NEURAL INFORMATION PROCESSING SYSTEMS

DiffBreak

Is Diffusion-Based Purification Robust?

Andre Kassis, Urs Hengartner & Yaoliang Yu

CYBER SECURITY AND PRIVACY INSTITUTE
UNIVERSITY OF WATERLOO

UNIVERSITY OF WATERLOO

★ Website: https://github.com/andrekassis/DiffBreak

NEURAL INFORMATION PROCESSING SYSTEMS

DiffBreak

Is Diffusion-Based Purification Robust? NO!!

Andre Kassis, Urs Hengartner & Yaoliang Yu

CYBER SECURITY AND PRIVACY INSTITUTE
UNIVERSITY OF WATERLOO

UNIVERSITY OF WATERLOO

Website: https://github.com/andrekassis/DiffBreak

Diffusion-Based Purification (DBP)

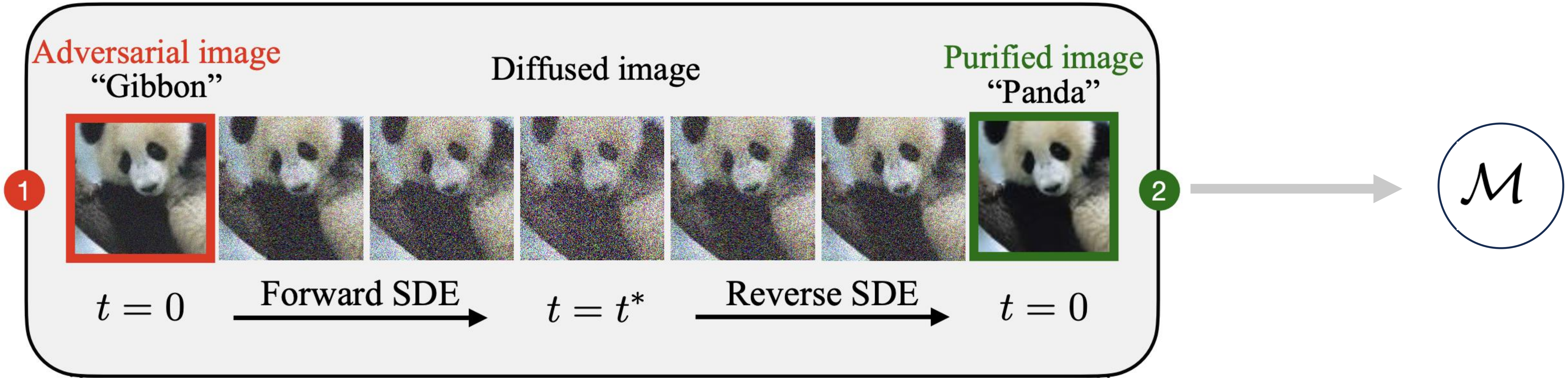
A defense against adversarial examples (AEs) that purifies AEs via diffusion models.

The only defense that remains robust to date.

Diffusion-Based Purification (DBP)

A defense against adversarial examples (AEs) that purifies AEs via diffusion models.

The only defense that remains robust to date.



A principled evaluation framework that challenges *DBP*'s robustness.

1) Theoretical scrutiny: exposes inherent adversarial vulnerability.

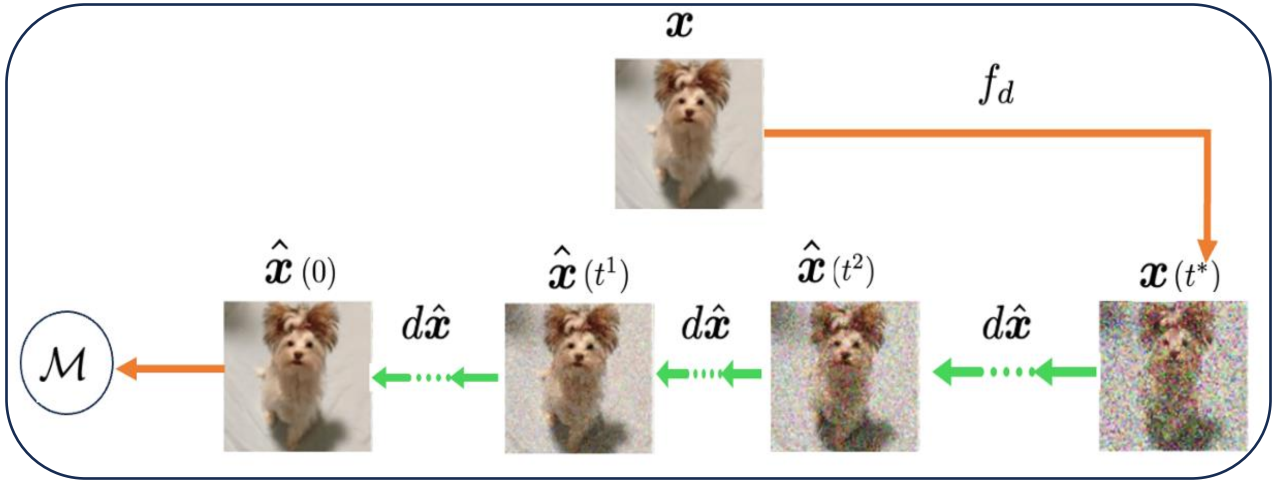
2) Majority-vote protocol: statistically-sound evaluation setup.

3) Reliable gradient module: fixes backprop issues, significantly degrading robustness.

4) Low-frequency attacks: structured *AE*s that completely break *DBP*.

1) Theoretical Scrutiny Reveals Fundamental Vulnerability

Diffusion models learn to reverse a process that gradually turns real data in $\boldsymbol{p} \subseteq \mathbb{R}^d$ into random noise.



DBP

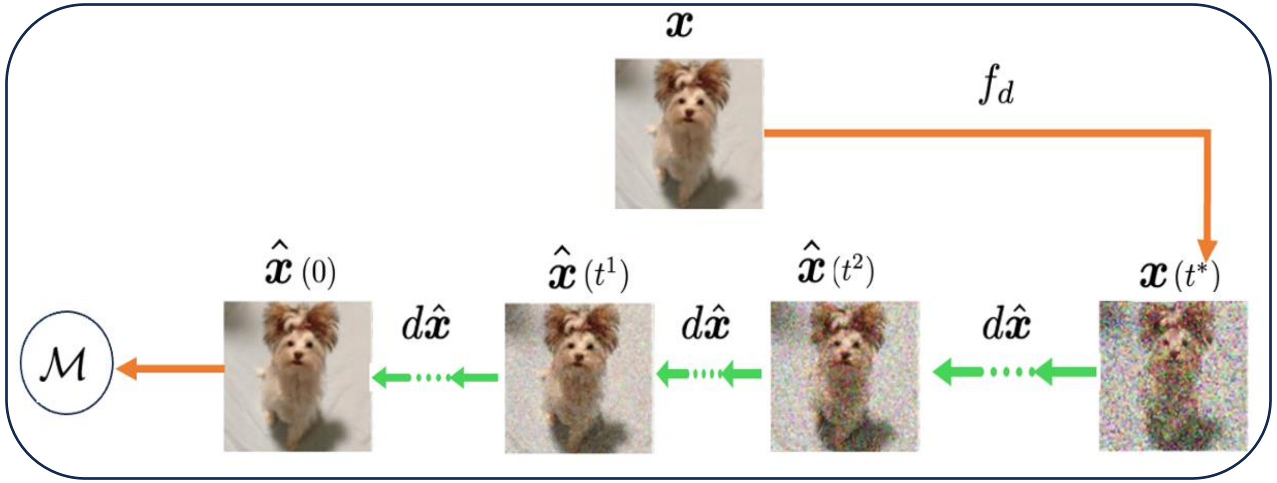
Guarantee: $\hat{x}(0) \sim p$

→ Off-manifold *AE*s are highly unlikely.

$$\Pr(\hat{x}(0)|x) \propto p(\hat{x}(0)) \cdot e^{-\frac{a(t^*)||\hat{x}(0)-x||_2^2}{2(1-a(t^*))}}$$

1) *Theoretical Scrutiny Reveals Fundamental Vulnerability*

Diffusion models learn to reverse a process that gradually turns real data in $\boldsymbol{p} \subseteq \mathbb{R}^d$ into random noise.



DBP

Guarantee: $\hat{x}(0) \sim p$

→ Off-manifold *AE*s are highly unlikely.

$$\Pr(\hat{x}(0)|x) \propto p\big(\hat{x}(0)\big) \cdot e^{-\frac{a(t^*)||\hat{x}(0)-x||_2^2}{2(1-a(t^*))}}$$

Overlooked caveat: Diffusion models require a pretrained neural network S_θ

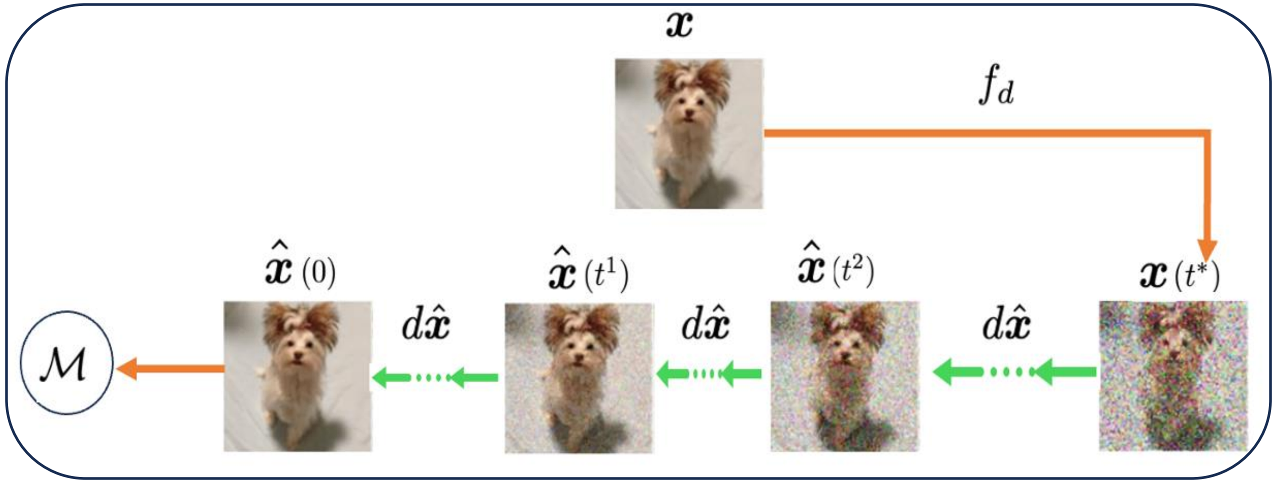
→ S_θ is *not* an oracle — it's an exploitable **ML algorithm**

Skilled adversary:

$$\max_{\{\theta_x^t\}_{t \le t^*}} \mathbb{E}_{\hat{x}(0) \sim DBP^{\{\theta_x^t\}}} \Pr(\neg y|\hat{\boldsymbol{x}}(\boldsymbol{0}))$$

1) Theoretical Scrutiny Reveals Fundamental Vulnerability

Diffusion models learn to reverse a process that gradually turns real data in $p \subseteq \mathbb{R}^d$ into random noise.



DBP

Guarantee: $\hat{x}(0) \sim p$

→ **Off-manifold *AE*s are highly unlikely.**

$$\Pr(\hat{x}(0)|x) \propto p(\hat{x}(0)) \cdot e^{-\frac{a(t^*)||\hat{x}(0)-x||_2^2}{2(1-a(t^*))}}$$

Overlooked caveat: Diffusion models require a pretrained neural network S_θ

→ S_θ is *not* an oracle — it's an exploitable **ML algorithm**

Skilled adversary:

$$\max_{\{\theta_x^t\}_{t\le t^*}} \mathbb{E}_{\hat{x}(0)\sim DBP^{\{\theta_x^t\}}} \Pr(\neg y|\hat{\boldsymbol{x}}(\mathbf{0}))$$

Problem: This requires the probability gradients of the different purification trajectories.

HARD TO COMPUTE DIRECTLY!

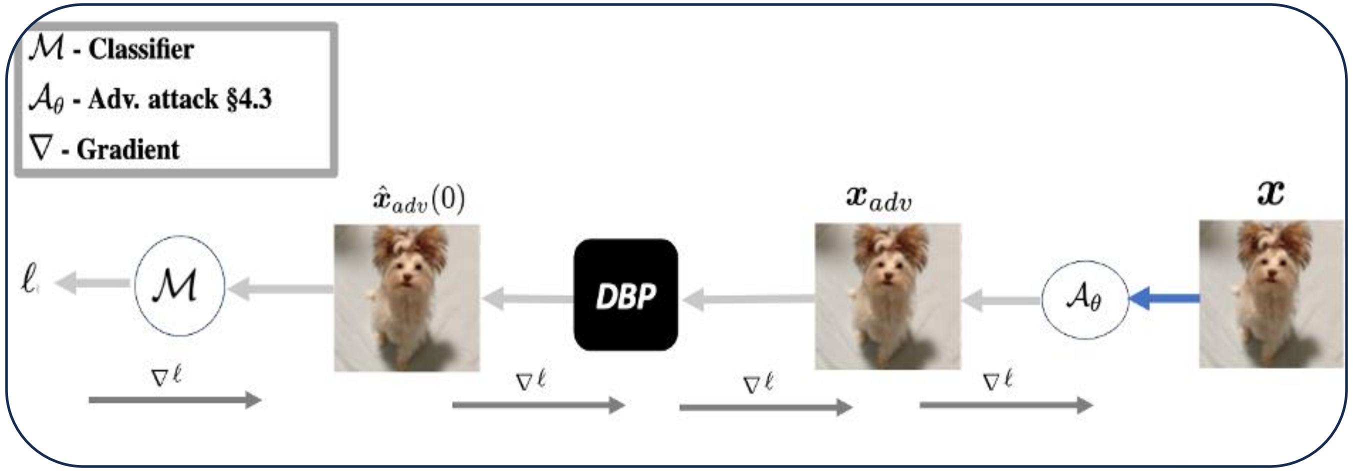
In practice:

Attackers target the classifier and propagate gradients through *DBP*.

Skilled adversary:

$$\max_{\{\theta_x^t\}_{t\le t^*}} \mathbb{E}_{\hat{x}(0)\sim DBP^{\{\theta_x^t\}}} \Pr(\neg y|\hat{\boldsymbol{x}}(\mathbf{0}))$$

Standard Adaptive Attack



Our key insight: The two attacks are equivalent! → *DBP's* robustness claims become invalid.

Exposing DBP's Shortcomings

Our theoretical analysis proves *DBP*'s vulnerability to gradient-based attacks

→Why did previous work fail to undermine DBP's robustness?

Exposing DBP's Shortcomings

Our theoretical analysis proves *DBP*'s vulnerability to gradient-based attacks

→Why did previous work fail to undermine DBP's robustness?

Our findings

1. Previous works mainly evaluated single random purification for the AE upon attack termination

→ Ignores stochasticity and resubmission risk and overstates robustness

→ Worst-case evaluations over N purified copies understate robustness (variance fragile)

Our theoretical analysis proves *DBP*'s vulnerability to gradient-based attacks

→Why did previous work fail to undermine DBP's robustness?

Our findings

1. Previous works mainly evaluated single random purification for the AE upon attack termination

→ Ignores stochasticity and resubmission risk and overstates robustness

→ Worst-case evaluations over N purified copies understate robustness (variance fragile)

2) Majority-vote protocol ✓

Exposing DBP's Shortcomings

Our theoretical analysis proves *DBP*'s vulnerability to gradient-based attacks

→Why did previous work fail to undermine DBP's robustness?

Our findings

1. Previous works mainly evaluated single random purification for the AE upon attack termination

→ Ignores stochasticity and resubmission risk and overstates robustness

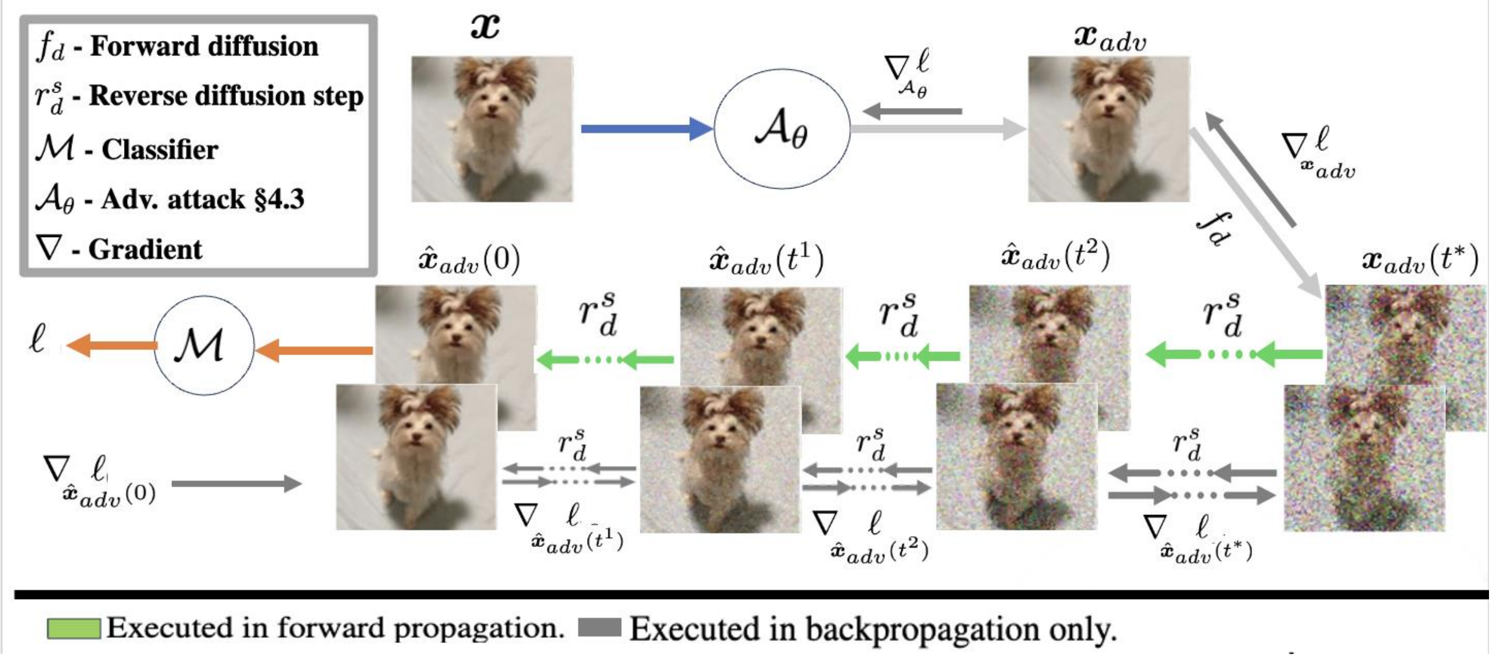
→ Worst-case evaluations over N purified copies understate robustness (variance fragile)

2) Majority-vote protocol ✓

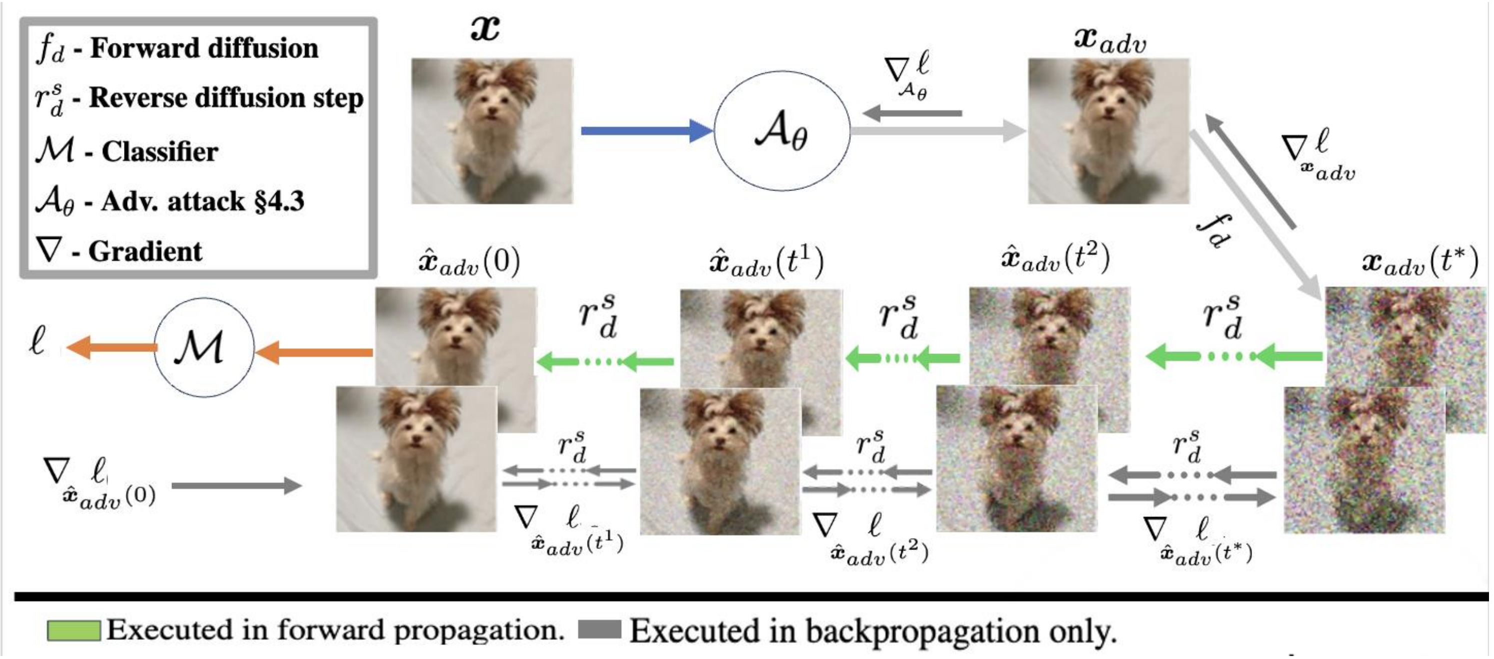
2. DBP's memory intensive nature requires gradient checkpointing for backpropagation

→ Prior checkpointing implementations contained subtle issues

3) DiffGrad: A Reliable Gradient module



3) DiffGrad: A Reliable Gradient module

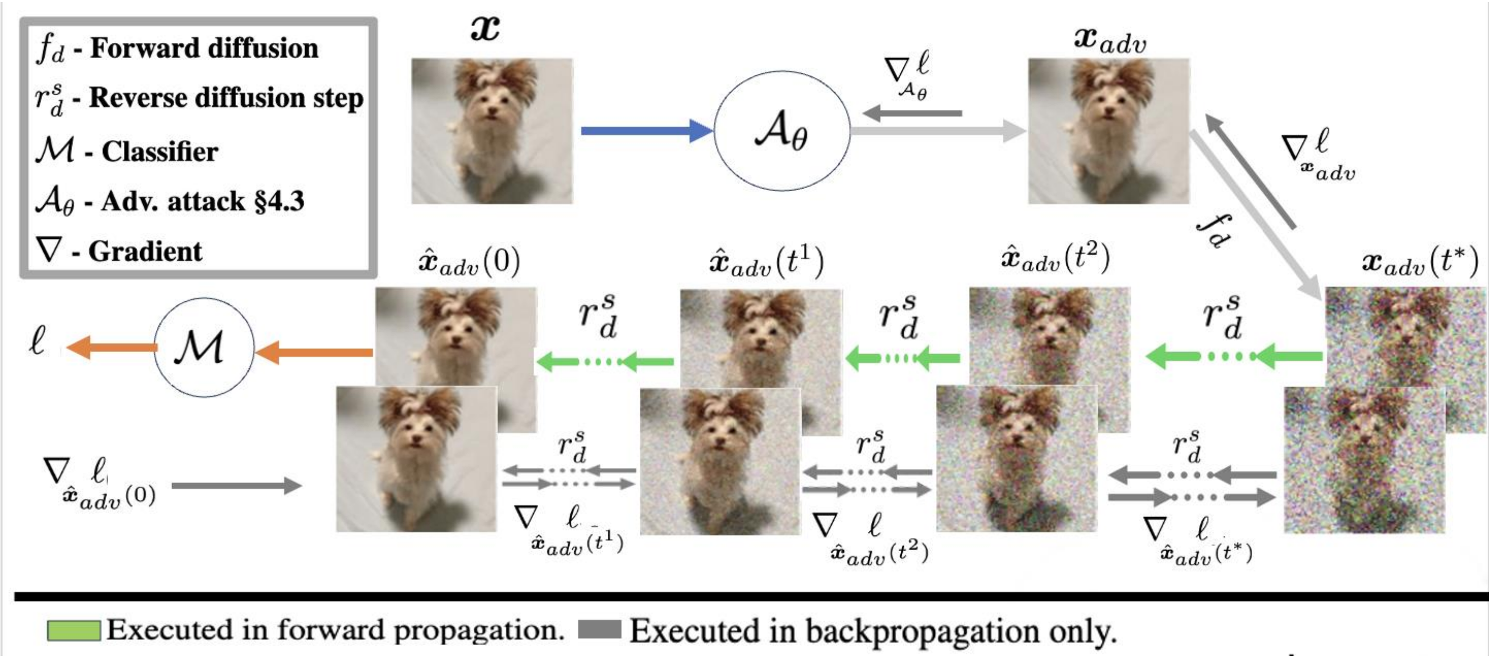


Variance Reduction ✓

Time Consistency ✓

Guidance Gradients ✓

3) DiffGrad: A Reliable Gradient module



Experimental Findings (*AutoAttack* on *CIFAR10 & ImageNet*)

***DBP*'s robustness is drastically degraded!!**

1. Worst-case robustness nearly vanishes with *DiffGrad.*

2. Majority Vote proves far superior but remains only partially robust (**≤39.45**).

Variance Reduction ✓

Time Consistency ✓

Guidance Gradients ✓

4) Low-Frequency Attacks Defeat DBP

Can DBP be degraded further under Majority Vote?

Can DBP be degraded further under Majority Vote?

Traditional adversarial attacks introduce high-frequency disruptions

→ Significantly limits their magnitudes.
→ Causes them to fail against DBP's stochasticity.

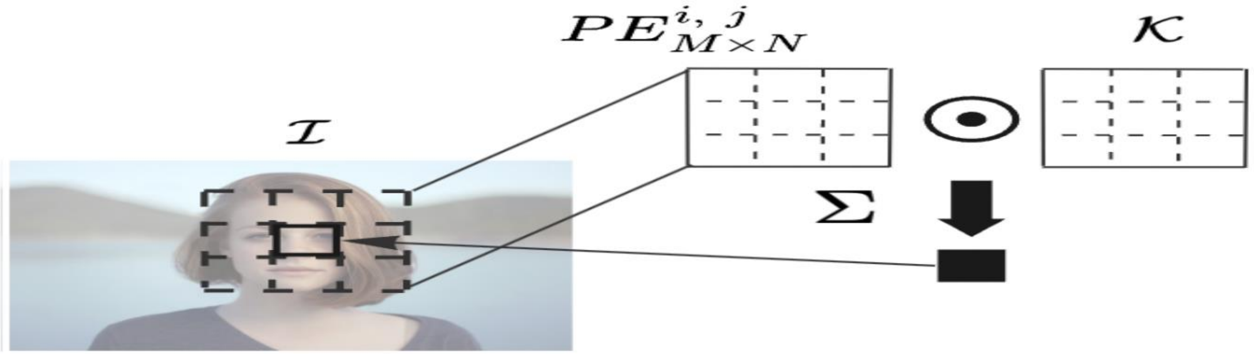
Can DBP be degraded further under Majority Vote?

Traditional adversarial attacks introduce high-frequency disruptions

→ Significantly limits their magnitudes.

→ Causes them to fail against DBP's stochasticity.

Low-frequency (LF) Attack: Chain of ***Optimizable Filters***

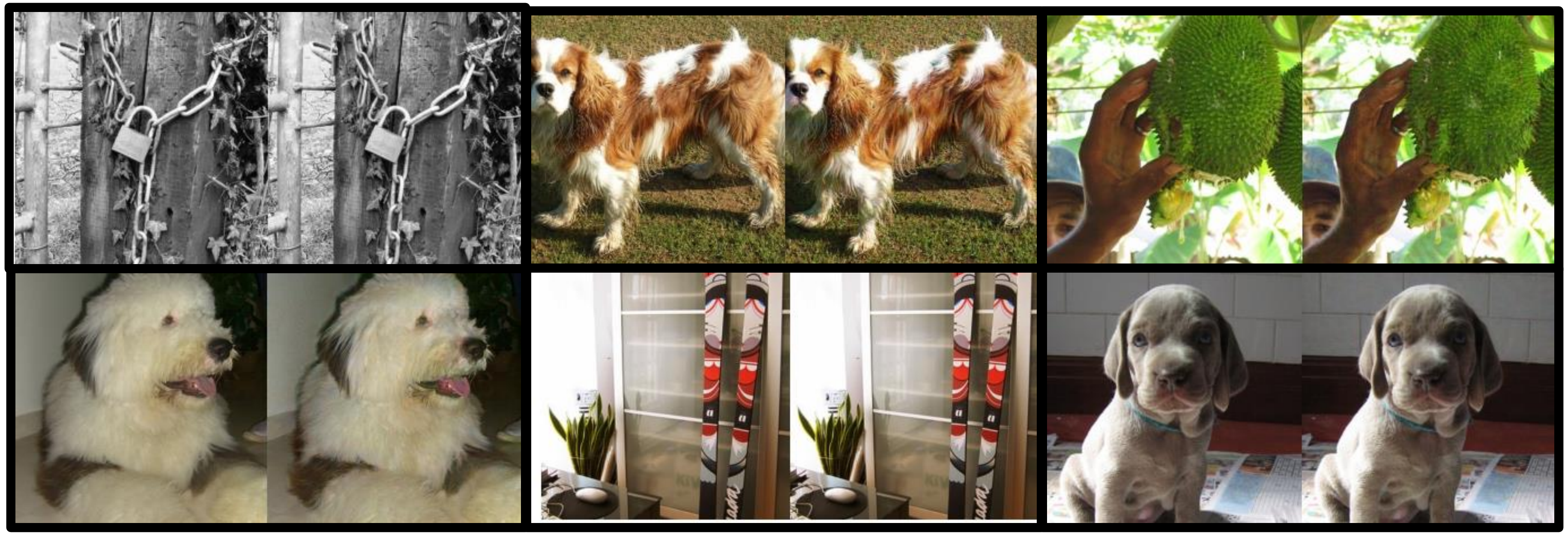


4) Low-Frequency Attacks Defeat DBP

Pur.	Dataset	Models	Cl-Acc %	Rob-Acc %
DiffPure [30]	***ImageNet***	ResNet-50	72.54	**0.00**
		WideResNet-50-2	77.02	**0.00**
		DeiT-S	77.34	**0.00**
	CIFAR-10	WideResNet-28-10	92.19	**2.73**
		WideResNet-70-16	92.19	**3.13**
GDMP [40]	***ImageNet***	ResNet-50	73.05	**0.39**
		WideResNet-50-2	71.88	**0.00**
		DeiT-S	75.00	**0.39**
	CIFAR-10	WideResNet-28-10	93.36	**0.00**
		WideResNet-70-16	92.19	**0.39**

MV robustness under LF

4) Low-Frequency Attack Samples



Right → original sample. Left → adversarial sample.

Insights

- **Theoretical assumptions fail:** *DBP*'s claimed robustness "by construction" collapses once gradient inconsistencies are resolved.
- **Evaluation variance matters:** Majority-vote testing reconciles prior over- and under-estimations of robustness.
- **Low-frequency attacks prevail:** Structured *AE*s bypass *DBP*'s stochastic defenses across datasets.

Insights

- **Theoretical assumptions fail:** *DBP*'s claimed robustness "by construction" collapses once gradient inconsistencies are resolved.
- **Evaluation variance matters:** Majority-vote testing reconciles prior over- and under-estimations of robustness.
- **Low-frequency attacks prevail:** Structured *AE*s bypass *DBP*'s stochastic defenses across datasets.

Takeaway: Current *DBP* is *not* a viable defense against adversarial examples—highlighting the need for more powerful alternatives.

Insights

- **Theoretical assumptions fail:** *DBP*'s claimed robustness "by construction" collapses once gradient inconsistencies are resolved.
- **Evaluation variance matters:** Majority-vote testing reconciles prior over- and under-estimations of robustness.
- **Low-frequency attacks prevail:** Structured *AE*s bypass *DBP*'s stochastic defenses across datasets.

Takeaway: Current *DBP* is *not* a viable defense against adversarial examples—highlighting the need for more powerful alternatives.

akassis@uwaterloo.ca

andrekassis.github.io

andrekassis7

Website: https://github.com/andrekassis/DiffBreak