中国科学院大学
University of Chinese Academy of Sciences

ICML
International Conference On Machine Learning

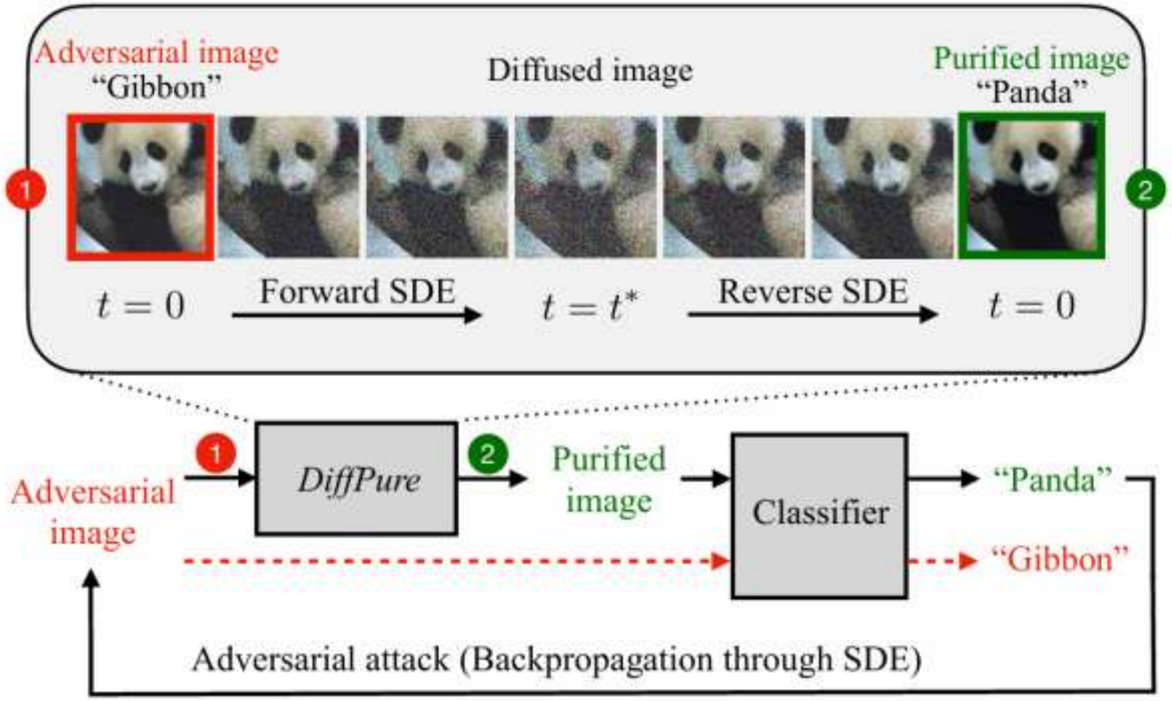
Diffusion-based Adversarial Purification from the Perspective of the Frequency Domain

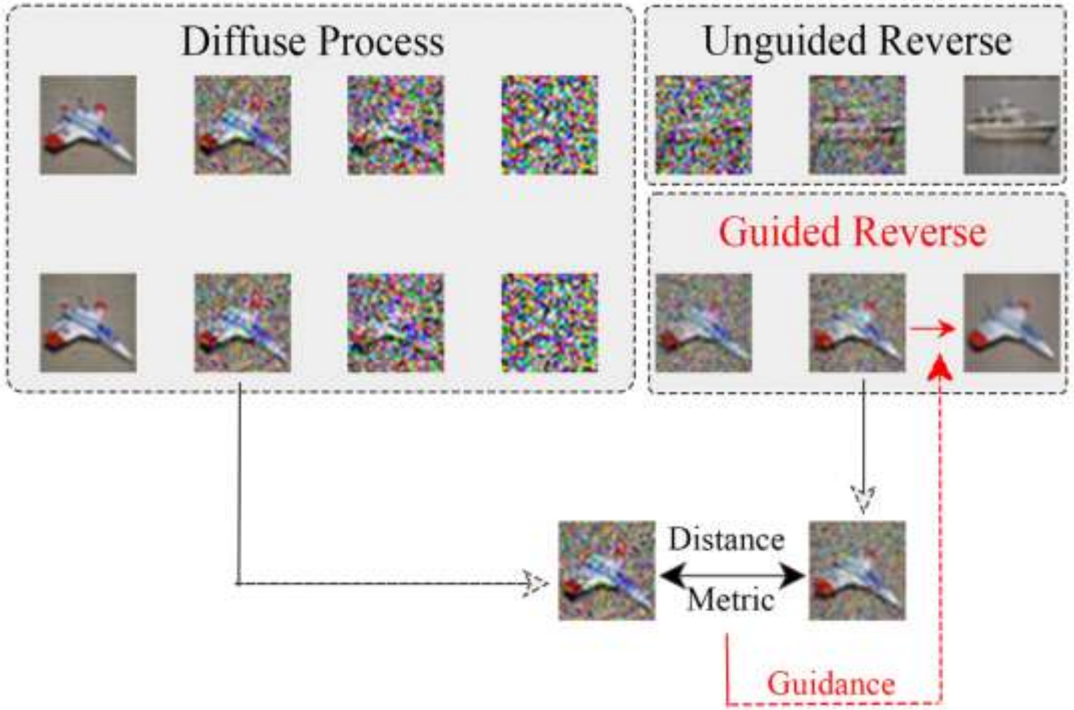
Gaozheng Pei, Ke Ma*, Yingfei Sun,
Qianqian Xu, Qingming Huang*

Arxiv Page

Github Page

Background





Large Noise-Step Original without Guidance Damages Semantic Information

Large Noise-Step Damages with Guidance Introduces Adversarial Perturbations

Can we use the **clean semantic information** of adversarial samples for **guidance** while **avoiding** the introduction of **adversarial perturbations**?

Motivation

we can **decompose** adversarial samples into clean semantic information and corrupted semantic information

Adversarial Image = **Clean Semantic Information** + **Damaged Semantic Information**

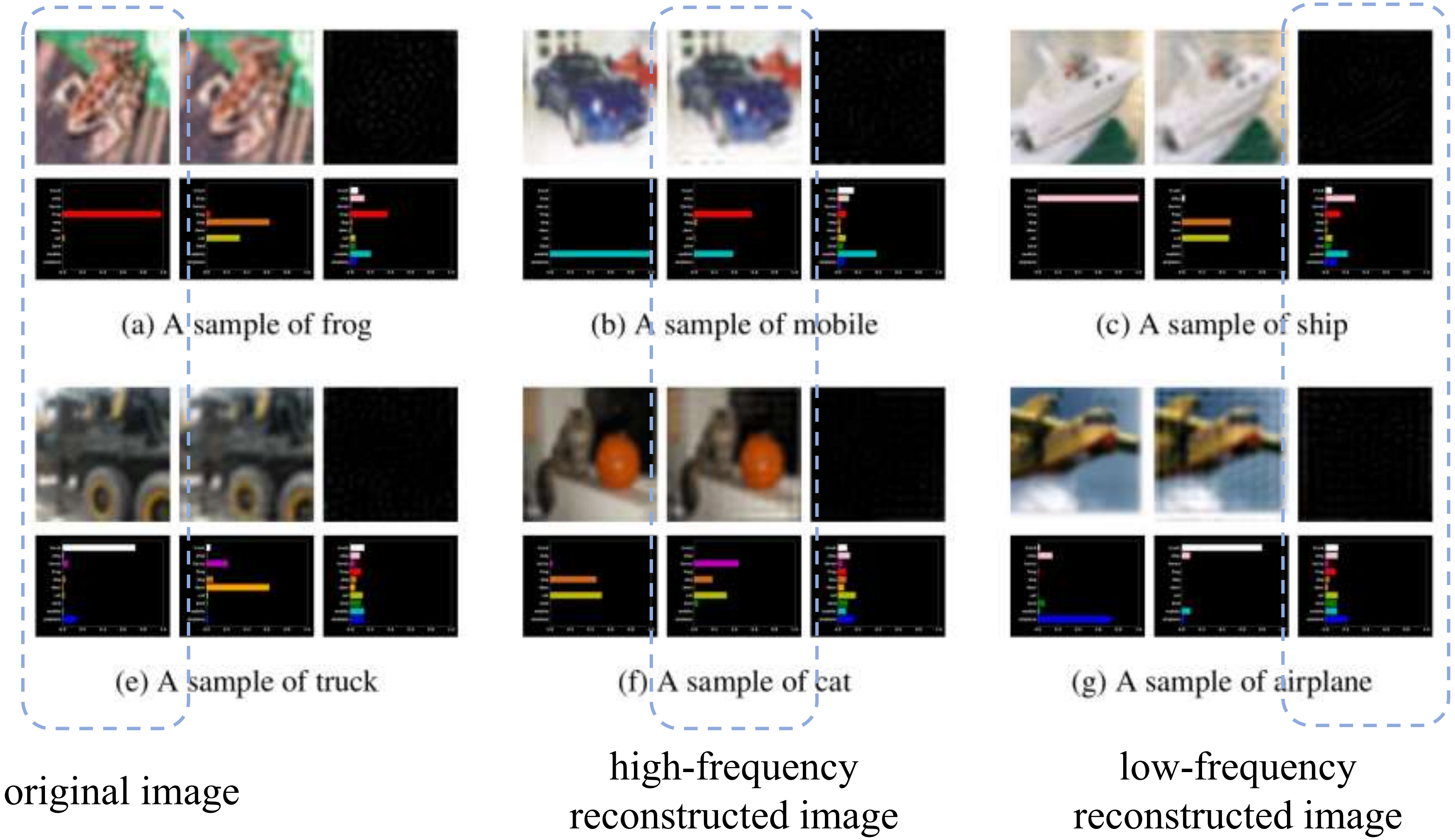
Clean Semantic Information → For the **Guidance** of the **Reverse Process**

Damaged Semantic Information → **Discard**

We can use the **clean semantic information** to **guide** the reverse process of diffusion model.

Motivation

Different frequency components have varying degrees of influence on the neural network's predictions.



This motivates us that we can investigate the distribution of adversarial perturbations from the perspective of frequency domain.

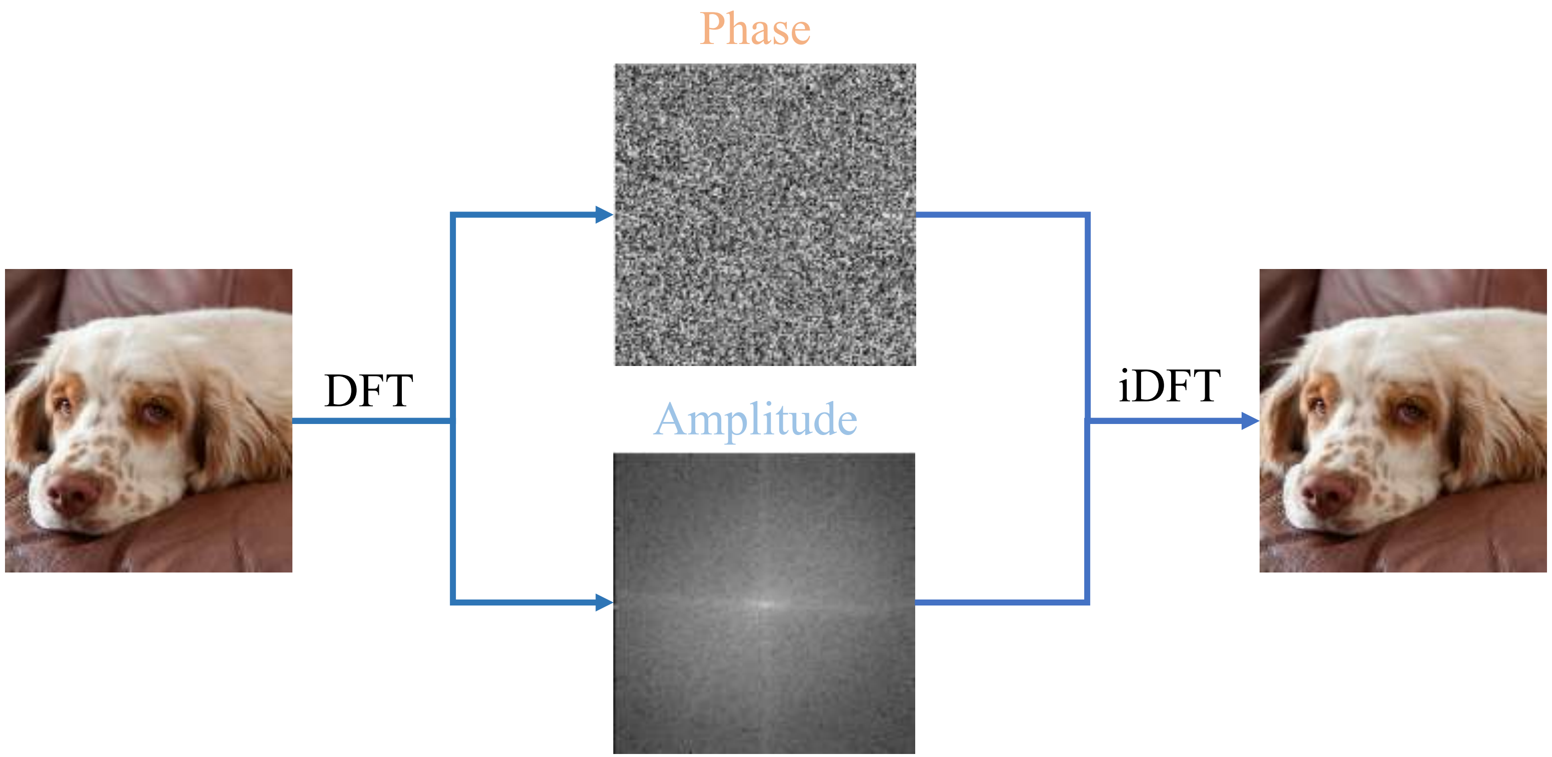
Preliminary

Discrete Fourier transform (DFT)

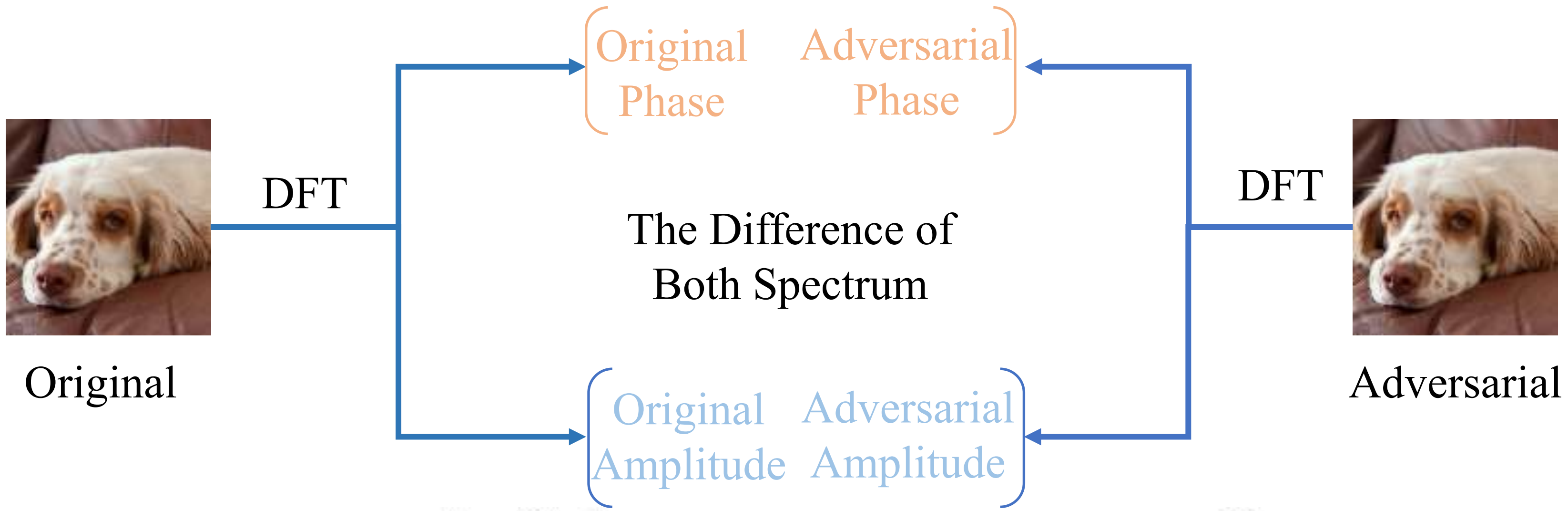
$$\mathbf{x}_0(u,v) = DFT(\mathbf{x}_0) = \underbrace{\boxed{|\mathbf{x}_0(u,v)|}}_{\text{Amplitude}} e^{\overbrace{\boxed{i\phi_{\mathbf{x}_0}(u,v)}}^{}}$$

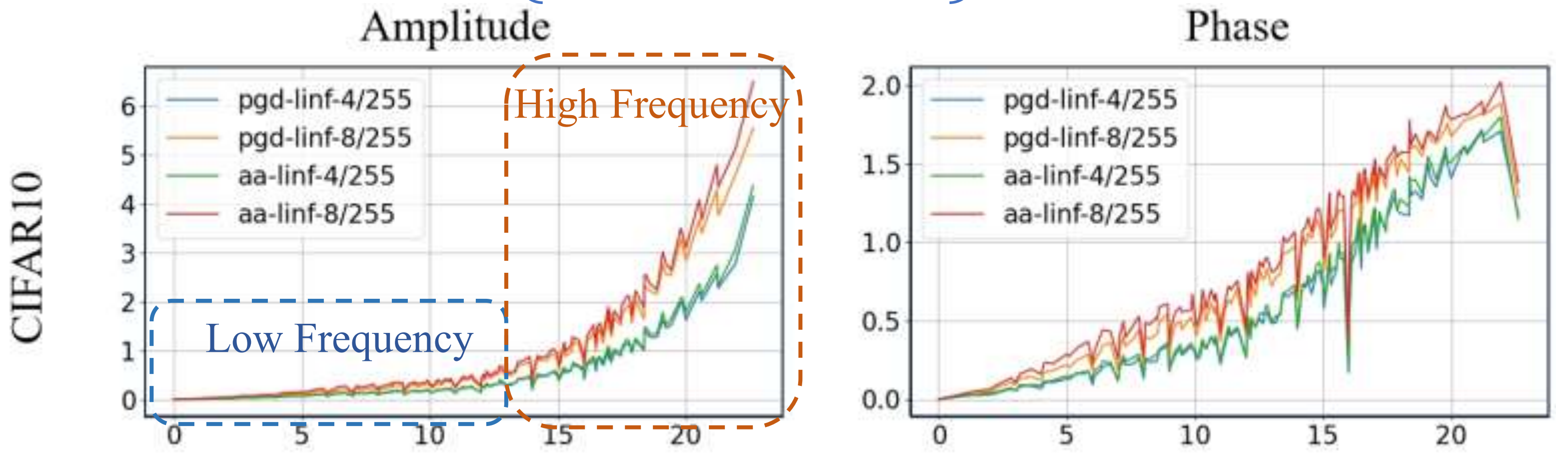
Amplitude Phase

Frequency of coordinate (u,v) $\quad D(u,v) = [(u - H/2)^2 + (v - W/2)^2]^{\frac{1}{2}}$



Experimental Phenomenon



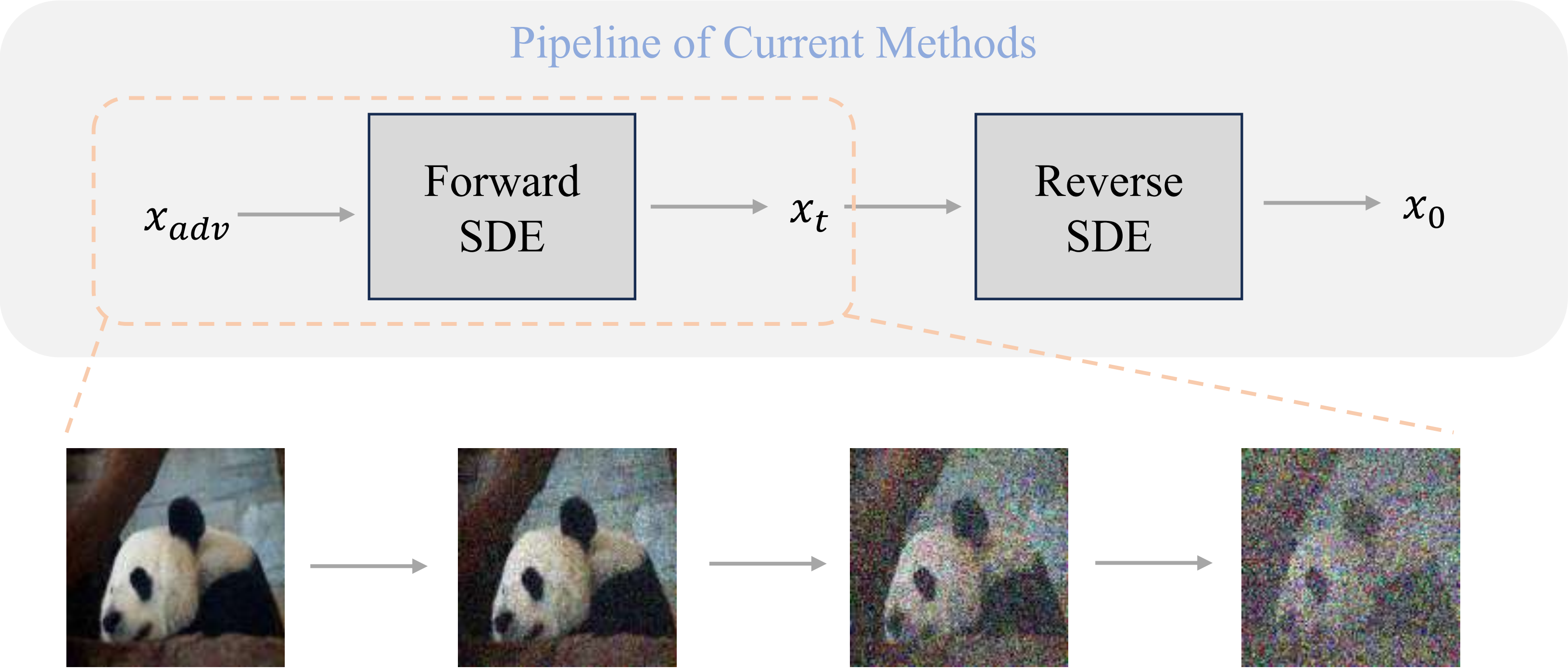


For Both the Amplitude Spectrum and the Phase Spectrum, the Degradation Caused by Adversarial Perturbations Exhibits an Approximately Monotonically Increasing Trend with Frequency

Experimental Phenomenon

Low-frequency components represent **Clean Semantic Information**, while high-frequency components represent **Damaged Semantic Information**.

Why Current Methods Fail ?



Intuitively, the forward process involves both high-frequency and low-frequency information. Here, we provide a rigorous **theoretical proof** of this.

Theoretical Analysis

Theorem 3.1 The variance of the first-order approximation of the difference of phase between clean image $\mathbf{x}_0$ and noisy image $\mathbf{x}_t$ at arbitrary coordinates (u, v) at frequency domain is as follows:

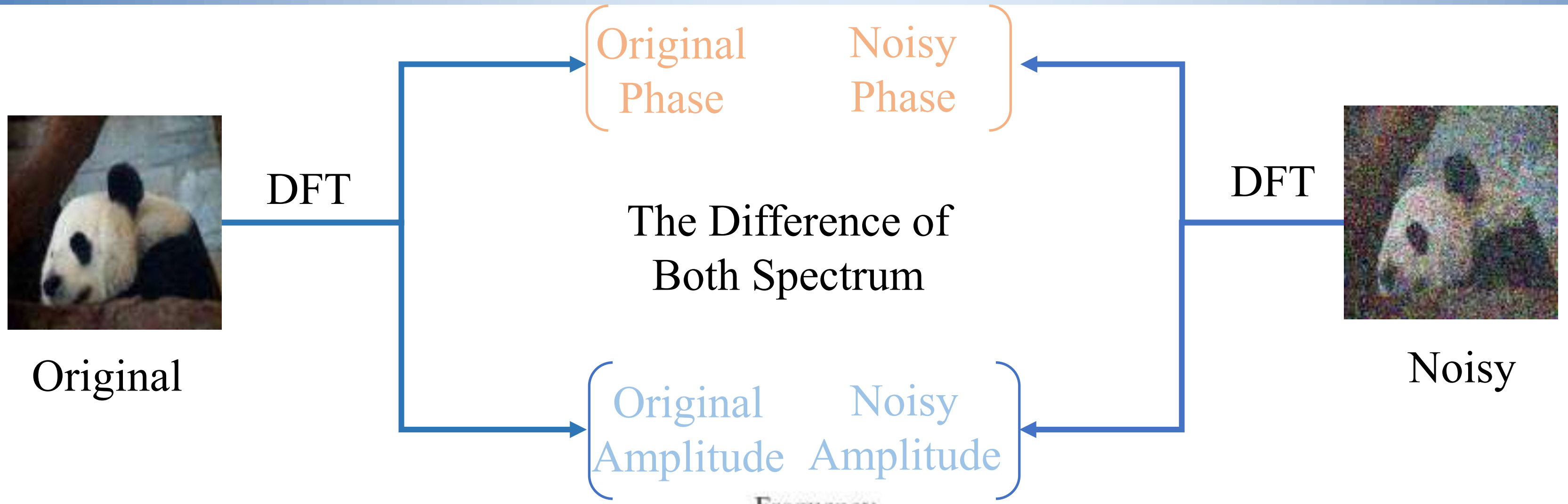
$$Var(\Delta\theta_t(u,v)) = \frac{1}{\sqrt{1 - \frac{1}{SNR_t^2(u,v)}}} - 1,$$

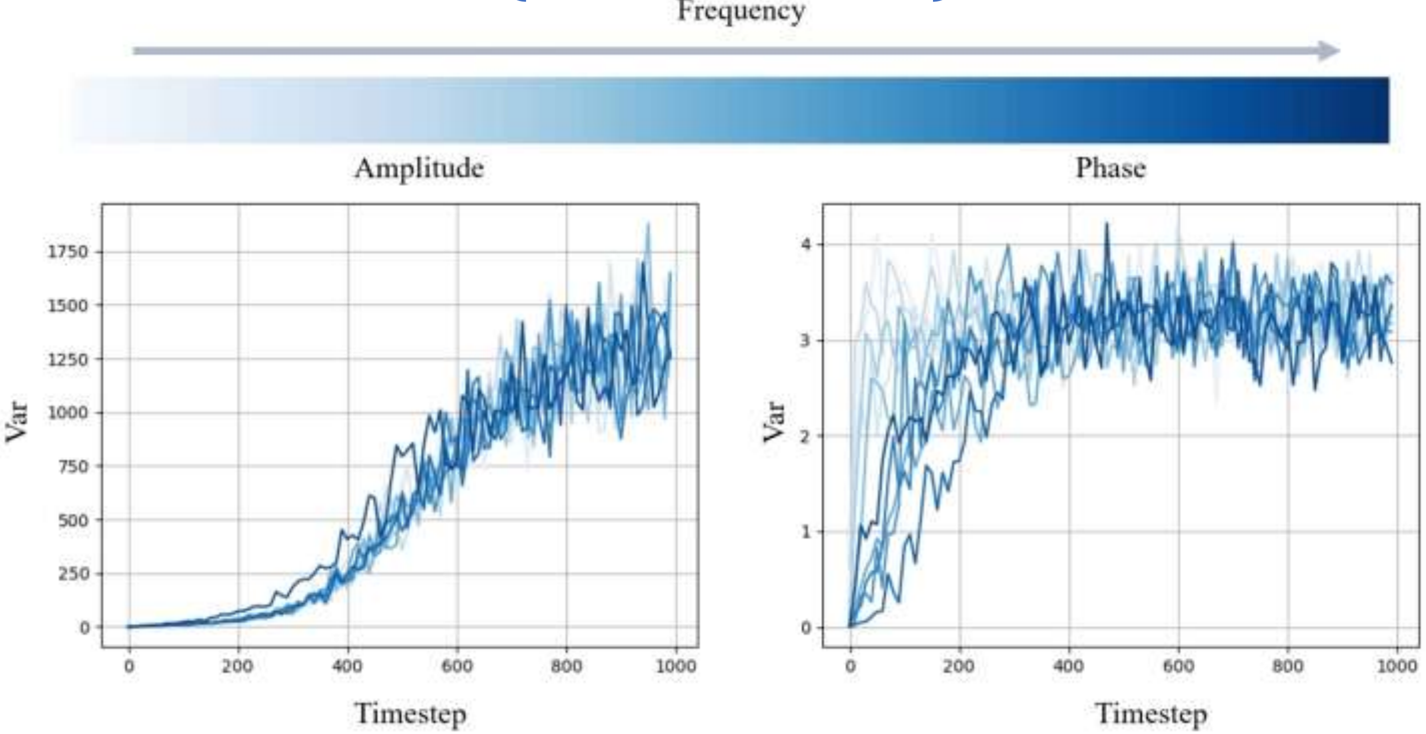
Theorm 3.2 The variance of the difference of amplitude at time-step t between clean image $\mathbf{x}_0$ and noisy image $\mathbf{x}_t$ at arbitrary coordinates (u, v) at frequency domain is as follows:

$$Var(\Delta A_t(u,v)) \approx \frac{1-\overline{\alpha}_t}{2} - \frac{(1-\overline{\alpha}_t)^2}{16|\mathbf{x}_0(u,v)|\overline{\alpha}_t}.$$

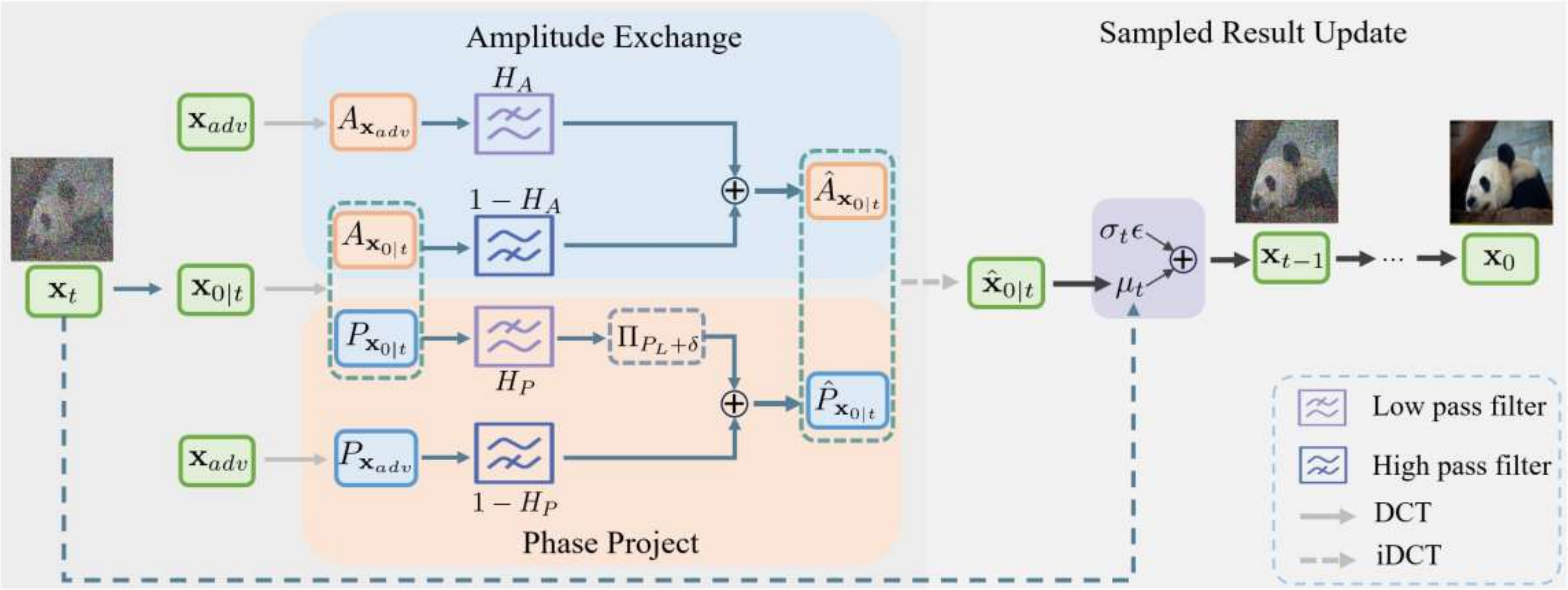
For both the **phase spectrum** and the **amplitude spectrum**, the forward process causes a **monotonically increasing** degradation of all frequency components with respect to t.

Theorem Validation





Framework



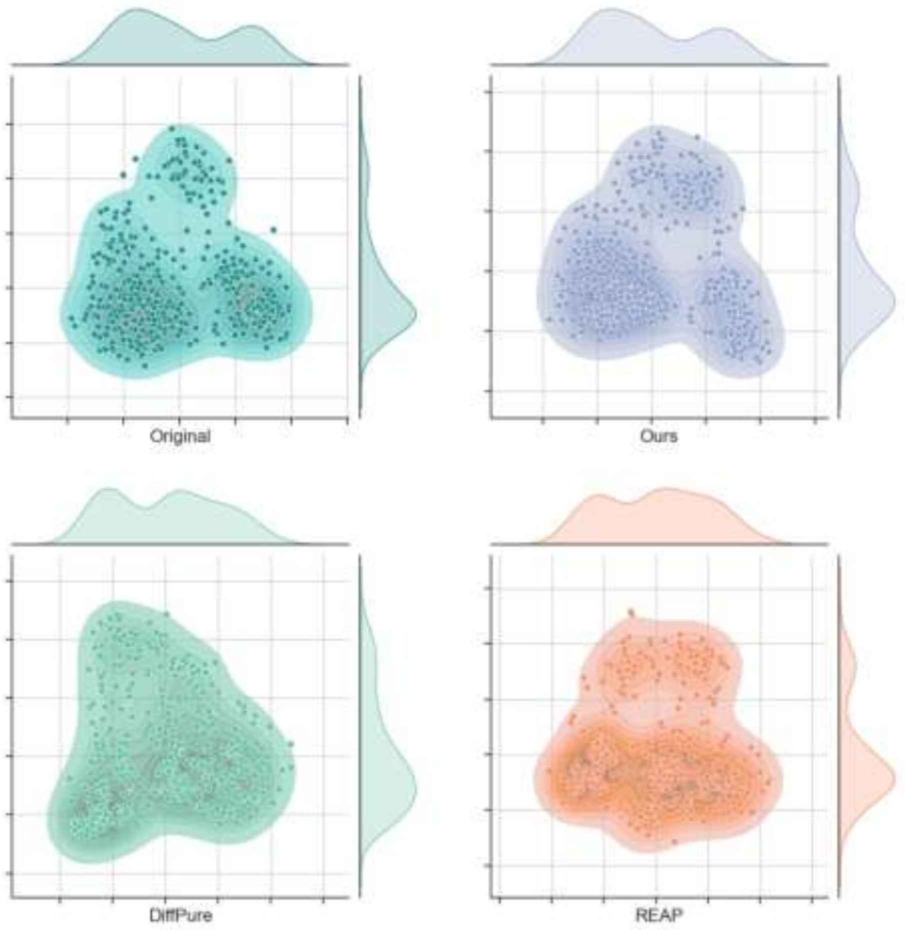
The core of our method is to preserve, at each time-step of the reverse process, the amplitude and phase spectrum information of the original clean samples extracted from adversarial images as a prior.

Experimental Result

Table 1. Standard and robust accuracy of different Adversarial Training (AT) and Adversarial Purification (AP) methods against PGD and AutoAttack $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10. * utilizes half number of iterations for the attack due to the high computational cost. † indicates the requirement of extra data. The result with an underline indicates the second highest.

Type	Method	Standard Acc.	Robust Acc. PGD	Robust Acc. AutoAttack
		WideResNet-28-10		
AT	(Gowal et al., 2021)	88.54	65.93	63.38
	(Gowal et al., 2020)†	87.51	66.01	62.76
	(Pang et al., 2022)	88.62	64.95	61.04
AP	(Yoon et al., 2021)	85.66±0.51	33.48±0.86	59.53±0.87
	(Nie et al., 2022)	90.07±0.97	<u>56.84±0.59</u>	63.60±0.81
	(Lee & Kim, 2023)	90.16±0.64	55.82±0.59	70.47±1.53
	(Bai et al., 2024)	<u>91.41</u>	49.22*	<u>77.08</u>
	(Zollicoffer et al., 2025)	84.20	-	59.14
	(Lin et al., 2024)	90.62	-	72.85
	Ours	**92.19 ±0.33**	**59.39±0.79**	**77.35±2.14**

Experimental Result on WideResNet28-10



Distribution of Purified Natural Examples