

- Let $\mathcal{X} \subset \mathbb{R}^d$ be a separable metric space and p be Borel probability measure on \mathcal{X} . Given a perturbation process transition distribution $p_{0t}(\mathbf{x}_t | \mathbf{x}_0)$ the expected perturbation score (EPS) of a sample $\mathbf{x} \sim p$ is:

$$S(\mathbf{x}) = \mathbb{E}_{t \sim U(0, T)} \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$$

where $p_t(\mathbf{x})$ is the marginal probability distribution of \mathbf{x}_t with $p_0(\mathbf{x}) := p(\mathbf{x})$

- The perturbation transition distribution $p_{0t}(\mathbf{x}_t | \mathbf{x}_0)$ can be any distribution, e.g., Gaussian distribution
- $S(\mathbf{x})$ incorporates multiple levels of noises instead of a single one at different timesteps
- Estimation for EPS with a score model: $S(\mathbf{x}) = \mathbb{E}_{t \sim U(0, T)} s_{\theta}(\mathbf{x}_t, t) \approx \mathbb{E}_{t \sim U(0, T)} \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$

Expected Perturbation Score (EPS)

■ **Theorem 1** Assuming that the distribution of natural data $p(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_x, \sigma_x^2 \mathbf{I})$, where \mathbf{I} is an identity matrix, given a perturbation transition kernel $p_{0t}(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\gamma_t \mathbf{x}_0, \sigma_t^2 \mathbf{I})$ with γ_t and σ_t being the time-dependent noise schedule, then the following three conclusions for $S(\mathbf{x}) = \mathbb{E}_{t \sim U(0,T)} \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ hold:

- For $\forall \mathbf{x} \sim p(\mathbf{x}), S(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \sigma_S^2 \mathbf{I})$
- For $\forall \mathbf{y} \sim p(\mathbf{x})$ and adversarial sample $\hat{\mathbf{y}} = \mathbf{y} + \boldsymbol{\varepsilon}$, $S(\hat{\mathbf{y}}) \sim \mathcal{N}(-\boldsymbol{\mu}_S, \sigma_S^2 \mathbf{I})$
- For $\forall \mathbf{x}, \mathbf{y} \sim p(\mathbf{x})$ and adversarial sample $\hat{\mathbf{y}} = \mathbf{y} + \boldsymbol{\varepsilon}$,

$$S(\mathbf{x}) - S(\mathbf{y}) \sim \mathcal{N}(\mathbf{0}, 2\sigma_S^2 \mathbf{I}), \quad S(\mathbf{x}) - S(\hat{\mathbf{y}}) \sim \mathcal{N}(\boldsymbol{\mu}_S, 2\sigma_S^2 \mathbf{I})$$

where $\boldsymbol{\mu}_S = \mathbb{E}_{t \sim U(0,T)} \boldsymbol{\mu}_t$ with $\boldsymbol{\mu}_t = \frac{\boldsymbol{\varepsilon}}{\gamma_t^2 \sigma_x^2 + \sigma_t^2}$ and $\sigma_S^2 = \mathbb{E}_{t \sim U(0,T)} \zeta_t^2$ with $\zeta_t^2 = \frac{1}{\gamma_t^2 \sigma_x^2 + \sigma_t^2}$

- Discrepancy between the adversarial sample and the natural sample is obvious due to the term $\boldsymbol{\mu}_S$

□ Why consider multiple scores in EPS?

One score of some unique timestep t is difficult to find a good solution: smaller variance σ_S^2 and larger mean $\|\boldsymbol{\mu}_S\|^2$ are required for good adversarial detection, but they decrease as the timestep t increases



Taking expectation on multiple scores makes the discrepancy more stable

Exploring EPS for Adversarial Detection

Nature samples $\mathbb{P}_X = \{\mathbf{x}^{(i)}\}_{i=1}^n$ A test sample $\mathbb{Q}_Y = \{\hat{\mathbf{x}}\}$

$$\widehat{\text{MMD}}^2[\mathbb{P}_X, \mathbb{Q}_Y; \mathcal{H}_k] = \frac{1}{n^2} \sum_{i,j=1}^n k(S(\mathbf{x}^{(i)}), S(\mathbf{x}^{(j)})) - \frac{2}{n} \sum_{i=1}^n k(S(\mathbf{x}^{(i)}), S(\mathbf{y})) + k(S(\mathbf{y}), S(\mathbf{y}))$$

■ **Corollary 1** Considering the Gaussian kernel $k(\mathbf{a}, \mathbf{b}) = \exp(-\|\mathbf{a} - \mathbf{b}\|^2 / (2\sigma^2))$ and the assumption in Theorem 1, for $\forall 0 < \eta < 1$, the probability of $P\{k(S(\mathbf{x}), S(\hat{\mathbf{y}})) > \eta\}$ is given by :

$$P\{k(S(\mathbf{x}), S(\hat{\mathbf{y}})) > \eta\} = \int_0^C \chi_d^2(z) dz$$

where $z = \|\boldsymbol{\mu}_S\|^2$ with $\boldsymbol{\mu}_t$ being the mean of $S(\mathbf{x}) - S(\mathbf{y})$, C is a constant for given η and σ

Smaller $\|\boldsymbol{\mu}_S\|^2$ leads to larger $k(S(\mathbf{x}), S(\hat{\mathbf{y}}))$ given an η

MMD between EPSs of the natural samples is smaller than that between natural and adversarial samples

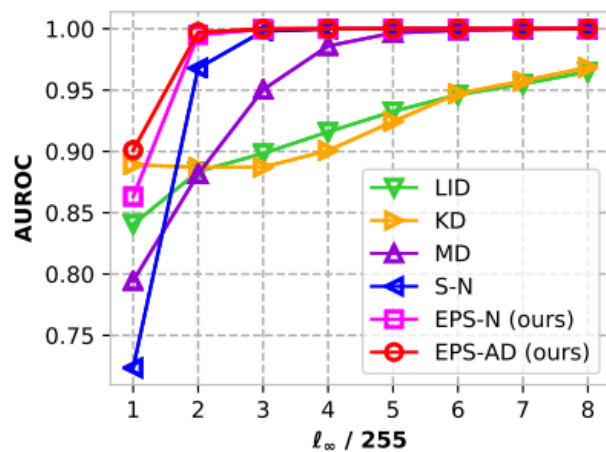
Contents

- 01 Background
- 02 Expected Perturbation Score for Adversarial Detection
 - Expected Perturbation Score (EPS)
 - Exploring EPS for Adversarial Detection
- 03 Experimental Results**
- 04 Conclusion

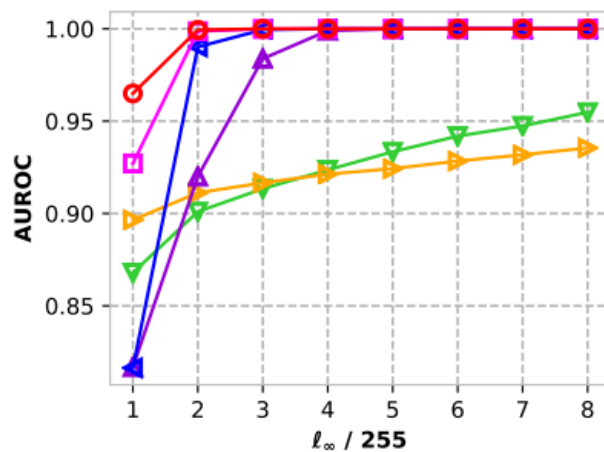


Detecting Known Attacks

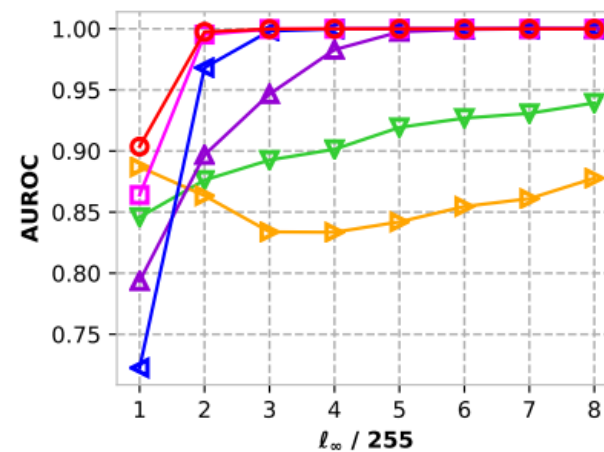
□ EPS-AD achieves the best detection performance over CIFAR-10 in terms of AUROC



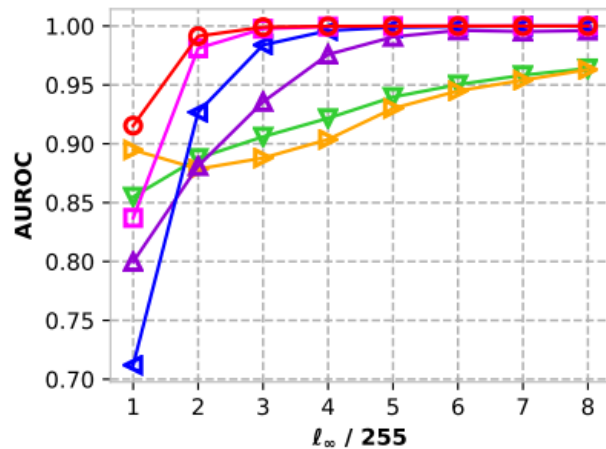
(a) Attack Method: PGD



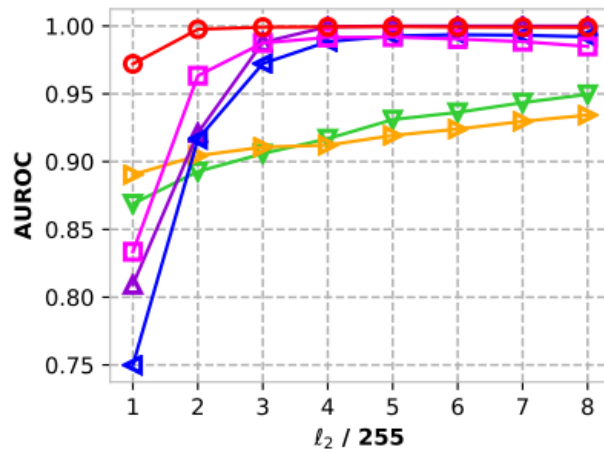
(b) Attack Method: FGSM



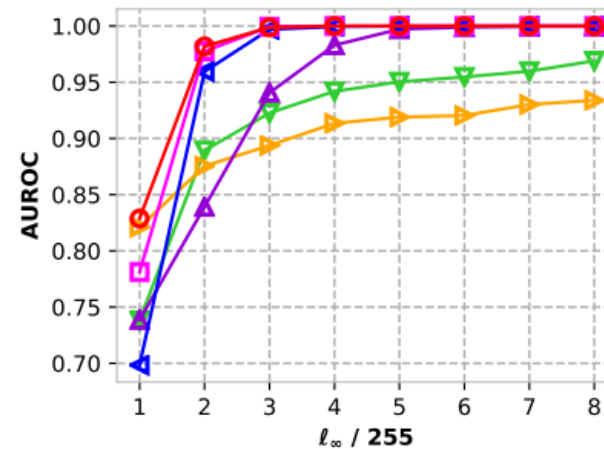
(c) Attack Method: CW



(d) Attack Method: BIM



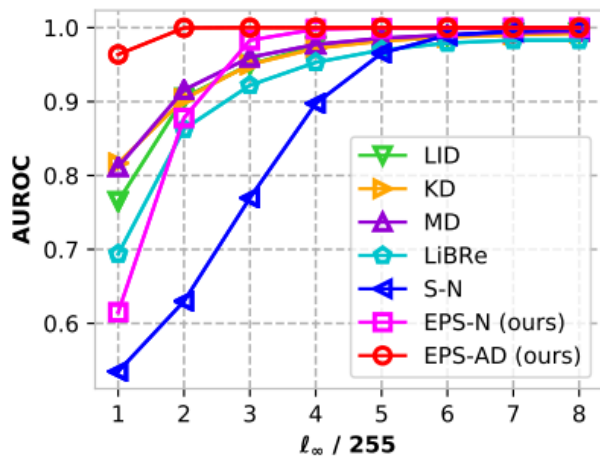
(e) Attack Method: FGSM- l_2



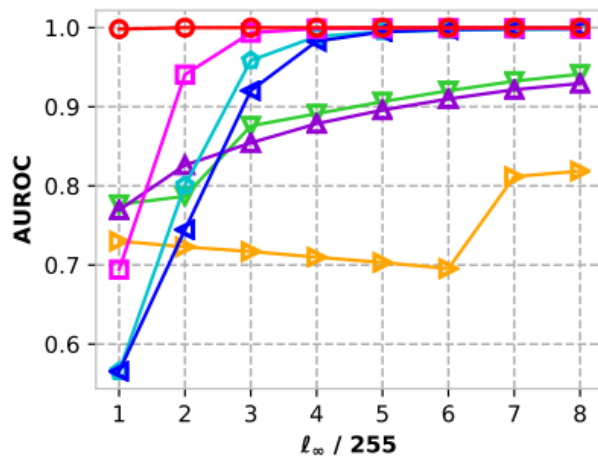
(f) Attack Method: AA Attack

Detecting Known Attacks

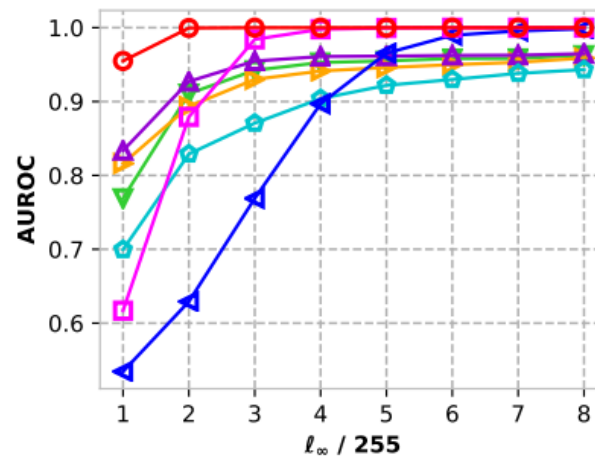
□ EPS-AD consistently achieves the best detection performance over ImageNet, especially for low ϵ



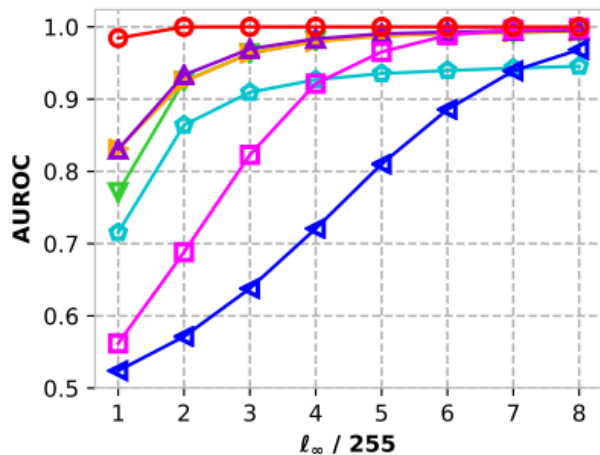
(a) Attack Method: PGD



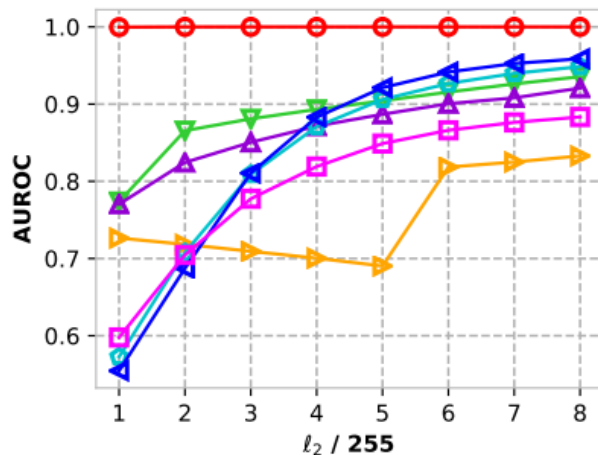
(b) Attack Method: FGSM



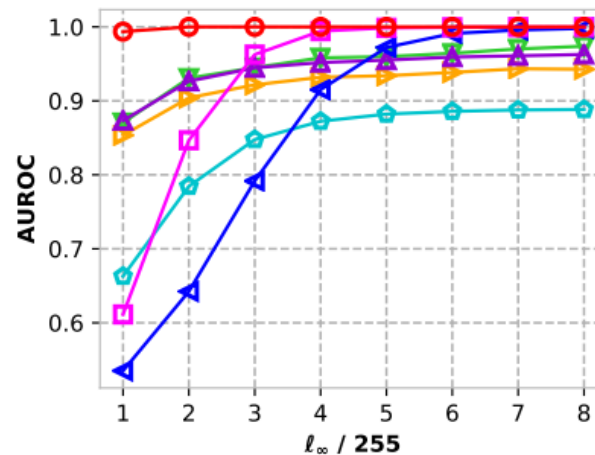
(c) Attack Method: CW



(d) Attack Method: BIM



(e) Attack Method: FGSM- l_2



(f) Attack Method: AA Attack

Detecting Unseen and Transferable Attacks

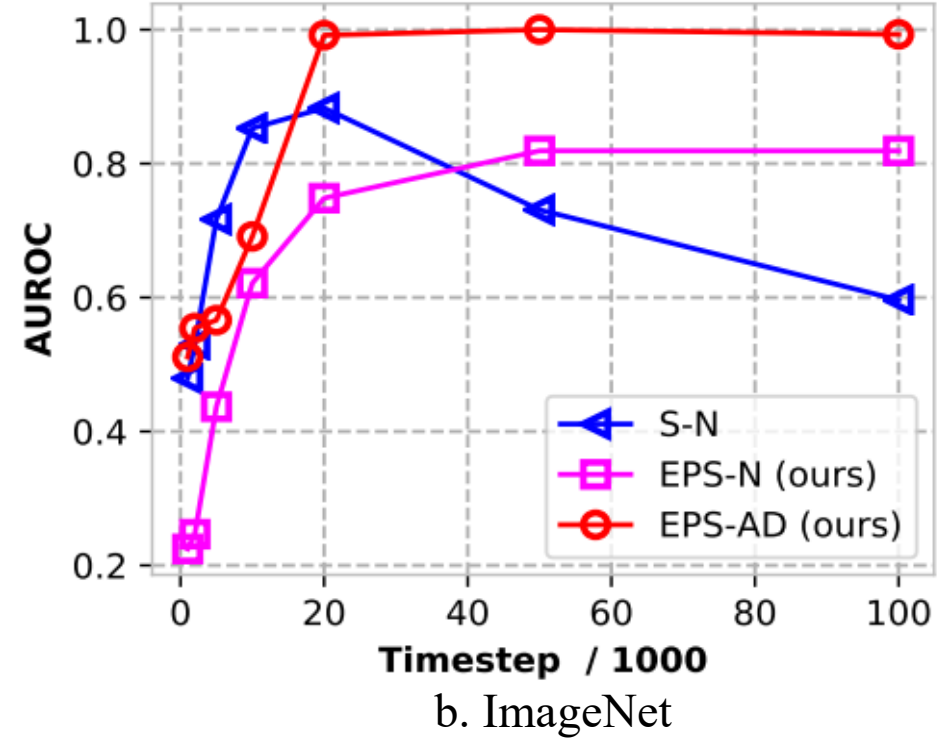
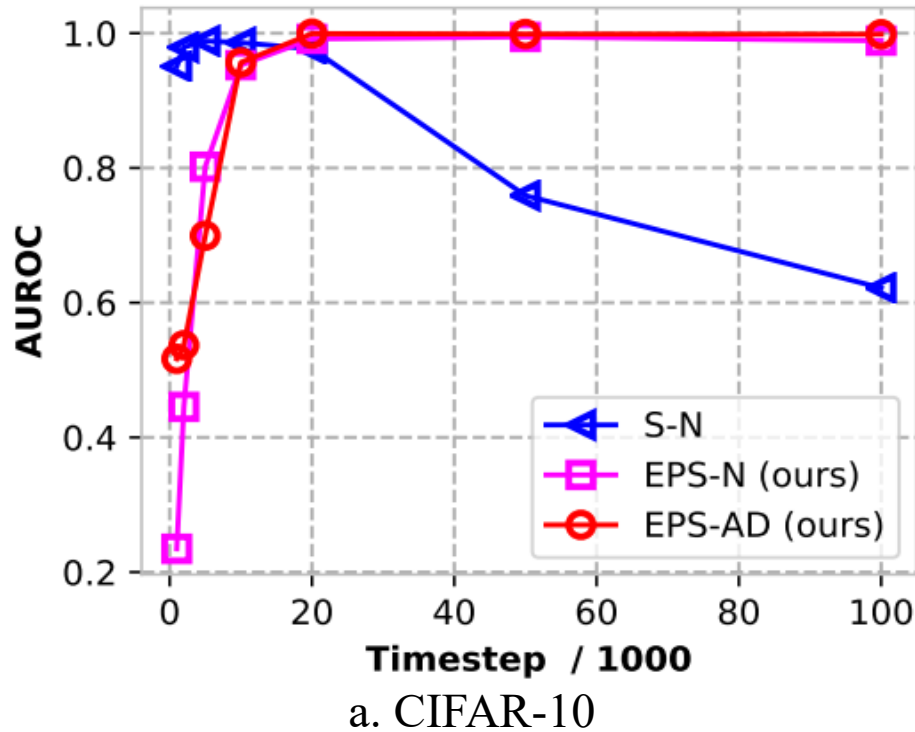
- Unseen attacks on CIFAR-10: several baselines (e.g., MD and LID) **worsens**, while diffusion-based methods **keep superior performance** since they focus more on modeling natural distribution

AUROC	FGSM(seen)	PGD	BIM	CW	FGSM- l_2 (seen)	BIM- l_2	AA
KD	0.9213	0.9007	0.9082	0.8339	0.9146	0.9146	0.9135
LID	0.9236	0.8964	0.9028	0.8828	0.9160	0.8984	0.9253
MD	0.9990	0.9855	0.9742	0.9835	0.9992	0.9503	0.9820
S-N	1.0000	0.9998	0.9961	0.9998	0.9885	0.9674	0.9995
EPS-N (Ours)	1.0000	1.0000	0.9996	1.0000	0.9916	0.9883	1.0000
EPS-AD (Ours)	1.0000	1.0000	0.9998	1.0000	0.9995	0.9991	1.0000

- Transferable attacks on ImageNet: non-diffusion-based methods **drop performance significantly**, while our EPS-AD achieves **significantly better transferability** since it does not rely on specific classifiers

AUROC	FGSM	PGD	BIM	CW	FGSM- l_2	BIM- l_2	AA
KD	0.7754	0.5999	0.5847	0.7632	0.7906	0.7756	0.7698
LID	0.8467	0.7627	0.7663	0.7704	0.8520	0.7925	0.7967
MD	0.8467	0.7698	0.7684	0.7665	0.8067	0.7759	0.7880
LiBRe	0.9849	0.8414	0.7161	0.8286	0.8489	0.7250	0.8485
S-N	0.9816	0.8965	0.7166	0.8963	0.8764	0.6705	0.9106
EPS-N (Ours)	0.9983	0.9975	0.9178	0.9979	0.8235	0.7215	0.9930
EPS-AD (Ours)	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	1.0000

Ablation Study on Impact of Timestep



- Our EPS-AD and EPS-N are **not sensitive** to the total timestep T while S-N **fluctuates greatly** with the timestep t
- As the total timestep T increases, our EPS-AD and EPS-N exhibit **progressively better** performance, however, this gain **gradually decreases** when T exceeds the optimal value

Contents

- 01 Background
- 02 Expected Perturbation Score for Adversarial Detection
 - Expected Perturbation Score (EPS)
 - Exploring EPS for Adversarial Detection
- 03 Experimental Results
- 04 **Conclusion**



Conclusion

- **A novel statistic EPS.** We find that the traditional score of one sample is sensitive in identifying adversarial samples due to insufficient information from a single sample only. To address this, we propose EPS by perturbing the sample with various noises
- **A novel adversarial detection method EPS-AD.** Based on EPS, we develop a novel single-sample adversarial detection method called EPS-AD relying on Maximum Mean Discrepancy
- **Theoretical justifications.** We theoretically analyze that the EPS of the natural sample is closer to those of other natural samples compared to adversarial samples under mild conditions and the EPS-based MMD between natural and adversarial samples is larger than that among natural samples

Code is available at <https://github.com/ZSHsh98/EPS-AD.git>





Thank you for your attention!
