AMRM-Pure: Semantic-Preserving Adversarial Purification

Zhihao Dou, Zhiqiang Gao, Dongfei Cui, Weida Wang, Qinjian Zhao, Dinggen Zhang, Jun Yan, Zeke Xie, Shufei Zhang

Case Western Reserve University, Wenzhou-Kean University, Shanghai Artificial Intelligence Laboratory

Motivation

1. Most existing methods rely on powerful generative models, such as diffusion models, and mainly focus on aligning adversarial samples with clean samples in the feature/distribution space.

However, they often overlook an important question:

How do adversarial perturbations affect semantic relationships among image patches within the same image?

An interesting phenomenon

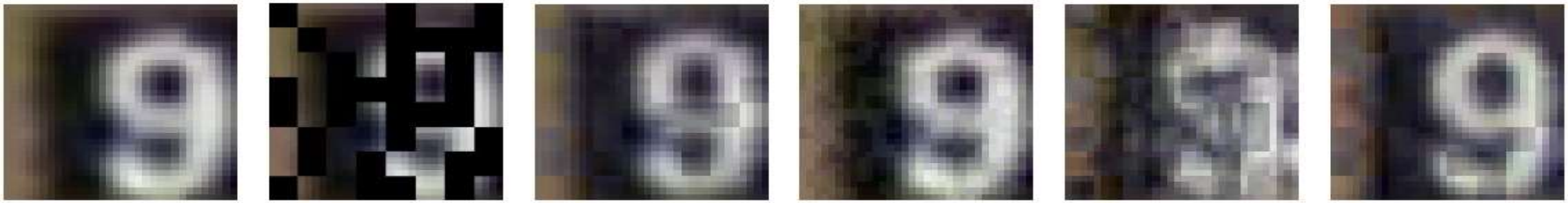
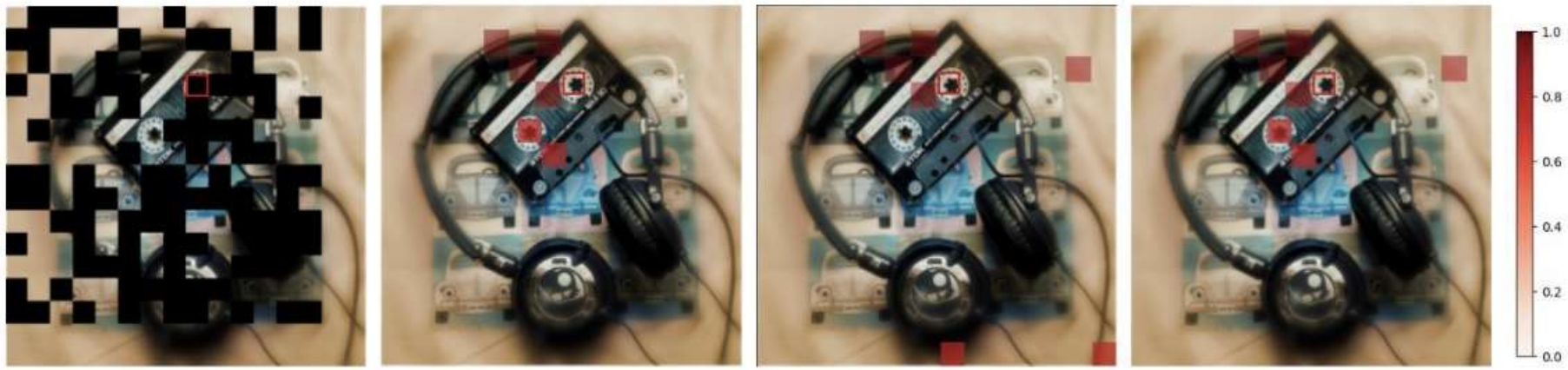


Figure 1: (a) Original image, (b) Masked image, (c) Clean image reconstruction from MAE, (d) Adversarial example under AutoAttack, (e) Reconstruction of adversarial example under AutoAttack from MAE, (f) Reconstruction of the denoised image under AutoAttack from MAE (denoised by our AMRM-Pure$_{\mathrm{MAE}}$).

Attention matrix shift (AMV)



(a) (b) (c) (d)

Figure 3: The first column, Figure (a) represents the Mask Matrix. The second column, Figure (b) illustrates the Attention Weights for clean samples. The third column, Figure (c) depicts the Attention Weights for adversarial examples. The fourth column, Figure (d) showcases the Attention Weights for denoised samples (by our AMRM-Pure$_{\text{MAE}}$). Patches with a deeper red color mean the elements with more attention. The data is sampled from the ImageNet dataset Deng et al. (2009).

Theoretical Analysis

Theorem 3.1. *Let $\delta_t = \boldsymbol{Z}^t_{adv} - \boldsymbol{Z}^t$ denotes the latent feature shift caused by the adversarial perturbation at layer t in MAE. With a set $\{\omega_i\}_{i=0}^k$ and kernel coefficient $\omega_i \in \mathcal{N}(0, \mathbf{I}_d)$, it holds that:*

$$||\mathbf{A}^t_{adv} - \mathbf{A}^t||_2 \geq \gamma \left\| \left[\left(\mathbf{Y} - \mathbf{B}\mathbf{Q}^t\right)^\top \mathbf{W}^t_Q + \left(\mathbf{Y} - \mathbf{B}\mathbf{K}^t\right)^\top \mathbf{W}^t_K \right] \delta_t \right\|_2,$$

$$\mathbf{B} = \sum_{i=0}^{k} \exp\left(\omega_i^\top (\mathbf{Q}^t + \mathbf{K}^t)\right), \quad \mathbf{Y} = \sum_{i=0}^{k} \exp\left(\omega_i^\top (\mathbf{Q}^t + \mathbf{K}^t)\right) \omega_i,$$

$$\gamma = \frac{\exp\left(-\frac{\|\mathbf{Q}^t\|^2 + \|\mathbf{K}^t\|^2}{2}\right)}{m}.$$

Theorem 3.2. *Let $\mathbf{A}^{\text{dec}}_{i,t}$ denote the attention matrix at the t-th layer of the MAE decoder for the i-th sample in the dataset, and let $\mathbf{A}^{\text{dec}}_{\text{adv},i,t}$ denote the corresponding attention matrix for the adversarial examples. With ratio constants C_A and H, it holds that:*

$$\mathcal{L}^{\text{adv}}_{rec} \geq \frac{1}{2}\mathcal{L}_{rec} + \frac{1}{2NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \left[HC_A \left\| \mathbf{A}^{\text{dec}}_{\text{adv},i,t} - \mathbf{A}^{\text{dec}}_{i,t} \right\|^2 - c_{rec} \right]$$

c_{rec} is the reconstruction bias, which symbolizes the disparity between the output of MAE and the original, unmasked image. The definition of c_{rec} can be found in Appendix I.1.

Empirical Validation

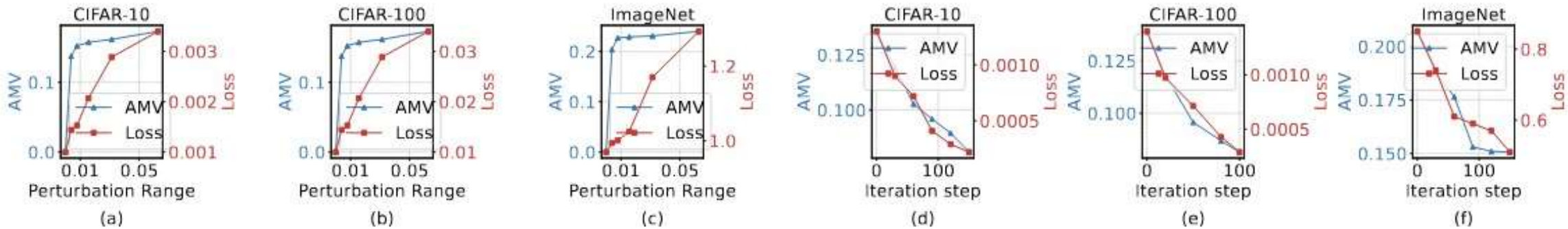
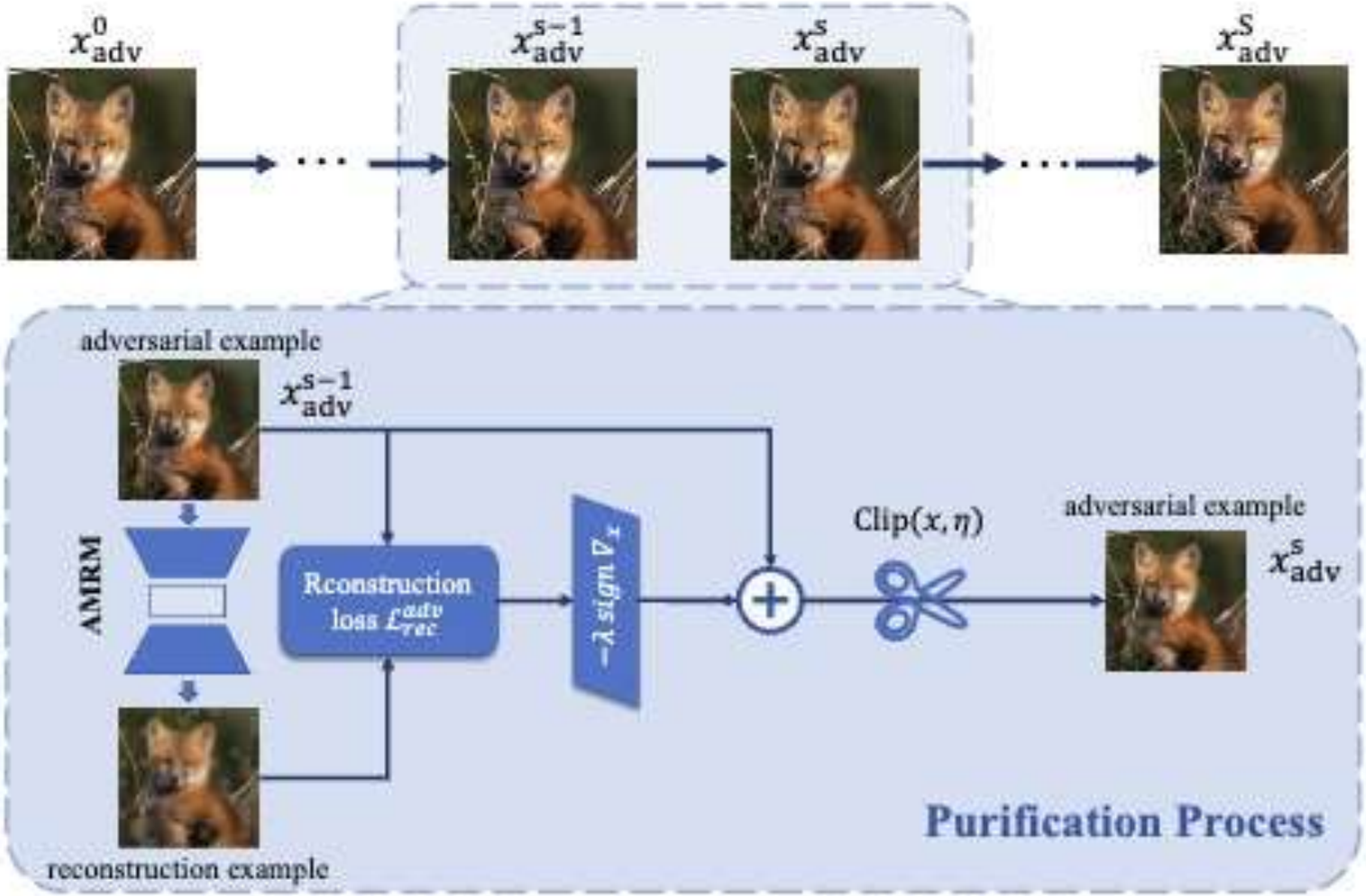


Figure 4: Trends of MAE reconstruction loss and attention matrix variation under AutoAttack and during purification across multiple datasets. (a–c): Under AutoAttack on CIFAR-10, CIFAR-100, and ImageNet with different attack budget. (d–f): During the purification process with AMRM-Pure$_{\mathrm{MAE}}$ on CIFAR-10, CIFAR-100, and ImageNet.

AMRM-Pure


x_{adv}^{0}
x_{adv}^{s-1}
x_{adv}^{s}
x_{adv}^{S}
adversarial example
x_{adv}^{s-1}
AMRM
Reconstruction loss $\mathcal{L}_{rec}^{adv}$
$-\lambda\, sign\, \nabla_x$
Clip(x, η)
adversarial example
x_{adv}^{s}
reconstruction example
Purification Process

Results

Table 1: Clean and robust accuracy (%) on CIFAR-10 obtained by different purification methods. WideResNet is commonly abbreviated as WRN.

Method	Classifier	Std Acc	Robust Acc ℓ_∞	Robust Acc ℓ_2
Shi et al. Shi et al. (2021)	WRN-28-10	91.89	4.56	7.25
Yoon et al. Yoon et al. (2021)	WRN-70-16	87.93	37.65	57.81
Zhang et al. Zhang et al. (2023)	WRN-70-16	**93.16**	22.07	35.74
Diffpure Nie et al. (2022)	WRN-70-16	92.50	42.20	60.80
COUP Zhang et al. (2024)	WRN-28-10	90.33	41.72	57.25
ADBM Li et al. (2025)	WRN-70-16	91.90	47.70	63.30
$\text{ADDT}_{w/\text{Diffpure}}$ Liu et al. (2025)	WRN-28-10	89.94	55.76	-
$\text{AMRM-Pure}_{\text{MAE}}$	WRN-28-10	88.57	40.53	53.50
$\text{AMRM-Pure}_{\text{MaskDiT}}$	WRN-28-10	92.03	50.57	64.53
$\text{RAMRM-Pure}_{\text{MAE}}$	WRN-28-10	90.09	45.15	60.72
$\text{RAMRM-Pure}_{\text{MaskDiT}}$	WRN-28-10	93.11	**62.13**	**73.57**

Table 2: Clean and robust accuracy (%) on CIFA 100 obtained by different purification methods. Tl experiment are implemented on WideResNet-28-10.

Method	Std Acc	Robust Acc ℓ_∞	Robust Acc ℓ_2
Diffpure Nie et al. (2022)	45.23	11.57	31.53
COUP Zhang et al. (2024)	65.71	15.22	34.28
$\text{ADDT}_{w/\text{DDPM}}$ Liu et al. (2025)	66.02	18.85	36.57
$\text{AMRM-Pure}_{\text{MAE}}$	65.34	14.28	29.29
$\text{AMRM-Pure}_{\text{MaskDiT}}$	**70.03**	24.39	36.51
$\text{RAMRM-Pure}_{\text{MAE}}$	66.28	19.53	31.58
$\text{RAMRM-Pure}_{\text{MaskDiT}}$	69.87	**29.91**	**43.27**

Table 3: Clean and robust accuracy (%) on SVH obtained by different purification methods. The exp iment are implemented on WideResNet-28-10.

Method	Std Acc	Robust Acc ℓ_∞	Robust Acc ℓ_2
Diffpure Nie et al. (2022)	93.90	39.70	63.30
COUP Zhang et al. (2024)	92.07	41.62	63.97
ADBM Li et al. (2025)	93.50	47.90	65.70
$\text{AMRM-Pure}_{\text{MAE}}$	94.54	27.59	55.29
$\text{AMRM-Pure}_{\text{MaskDiT}}$	94.91	46.57	66.38
$\text{RAMRM-Pure}_{\text{MAE}}$	94.47	39.15	60.51
$\text{RAMRM-Pure}_{\text{MaskDiT}}$	**95.39**	**55.90**	**70.18**

Thank You