Robust Evaluation of Diffusion-Based Adversarial Purification

Minjong Lee
CSED POSTECH
minjong.lee@postech.ac.kr

Dongwoo Kim
CSED, GSAI POSTECH
dongwoo.kim@postech.ac.kr

Abstract

We question the current evaluation practice on diffusion-based purification methods. Diffusion-based purification methods aim to remove adversarial effects from an input data point at test time. The approach gains increasing attention as an alternative to adversarial training due to the disentangling between training and testing. Well-known white-box attacks are often employed to measure the robustness of the purification. However, it is unknown whether these attacks are the most effective for the diffusion-based purification since the attacks are often tailored for adversarial training. We analyze the current practices and provide a new guideline for measuring the robustness of purification methods against adversarial attacks. Based on our analysis, we further propose a new purification strategy improving robustness compared to the current diffusion-based purification methods.

1. Introduction

Adversarial attacks [21, 6] can cause deep neural networks (DNNs) to produce incorrect outputs by adding imperceptible perturbations to inputs. While various adversarial defenses have been proposed, adversarial training [41, 13] has shown promising results in building robust DNNs. Since adversarial training feeds the model both normal and adversarial examples during training time, one needs to pre-determine which attack method is used to generate the adversarial examples. On the other hand, adaptive test-time defense [19, 16] has recently gained increasing attention since it adaptively removes the adversarial effect at test time without adversarial training. *Adversarial purification* [31, 24], one of the adaptive test-time defenses, uses generative models to restore the clean examples from the adversarial examples.

Diffusion-based generative models [17, 35] has been suggested as a potential solution for adversarial purification [24, 36]. Diffusion models learn transformations from data distributions to well-known simple distributions such as the Gaussian and vice versa through forward and reverse processes, respectively. When applied to the purification, the forward process gradually adds noise to the input, and the reverse process gradually removes the noises to uncover the original image without imperceptible adversarial noise. With a theoretical guarantee, the recent success of DiffPure [24] against many adversarial training methods shows the potential of using diffusion processes for improving the robustness against adversarial attacks.

Evaluating the robustness of adaptive test-time defenses is, however, known to be difficult due to their complex defense algorithms and properties. Croce et al. [8] shows that finding worse-case perturbation is important to measure the robustness of the defenses. Randomness and iterative calls of adaptive test-time defenses, however, make their gradients obfuscated. Therefore, the gradient-based attack methods [21, 6] might be inappropriate for measuring the robustness under such obfuscation. New algorithms, such as Backward Pass Differentiable Approximation (BPDA) [3], and additional recommendations [8] have been proposed to evaluate their robustness accurately. However, it is unclear whether these algorithms and recommendations can still be used to evaluate diffusion-based purification.

In the first part of this work, we analyze the existing evaluation methods for diffusion-based purification. We find that the adjoint method, often used to compute a full gradient of the iterative process, relies on the performance of an underlying numerical solver. Tailored to the diffusion models, we propose a surrogate process, an alternative method to approximate the gradient from the iterative procedure and show the strong robustness in recent work can be weaker than claimed with the surrogate process. We then compare two gradient-based attack methods, AutoAttack [6] and PGD [21], with the surrogate process and find that PGD is more effective for the diffusion-based purification. To this end, we propose a practical recommendation to evaluate the robustness of diffusion-based purification.

In the second part of this work, we analyze the importance of the hyperparameter for successive defenses with purification. The diffusion models are trained without adversarial examples. Thus, proper validation of hyperparameters is impossible in general. Instead, we empirically an-

alyze the influence of different hyperparameter selections from the attacker's and defender's perspectives. Based on our analysis, we propose a gradual noise-scheduling for multi-step purification. We show that our defense strategy highly improves robustness compared to the current diffusion-based purification methods under our proposed evaluation scheme.

We summarize our contributions as follows:

- We analyze the current evaluation of diffusion-based purification and provide a recommendation for robust evaluation.
- We investigate the influence of hyperparameters on the robustness of diffusion-based purification.
- We propose a gradual noise-scheduling strategy for diffusion-based purification, improving the robustness of diffusion-based purification.

2. Preliminary

We provide the background on the adversarial attack, diffusion models, and adversarial purification in this section.

2.1. Adversarial Attacks

Adversarial attacks aim to manipulate or trick machine learning models by adding imperceptible perturbations to input data that can cause the model to misclassify or produce incorrect outputs. The adversarial attacks can be categorized into black-box, grey-box, and white-box attacks. The black-box attack assumes that the attacker knows nothing about the internal structure of the classifier and defender. The white-box attack assumes that the attacker can obtain any information about the defender and the target classifier, including the architecture and parameter weights. The grey-box lies between the white- and black-box attacks, where the attacker partially knows the target model. In this work, we only focus on the performance of purification in the white-box attack since the white-box attack is the most difficult to defend from the defender's perspective.

The Projected Gradient Descent (PGD) [21] method is a common white-box attack. PGD is a gradient-based attack that iteratively updates an adversarial example using the following rule

$$\mathbf{x}_{i+1} = \Pi_{\mathcal{X}} \left(\mathbf{x}_i + \alpha_i \text{sign} \nabla_{\mathbf{x}} \mathcal{L}(f_\phi(\mathbf{x}), y)|_{\mathbf{x}=\mathbf{x}_i} \right), \quad (1)$$

where f_ϕ represents a classifier, and $\Pi_{\mathcal{X}}$ indicates a projection operation onto $\mathcal{X}$. PGD can only be applied for the differentiable defense methods. For non-differentiable defense methods, the Backward Pass Differentiable Approximation (BPDA) [3] is widely used, which computes the gradient of the non-differentiable function by using a differentiable approximation. Expectation over Transformation (EOT) [2] can be additionally employed for randomized defenses, which optimizes the expectation of the randomness. AutoAttack [6] is an ensemble of four different types of attacks. In this work, we measure the robustness of purification methods against these attack methods.

2.2. Diffusion Models

Recently, diffusion-based models [17, 35] have gained increasing attention in generative models. Unlike the VAEs and GANs, the diffusion-based models produce samples by gradually removing noise from random noise. The training of diffusion-based models consists of two processes, the forward process, and the reverse denoising process. The forward process adds Gaussian noise over T steps to the observed input $\mathbf{x}_0$ with a predefined variance scheduler β_t, whose joint distribution is defined as

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad (2)$$

where $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ is a Gaussian transition kernel from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}). \quad (3)$$

The reverse process denoises the random noise $\mathbf{x}_T$ over T times, whose joint distribution is defined as

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t). \quad (4)$$

The transition distribution from $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$ is often modeled by Gaussian distribution

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}), \quad (5)$$

where σ_t is a variance, and $\boldsymbol{\mu}_\theta$ is a predicted mean of $\mathbf{x}_{t-1}$ derived from a learnable denoising model $\boldsymbol{\epsilon}_\theta$. The denoising model is often trained by predicting a random noise at each time step via following objective

$$L(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2 \right], \quad (6)$$

where $\boldsymbol{\epsilon}$ is a Gaussian noise, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. The model $\boldsymbol{\epsilon}_\theta$ takes the noisy input $\mathbf{x}_t$ and the time step t to predict the actual noise $\boldsymbol{\epsilon}$ at time t. In the Denoising Diffusion Probabilistic Model (DDPM) [17], the reverse denoising process is performed over T steps through random sampling, resulting in a slower generation of samples compared with GANs and VAEs.

Based on the fact that the multiple denoising steps can be performed at a single step via a non-Markovian process, Song et al. [33] proposes a new sampling strategy, which we call Denoising Diffusion Implicit Model (DDIM) sampler, to accelerate the reverse denoising process. In this work, we compare the performances of DDPM and DDIM samplers in the diffusion-based purification approach.

2.3. Adversarial Purification

Adversarial purification via generative models is a technique used to improve the robustness of machine learning models against adversarial attacks [31]. The idea behind this technique is to use a generative model to learn the underlying distribution of the clean data and use it to purify the adversarial examples.

Diffusion-based generative models can be used as a purification process if we assume that the imperceptible adversarial signals as noise [24]. To do so, the purification process adds noise to the adversarial example via the forward process with t^* steps, and it removes noises via the denoising process. The choice of the number of forward steps t^* is essential since too much noise can remove the semantic information of the original example, or too little noise cannot remove adversarial perturbation. In theory, as we add more noise to the adversarial example, the distributions over the noisy adversarial example and the true example become close to each other [24]. Therefore, the denoised examples are likely to be similar.

3. Evaluation for Diffusion-Based Purification

In this section, we first review the current practices in evaluating diffusion-based purification methods. We then curate three research questions to address their potential limitations and provide our answers to these questions through empirical evaluations.

3.1. Current Practices and Research Questions

Evaluation of the diffusion-based purification against gradient-based white-box attacks is non-trivial due to many function calls on the denoising process. Multiple function calls in the denoising step often require an impractical amount of memory, making it unfeasible to compute the gradient of the full defense process. Because of this problem, most defenses [39, 36, 16] consider BPDA the strongest adaptive white-box attack in the current practice since it does not rely on the gradients of defense methods. However, the vulnerability of diffusion-based purification on white-box attacks has yet to be fully identified. The importance of testing in adaptive white-box attacks of purification has been recognized only recently by the work of DiffPure [24].

DiffPure calculates the full gradients of their defense process using an adjoint method. The adjoint method is employed to avoid the extensive use of memory while obtaining the full gradient. DiffPure is evaluated on AutoAttack, a de facto evaluation method in adversarial training. Although their evaluation framework is more robust than the previous work, the design choices of their evaluation still raise questions since 1) the adjoint method relies on the performance of an underlying numerical solver [42], and 2) there is no comprehensive comparison between different attacks using the full gradient.

Based on our observation, we carefully curate the following three research questions to address the robustness of the current evaluation framework in diffusion-based purification:

- **RQ1.** Is the adjoint method the best way to generate adversarial examples with full gradients? Is there any alternative to the adjoint method?
- **RQ2.** Is AutoAttack still better than the other gradient-based attacks, such as PGD, when the alternative is available?
- **RQ3.** Is BPDA still more effective than the best combination of full-gradient attacks?

In the next section, we re-evaluate the existing purification methods to answer these questions.

3.2. Experimental Results & Analysis

We evaluate the performance of three diffusion-based purification methods: ADP[1] [39], DiffPure [24], GDMP [36]. We additionally evaluate two non-diffusion-based adaptive test-time defenses: SODEF [19], and DISCO [16] to address whether our findings still hold for the non-diffusion-based purification methods. We evaluate their robustness on CIFAR-10 against three gradient-based attacks, including PGD, BPDA, and AutoAttack, with a maximum attack strength of $\ell_\infty(\epsilon = 8/255)$. A comprehensive description of evaluation configurations is provided in Appendix A.

Surrogate process and its gradient. The adjoint method can compute the exact gradient in theory, but in practice, the adjoint relies on the performance of the numerical solver, whose performance becomes problematic in some cases as reported by Zhuang et al. [42]. To answer the RQ1, we compare the adjoint method against the full gradient obtained from back-propagation if possible, and if not due to the memory issue, we use the approximated gradient obtained from a *surrogate process*. The surrogate process utilizes the fact that given the total amount of noise, we can denoise the same amount of noise with different numbers of denoising steps [33]. Therefore, instead of using the entire denoising steps, we can mimic the original denoising process with fewer function calls, whose gradients can be obtained by back-propagating the forward and denosing process directly.

The gradients obtained from the surrogate process differ from the exact gradients. However, if the accumulated

[1]Although ADP uses a score-based model and Langevin dynamics, since the concept is similar to the diffusion model, we consider ADP diffusion-based purification.

Defense	Gradient of Def	Robust Accuracy (%)
DiffPure [24]	Adjoint	74.38±1.03
	Surrogate	46.84±1.44
GDMP [36]	BPDA	75.59±1.26
	Surrogate	24.06±0.47
SODEF [19]	w/o	53.69
	Adjoint	57.76
	Full	49.28

Table 1: Robust accuracy of DiffPure, GDMP, and SODEF against attacks ($\ell_\infty(\epsilon = 8/255)$) on CIFAR-10. We use PGD+EOT for DiffPure and GDMP and AutoAttack for SODEF. *Adjoint* calculates full gradients using the adjoint method, and *Surrogate* (or *Full*) calculates approximated (or full) gradients using direct back-propgation. *w/o* is the performance of the underlying classifier without the SODEF.

denoising steps can be approximated with fewer denoising steps, we can use the approximated gradients as a proxy of the exact gradients. The surrogate process can also relax the randomness occurring in multiple denoising steps.

RQ1: Is the adjoint method the best way to generate adversarial examples with full gradients? We compare the adjoint method with the full gradient obtained from direct back-propagation of the defense process with the original or surrogate processes.

Table 1 shows that the robust accuracy of DiffPure [24] with the direct back-propagation is 46.84% on PGD+EOT attack, which is 27.54% lower than the reported accuracy with the adjoint method. The results show that direct back-propagation is more effective than the adjoint method. Furthermore, we use a surrogate process for GDMP [36], and the robust accuracy is 24.06%, which is 51.53% lower than the reported accuracy against the BPDA attack. It can be concluded that, in cases where the gradients of the defense process are unavailable to calculate, the surrogate process can be an alternative to generate adversarial examples.

SODEF [19], a non-diffusion-based purification, originally uses the adjoint method to generate adversarial examples. We evaluate SODEF with direct back-propagation for the attack and observe 49.28% robust accuracy against AutoAttack, which is lower than 53.69% of the underlying model without defense. This result suggests that the use of a numerical solver would not be effective for the adversarial attack.

RQ2: Is AutoAttack still better than the other gradient-based attacks, such as PGD, when the alternative is available? AutoAttack has recently been used as a standard method to evaluate defenses due to its robustness against defenses. Although AutoAttack may not be an ideal choice for randomized defenses[2], still many purification methods, such as DiffPure, rely on AutoAttack. However, as shown in Table 2, AutoAttack has a lower success rate than PGD+EOT against diffusion-based purification methods. For the ℓ_∞ threat model ($\epsilon = 8/255$), PGD+EOT shows 16.76% and 26.05% more attack success rate than AutoAttack against DiffPure and ADP, respectively. We observe a similar result with the ℓ_2 threat model ($\epsilon = 0.5$). Therefore, evaluation with PGD+EOT for diffusion-based purification can be useful to evaluate their robustness. Further results of the difference between PGD+EOT against DiffPure with additional settings can be found in Appendix B.

Threat Model	Defense	Attack	Robust Accuracy (%)
$\ell_\infty(\epsilon = 8/255)$	ADP [39]	PGD+EOT	33.48±0.86
		AutoAttack	59.53±0.87
	DiffPure [24]	PGD+EOT	46.84±1.44
		AutoAttack	63.60±0.81
$\ell_2(\epsilon = 0.5)$	ADP [39]	PGD+EOT	73.32±0.76
		AutoAttack	79.57±0.38
	DiffPure [24]	PGD+EOT	79.45±1.16
		AutoAttack	81.70±0.84

Table 2: Robust accuracy of DiffPure and ADP against PGD+EOT and AutoAttack ($\ell_\infty(\epsilon = 8/255)$) on CIFAR-10.

Defense	Type	BPDA	Ours
ADP [39]	DSM+LD	66.91±1.75	33.48±0.86
DiffPure [24]	Diffusion	81.45±1.51	46.84±1.44
GDMP [36]	Diffusion	75.59±1.26	24.06±0.47
DISCO [16]	Implicit function	47.18	0.00

Table 3: Robust accuracy of defenses against BPDA and our full-gradient based attacks ($\ell_\infty(\epsilon = 8/255)$) on CIFAR-10. We report the lowest robust accuracy between PGD and AutoAttack.

RQ3: Is BPDA still more effective than the best combination of full-gradient attacks? BPDA [3] has been widely used to evaluate defenses that can cause gradient obfuscation. Because multiple function calls can cause gradient obfuscation, ADP, GDMP, and DISCO have been evaluated on BPDA as the strongest adaptive white-box attack. However, our evaluation shows that BPDA has a

[2] https://github.com/fra31/auto-attack/blob/master/flags_doc.md

lower attack success rate than the attacks using direct gradients of the defense process, as shown in Table 3. Against PGD+EOT using direct gradients of defense process, ADP and GDMP show robust accuracy of 33.48% and 24.06%, respectively, significantly lower than the reported accuracy with BPDA [39, 36]. DISCO even has 0% robust accuracy. From the results, we suggest that the direct gradients of the defense process need to be tested to check the robustness.

Recommendation. We propose an overall guideline for evaluating diffusion-based purifications as follows. We recommend using PGD+EOT rather than AutoAttack. When calculating gradients, it is best to directly back-propagate the full defense process. If this is unavailable due to memory constraints, using the surrogate process rather than the adjoint method is recommended. Note that our recommendation generally follows the suggestions made by Croce et al. [8] but is more tailored for the diffusion-based purification.

4. Analysis of Hyperparameters

The performance of diffusion-based purification is significantly influenced by varying hyperparameter configurations. In this section, we explore the importance of hyperparameters in defense processes.

4.1. Experimental Settings

Understanding the importance of hyperparameters can help build a better defense mechanism. We investigate the effect of various hyperparameters of diffusion-based purification methods to determine the most robust configuration for the adaptive attack. Specifically, the following three hyperparameters are evaluated 1) the number of forward steps, 2) the number of denoising steps, and 3) the number of purification steps. In addition, we re-evaluate the efficiency of several techniques proposed in previous works under our defense scheme.

We evaluate the purification against PGD+EOT on CIFAR-10. We provide the additional results on CIFAR-10 and ImageNet in Appendix C. Although we do not report ImageNet results in the main text, the overall findings are similar to those from CIFAR-10. We use a naturally pretrained WideResNet-28-10 [40] as an underlying classifier provided by Robustbench [7]. For a diffusion model, we use pretrained DDPM++ [35]. The variances for the diffusion model are linearly increasing from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ when $T = 1000$ [17]. We use two different denoising models: DDPM [17] and DDIM [33].

For all experiments, we report the mean and standard deviation over five runs to measure the standard and robust accuracy. PGD uses 200 update iterations. 20 samples are used to compute EOT. Following the settings in DiffPure [24], we use a fixed subset of 512 randomly sampled images. To calculate gradients, we use direct gradients of the entire process. If impossible, we compute the approximated gradients from a surrogate process. In each experiment, we explain the defense process and the surrogate process in more detail.

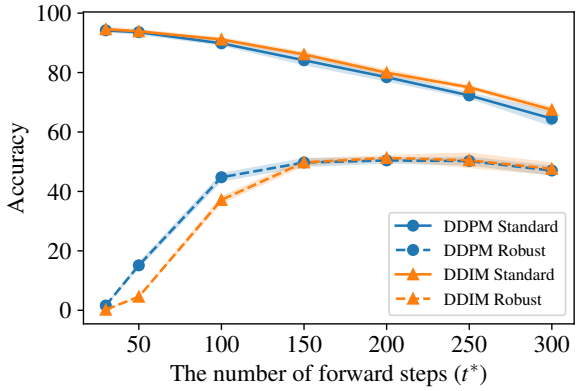

Figure 1: Standard and robust accuracy as we change the number of forward steps against PGD+EOT $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10. Five denoising steps for both attack and defense are used.

4.2. The Number of Forward Steps

We explore the effect of forward noising steps on robustness by varying the number of forward steps from 30 to 300, resulting in the changes of total variance ranged from 0.012 to 0.606. The same number of forward steps are used for both attack and defense, and we set five denoising steps for attack and defense for all experiments.

As shown in Figure 1, the standard accuracy continuously decreases as the number of forward steps increases since more forward steps induce more noise. The robust accuracy increases first and decreases after 200 forward steps, i.e., $t^* = 200$. When the number of forward steps is small, the DDPM is more robust than the DDIM. However, DDIM shows better accuracy for both standard and robust than DDPM after 200 forward steps.

4.3. The Number of Denoising Steps

Defenders may use fewer denoising steps to accelerate the defense process. From the other perspective, attackers may want to use fewer denoising steps than those used in the defense due to memory constraints. We explore the influence of the number of denoising steps through the following three experimental settings:

(a) The number of denoising steps in attack is set to five, and the number of denoising steps in defense is ranged from one to the maximum number of denoising steps.

(b) The number of denoising steps in both the attack and defense are the same, ranging from one to 20.

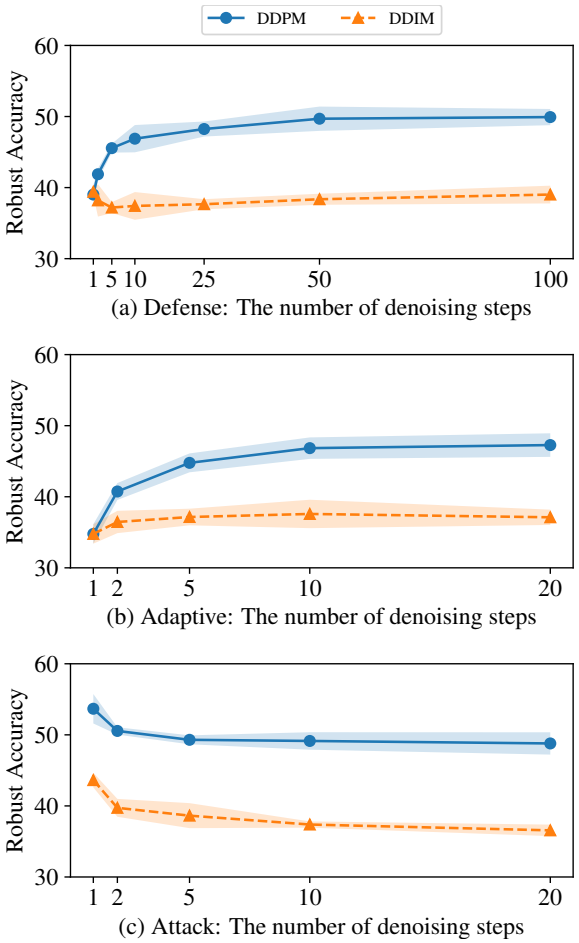

Figure 2: Robust accuracy as we change the number of denoising steps against PGD+EOT $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10. We change the number of denoising steps in (a) defense, (b) both, and (c) attack for each experiment with the other hyperparameters fixed.

(c) The number of denoising steps in defense is set to the maximum number of denoising steps, and the number of denoising steps in attack is ranged from one to 20.[3]

The results are displayed in Figure 2. From the defense perspective, the results of (a) and (b) demonstrate that more denoising steps can improve robustness. DDPM gains more advantage from having more denoising steps than DDIM. (c) shows the effect of the number of denoising steps in the attack. As the number of denoising steps increases, the attack success rate slightly increases. However, we also find that increasing the number of denoising steps in an attack can decrease the attack success rate when the number of forward steps is 200 (i.e., $t^* = 200$).

[3]The 20 denoising steps is the maximum limit of 40GB of memory.

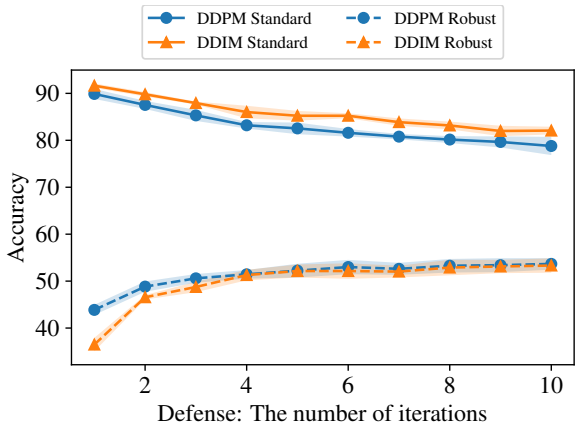

Figure 3: The number of purification steps in defense and its influence to the standard and robust accuracy against PGD+EOT $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10. The number of forward steps is 100 (i.e., $t^* = 100$). The reported robust accuracy is the lowest performance of all settings of the number of purification steps of the attack.

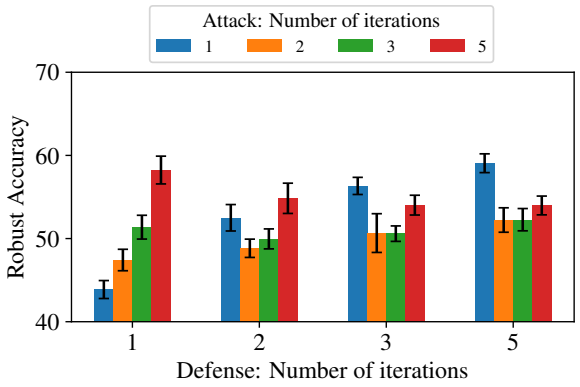

Figure 4: The number of purification steps during attacks and its influence on the robust accuracy against PGD+EOT $\ell_\infty(\epsilon = 8/255)$ with CIFAR-10. The number of forward steps is fixed as 100 (i.e., $t^* = 100$).

4.4. The Number of Purification Steps

Although a single forward and reverse process can purify the input image, one can apply the purification process multiple times as proposed in Wang et al. [36]. We denote the number of forward and denoising processes as the number of *purification step*. Similar to the case of the denoising step, computing the gradients of multiple purification steps is impossible due to memory constraints in most cases.

The number of purification steps can also differ between attack and defense. Through experiments, we measure the changes in robust accuracy with the different number of purification steps in the defense and attack. For all experiments, we fixed the number of forward steps to 100

Guidance	Accuracy (%)		
	Standard	BPDA	PGD+EOT
No guide	87.70±0.46	75.23±0.61	38.44±0.59
MSE	89.96±0.40	75.59±1.26	24.06±0.47
SSIM	93.75±0.39	74.02±1.17	6.88±0.21

Table 4: Standard and robust accuracy of GDMP [36] against BPDA and PGD+EOT $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10. We compare two types of guidance, MSE and SSIM, and the defense without guidance.

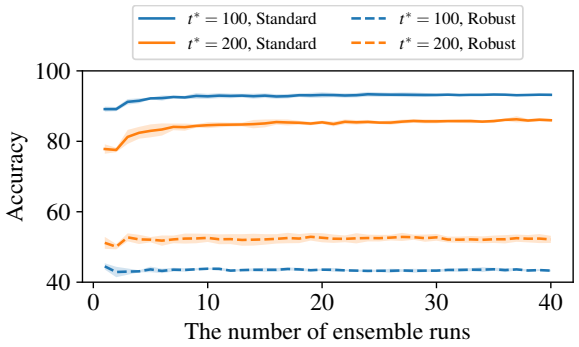

Figure 5: Standard and robust accuracy against PGD+EOT $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10 when using an ensemble with a different number of purification runs. Five denoising steps are used for the surrogate process of the attack.

($t^* = 100$), and the number of denoising steps is set to five.

Figure 3 shows the standard and robust accuracy with a varying number of purification steps in defense. The robust accuracy increases as the number of purification steps increases while the standard accuracy steadily decreases. Figure 4 shows the effect of the number of purification steps in the attack. When the number of purification steps in defense is one or two, the same number of purification steps in attack is the most effective. However, as we set the number of purification steps in defense to three and five, two and three purification steps in attack show a better attack success, respectively.

4.5. Other Techniques

We evaluate several other techniques proposed in earlier work [39, 24, 36] within our new evaluation framework.

Guidance. GDMP [36] proposes to use gradients of a distance between an original input example and a target example to preserve semantic information while denoising. They show guidance can improve robustness against preprocessor-blind attacks. However, as shown in Table 4, when the gradients of the surrogate process are used in the attack, the guidance of GDMP decreases the robust accuracy. Specifically, the defense with guidance using the SSIM similarity has 6.88% robust accuracy, which is 31.56% lower than the defense without guidance.

Underlying Classifier	t^*	Robust Accuracy (%)
TRADES [41]	0	55.32
	100	54.02±0.98
	200	51.52±1.96
Gowal et al. [13]	0	69.03
	100	58.24±0.49
	200	52.97±1.38

Table 5: Combination of diffusion models with adversarial training evaluated on CIFAR-10 against PGD+EOT $\ell_\infty(\epsilon = 8/255)$. Five denoising steps are used for both attack and defense.

t^*	Defense	Attack	Robust Accuracy (%)
100	DDPM	DDPM	44.77±1.48
		DDIM	46.68±1.25
	DDIM	DDPM	40.51±1.01
		DDIM	37.15±1.31
200	DDPM	DDPM	50.43±1.11
		DDIM	52.15±1.88
	DDIM	DDPM	53.63±1.11
		DDIM	51.29±1.00

Table 6: Robust accuracy against PGD+EOT $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10 when using a denoising model in attack different from the denoising model in defense.

Ensemble of multiple purification runs. ADP [39] uses the ensemble of multiple purification runs as the predicted label to mitigate the randomness in the defense. For diffusion-based purification methods, as shown in Figure 5, multiple purification runs especially can help improve standard accuracy while the robust accuracy keeps the same level. In particular, for $t^* = 200$ with 40 purification runs, standard accuracy is 8.17% higher than the case without ensemble.

Combination with adversarial training. An adjoint method based DiffPure [24] shows robustness can be improved by using diffusion models together with adversarial training. However, as shown in Table 5, the adversarial training with purification shows lower robustness than the classifier without purification.

Transferability of gradients from different samplers in the attack. One may employ a sampler of diffusion mod-

	Type	Method	Standard	PGD	AutoAttack
WRN-28-10	AT	Gowal et al. [13]	87.51	66.01	63.38
		Gowal et al. [12]*	88.54	65.93	62.76
		Pang et al. [25]	88.62	64.95	61.04
	AP	Yoon et al. [39]	85.66±0.51	33.48±0.86	59.53±0.87
		Nie et al. [24]	90.07±0.97	46.84±1.44	63.60±0.81
		Ours	90.16±0.64	55.82±0.59	70.47±1.53
WRN-70-16	AT	Rebuffi et al. [29]*	92.22	69.97	66.56
		Gowal et al. [13]	88.75	69.03	66.10
		Gowal et al. [12]*	91.10	68.66	65.87
	AP	Yoon et al. [39]	86.76±1.15	37.11±1.35	60.86±0.56
		Nie et al. [24]	90.43±0.60	51.13±0.87	66.06±1.17
		Ours	90.53±0.14	56.88±1.06	70.31±0.62

	Type	Method	Standard	PGD	AutoAttack
WRN-28-10	AT	Rebuffi et al. [29]*	91.79	85.05	78.80
		Augustin et al. [4]†	93.96	86.14	78.79
		Sehwag et al. [32]†	90.93	83.75	77.24
	AP	Yoon et al. [39]	85.66±0.51	73.32±0.76	79.57±0.38
		Nie et al. [24]	91.41±1.00	79.45±1.16	81.70±0.84
		Ours	90.16±0.64	83.59±0.88	86.48±0.38
WRN-70-16	AT	Rebuffi et al. [29]*	95.74	89.62	82.32
		Gowal et al. [12]*	94.74	88.18	80.53
		Rebuffi et al. [29]	92.41	86.24	80.42
	AP	Yoon et al. [39]	86.76±1.15	75.66±1.29	80.43±0.42
		Nie et al. [24]	92.15±0.72	82.97±1.38	83.06±1.27
		Ours	90.53±0.14	83.75±0.99	85.59±0.61

Table 7: Standard and robust accuracy against PGD+EOT (left: $\ell_\infty(\epsilon = 8/255)$, right: $\ell_2(\epsilon = 0.5)$) on CIFAR-10. Adversarial Training (AT) and Adversarial Purification (AP) methods are evaluated. † This method uses WideResNet-34-10 as a classifier. * This method is trained with extra data.

Type	Method	Accuracy (%) Standard	Robust
AT	Salman et al. [30]	63.86	39.11
	Engstrom et al. [11]	62.42	33.20
	Wong et al. [37]	53.83	28.04
AP	Nie et al. [24]	75.48±9.18	38.71±0.96
	Ours	66.21±1.00	43.05±1.09

Table 8: Standard and robust accuracy against PGD+EOT $\ell_\infty(\epsilon = 4/255)$ on ImageNet. ResNet-50 is used as a classifier.

Type	Defense	Accuracy (%) Standard	Robust
AT	Rade and Moosavi-Dezfooli [27]	93.08	52.83
	Gowal et al. [12]	92.87	56.83
	Gowal et al. [13]	94.15	60.90
AP	Nie et al. [24]	97.85±0.53	34.30±0.41
	Ours	95.55±0.40	49.65±1.06

Table 9: Standard and robust accuracy against attacks $\ell_\infty(\epsilon = 8/255)$ on SVHN. Adversarial training methods are evaluated on AutoAttack, and adversarial purification methods are evaluated on PGD+EOT. WideResNet-28-10 is used as a classifier except for Rade and Moosavi-Dezfooli [27], which uses ResNet-18.

els to generate adversarial examples different from those used in defense. For example, attacks using gradients from DDPM could be transferred to the defense using DDIM and vice-versa. We test whether the gradients from a different sampler of denoising models can improve the attack success rate. As shown in Table 6, although the transferred attack is valid, the attack success rates using different samplers are slightly lower than those using the original samplers.

5. Gradual Noise-Scheduling for Multi-Step Purification

In this section, we propose a new sampling strategy for diffusion-based purification and compare the performance with other state-of-the-art defenses.

Gradual noise-scheduling strategy. As highlighted in Section 4, selecting appropriate hyperparameter values is essential to improve robustness. Thus, we conduct an extensive exploration of hyperparameter settings to maximize robust accuracy. In particular, we mainly focus on the fact that each purification step can contain a different number of forward steps. We empirically find that fewer forward steps in the first few purification steps can improve robustness.

Based on this observation, for CIFAR-10, we set the number of forward steps as $\{30 \times 4, 50 \times 2, 125 \times 2\}$ for eight purification steps. For ImageNet and SVHN, we set the number of forward steps as $\{30 \times 4, 50 \times 2, 200 \times 2\}$ and $\{30 \times 4, 50 \times 2, 80 \times 2\}$, respectively. We set the number of denoising steps to equal the number of forward steps for all purification steps. We use the DDPM and an ensemble of ten purification runs.

Experimental settings. We conduct evaluations on three datasets, CIFAR-10 [20], ImageNet [9], and SVHN [23]. We use three diffusion model architectures, DDPM++ [35], Guided Diffusion [10], and DDPM [17] for each dataset. We use pretrained models for CIFAR-10 and ImageNet, but we trained a model for SVHN. Pretrained WideResNet-28-10, WideResNet-70-16, and ResNet-50 [40, 14] are served as baseline classifiers. We compare our method with adversarial training and diffusion-based purification methods. We evaluate diffusion-based purification methods on the PGD+EOT attack with 200 update iterations, except for ImageNet, which uses 20 iterations. We set the number of EOT to 20. For ad-

Method	Purification	Accuracy (%)	
		Standard	Robust
Song et al. [34]	Gibbs Update	95.00	9.00
Yang et al. [38]	Mask+Recon	94.00	15.00
Hill et al. [15]	EBM+LD	84.12	54.90
Yoon et al. [39]	DSM+LD	85.66±0.51	66.91±1.75
Nie et al. [24]	Diffusion	90.07±0.97	81.45±1.51
Ours	Diffusion	90.16±0.64	88.40±0.88

Table 10: Standard and robust accuracy against BPDA+EOT $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10. WideResNet-28-10 is used as the underlying classifier architecture.

versarial training methods, we use 20 update iterations for the PGD attack. For our method, we report the worst robust accuracy with surrogate processes. We explain the detailed settings in Appendix D.

Results. Table 7 shows the defense performance against $\ell_\infty(\epsilon = 8/255)$ and $\ell_2(\epsilon = 0.5)$ threat models on CIFAR-10, respectively. Our method outperforms other diffusion-based purification methods. Specifically, compared to DiffPure on ℓ_∞ PGD attack, our method improves robust accuracy by 8.98% with WideResNet-28-10 and by 5.75% with WideResNet-70-16, respectively. Despite the improvement in robustness, the purification methods perform worse than the adversarial training methods. Table 8 shows the performance against $\ell_\infty(\epsilon = 4/255)$ threat model on ImageNet. Our method outperforms both adversarial training and purification methods. Compared to DiffPure and Salman et al. [30], our method improves robust accuracy by 4.34% and 3.94%, respectively. Results on SVHN against threats model $\ell_\infty(\epsilon = 8/255)$ are similar with CIFAR-10. Although our method improves robust accuracy by 15.35% compared to the DiffPure framework, which uses $t^* = 0.075$, our method performs worse than the adversarial training methods.

We additionally compare the robustness of our defense strategy with other adversarial purification methods against BPDA ($\ell_\infty(\epsilon = 8/255)$). As shown in Table 10, our proposed method outperforms all other adversarial purification methods, achieving a robust accuracy of 88.40%, 6.95% greater than the robust accuracy of DiffPure. Furthermore, Table 11 shows our robustness against other attacks, including the Square attack [1], a black-box attack. Our defense shows strong robustness higher than 80% against all attacks.

6. Related Work

Adversarial training [21, 41] is one of the most successful adversarial defense methods. These methods train a classifier with adversarial examples in a training phase. Zhang et al. [41] and Pang et al. [25] propose loss functions that can effectively utilize the trade-off between robustness and accuracy. Huang et al. [18] analyze architectural factors with respect to robustness. Rebuffi et al. [28] and Gowal et al. [13] improve robustness by utilizing data augmentations.

Attack	Robust Accuracy (%)
Square [1]	89.38±0.26
FAB [5]	89.18±0.60
Deep Fool [22]	82.32±0.14
FMN Attack [26]	80.86±1.80

Table 11: Robust accuracy of our defense strategy against several threat models $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10. Square attack is a black-box attack, and the others are white-box attacks.

Adaptive test-time defenses purify adversarial examples using extra neural networks that utilize techniques from other domains. ADP [39] jointly uses denoising score matching and Langevin dynamics for purification. DiffPure [24] demonstrates from the stochastic differential equations (SDE) perspective that diffusion models can purify adversarial examples. GDMP [36] uses the guidance of diffusion models to recover adversarial examples as similar as possible to the original examples. SODEF [19] uses a Lyapunov-stable ODE block so that the input converges to a stable point that can be correctly classified. DISCO [16] is one of the denoising models that predict clean RGB value using local implicit functions.

7. Conclusion

Throughout the paper, we first analyze the current evaluation methods for diffusion-based adversarial purification and then propose a recommendation for the reliable evaluation of the robustness of adversarial purification. We further investigate the influence of hyperparameters of the diffusion model on the robustness of the purification. Based on our analysis, we propose a new strategy to maximize the benefit of the purification methods.

Acknowledgements This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01906, Artificial Intelligence Graduate School Program (POSTECH)) and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00217286) and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2021R1C1C1011375)

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. *ArXiv*, abs/1912.00049, 2019. 9

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning*, 2017. 2

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018. 1, 2, 4

[4] Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in- and out-distribution improves explainability. In *European Conference on Computer Vision*, 2020. 8

[5] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, 2019. 9

[6] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *ArXiv*, abs/2003.01690, 2020. 1, 2

[7] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 5

[8] Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, and Taylan Cemgil. Evaluating the adversarial robustness of adaptive test-time defenses. In *International Conference on Machine Learning*, pages 4421–4435. PMLR, 2022. 1, 5

[9] Jia Deng, R. Socher, Li Fei-Fei, Wei Dong, Kai Li, and Li-Jia Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, volume 00, pages 248–255, 06 2009. doi: 10.1109/CVPR.2009.5206848. 8

[10] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. 8

[11] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. 8

[12] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy A. Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *ArXiv*, abs/2010.03593, 2020. 8

[13] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy Mann. Improving robustness using generated data. In *Neural Information Processing Systems*, 2021. 1, 7, 8, 9

[14] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2015. 8

[15] Mitch Hill, Jonathan Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. *ArXiv*, abs/2005.13525, 2020. 9

[16] Chih-Hui Ho and Nuno Vasconcelos. DISCO: Adversarial defense with local implicit functions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 3, 4, 9, 12, 13

[17] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. 1, 2, 5, 8, 12

[18] Hanxun Huang, Yisen Wang, Sarah Monazam Erfani, Quanquan Gu, James Bailey, and Xingjun Ma. Exploring architectural ingredients of adversarially robust deep neural networks. In *Neural Information Processing Systems*, 2021. 9

[19] Qiyu Kang, Yang Song, Qinxu Ding, and Wee Peng Tay. Stable neural ode with lyapunov-stable equilibrium points for defending against adversarial attacks. *Advances in Neural Information Processing Systems*, 34:14925–14937, 2021. 1, 3, 4, 9, 12, 13

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. 8

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2017. 1, 2, 9

[22] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2015. 9

[23] Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. 2011. 8

[24] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *ArXiv*, abs/2205.07460, 2022. 1, 3, 4, 5, 7, 8, 9, 12, 13, 14

[25] Tianyu Pang, Min Lin, Xiao Yang, Junyi Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, 2022. 8, 9

[26] Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast minimum-norm adversarial attacks through adaptive norm constraints. In *Neural Information Processing Systems*, 2021. 9

[27] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*, 2022. 8

[28] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Data augmentation can improve robustness. In *Neural Information Processing Systems*, 2021. 9

[29] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. Fixing data augmentation to improve adversarial robustness. *ArXiv*, abs/2103.01946, 2021. 8

[30] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *ArXiv*, abs/2007.08489, 2020. 8, 9

[31] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *ArXiv*, abs/1805.06605, 2018. 1, 3

[32] Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2021. 8

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. 2, 3, 5, 12

[34] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *ArXiv*, abs/1710.10766, 2017. 9

[35] Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020. 1, 2, 5, 8, 13

[36] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *ArXiv*, abs/2205.14969, 2022. 1, 3, 4, 5, 6, 7, 9, 13

[37] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. *ArXiv*, abs/2001.03994, 2020. 8

[38] Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. Me-net: Towards effective adversarial robustness with matrix estimation. In *International Conference on Machine Learning*, 2019. 9

[39] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, 2021. 3, 4, 5, 7, 8, 9, 12

[40] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *ArXiv*, abs/1605.07146, 2016. 5, 8

[41] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019. 1, 7, 9

[42] Juntang Zhuang, Nicha C. Dvornek, Xiaoxiao Li, Sekhar Chandra Tatikonda, Xenophon Papademetris, and James S. Duncan. Adaptive checkpoint adjoint method for gradient estimation in neural ode. *Proceedings of machine learning research*, 119:11639–11649, 2020. 3