Leveraging Distributional Discrepancies For Accuracy-robustness Trade-off

Jiacheng Zhang

School of Computing and Information Systems

The University of Melbourne

25 June 2025

Outline

- Background
- **ICML 2025:** Sample-specific Noise Injection for Diffusion-based Adversarial Purification
- **ICML 2025:** One Stone, Two Birds: Enhancing Adversarial Defense Through the Lens of Distributional Discrepancy

What is an adversarial example (attack)?

- **Left-or-right challenge:** Guess which one is the adversarial example?

What is an adversarial example (attack)?

99% Guacamole

88% Tabby Cat

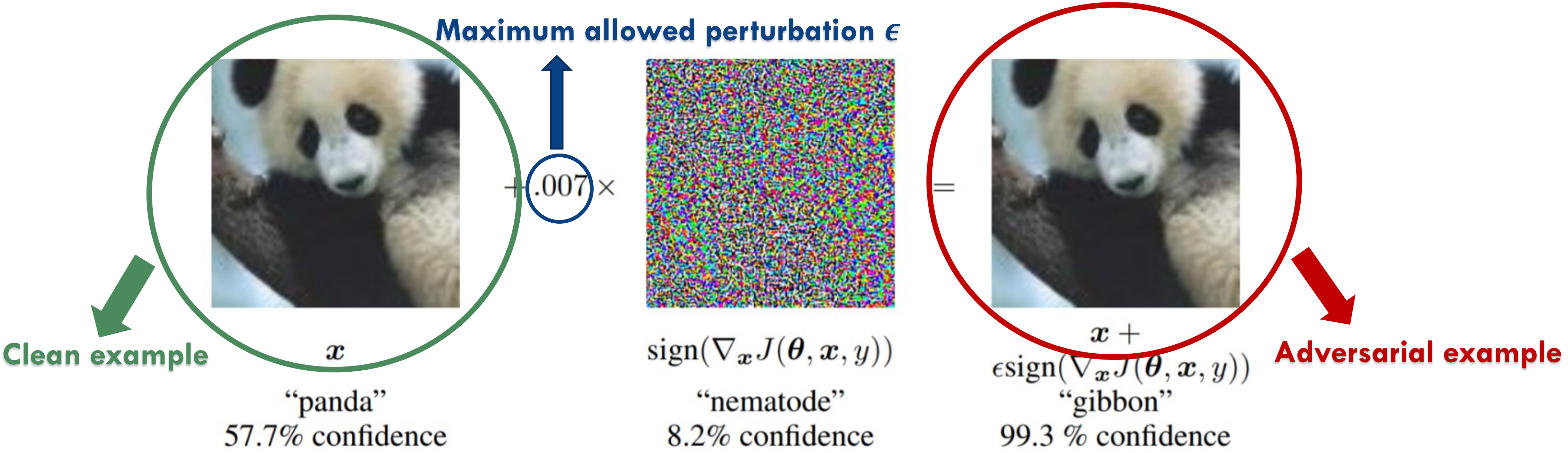
What is an adversarial example (attack)?

Adversarial examples can significantly drop the classification accuracy to **0%.**

<u>How</u> it works?

What is an adversarial example (attack)?

- Adding **imperceptible, non-random perturbations** to input data.

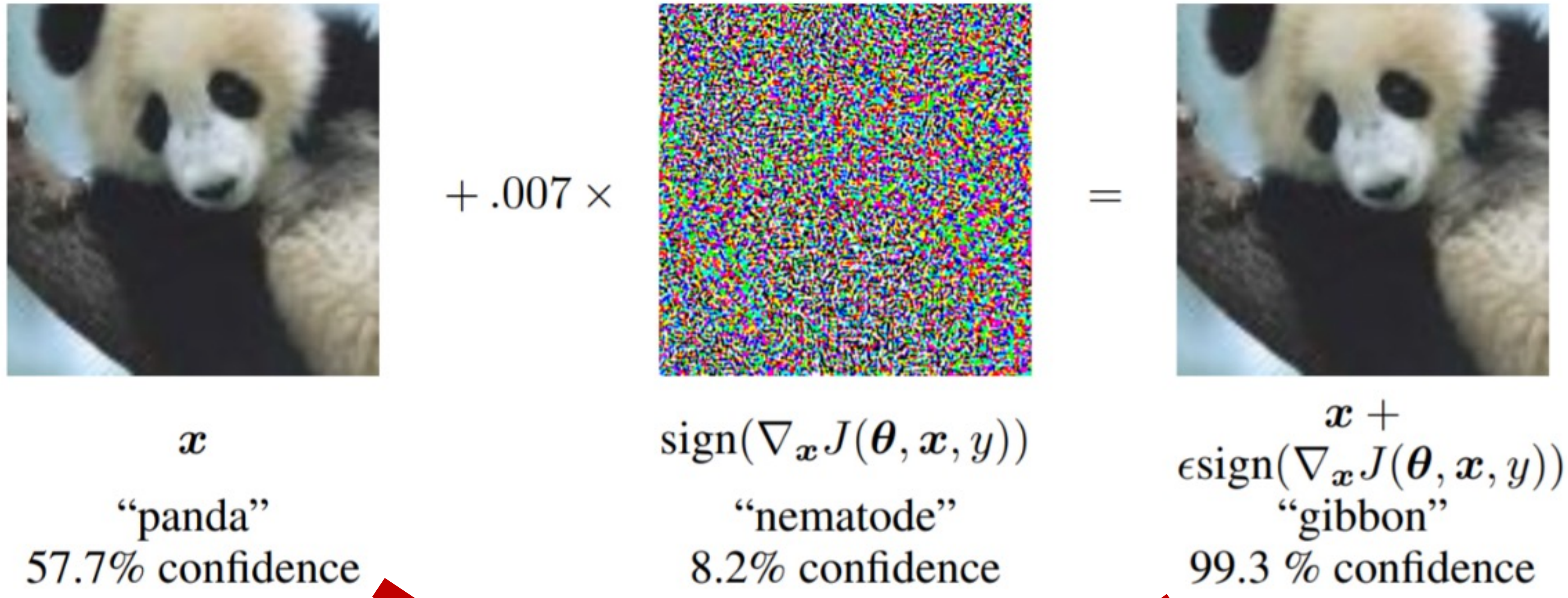


- Cannot fool human eyes but **can easily fool** state-of-the-art neural networks.

Explaining and Harnessing Adversarial Examples. In *ICLR*, 2015.

Why it works?

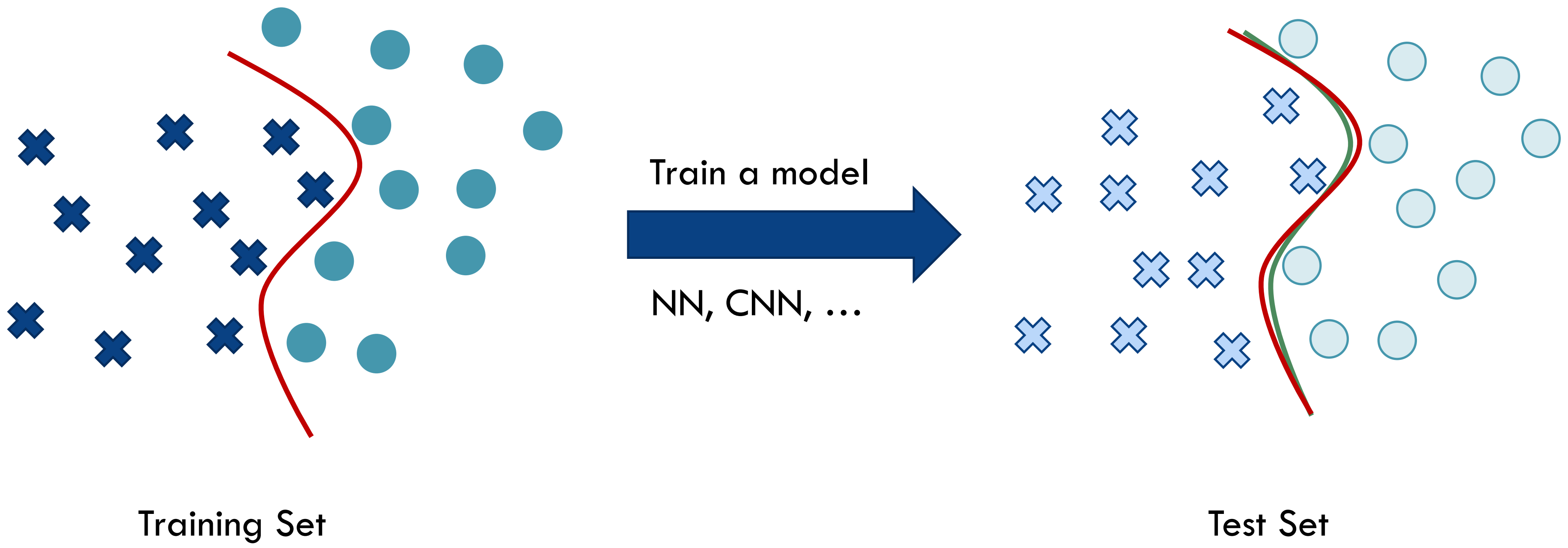
Why adversarial attack can be successful?



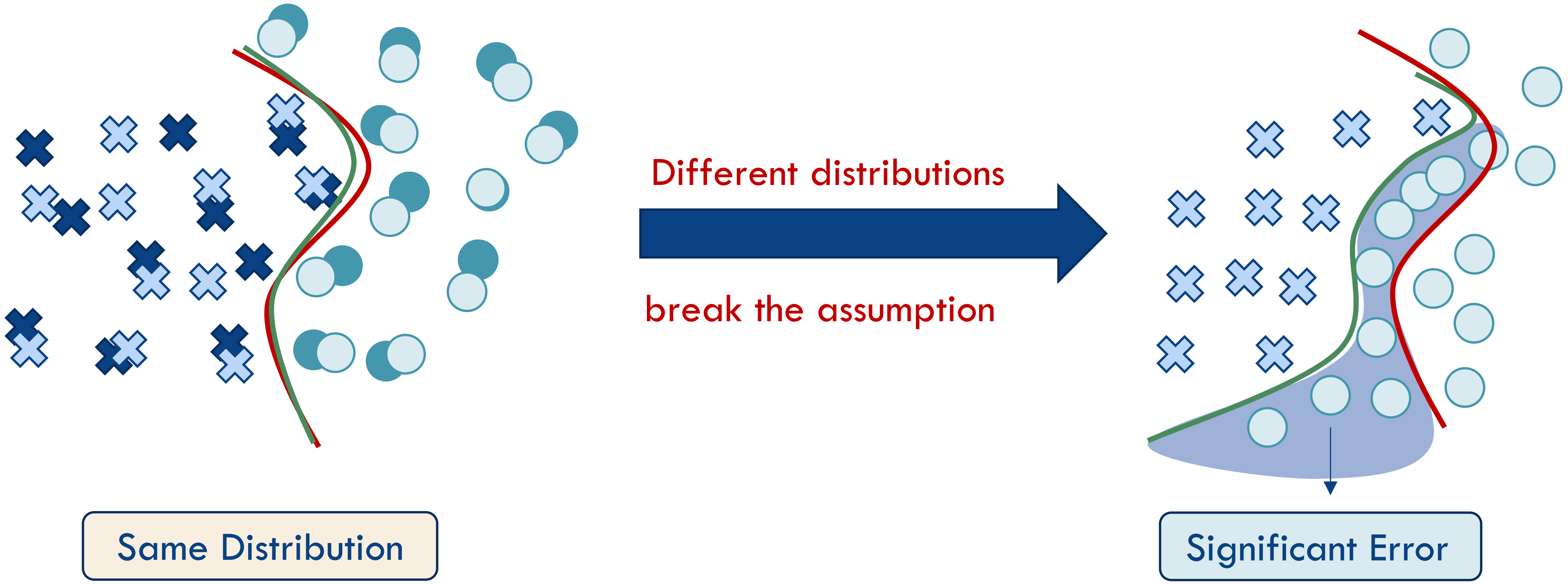
Different Distributions → Significant Error

Maximum Mean Discrepancy Test is Aware of Adversarial Attacks. In *ICML*, 2021.

Basic assumption in machine learning



Basic assumption in machine learning

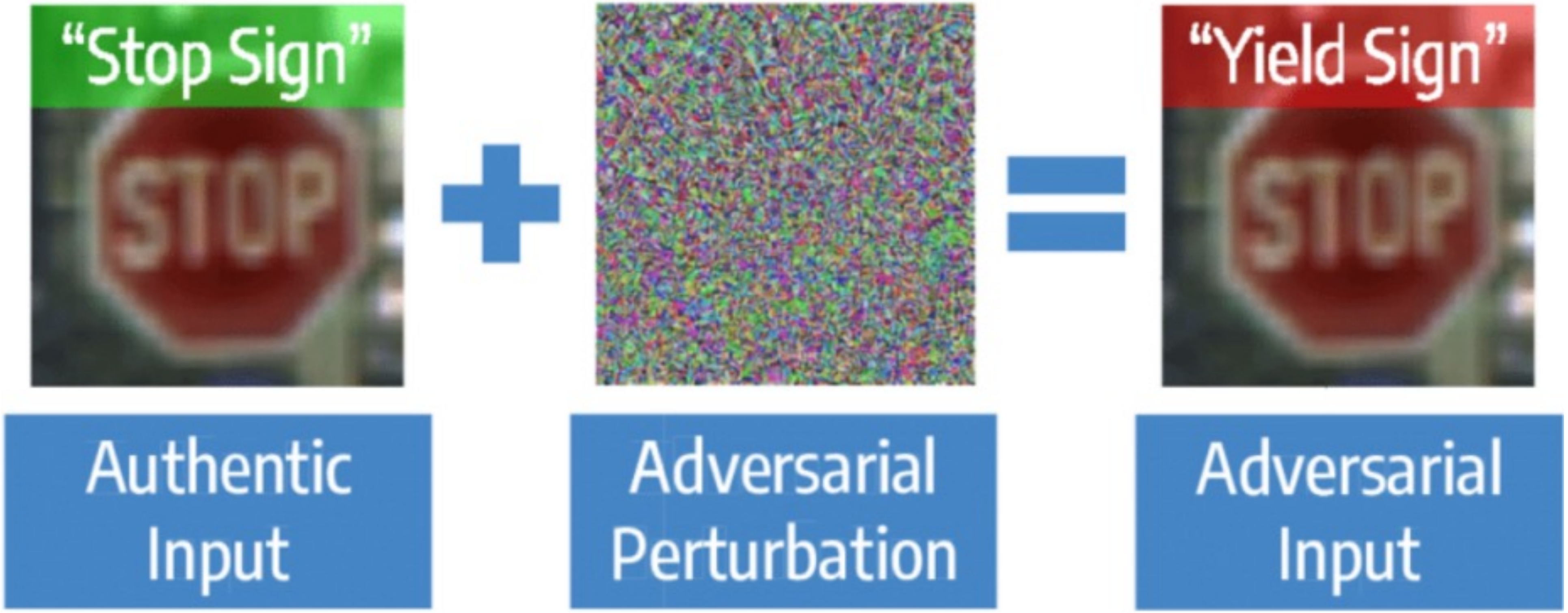


Basic assumption in machine learning

<u>Why</u> do we care?

Why do we care?

- Cause **security and reliability issues** in the deployment of machine learning systems.
- E.g., mislead the autonomous driving system to recognize **a stop sign** into **something else**.



Black-box adversarial attacks in autonomous vehicle technology. In 49th IEEE Applied Imagery Pattern Recognition Workshop, 2020.

Why do we care?

- ❑ Adding **adversarial examples** on T-shirts can bypass the AI detection system.
- ❑ Let you be invisible to the AI detection system!
- ❑ It's cool but it can cause **security and reliability issues.**

Adversarial T-shirt! Evading Person Detectors in A Physical World. ArXiv 2019.

How to defend against it?

Defend against adversarial attacks

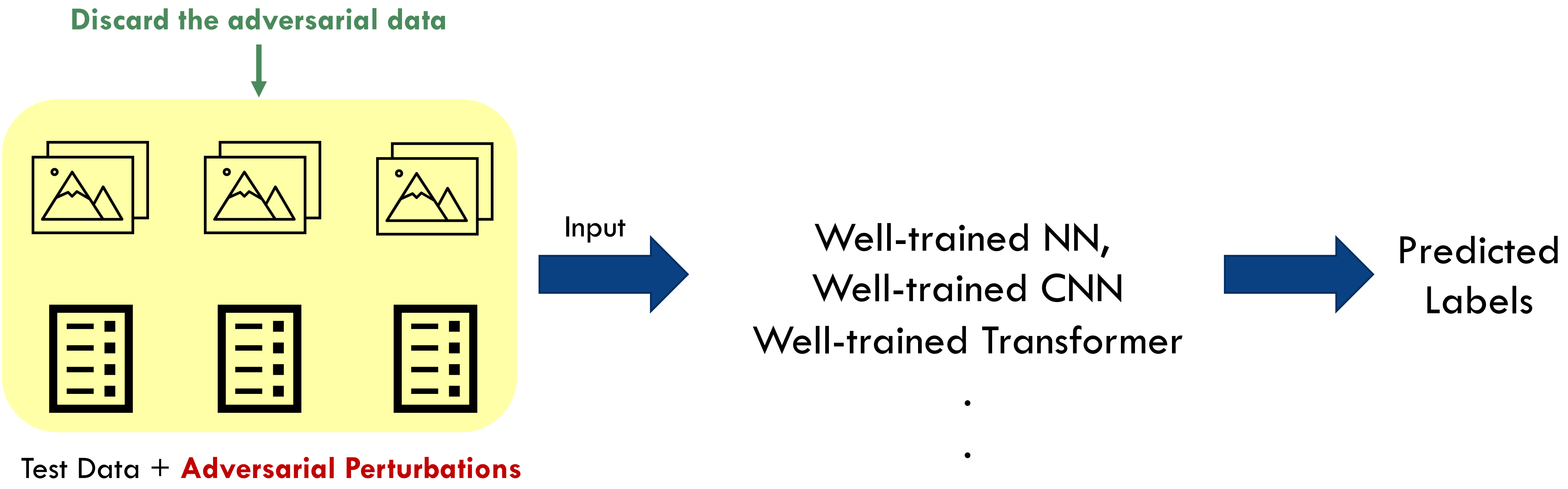
Trustworthy Machine Learning Under Adversarial Data

- Actively Handle Adversarial Data
  - Model Perspective — **Adversarial Training**
  - Data Perspective — **Adversarial Purification**
- Passively Handle Adversarial Data — **Adversarial Detection**

Today's Focus

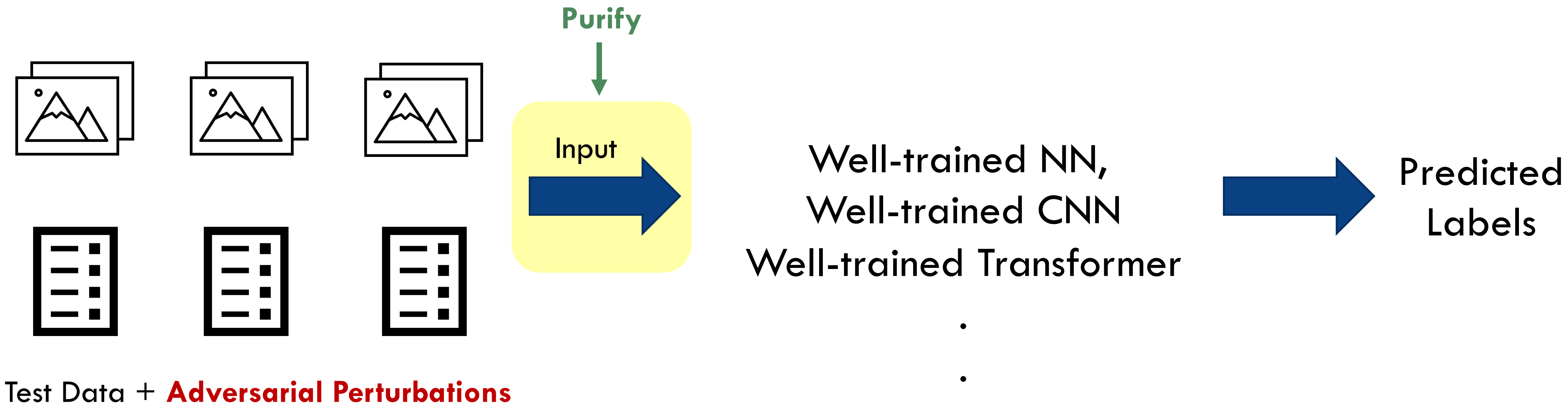
Adversarial detection

- *Adversarial Detection* (AD): aims to detect and discard AEs.



Adversarial purification

- ❑ *Adversarial Purification* (AP): aims to shift AEs back towards their natural counterparts.



Sample-specific Noise Injection for Diffusion-based Adversarial Purification

Yuhao Sun^, Jiacheng Zhang^, Zesheng Ye^, Chaowei Xiao, Feng Liu*

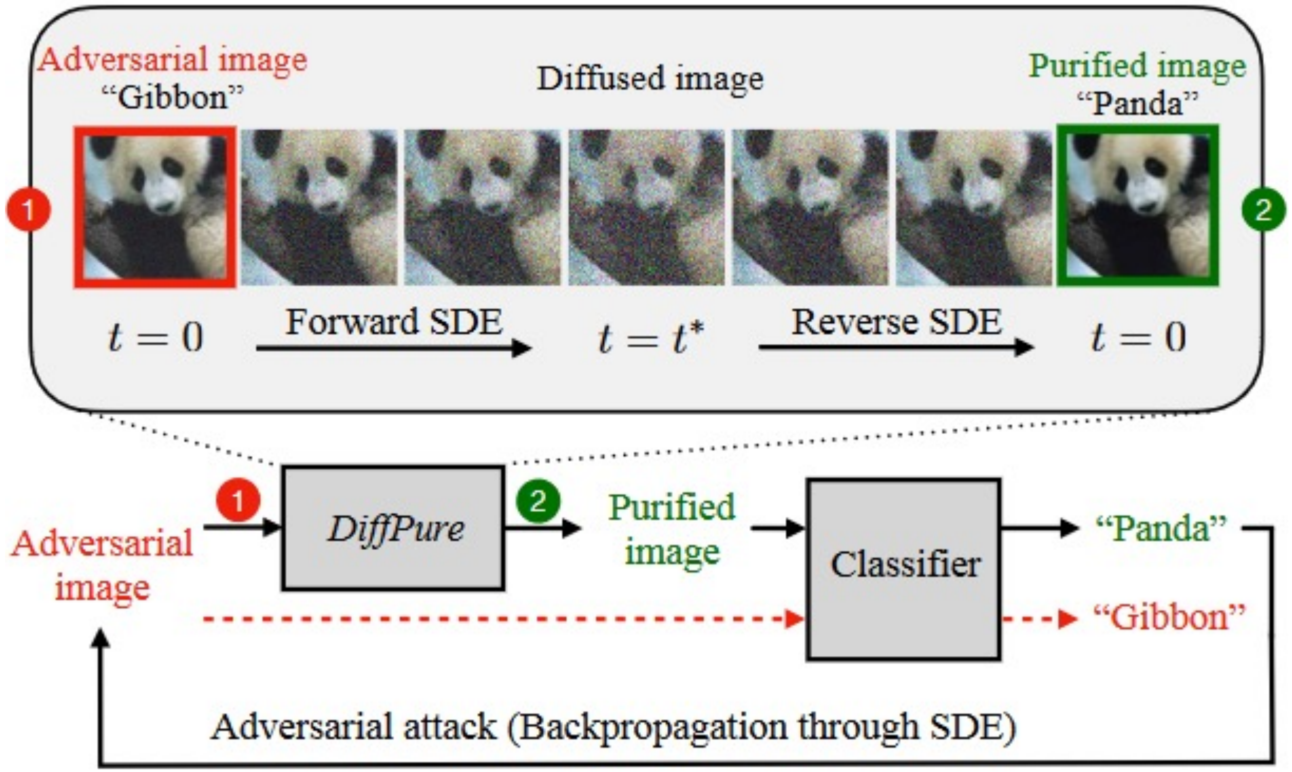
(^ Co-first authors, * Corresponding authors)

In *ICML*, 2025.

Paper

Code

Preliminary: diffusion-based adversarial purification

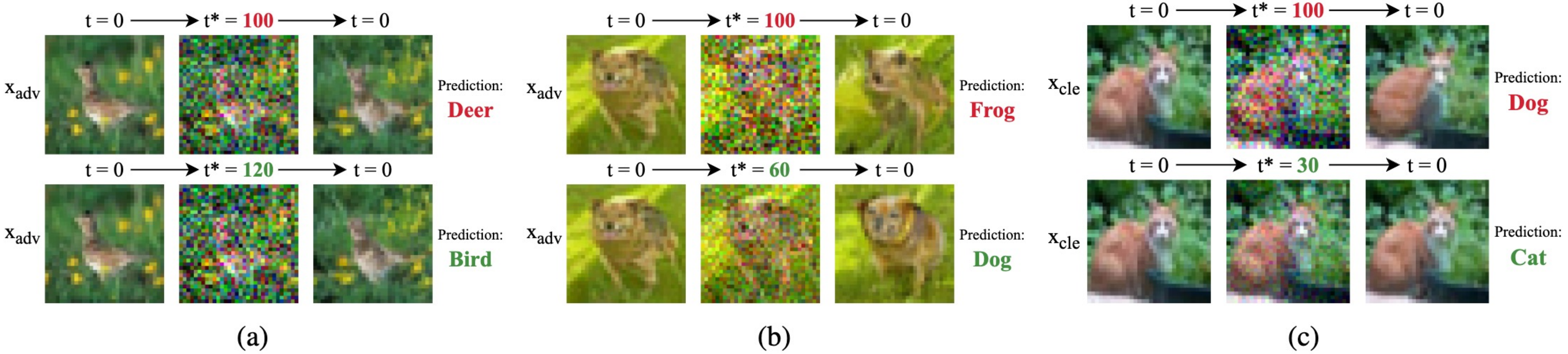


A Key Challenge: The Choice of t

- ❑ **If t is too small,** then adversarial noise cannot be fully removed.
- ❑ **If t is too large,** then the purified image may have a different semantic meaning.
- ❑ **Research gap:** current methods empirically select a *fixed* timestep t for all images, which is counterintuitive.

Diffusion Models for Adversarial Purification. In *ICML*, 2022.

Motivation

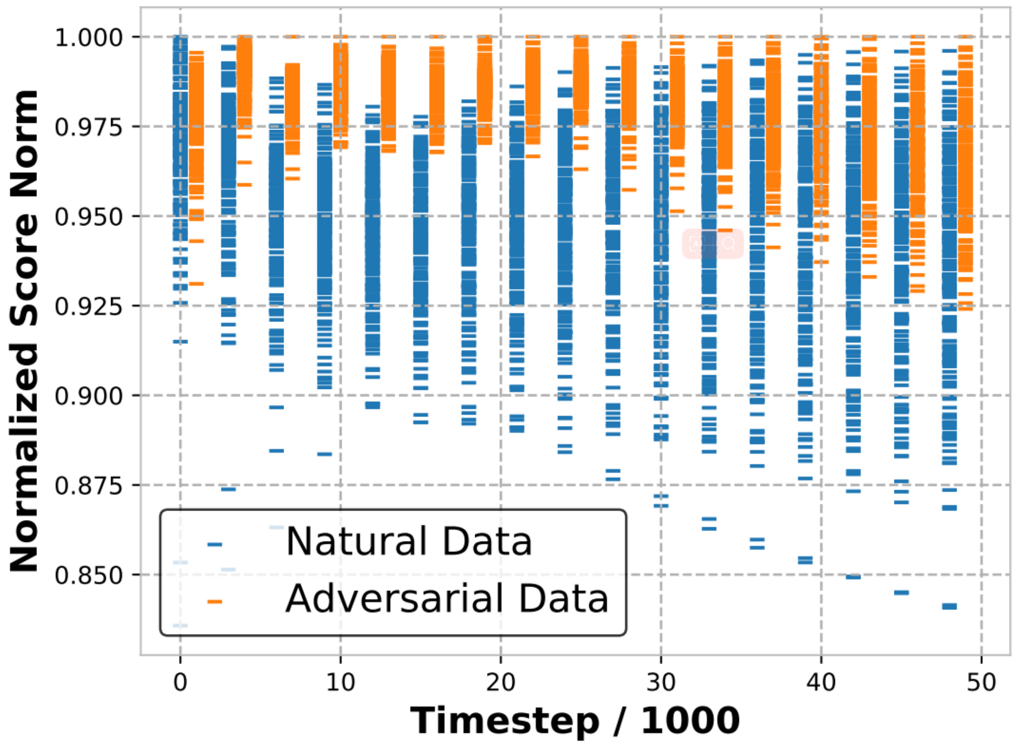


- ❑ Sample-shared noise level *fail* to address diverse adversarial perturbations.
- ❑ These findings *highlight* the need for sample-specific noise injection levels.

<u>What</u> is the metric?

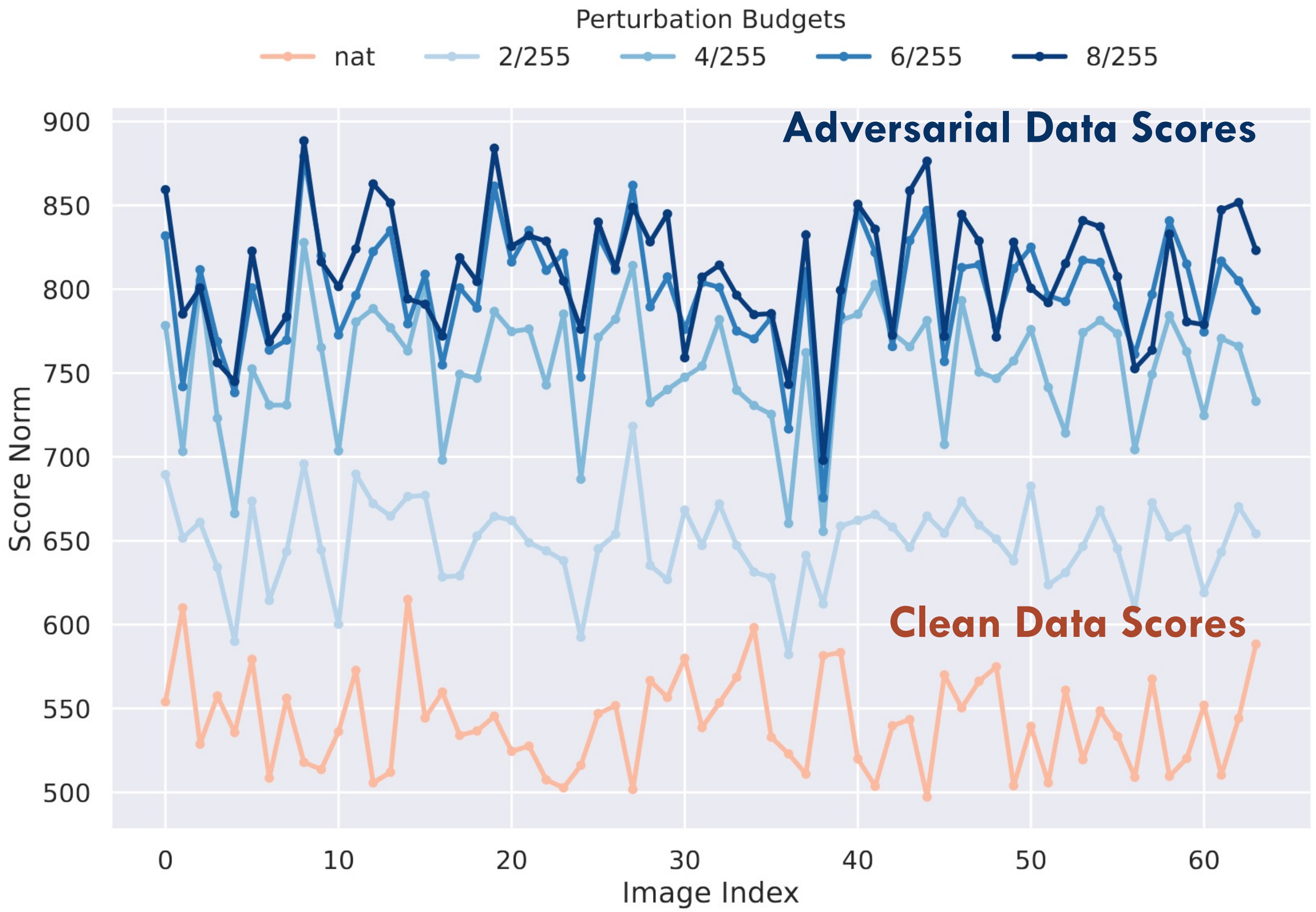
Intuition from score function

- 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 distribution (Song et al., 2019)
  - A lower score norm $||\nabla_{\mathbf{x}} \log p(\mathbf{x})||$ indicates the sample is closer to the high-density areas of natural data distribution



Detecting Adversarial Data by Probing Multiple Perturbations Using Expected Perturbation Score. In *ICML*, 2023.

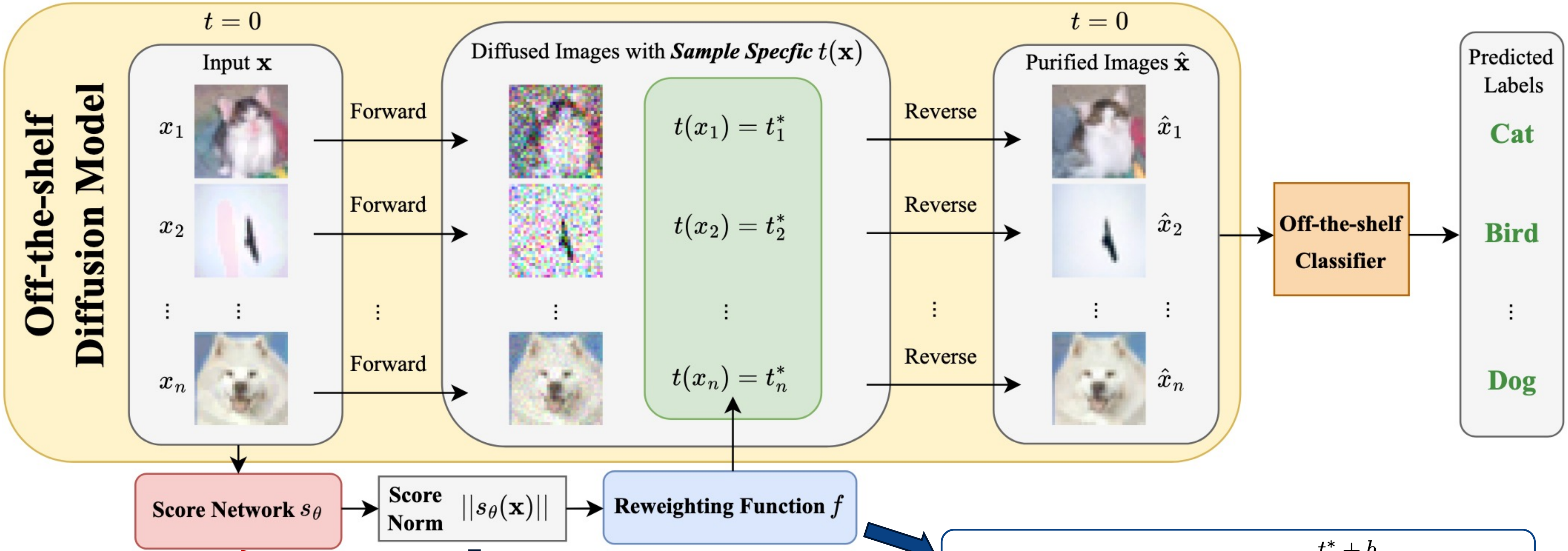
Score norms vs perturbation budgets



- We further find that score norms *scale directly* with perturbation budgets.
- Score norms can act as *proxies* for estimating the sample-specific noise level.

Sample-specific Noise Injection for Diffusion-based Adversarial Purification. In *ICML*, 2025.

Sample-specific Score-aware Noise Injection (SSNI)



estimate

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

$$f_\sigma(\|\mathrm{EPS}(\mathbf{x})\|\,,\ t^*) = \frac{t^* + b}{1 + \exp\{-(\|\mathrm{EPS}(\mathbf{x})\| - \mu)/\tau\}}$$

Sample-specific Noise Injection for Diffusion-based Adversarial Purification. In *ICML*, 2025.

Main results: CIFAR10

	PGD+EOT ℓ_∞ ($\epsilon = 8/255$)		
	DBP Method	Standard	Robust
WRN-28-10	Nie et al. (2022)	89.71±0.72	47.98±0.64
	+ *SSNI-N*	**93.29±0.37** (+3.58)	**48.63±0.56** (+0.65)
	Wang et al. (2022)	92.45±0.64	36.72±1.05
	+ *SSNI-N*	**94.08±0.33** (+1.63)	**40.95±0.65** (+4.23)
	Lee & Kim (2023)	90.10±0.18	56.05±1.11
	+ *SSNI-N*	**93.55±0.55** (+2.66)	**56.45±0.28** (+0.40)
WRN-70-16	Nie et al. (2022)	90.89±1.13	52.15±0.30
	+ *SSNI-N*	**94.47±0.51** (+3.58)	**52.47±0.66** (+0.32)
	Wang et al. (2022)	93.10±0.51	43.55±0.58
	+ *SSNI-N*	**95.57±0.24** (+2.47)	**46.03±1.33** (+2.48)
	Lee & Kim (2023)	89.39±1.12	56.97±0.33
	+ *SSNI-N*	**93.82±0.24** (+4.44)	**57.03±0.28** (+0.06)

	PGD+EOT ℓ_2 ($\epsilon = 0.5$)		
	DBP Method	Standard	Robust
WRN-28-10	Nie et al. (2022)	91.80±0.84	**82.81±0.97**
	+ *SSNI-N*	**93.95±0.70** (+2.15)	82.75±1.01 (-0.06)
	Wang et al. (2022)	92.45±0.64	82.29±0.82
	+ *SSNI-N*	**94.08±0.33** (+1.63)	**82.49±0.75** (+0.20)
	Lee & Kim (2023)	90.10±0.18	83.66±0.46
	+ *SSNI-N*	**93.55±0.55** (+3.45)	**84.05±0.33** (+0.39)
WRN-70-16	Nie et al. (2022)	92.90±0.40	82.94±1.13
	+ *SSNI-N*	**95.12±0.58** (+2.22)	**84.38±0.58** (+1.44)
	Wang et al. (2022)	93.10±0.51	**85.03±0.49**
	+ *SSNI-N*	**95.57±0.24** (+2.47)	84.64±0.51 (-0.39)
	Lee & Kim (2023)	89.39±1.12	84.51±0.37
	+ *SSNI-N*	**93.82±0.24** (+4.43)	**84.83±0.33** (+0.32)

Sample-specific Noise Injection for Diffusion-based Adversarial Purification. In *ICML*, 2025.

Main results: ImageNet-1K

	PGD+EOT ℓ_∞ ($\epsilon = 4/255$)		
	DBP Method	Standard	Robust
RN-50	Nie et al. (2022)	68.23±0.92	30.34±0.72
	+ *SSNI-N*	**70.25±0.56** (+2.02)	**33.66±1.04** (+3.32)
	Wang et al. (2022)	74.22±0.12	0.39±0.03
	+ *SSNI-N*	**75.07±0.18** (+0.85)	**5.21±0.24** (+4.82)
	Lee & Kim (2023)	70.18±0.60	42.45±0.92
	+ *SSNI-N*	**72.69±0.80** (+2.51)	**43.48±0.25** (+1.03)

Sample-specific Noise Injection for Diffusion-based Adversarial Purification. In *ICML*, 2025.

Limitations of DBP framework & SSNI

- **Limitation 1:** Having a pre-trained diffusion model is not always feasible, training a diffusion model is resource-consuming.
- **Limitation 2:** The inference speed of DBP-based methods is slow.
- **Limitation 3:** SSNI still injects noise to clean samples, which cannot fully preserve the utility (i.e., clean accuracy) of the model.

Can we do better?

Sample-specific Noise Injection for Diffusion-based Adversarial Purification. In *ICML*, 2025.

One Stone, Two Birds: Enhancing Adversarial Defense Through the Lens of Distributional Discrepancy

Jiacheng Zhang, Benjamin I. P. Rubinstein, Jingfeng Zhang, Feng Liu*

(* Corresponding authors)

In *ICML*, 2025.

Paper

Code

Distributional discrepancy minimization improves robustness

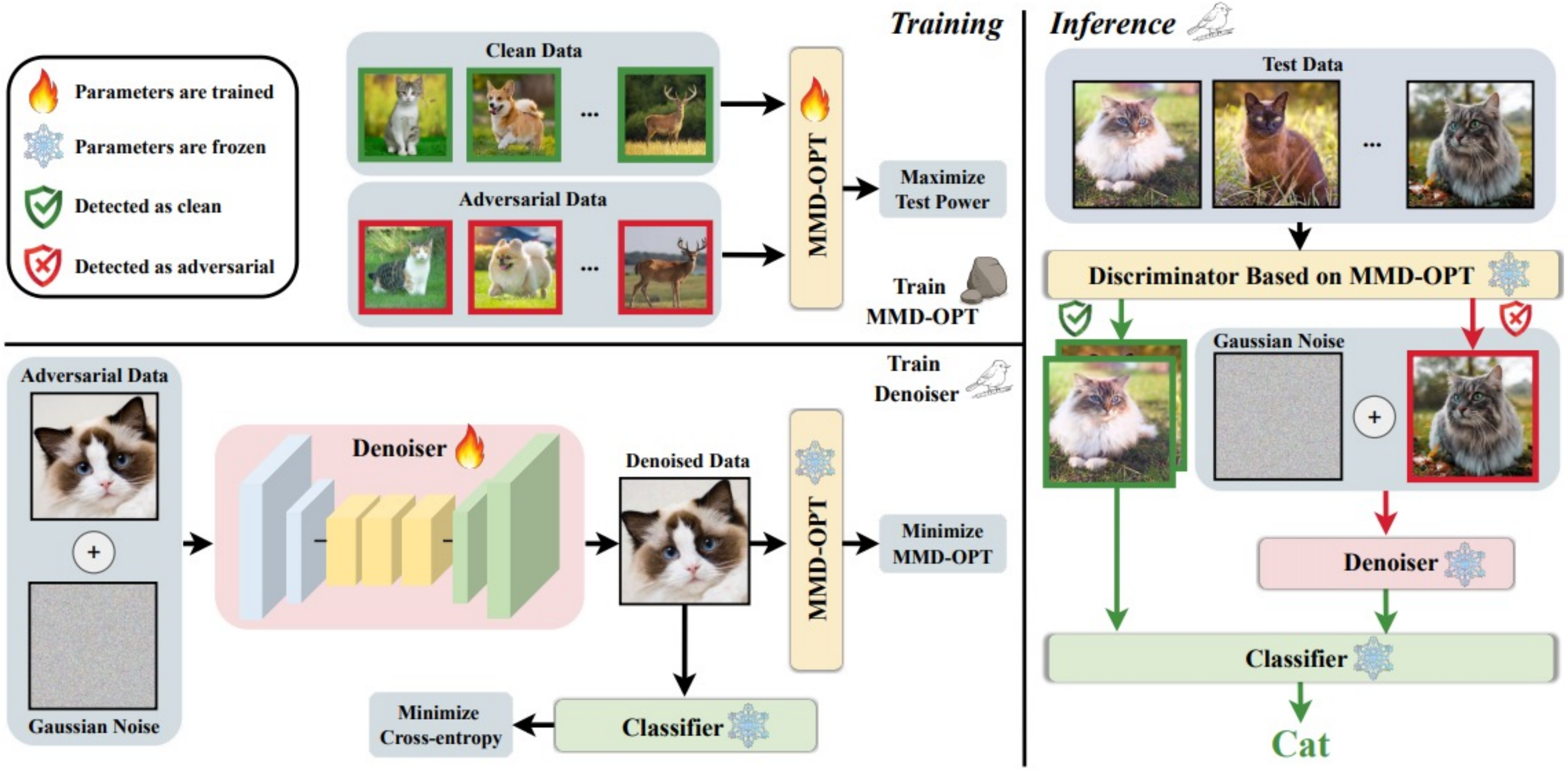
Theorem 1. *For a hypothesis $h \in \mathcal{H}$ and a distribution $\mathcal{D}_{\mathcal{A}} \in \mathbb{D}$:*

$$R(h, f_{\mathcal{A}}, \mathcal{D}_{\mathcal{A}}) \leq R(h, f_{\mathcal{C}}, \mathcal{D}_{\mathcal{C}}) + d_1(\mathcal{D}_{\mathcal{C}}, \mathcal{D}_{\mathcal{A}}).$$

risk on adversarial data | risk on clean data | distributional discrepancy

One Stone, Two Birds: Enhancing Adversarial Defense Through the Lens of Distributional Discrepancy. In *ICML*, 2025.

Distributional-discrepancy-based Adversarial Defense (DAD)



One Stone, Two Birds: Enhancing Adversarial Defense Through the Lens of Distributional Discrepancy. In *ICML*, 2025.

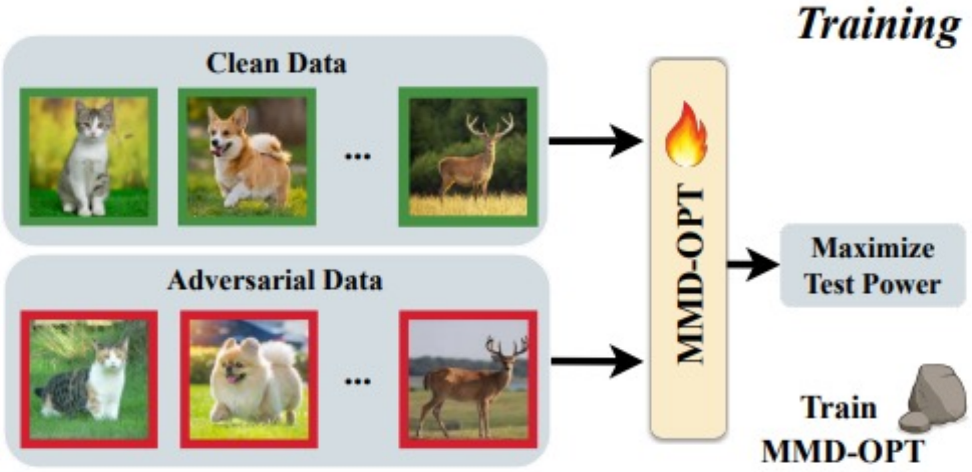
One stone: optimized MMD

$$\text{MMD-OPT}(S'_X, S'_Z) = \widehat{\text{MMD}}^2_u(S'_X, S'_Z; k^*_\omega)$$

MMD values

0 if same distribution

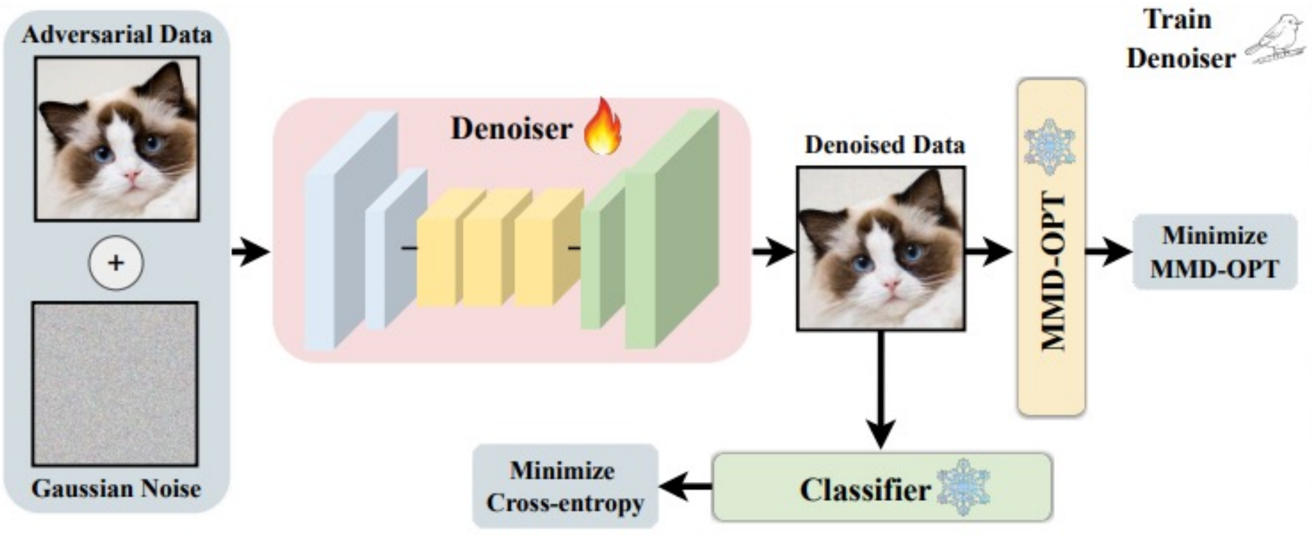
1 if different



Algorithm 1 Optimizing MMD (Liu et al., 2020).

Input: clean data $S_{\mathcal{C}}^{\text{train}}$, adversarial data $S_{\mathcal{A}}^{\text{train}}$, learning rate η, epoch T;
Initialize $\omega \leftarrow \omega_0$; $\lambda \leftarrow 10^{-8}$;
for epoch $= 1, ..., T$ **do**
  $S'_{\mathcal{C}} \leftarrow$ minibatch from $S_{\mathcal{C}}^{\text{train}}$;
  $S'_{\mathcal{A}} \leftarrow$ minibatch from $S_{\mathcal{A}}^{\text{train}}$;
  $k_\omega \leftarrow$ kernel function with parameters ω using Eq. (3);
  $M(\omega) \leftarrow \widehat{\text{MMD}}^2_u(S'_{\mathcal{C}}, S'_{\mathcal{A}}; k_\omega)$ using Eq. (2);
  $V_\lambda(\omega) \leftarrow \hat{\sigma}_\lambda(S'_{\mathcal{C}}, S'_{\mathcal{A}}; k_\omega)$ using Eq. (5);
  $\hat{J}_\lambda(\omega) \leftarrow M(\omega)/\sqrt{V_\lambda(\omega)}$ using Eq. (4);
  $\omega \leftarrow \omega + \eta \nabla_{\text{Adam}} \hat{J}_\lambda(\omega)$;
end for
Output: k^*_ω

First bird: MMD-OPT-based denoiser



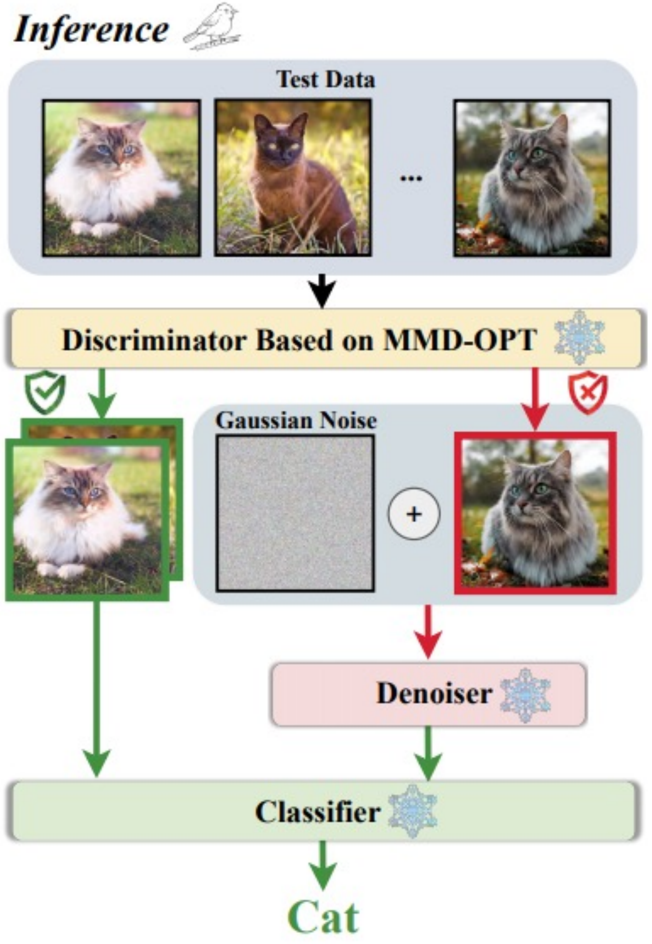
$\mathrm{MMD\text{-}OPT}(S_{\mathcal{C}}, g_{\boldsymbol{\theta}}(S_{\text{noise}}))$

$\mathcal{L}_{\text{ce}}(\widehat{h^*_{\mathcal{C}}}(g_{\boldsymbol{\theta}}(S_{\text{noise}})), Y_{\mathcal{C}})$

$$g_{\boldsymbol{\theta}^*} = \arg\min_{\boldsymbol{\theta}} \mathrm{MMD\text{-}OPT}(S_{\mathcal{C}}, g_{\boldsymbol{\theta}}(S_{\text{noise}})) + \alpha \cdot \mathcal{L}_{\text{ce}}(\widehat{h^*_{\mathcal{C}}}(g_{\boldsymbol{\theta}}(S_{\text{noise}})), Y_{\mathcal{C}})$$

One Stone, Two Birds: Enhancing Adversarial Defense Through the Lens of Distributional Discrepancy. In *ICML*, 2025.

Second bird: MMD-OPT-based discriminator



$$\mathrm{MMD\text{-}OPT}(S_{\mathcal{V}}, S_{\mathcal{T}}) = \widehat{\mathrm{MMD}}_{\mathrm{u}}^{2}(S_{\mathcal{V}}, S_{\mathcal{T}}; k_{\omega}^{*})$$

0 → **Classifier**

1 → **Denoiser** → **Classifier**

One Stone, Two Birds: Enhancing Adversarial Defense Through the Lens of Distributional Discrepancy. In *ICML*, 2025.

Main results: CIFAR-10

ℓ_∞ ($\epsilon = 8/255$)			
Type	Method	Clean	Robust
WRN-28-10			
AT	Gowal et al. (2021)	87.51	63.38
	Gowal et al. (2020)*	88.54	62.76
	Pang et al. (2022a)	88.62	61.04
AP	Yoon et al. (2021)	85.66	33.48
	Nie et al. (2022)	90.07	46.84
	Lee & Kim (2023)	90.16	55.82
Ours	DAD	**94.16 ± 0.08**	**67.53 ± 1.07**
WRN-70-16			
AT	Rebuffi et al. (2021)*	92.22	66.56
	Gowal et al. (2021)	88.75	66.10
	Gowal et al. (2020)*	91.10	65.87
AP	Yoon et al. (2021)	86.76	37.11
	Nie et al. (2022)	90.43	51.13
	Lee & Kim (2023)	90.53	56.88
Ours	DAD	**93.91 ± 0.11**	**67.68 ± 0.87**

ℓ_2 ($\epsilon = 0.5$)			
Type	Method	Clean	Robust
WRN-28-10			
AT	Rebuffi et al. (2021)*	91.79	78.80
	Augustin et al. (2020)†	93.96	78.79
	Sehwag et al. (2022)†	90.93	77.24
AP	Yoon et al. (2021)	85.66	73.32
	Nie et al. (2022)	91.41	79.45
	Lee & Kim (2023)	90.16	83.59
Ours	DAD	**94.16 ± 0.08**	**84.38 ± 0.81**
WRN-70-16			
AT	Rebuffi et al. (2021)*	**95.74**	82.32
	Gowal et al. (2020)*	94.74	80.53
	Rebuffi et al. (2021)	92.41	80.42
AP	Yoon et al. (2021)	86.76	75.66
	Nie et al. (2022)	92.15	82.97
	Lee & Kim (2023)	90.53	83.57
Ours	DAD	93.91 ± 0.11	**84.03 ± 0.75**

One Stone, Two Birds: Enhancing Adversarial Defense Through the Lens of Distributional Discrepancy. In *ICML*, 2025.

Main results: ImageNet-1K

ℓ_∞ ($\epsilon = 4/255$)			
Type	Method	Clean	Robust
RN-50			
AT	Salman et al. (2020a)	64.02	34.96
	Engstrom et al. (2019)	62.56	29.22
	Wong et al. (2020)	55.62	26.24
AP	Nie et al. (2022)	71.48	38.71
	Lee & Kim (2023)	70.74	42.15
Ours	DAD	**78.61 ± 0.04**	**53.85 ± 0.23**

Strength of DAD

- **Strength 1:** DAD can largely preserve the original utility (i.e., clean accuracy of the classifier).

- **Strength 2:** Compared to DBP methods that reply on density estimation, learning distributional discrepancies is a simpler and more feasible task.

- **Strength 3:** DAD is efficient in both training and inferencing.

Thank You!