ADBM: Adversarial diffusion bridge model for reliable adversarial purification

Xiao Li, Wenxuan Sun, Huanran Chen, Qiongxiu Li,
Yining Liu, Yingzhe He, Jie Shi, Xiaolin Hu

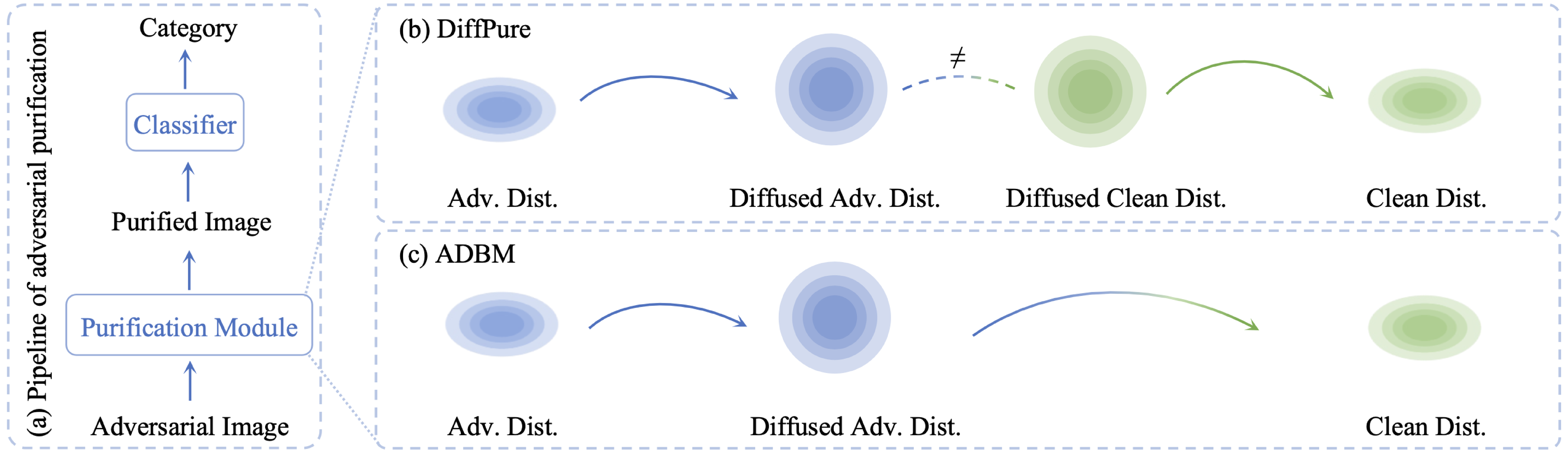
Paper:

Summary

- ***Problem***: Diffusion-Based Purification (DBP) methods like DiffPure have shown promise in defending against adversarial examples. However, they often suffer from a trade-off between effective noise removal and accurate data recovery. Additionally, evaluations of such methods have been questioned due to reliance on weak adaptive attacks.
- ***Contribution***: We first build a **reliable** evaluation for the robustness of DBP. Then we introduce the Adversarial Diffusion Bridge Model (ADBM), which constructs a **direct reverse bridge** from diffused adversarial example back to clean examples.
- ***Results***: Through theoretical analysis and extensive experiments, ADBM demonstrates superior defense performance across various scenarios, highlighting its reliability and potential for practical applications.

DiffPure v.s. ADBM

- DiffPure relies on the assumption that the two diffused distributions are sufficiently close, such that the original reverse process can recover the diffused adversarial data distribution.
- Unlike original diffusion models relying on the similarity between the diffused distributions of clean and adversarial examples for a balanced trade-off, ADBM constructs a direct reverse process (or "bridge") from the diffused adversarial data distribution to the distribution of clean examples



Theoretical Analysis

Theorem 1. *Given an adversarial example $\mathbf{x}_0^a$ and assuming the training loss $L_b \le \delta$, the distance between the purified example of ADBM and the clean example $\mathbf{x}_0$, denoted as $\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|$, is bounded by δ (constant omitted) in expectation when using a one-step DDIM sampler. Specifically, we have $\mathbb{E}_{\epsilon}\left[\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|^2\right] \le \frac{(1-\bar{\alpha}_T)T}{\bar{\alpha}_T}\delta$, where $\frac{(1-\bar{\alpha}_T)T}{\bar{\alpha}_T}$ is the constant.*

- Theorem 1 implies that if the training loss of ADBM converges to zero, it can **perfectly** remove adversarial noises by employing a one-step DDIM sampler.

Theorem 2. *Denote the probability of reversing the adversarial example to the clean example using ADBM and DiffPure as $P(B)$ and $P(D)$, respectively. Then $P(\cdot) = \int \mathbb{1}_{\{\mathbf{x}_0 \notin \mathbb{D}_a\}} p(\mathbf{x}_0|\hat{\mathbf{x}}_t)\mathrm{d}\mathbf{x}_0$, where $\mathbb{D}_a$ denotes the set of adversarial examples. If the timestep is infinite, the following inequality holds:*

$$P(B) > P(D),$$

wherein

$$\text{for } P(B) : p(\mathbf{x}_0|\hat{\mathbf{x}}_t) \propto \exp\left(-\frac{\|\mathbf{x}_t^d - \sqrt{\bar{\alpha}_t}\mathbf{x}_0^a\|^2}{2(1-\bar{\alpha}_t)}\right), \quad (10)$$

$$\text{for } P(D) : p(\mathbf{x}_0|\hat{\mathbf{x}}_t) \propto \exp\left(-\frac{\|\mathbf{x}_t^a - \sqrt{\bar{\alpha}_t}\mathbf{x}_0\|^2}{2(1-\bar{\alpha}_t)}\right). \quad (11)$$

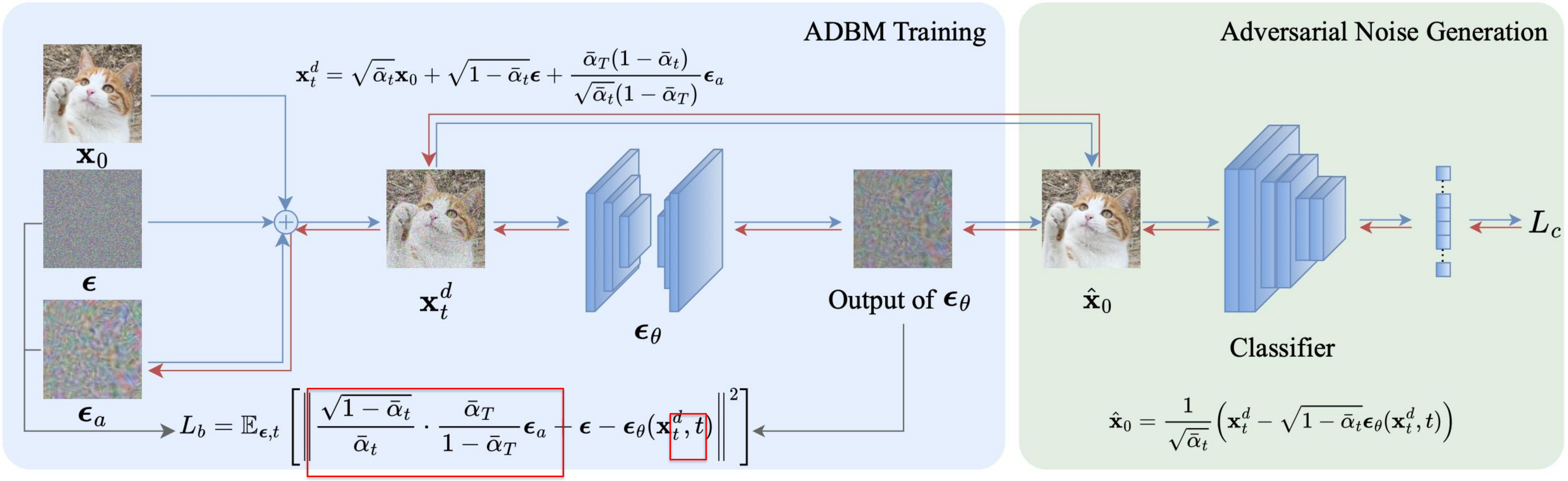
- Theorem 2 indicates that with infinite reverse timesteps, adversarial examples purified with ADBM **are more likely** to align with the clean data distribution than those with DiffPure.

Reliable Evaluation on DBP

Evaluation	Clean Acc	Robust Acc
BPDA (Athalye et al., 2018)	90.49 ± 0.97	81.40 ± 0.16
Nie et al. (2022)	90.07 ± 0.97	71.29 ± 0.55
Chen et al. (2023a)	90.97	53.52
Lee & Kim (2023)	90.43 ± 0.60	51.13 ± 0.87
Ours (EOT=20, steps=200)		45.83 ± 1.27
Ours (EOT=40, steps=200)	90.49 ± 0.97	**45.64** ± 1.14
Ours (EOT=20, steps=400)		46.16 ± 1.33

- We first develop a simple yet reliable adaptive attack evaluation method for DBP, achieving the SOTA attack success rate for DiffPure.
- ***Evaluation:*** PGD-200+EOT-20 w/ full gradient

ADBM Framework



where $\mathbf{x}_t^d = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} + \frac{\bar{\alpha}_T(1-\bar{\alpha}_t)}{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_T)}\boldsymbol{\epsilon}_a$. ϵ_a denotes the adversarial noise can can be generated by $\frac{\partial L_c}{\partial \epsilon_a}$.

ADBM injects scaled adversarial noise into the training compared with the original training objective of diffusion models

Experimental Results

Method	Type	Clean Acc	Robust Acc			
			l_∞ norm	l_1 norm	l_2 norm	Average
Vanilla	-	97.02	0.00	0.00	0.00	0.00
[41]	AT	91.10	65.92	8.26	27.56	33.91
[42]	AT	88.54	64.26	12.06	32.29	36.20
Augment w/ Diff [24]	AT	88.74	66.18	9.76	28.73	34.89
Augment w/ Diff [25]	AT	93.25	**70.72**	8.48	28.98	36.06
[13]	AP	91.89	4.56	8.68	7.25	6.83
[14]	AP	87.93	37.65	36.87	57.81	44.11
DiffPure+Guide [26]	AP	93.16	22.07	28.71	35.74	28.84
Diff+ScoreOpt [28]	AP	91.41	13.28	10.94	28.91	17.71
DiffPure+Langevin [27]	AP	92.18	43.75	39.84	55.47	46.35
DiffPure [16]	AP	92.5 ± 0.5	42.2 ± 2.1	44.3 ± 1.3	60.8 ± 2.3	49.1 ± 1.7
ADBM (Ours)	AP	91.9 ± 0.8	47.7 ± 2.2	**49.6** ± 2.2	**63.3** ± 1.9	**53.5** ± 2.1

- ADBM achieved better robustness than DiffPure under reliable attacks under 1 reverse step.
- ADBM exhibited much better robustness than AT methods when facing unseen threats.