MimicDiffusion: Purifying Adversarial Perturbation via Mimicking Clean Diffusion Model

Supplementary Material

Sampling	Time(s)
	268
√	156

Table 6. Alabation studying of time cost for purifying one image in ImageNet, where sampling is the sampling strategy.

7. Extra Experiment

7.1. ImageNet

To further prove the validity of MimicDiffusion, we report the extra experimental results on ImageNet [29] shown in Table. 7 following the experimental setting in Nie *et al.* [25]. According to the results, we still get a large improvement, improving the average robust accuracy by 17.64%

7.2. Additional Attack Method

We report the additional experimental results on PGD and C&W attack based on CIFAR-10 compared with the latest works for adversarial purification, including GDPM, [25], and [22]. To make sure of a fair comparison, we test our method following the advice of [22], and the results are shown in Table. 8 and Table. 9. It can be noticed that our method achieves the best performance under average robust accuracy against PGD and C&W attacks. Concretely, when against the C&W attack, we improved almost 35.76% average robust accuracy in the worst condition. When against the PGD attack, we improved almost 1% average robust accuracy. Meanwhile, MimicDiffusion significantly reduces the gap between standard accuracy and robust accuracy. In this way, we prove that the proposed MimicDiffusion could reduce the negative influence of adversarial perturbation and avoid adding too much extra noise. Meanwhile, the performance of MimicDiffusion is stable against different attack methods based on the same setting.

7.3. Surrogate Process

Following the advice of Li *et al.* [22], we report the experimental results based on the surrogate process and focus on the most effective attacks: PGD-EOT and BPDA+EOT compared with Li shown in Table. 10-12. MimicDiffusion still achieves the SOTA results.

7.4. Visualization

We report the purified images shown in Fig. 3 to show the purification ability. It can be noticed that MimicDiffusion could successfully purify the adversarial perturbation and keep the label semantic as much as possible.

7.5. Time Cost

To further prove the necessity of the sampling strategy, we make an ablation study for the time cost of using the sampling strategy shown in Table. 6. It can be noticed that the sampling strategy reduces half of the time cost compared with implementing the guided method in the entire reverse process.

Table 7. Standard accuracy and robust accuracy against AutoAttack $\ell_\infty(\epsilon = 8/255)$ on ImageNet

Method	Classifier	Standard Accuracy(%)	Robust Accuracy(%)
Wang *et al.* [6]	ResNet50	62.56	31.06
Wong *et al.* [40]	ResNet50	55.62	26.95
Salman *et al.* [30]	ResNet50	64.02	37.89
Bai *et al.* [2]	ResNet50	67.38	35.51
Nie *et al.*	ResNet50	**68.22**	43.89
MimicDiffusion (Our)	ResNet50	66.92 $\pm$ 10.44	**61.53** $\pm$ **9.7**

Table 8. Standard accuracy and robust accuracy against PGD $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10

Method	Classifier	Standard Accuracy(%)	Robust Accuracy(%)
Nie *et al.*	WideResNet-70-16	91.03	57.69
Wang *et al.* [22]	WideResNet-70-16	90.67	63.52
MimicDiffusion (Our)	WideResNet-70-16	**92.05** $\pm$ 6.02	**91.55** $\pm$ 6.84
GDPM	WideResNet-28-10	93.50	90.10
Nie *et al.*	WideResNet-28-10	91.00	54.92
Wang *et al.* [22]	WideResNet-28-10	90.70	62.15
MimicDiffusion (Our)	WideResNet-28-10	**91.93** $\pm$ 6.00	**91.88** $\pm$ 6.01

Table 9. Standard accuracy and robust accuracy against C&W Attack $\ell_2(\epsilon = 8/255)$, EOT $= 50$ on CIFAR-10

Method	Classifier	Standard Accuracy(%)	Robust Accuracy(%)
Nie *et al.*	WideResNet-70-16	**92.35**	47.00
MimicDiffusion (Our)	WideResNet-70-16	91.85 $\pm$ 6.46	**91.67** $\pm$ **6.49**
GDPM	WideResNet-28-10	21.30	21.71
Nie *et al.*	WideResNet-28-10	**93.53**	47.65
MimicDiffusion (Our)	WideResNet-28-10	90.34 $\pm$ 6.2	**89.91** $\pm$ **6.5**

Table 10. Standard accuracy and robust accuracy against PGD+EOT on CIFAR-10 by using the surrogate process recommended by Lee *et al.* [22] with WideResNet-28-10 (WRN-28-10). Our (Bicubic) represents using Bicubic super-resolution operation, and Our (Bilinear) represents using Bilinear super-resolution operation. $*$ This method uses ResNet-110 as the classifier.

Method	Standard	$\ell_\infty(\epsilon = 8/255)$	$\ell_2(\epsilon = 0.5)$
Handi *et al.* [15]$*$	67.38	24.41	29.30
Zhang *et al.* [46]$*$	72.23	20.51	33.01
Lee *et al.*	90.53 $\pm$ 0.14	56.88 $\pm$ 1.06	83.75 $\pm$ 0.99
Our (Bicubic)	**91.41** $\pm$ **1.12**	**80.86** $\pm$ **1.48**	**89.26** $\pm$ **1.13**
Our (Bilinear)	**92.01** $\pm$ **1.20**	**81.16** $\pm$ **1.73**	**88.50** $\pm$ **1.45**

Table 11. Standard accuracy and robust accuracy against PGD+EOT $\ell_\infty(\epsilon = 4/255)$ on ImageNet using the surrogate process with ResNet-50.

Method	Standard	Robust
Lee *et al.*	**67.21**	44.14
Our	62.25 $\pm$ 3.87	**51.14** $\pm$ **5.15**

Table 12. Standard accuracy and robust accuracy against BPDA+EOT$\ell_\infty(\epsilon = 8/255)$ on CIFAR-10 using WRN-28-10.

Method	Standard	Robust
Lee *et al.*	90.16 ± 0.64	88.40 ± 0.88
Our	$\mathbf{92.97 \pm 0.81}$	$\mathbf{91.41 \pm 2.01}$

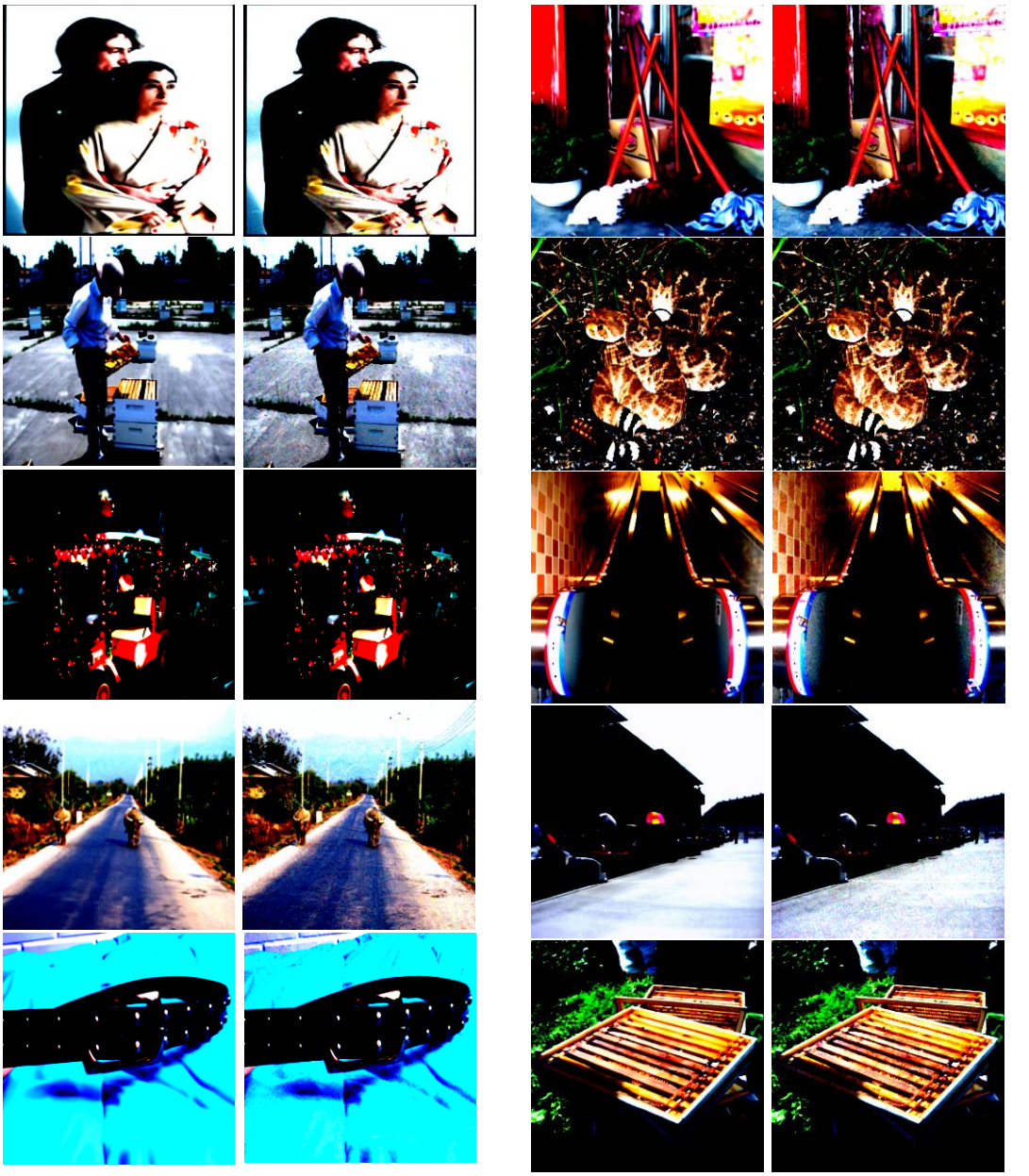

Figure 3. Visualization of MimicDiffusion against AutoAttack $\ell_\infty(\epsilon = 8/255)$