
Supplementary:

MAE-Pure: Semantic-Preserving Adversarial Purification

Anonymous Author(s)

Affiliation

Address

email

A Comparing with the advanced generative capability of the diffusion-based method

To explore the origins of the enhanced performance observed with MaskDiT-Pure and RMaskDiT-Pure, we conducted a series of controlled experiments. We began by evaluating the generative capabilities of the diffusion model using the standard reconstruction process of MaskDiT-Pure and RMaskDiT-Pure for image purification. This process, which includes both the forward (noise addition) and backward (denoising) steps, mirrors the approach used in DiffPure [1], where the reconstructed images serve as purified samples. We labeled these modified models as MaskDiT_{purification} and RMaskDiT_{purification}. We then compared their performance against our AMV-based methods, MaskDiT-Pure and RMaskDiT-Pure. By keeping the model architecture consistent across all variants, we enabled a direct comparison of the effectiveness of different denoising strategies.

Table 1: Ablation analysis on different denoising components across various datasets.

Method	Architecture	CIFAR10		CIFAR100		SVHN	
		Std acc	Robust acc	Std acc	Robust acc	Std acc	Robust acc
MaskDiT _{purification}	WRN-28-10	91.13	42.99	63.29	13.57	92.28	40.07
RMaskDiT _{purification}	WRN-28-10	90.57	47.57	64.15	18.55	93.03	45.19
MaskDiT-Pure	WRN-28-10	92.03	50.57	70.03	24.39	94.91	46.57
RMaskDiT-Pure	WRN-28-10	93.11	62.13	69.87	29.91	95.39	55.90

Table 1 displays the standard and robust accuracies under a classifier (WRN-28-10) across three datasets: CIFAR-10, CIFAR-100, and SVHN. MaskDiT-Pure and RMaskDiT-Pure significantly outperform their counterparts that rely on the reconstruction pipeline, MaskDiT_{purification} and RMaskDiT_{purification}, in both standard and robust accuracy. For example, RMaskDiT-Pure improves robust accuracy from 47.57% to 62.13% on CIFAR-10 and from 45.19% to 55.90% on SVHN. These results confirm two key points: (1) The advanced generative capability of the diffusion backbone indeed contributes to performance improvements; (2) Our AMV-based method is orthogonal to the advanced generation ability, indicating that diffusion models can be integrated into our approach to further boost performance.

References

- [1] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *International Conference on Machine Learning*, 2022.