SUPPLEMENTARY MATERIALS FOR EFFICIENT ADVERSARIAL DETECTION AND PURIFICATION WITH DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

A DETAIL EXPERIMENT SETTINGS

Our experiment is implemented with PyTorch on an NVIDIA GeForce RTX 3090 GPU. For the CIFAR10 dataset, we upscale the adversarial anti-aliased images with PyToch to 64×64 resolution for ResShift. We use ResShift default v3 parameter for our experiments.

B ADDITIONAL EXPERIMENTS

We further report additional experiments against various adversaries on CIFAR10 and ImageNet datasets, as shown in Table 1 and 2. Noted that DiffAttack Chen et al. (2023) adopt latent inversion from the validation set to generate adversarial examples, we only report the standard accuracy to the clean validation set.

Table 1: **The defense performance against AutoAttack ($\ell_2 = 0.5$) on the CIFAR10 dataset.**

Method	Target Model	Standard Accuracy(%)	Robust Accuracy(%)
Rony *et al.* (Rony et al., 2019)	WideResNet-28-10	89.05	66.41
Ding *et al.* (Ding et al., 2020)	WideResNet-28-10	88.02	67.77
Rebuffi *et al.* (Rebuffi et al., 2021)	WideResNet-28-10	91.79	78.32
Wang *et al.* (Wang et al., 2022)	WideResNet-28-10	92.00	75.28
Nie *et al.* (Nie et al., 2022)	WideResNet-28-10	91.38	78.98
Song *et al.* (Song et al., 2024)	WideResNet-28-10	**92.84**	81.52
$\text{Ours}_{\text{Detection}}$	WideResNet-28-10	*97.50 ± 2.15*	*92.95 ± 0.36*
$\text{Ours}_{\text{Purification}}$	WideResNet-28-10	92.54 ± 1.66	**84.90 ± 2.82**
Gowal *et al.* (Gowal et al., 2021)	WideResNet-70-16	90.90	74.03
Rebuffi *et al.* (Rebuffi et al., 2021)	WideResNet-70-16	92.41	80.86
Nie *et al.* (Nie et al., 2022)	WideResNet-70-16	93.24	81.17
Song *et al.* (Song et al., 2024)	WideResNet-70-16	92.51	83.60
$\text{Ours}_{\text{Detection}}$	WideResNet-70-16	*98.13 ± 1.94*	*94.57 ± 1.82*
$\text{Ours}_{\text{Purification}}$	WideResNet-70-16	**93.42 ± 1.51**	**87.60 ± 2.35**

Table 2: **The defense performance against DiffAttack ($\ell_{\text{inf}} = 8/255$) on the ImageNet dataset.**

Method	Target Model	Standard Accuracy(%)	Robust Accuracy(%)
Nie *et al.* Nie et al. (2022)	ResNet50	68.22	59.15
Song *et al.* (Song et al., 2024)	ResNet50	66.92	60.17
$\text{Ours}_{\text{Detection}}$	ResNet50	*88.30 ± 2.44*	*78.44 ± 1.95*
$\text{Ours}_{\text{Purification}}$	ResNet50	**75.28 ± 1.06**	**65.51 ± 1.33**

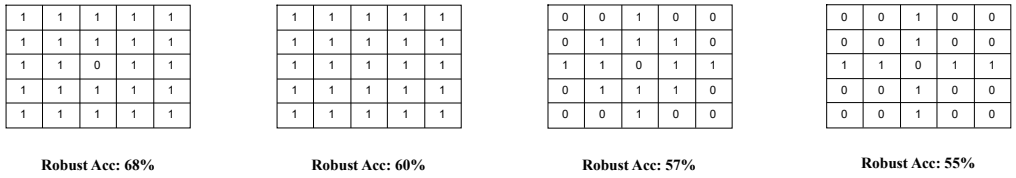

Figure 1: **The defense performance of various filter weights against AutoAttack ($\ell_{\text{inf}} = 8/255$) on the ImageNet dataset.** We use 1 for better understanding, while we set to the mean value according to the number of 1 blocks in the experiments.

C Filter Selection

We discuss the choice of filter settings in the ablation study. However, it is also critical to design the filter weight for the adversarial anti-aliasing. Figure 1 demonstrates that the selection of filter weight is empirical and achieves the best performance on setting the mean value except for the center.

D Limitation

Despite achieving a significantly higher time efficiency and better defense performance than previous diffusion-based adversarial purification, our defense still has several limitations. One drawback is there exists a gap between the adversarial detection rate and robust accuracy. Therefore, a stronger defense can be proposed to increase the robust accuracy that focuses on defending against the detected adversarial examples. Another drawback is that the robust accuracy against UAEs is still not comparable to perturbation-based adversarial attacks. We aim to further improve it in future work.

References

Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *arXiv preprint arXiv:2305.08192*, 2023.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. MMA training: Direct input space margin maximization through adversarial training. In *Proceedings of the International Conference on Learning Representations*, 2020.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

Jérôme Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 4322–4330, 2019.

Kaiyu Song, Hanjiang Lai, Yan Pan, and Jian Yin. Mimicdiffusion: Purifying adversarial perturbation via mimicking clean diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24665–24674, 2024.

Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022.