MimicDiffusion: Purifying Adversarial Perturbation via Mimicking Clean Diffusion Model

Kaiyu Song, Hanjiang Lai*, Yan Pan, Jian Yin
Sun Yat-sen University
Guangdong, China
songky7@mail2.sysu.edu.cn, {laihanj3, panyan5, issjyin}@mail.sysu.edu.cn

Abstract

Deep neural networks (DNNs) are vulnerable to adversarial perturbation, where an imperceptible perturbation is added to the image that can fool the DNNs. Diffusion-based adversarial purification uses the diffusion model to generate a clean image against such adversarial attacks. Unfortunately, the generative process of the diffusion model is also inevitably affected by adversarial perturbation since the diffusion model is also a deep neural network where its input has adversarial perturbation. In this work, we propose MimicDiffusion, a new diffusion-based adversarial purification technique that directly approximates the generative process of the diffusion model with the clean image as input. Concretely, we analyze the differences between the guided terms using the clean image and the adversarial sample. After that, we first implement MimicDiffusion based on Manhattan distance. Then, we propose two guidance to purify the adversarial perturbation and approximate the clean diffusion model. Extensive experiments on three image datasets, including CIFAR-10, CIFAR-100, and ImageNet, with three classifier backbones including WideResNet-70-16, WideResNet-28-10, and ResNet-50 demonstrate that MimicDiffusion significantly performs better than the state-of-the-art baselines. On CIFAR-10, CIFAR-100, and ImageNet, it achieves 92.67%, 61.35%, and 61.53% average robust accuracy, which are 18.49%, 13.23%, and 17.64% higher, respectively. The code is available at https://github.com/psky1111/MimicDiffusion.

1. Introduction

Deep neural networks (DNNs) have achieved great success in various fields of computer vision, e.g., image detection [31], image classification [34]. However, DNNs are vulnerable to the *adversarial samples* [10], where the adversarial sample consists of the clean sample and an imperceptible adversarial perturbation.

To defend against adversarial attack, *adversarial training* [13, 32] has been proposed by leveraging the generated adversarial samples to train the classifier. For example, Bai *et al.* [2] used the adversarial samples as the training data to train the classifier directly. However, adversarial training may be ineffective when suffering from unknown attack methods [9].

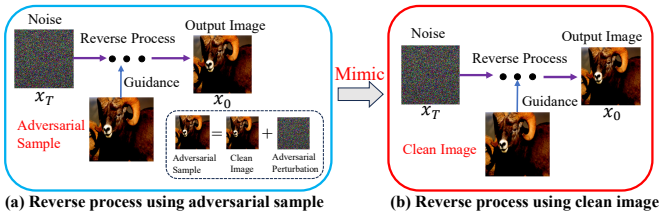
In contrast, another popular method for the adversarial attack is *adversarial purification* [21, 28, 37]. Given an adversarial sample as the input, adversarial purification methods aim to purify the adversarial perturbation from the adversarial sample and obtain clean samples. Then, generated clean samples are fed into the classifier.

As one of the popular generative models, the diffusion model [15] becomes a potential tool for adversarial purification due to generating high-quality images. Previous methods [21, 28] depended on finding an optimal time step in the forward process to cover the adversarial perturbation. Then, the reverse process tries to purify both the Gaussian noise and adversarial perturbation while keeping the label semantic simultaneously. Yong *et al.* [37] proposed the score-based method by using the property that the clear sample tends to be the lower value of the score function. Nie *et al.* [21] proposed the DiffPure based on a small step denoiser. Further, some improved methods are based on the iteration and guided methods [33]. For example, Wang *et al.* [33] alleviated the requirement for keeping the label semantic by incorporating a guidance [4].

Despite the success of diffusion-based adversarial purification methods, we argue that adversarial perturbations added to clean samples will still affect the generative process of the diffusion model, which deviates from the trajectory of the clean diffusion model. Thus, it will generate extra noise in the synthetic images and cause performance degradation (More details can be found in Section 3). The key is to remove the effects of the adversarial perturbation when performing the generative process.

As shown in Fig. 1, an intuitive approach is that if the

*Hanjiang Lai is the corresponding author.



(a) Reverse process using adversarial sample (b) Reverse process using clean image

Figure 1. An illustration of MimicDiffusion. By implementing the guidance method and using the adversarial sample (clean image + adversarial perturbation), we aim to alleviate the influence of the adversarial perturbation for the reverse process of the diffusion model such that it is similar to the reverse process with the input of the clean image.

input is not an adversarial sample but a clean image, the adversarial perturbation problem would disappear. Therefore, an interesting idea arises: *without knowing the clean inputs, can we mimic the trajectory of the diffusion model with clean inputs to reduce the effect of adversarial perturbations?*

In this work, we propose a novel MimicDiffusion to reduce the negative influence of adversarial perturbations. We use guided diffusion as a backbone, where the Gaussian noise is used as input, the adversarial sample is composed of a clean sample, and adversarial perturbation is used as guidance. The main problem is that the guided term also includes adversarial perturbation. Fortunately, adversarial perturbations are imperceptible. Under this assumption, we first use Manhattan distance (ℓ_1 distance) instead of Euclidean distance (ℓ_2 distance). Thus, we reduce the range of the derivative to +1 or -1. We further show that using Manhattan distance can be divided into two cases, short range and long range, to compare the difference in derivatives of the adversarial sample and the clean sample. Concretely, 1) when the Manhattan distance between the generated image and the clean image is larger than the maximum value in adversarial perturbations, called the *long-range distance*, we show that the gradients of the two guidance are the same. That is what we want. 2) When the Manhattan distance between the generated image and the clean image is smaller than the maximum value in adversarial perturbations, called the *short-range distance*, the gradients of the two guidance may or may not be equal.

According to the above observations, we propose two guidance: one for the long-range distance and another for the short-range distance. In the long-range guidance, we can use the adversarial sample as the guidance since the gradients are the same. In the short-range guidance, we propose a non-linear transform operation inspired by super-resolution measurement [4]. This method involves projecting the generated image and the adversarial image onto a higher dimensional space, effectively increasing the Manhattan distance between them beyond the maximum value of adversarial perturbations. Hence, it may be the same as the case for the long-range distance.

Ultimately, we compare our method to the latest adversarial training and adversarial purification methods on various strong adaptive attack benchmarks. Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet across various classifiers, such as WideResNet-28-10 and WideResNet-70-16, show that MimicDiffusion achieves state-of-the-art performance. Compared with the latest adversarial purification methods [37], e.g., AutoAttck ($\ell_\infty, \epsilon = 8/255$) [5], we show the absolute improvement of $+18.49\%$, $+13.23\%$ and 17.64% in average robust accuracy on CIFAR-10, CIFAR-100 and ImageNet with WideResNet-28-10 respectively.

To sum up, the main contributions of this paper are:

- We propose a new perspective for diffusion-based adversarial purification methods, which mimic the generative process of the diffusion model with the clean image as input to reduce the negative influence of adversarial perturbation.
- We propose a novel MimicDiffusion, where we use Manhattan distance and propose long-range and short-range guidance to bridge the gap between clean and adversarial samples.
- The experimental results show that our model achieves state-of-the-art performance on various adaptive attack benchmarks.

2. Related Work

Adversarial training uses the adversarial samples to train the classifier [20]. Thus, the classifier can also correctly recognize the adversarial samples. For example, Kang *et al.* [16] used ordinary differential equations to re-sample the feature point from the Lyapunov-stable equilibrium points. Sven *et al.* [11] proposed the data augmentation method to generate many adversarial samples and improve robust prior knowledge.

Adversarial purification uses the generative models to generate clean images and could be regarded as the classifier-agnostic method. These methods try to remove the adversarial perturbation in the adversarial sample. Thus, the classifier can be fixed. Pouya *et al.* [26] used the generative adversarial network (GANs) to purify the perturbation. Meanwhile, the score-based match model [30, 37] was proposed to eliminate the influence of the perturbation and recover a clean image based on the score-based match network. We empirically compare our method with the previous works, and the experimental results show that our method can achieve significant improvements.

Diffusion model based adversarial purification. Recently, Nie *et al.* [21] proposed the diffusion-based adversarial purification method, which proved that generated

images from the diffusion model tend to be clean images. Therefore, one entire diffusion process step could purify the adversarial perturbation. However, limited by the different designs in different types of diffusion models [17], it is difficult to reach the optimal performance under adversarial purification. To alleviate this, GDPM [33] proposed an iteration-based method and incorporated the guided method [4] to keep the label semantic under the iteration process. Proven by Lee *et al.* [19], GDPM and Nie *et al.* rely on finding an optimal hyperparameter setting, e.g., the optimal time step. Eventually, adversarial perturbation still influences the adversarial purification methods.

3. Preliminary

We first give some definitions. In the adversarial purification, we have a diffusion model, e.g., the score function $s_\theta(x_t) = \nabla_x \log p(x;t)$, that is trained on the original dataset. Now given the adversarial sample denoted as x^{adv}, where $x^{adv} = x^{ori} + \phi$ and x^{ori} is the clean image (unknown) and ϕ is the adversarial perturbation (unknown) generated by adversarial attack methods, adversarial purification aims to recover the clean image x^{ori} from the inputs of the adversarial sample x^{adv}.

Diffusion model based adversarial purification. The idea is to remove both the adversarial purification and Gaussian noise in the reverse process of the diffusion model. First, it finds a time step t^* such that:

$$\begin{aligned} x_{t^*} &= \sqrt{\sigma(t^*)}x^{adv} + \sqrt{(1-\sigma(t^*))}\epsilon \\ &= \sqrt{\sigma(t^*)}(x^{ori} + \phi) + \sqrt{(1-\sigma(t^*))}\epsilon, \end{aligned} \tag{1}$$

where it is the forward process of diffusion model [15], x_{t^*} is the state in the t^* time, $\sigma(*)$ is the noise schedule related to the time step t, and $\epsilon \sim \mathcal{N}(0,1)$ is the Gaussian noise. Then, the reverse process of diffusion model [15] is performed on x_{t^*} to generate the clean image $\hat{x}^{ori}$.

In the above approaches, the key to success is finding an optimal t^* [21], and thus the performance is sensitive to the value of t^*.

Guided-diffusion based adversarial purification. One of the state-of-the-art methods is the guided diffusion model [33]. In the guided diffusion methods, the adversarial sample x^{adv} is used as guidance, and it starts from the pure Gaussian noise x_T in the reverse process. In the t time step, the guided generating process is formulated as follows:

$$\nabla_x \log p(x_t|x^{adv};t) = \underbrace{\nabla_x \log p(x_t;t)}_{\text{Score Function}} + \underbrace{\nabla_x \log p(x^{adv}|x_t;t)}_{\text{Guidance Term}}. \tag{2}$$

The score function is already known, and the guidance term can be approximated as [4]:

$$\begin{aligned} \nabla_x \log p(x^{adv}|x_t;t) &= -R_t \nabla_{x_t} d(\hat{x}_t, x^{adv}), \\ \hat{x}_t &= \frac{x_t - \sqrt{1-\sigma(t)}s_\theta(x_t)}{\sqrt{\sigma(t)}}, \end{aligned} \tag{3}$$

where $s_\theta(x_t)$ is the known score function [30] with the parameter θ for x_t in the t time, $\hat{x}_t$ is the estimation for x_0 in the t time, R_t is the guided factor related to the t time, and $d(*,*)$ is the ℓ_2 norm distance metric.

Motivation. However, we argue that adversarial perturbation will influence the trajectory of the guided method. To show this, considering the Eq. 3 with the ℓ_2 distance norm, we have

$$\begin{aligned} -R_t \nabla_{x_t} d(\hat{x}_t, x^{adv}) &= -R_t \nabla_{x_t} ||\hat{x}_t - x^{adv}||_2^2 \\ &= -R_t \nabla_{x_t} ||\hat{x}_t - x^{ori} - \phi||_2^2 \\ &= -R_t \frac{\partial ||\hat{x}_t - x^{ori} - \phi||_2^2}{\partial \hat{x}_t} \frac{\partial \hat{x}_t}{\partial x_t}. \end{aligned} \tag{4}$$

Then, the Jacobi matrix for the partial part is:

$$\begin{aligned} \frac{\partial ||\hat{x}_t - x^{ori} - \phi||_2^2}{\partial \hat{x}_t} &= \nabla_{x_t} J((\hat{x}_t - x^{ori} - \phi)^2) \\ &= 2J(\hat{x}_t - x^{ori} - \phi), \end{aligned} \tag{5}$$

where $J(*)$ is the operation to calculate the Jacobi matrix. From Eq. 5, we can see that the gradient of the guidance term also includes the adversarial perturbation ϕ. Hence, the adversarial perturbation still influences the generative process of the guided diffusion model, which would cause the generated trajectory to deviate from the correct direction.

Suppose that we can replace the adversarial input to the clean sample x^{ori}, according to Eq. 4 - Eq. 5, we have

$$R_t \nabla_{x_t} d(\hat{x}_t, x^{ori}) \propto 2R_t J((\hat{x}_t - x^{ori})\frac{\partial \hat{x}_t}{\partial x_t}). \tag{6}$$

Hence, if we can mimic the diffusion model with clean images as inputs, we can remove the negative influence of the adversarial perturbation.

4. Method

In this paper, we aim to remove the influence of the adversarial perturbation on the guided diffusion model. According to our motivation, we want to approximate the gradients of the guided terms with the adversarial sample and clean sample as inputs:

$$\nabla_{x_t} d(\hat{x}_t, x^{adv}) \approx \nabla_{x_t} d(\hat{x}_t, x^{ori}). \tag{7}$$

Manhattan distance. It is difficult to approximate the derivatives for different inputs for Euclidean distance. Fortunately, a common distance metric, i.e., Manhattan distance, might have the same gradients for different inputs.

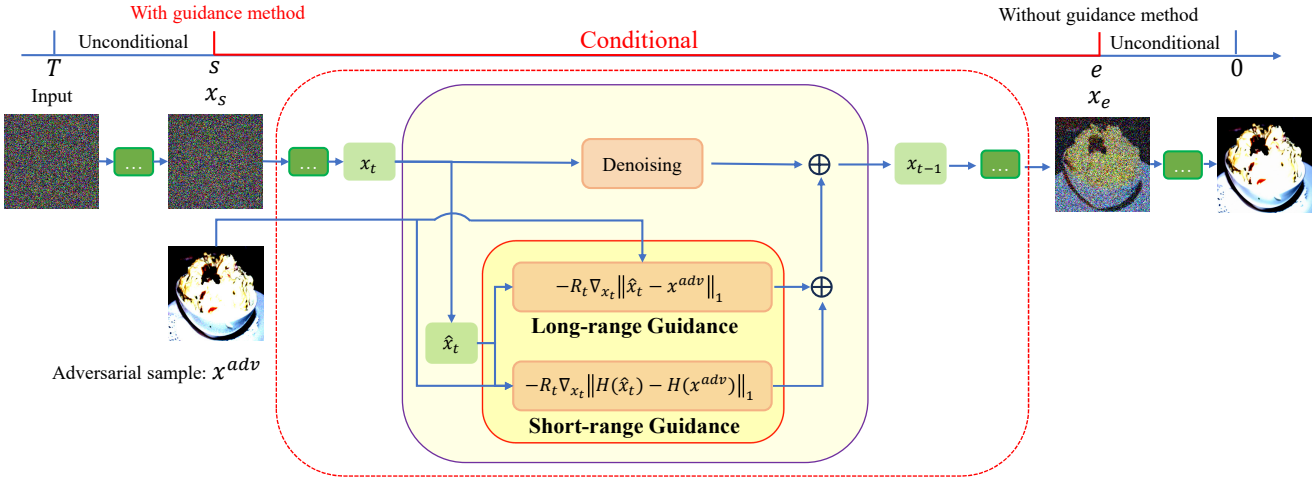


Figure 2. An overview of the proposed MimicDiffusion, where x_T is the pure Gaussian noise, $[s, e]$ is the interval to implement the guidance method noted as conditional by using x^{adv} as the measurement, the other time step without the guidance method noted as unconditional, denoise is one step reverse process, long-range guidance is used to eliminate the adversarial perturbation in the long-range condition. Short-range guidance is used to alleviate the adversarial perturbation in the short-range condition.

We denote $||x||_{\min} = \min(|x_1|, |x_2|, \cdots, |x_n|)$, where n is the number of values in x.

Lemma 1. *Let $||\phi||_\infty < \xi$ and $x^{adv} = x^{ori} + \phi$, then we have the following relations for any x_t: 1) when $||x_t - x^{ori}||_{\min} > \xi$, we have $\nabla_{x_t}||x_t - x^{adv}||_1 = \nabla_{x_t}||x_t - x^{ori}||_1$; 2) when $||x_t - x^{ori}||_{\min} \leq \xi$, we have $\nabla_{x_t}||x_t - x^{adv}||_1 \overset{Unknown}{\longleftrightarrow} \nabla_{x_t}||x_t - x^{ori}||_1$.*

Proof. The proof is simple based on the derivative of ℓ_1, which can only be +1 or -1. We have:

$$\begin{aligned} \nabla_{x_t}||x_t - x^{adv}||_1 &= \nabla_{x_t}||x_t - x^{ori} - \phi||_1 \\ &= \text{Sign}(x_t - x^{ori} - \phi), \end{aligned} \tag{8}$$

where $\text{Sign}(x) = 1$ if $x > 0$ otherwise $\text{Sign}(x) = -1$. When $||x_t - x^{ori}||_{\min} > \xi$, we have $(x_t - x^{ori}) > \xi$ or $(-x_t + x^{ori}) > \xi$. In the two cases, it is easy to verify that

$$\text{Sign}(x_t - x^{ori} - \phi) = \text{Sign}(x_t - x^{ori}). \tag{9}$$

Hence, we have $\nabla_{x_t}||x_t - x^{adv}||_1 = \text{Sign}(x_t - x^{ori} - \phi) = \text{Sign}(x_t - x^{ori}) = \nabla_{x_t}||x_t - x^{ori}||_1$.

When $||x_t - x^{ori}||_{\min} < \xi$, the distance between x_t and x^{ori} is too close, and we cannot clarify the relation between $\nabla_{x_t}||x_t - x^{adv}||$ and $\nabla_{x_t}||x_t - x^{ori}||$. □

Lemma 1 shows two interesting observations. When t is large and x_t will tend to Gaussian noise, thus it satisfies $||x_t - x^{ori}||_{\min} > \xi$, and the gradients for the absolute value of $|x_t - x^{adv}|$ and $|x_t - x^{ori}|$ are equal. On the contrary, when t is small and the distance between x_t and x^{adv} is too close, i.e., $||x_t - x^{ori}||_{\min} \leq \xi$, we cannot clarify the relation between $\nabla_{x_t}||x_t - x^{adv}||_1$ and $\nabla_{x_t}||x_t - x^{ori}||_1$.

Based on the two observations, we propose MimicDiffusion shown in Fig. 2 to achieve mimicking. Concretely, we start from the Gaussian noise directly, which can avoid adding additional perturbation from the adversarial sample. The adversarial sample is only used as the guidance. The Manhattan distance is proposed in the guidance term to reduce the negative influence of adversarial perturbation. When the generated x_t is far from the x^{ori}, we call it long-range distance; we can use the adversarial sample as guidance directly. Then, we propose to use a super-resolution operation to reduce the extra perturbation when x_t is close to x^{ori} called short-range distance. In the end, to better mimic the trajectory of the guidance with the x^{ori} input, we further propose a novel sampling strategy to implement guidance in a particular time interval to reduce the extra noise and time cost at the same time.

Long-range guidance. In the long-range distance situation, e.g., $||\hat{x}_t - x^{ori}||_{\min} > \xi$, based on Lemma 1, we apply the Manhattan distance:

$$\begin{aligned} \frac{\partial||\hat{x}_t - x^{ori} - \phi||_1}{\partial \hat{x}_t} &= \nabla_{x_t} J(|\hat{x}_t - x^{adv}|) \\ &= \nabla_{x_t} J(|\hat{x}_t - x^{ori}|) \\ &= \frac{\partial||\hat{x}_t - x^{ori}||_1}{\partial \hat{x}_t}, \end{aligned} \tag{10}$$

In this way, we have $\nabla_{x_t} d(\hat{x}_t, x^{adv}) = \nabla_{x_t} d(\hat{x}_t, x^{ori})$ that can eliminate ϕ, and thus avoid adding extra adversarial perturbation from the guidance term.

Hence, in the long-range distance, we can set the guidance $y^l = x^{adv}$, where the trajectory will be similar to the diffusion model with clear input. The long-range guidance can be formulated as:

$$\nabla_{x_t} \log p(y^l|x_t) = -R_t \nabla_{x_t} ||\hat{x}_t - x^{adv}||_1, \tag{11}$$

where $\hat{x}_t = \frac{x_t - \sqrt{1-\sigma(t)} s_\theta(x_t)}{\sqrt{\sigma(t)}}$ is the estimated image and $y^l = x^{adv}$.

Short-range guidance. Based on the definition of short-range distance, the unknown relation will eventually lead to the trajectory deviation in small time steps, e.g., the later phase of the reverse process. To alleviate this, we chose the super-resolution operation at the short-range distance. The super-resolution operation is a non-linear mapping operation. Under the transform operation of super-resolution, the non-linear transform could increase the Manhattan distance to change the short-range distance to the long-range distance. In this condition, based on the Eq.10, we eliminate the perturbation term and let the trajectory of the diffusion model go back to that with the clear input and achieve the mimicking.

Therefore, in the short-range guidance term, we use the super-resolution operation [4] to map the adversarial sample to the guidance sample $y^s = H(x^{adv})$. The estimated image $\hat{x}_t$ is also performed via the super-resolution operation, which is defined as:

$$\begin{aligned} &\nabla_{x_t} \log p(y^s|x_t) = \\ &\quad - R_t \nabla_{x_t} ||H(\hat{x}_t) - H(x^{adv})||_1, \end{aligned} \tag{12}$$

where $H(*)$ is the super-resolution (x4) operator [4] and is a non-linear transform to project images of low resolution onto high resolution, which could be calculated based on the Bicubic interpolation [4], the bias leads by Bicubic interpolation will increase the Manhattan distance, and thus achieve a change from the short-range distance to the long-range distance.

To apply two guidance, the guidance term in Eq. 2 could be re-defined as:

$$\nabla_{x_t} \log p(y^l|x_t) + \nabla_{x_t} \log p(y^s|x_t), \tag{13}$$

where the two guidance are independent since the long-range and short-range guidance are two independent cases. Eq. 13 could be calculated directly based on Eq. 11 and Eq. 12.

Following the advice of DPS [4], the guided factor is:

$$R_t = \frac{1}{\sigma(t)^2}. \tag{14}$$

Note that the proposed method tries to mimic the trajectory of the diffusion model with x^{ori} input, and there is no serious setting for MimicDiffusion.

Algorithm 1 The overall algorithm for MimicDiffusion

Input: x^{adv}, T ▷ Pre-trained diffusion model
Output: x_0
$s = 50\%T, e = 20\%T$ ▷ Initialize sampling strategy
$x_T \sim \mathcal{N}(0, 1)$
for t in $[T, T-1, ..., 1]$ **do**
  Calculate $\hat{x}_t$ and x_{t-1} by the Reverse process
  if $t \in [s, e]$ **then**
    Calculate R_t by Eq. 14
    $g^l \leftarrow \nabla_{x_t} \log p(y^l|x_t)$ by Eq. 11
    $g^s \leftarrow \nabla_{x_t} \log p(y^s|x_t)$ by Eq. 12
    $x_{t-1} \leftarrow x_{t-1} + g^l + g^s$
  else
    $x_{t-1} \leftarrow x_{t-1}$
  end if
end for
Return: x_0

Sampling strategy. It is difficult to let the entire reverse process be the long-range distance. Besides, calculating the gradient has a high computation cost. To alleviate these, we choose to implement the guided method in the middle phase of the reverse process of all time steps. Following the empirical finding [38], the whole generation time step is in $[T :: 0]$. We choose the middle phase of the reverse process, i.e., $[s :: e]$, to implement the guided method and vice versa, not to implement the guided method, where $s = 50\%T, e = 20\%T$. In this way, we avoid implementing the guided method in small time steps to avoid adding extra perturbation, reducing the time cost at the same time, and thus try to avoid implementing the guidance method on the short-range distance part.

Algorithm 1 summarizes the proposed MimicDiffusion. In our method, the hyperparameters include the R_t and the interval $[s, e]$. For the guided factor, it could be calculated directly without additional constraints. Due to the proposed multi-guidance's independent property, we use the same guided factor for both guidance terms. In the end, Yu *et al.* [38] showed that minor numerical fluctuation under the middle phase of the reverse process for s and e will have little influence on the performance of the guided method. Therefore, we successfully achieve the adversarial purification without the serious setting.

5. Experiment

5.1. Experimental Settings

Datasets and network architectures. We consider three datasets for evaluation: CIFAR-10, CIFAR-100 [18], and ImageNet [25]. Meanwhile, we compare various state-of-the-art defense methods reported by the standardized benchmark: RobustBench [6] on CIFAR-10 and CIFAR-

100 while comparing other adversarial purification methods on CIFAR-10. We consider two widely used backbones on RobustBench for classifiers: WideResNet-28-10 and WideResNet-70-16 [39]. For ImageNet, we consider the ResNet-50 as the backbone.

Adversarial attack methods. We evaluate our method with the common adversarial attack method: the strong adversarial attack method AutoAttack [5] with two settings: AutoAttack($\ell_\infty, \epsilon = 8/255$) and AutoAttack($\ell_2, \epsilon = 0.5$) respectively, projected gradient descent (PGD) attack($\ell_\infty, \epsilon = 8/255$) [20], and C&W attack [3]. Meanwhile, to make a fair comparison with other adversarial purification methods, we evaluate our method with the adaptive attack: Backward pass differentiable approximation (BPDA+EOT) [14]. Meanwhile, we use the adjoint method to get the gradient of the reverse process for white-box attacks. The experimental results for the C&W and PGD attacks are reported in the supplementary material. In the end, we also reports the performance following the surrogate process in Appendix [19]

Pre-trained diffusion model. We use the unconditional CIFAR-10 checkpoint of EDM offered by NVIDIA [17] for our method on CIFAR-10 datasets. We fine-tune the unconditional CIFAR-10 checkpoint based on CIFAR-100 for our method following the training method offered by NVIDIA [17]. For ImageNet, we use the pre-trained diffusion model offered by Nie *et al.* [21]. We evaluate our model on a single RTX4090 GPU with 24 GB memory. For CIFAR-10 and CIFAR-100, $T = 100$. For ImageNet, $T = 1000$.
Evaluation metrics. We use *standard accuracy* and *robust accuracy* as the evaluation metrics following the prior works [21]. Meanwhile, following the experimental setting [21] to reduce the computation cost of applying adaptive attacks, we evaluate the robust accuracy for all methods with BPDA+EOT attack on a fixed subset of 512 images randomly sampled from the test set. Meanwhile, the visualization for the purified images is reported in the supplementary material.

5.2. Experimental Results

We first report the results of MimicDiffusion compared with the state-of-the-art adversarial training method reported by RobustBench [6] against the ℓ_∞ and ℓ_2 threat models, respectively.

CIFAR-10. Table. 1 shows the robustness performance against the AutoAttack with ℓ_∞ and $\epsilon = 8/255$. Specifically, MimicDiffusion improves the average robust accuracy by 21.64% on WideResNet-28-10 and by 20.69% on WideResNet-70-16, respectively, compared with the best baseline method. Meanwhile, compared with the adversarial training methods that need extra data, MimicDiffusion improves the average robust accuracy by 29.97% on WideResNet-28-10 and by 23.97% on WideResNet-70-16, respectively. It should be noted that our method effectively narrows the gap between standard accuracy and robust accuracy, showcasing the efficacy of mimicking the diffusion model using clean images, where the gap between standard accuracy and robust accuracy is 0.65% and 1.1% on WideResNet-28-10 and WideResNet-70-16 respectively.

Table. 2 shows the robustness performance against the AutoAttack with ℓ_2 and $\epsilon = 0.5$. Specifically, MimicDiffusion improves the average robust accuracy by 13.28% on WideResNet-28-10 and by 11.09% on WideResNet-70-16, respectively, compared with the best baseline method. Meanwhile, MimicDiffusion outperforms the method with the extra data. Except for the improvement in the robust accuracy, the gap between the standard and robust accuracy is still reduced. These results confirm that MimicDiffusion effectively improves accuracy against the ℓ_2 threat. To sum up, the experimental results on CIFAR-10 show the effectiveness of MimicDiffusion in defending against ℓ_∞ and ℓ_2 threat models on CIFAR-10. Meanwhile, MimicDiffusion keeps a high level of standard accuracy, which demonstrates the validity of MimicDiffusion.

CIFAR-100. Table. 3 shows the robustness performance against the AutoAttack with ℓ_∞ and $\epsilon = 8/255$. It can be found that MimicDiffusion improves the average robust accuracy by 18.68% on WideResNet-28-10 and by 19.59% on WideResNet-70-16, respectively. Even with the minimum level of MimicDiffusion, we still improve the average robust accuracy by 13.23% on WideResNet-28-10 and by 15.28% on WideResNet-70-16. Meanwhile, based on the different datasets, our method tends to reduce the gap between standard and robust accuracy. These results prove that our model achieves the mimicking and significantly outperforms other baseline models on the CIFAR-100 dataset.

ImageNet. We report the extra experimental results on ImageNet [25] shown in the Supplementary. These results follow the experimental setting in Nie *et al.* [21]. According to the results, we still see a significant improvement, with the average robust accuracy increasing by 17.64%. Meanwhile, MimicDiffusion successfully reduces the gap between standard accuracy and robust accuracy.

Overall, experimental results demonstrate that MimicDiffusion achieves a significant improvement in defending the ℓ_∞ and ℓ_2 threat model on CIFAR-10 and defending the ℓ_∞ threat model on CIFAR-100 and ImageNet. This also demonstrates the effectiveness of the proposed method for adversarial purification.

Defense against unseen threats. To demonstrate the effectiveness of MimicDiffusion, we compare it with other adversarial purification methods using BPDA+EOT ($\ell_\infty, \epsilon = 8/255$) [14]. This attack, which is adaptive and stochastic, is specifically designed for purification methods, as some adversarial purification methods are not compati-

Table 1. Standard accuracy and robust accuracy against AutoAttack $\ell_\infty(\epsilon = 8/255)$ on CIFAR-10 (*methods use extra data)

Method	Backbone	Standard Accuracy(%)	Robust Accuracy(%)
Zhang *et al.* [40]*	WideResNet-28-10	89.36	59.96
Wu *et al.* [35]*	WideResNet-28-10	88.25	62.11
Gowal *et al.* [12]*	WideResNet-28-10	89.48	62.70
Wu *et al.* [35]	WideResNet-28-10	85.36	59.18
Gowal *et al.* [13]	WideResNet-28-10	87.33	61.72
Rebuffi *et al.* [23]	WideResNet-28-10	87.50	65.24
GDPM [33]	WideResNet-28-10	84.85	71.18
Nie *et al.* [21]	WideResNet-28-10	89.23	71.03
MimicDiffusion (Our)	WideResNet-28-10	**93.32 ± 2.94**	**92.67 ± 3.15**
Gowal *et al.* [12]*	WideResNet-70-16	91.10	66.02
Rebuffi *et al.* [23]*	WideResNet-70-16	92.23	68.56
Gowal *et al.* [12]	WideResNet-70-16	85.29	59.57
Rebuffi *et al.* [23]	WideResNet-70-16	88.54	64.46
Gowal *et al.* [13]	WideResNet-70-16	88.74	66.60
Nie *et al.* [21]	WideResNet-70-16	91.04	71.84
MimicDiffusion (Our)	WideResNet-70-16	**93.63 ± 2.67**	**92.53 ± 3.06**

Table 2. Standard accuracy and robust accuracy against AutoAttack $\ell_2(\epsilon = 0.5)$ on CIFAR-10 (*methods use extra data)

Method	Classifier	Standard Accuracy(%)	Robust Accuracy(%)
Augustin *et al.* [1]*	WideResNet-28-10	92.23	77.93
Rony *et al.* [24]	WideResNet-28-10	89.05	66.41
Ding *et al.* [8]	WideResNet-28-10	88.02	67.77
Wu *et al.* [35]*	WideResNet-28-10	88.51	72.85
Sehwag *et al.* [27]*	WideResNet-28-10	90.31	75.39
Rebuffi *et al.* [23]	WideResNet-28-10	91.79	78.32
GDPM	WideResNet-28-10	92.00	75.28
Nie *et al.* [21]	WideResNet-28-10	91.38	78.98
MimicDiffusion (Our)	WideResNet-28-10	**93.66 ± 3.22**	**92.26 ± 3.40**
Gowal *et al.* [12]*	WideResNet-70-16	94.74	79.88
Rebuffi *et al.* [23]*	WideResNet-70-16	95.74	81.44
Gowal *et al.* [12]	WideResNet-70-16	90.90	74.03
Rebuffi *et al.* [23]	WideResNet-70-16	92.41	80.86
Nie *et al.* [21]	WideResNet-70-16	**93.24**	81.17
MimicDiffusion (Our)	WideResNet-70-16	93.49 ± 2.77	**92.26 ± 2.78**

ble with AutoAttack. The experimental results are shown in Table. 4. It can be found that MimicDiffusion also gets the best performance in terms of robust accuracy, which improves the average robust accuracy by almost 4.35% in the worst performance. Meanwhile, we find that there is a smaller gap between standard accuracy and robust accuracy for mimic diffusion, which demonstrates its success.

5.3. Ablation Study

As shown in Table. 5, we list the different combinations among MimicDiffusion. To prove the necessity for two guidance, it can be found that using each guidance individually results in a significant decrease in robust accuracy. Then, Compared with the ℓ_2 norm, widely used in previous methods, ℓ_1 norm could increase 31.93% robust accuracy and achieve a large improvement, which proves the effectiveness of Lemma 1. Meanwhile, the super-resolution guidance should change the short-range to the long-range. This means that using the g^s and g^l significantly improves the robust accuracy by almost 66.32%

Table 3. Standard accuracy and robust accuracy against AutoAttack $\ell_\infty(\epsilon = 8/255)$ on CIFAR-100 (*methods use extra data)

Method	Classifier	Standard Accuracy(%)	Robust Accuracy(%)
Debenedetti *et al.* [7]*	XCiT-M12	69.21	34.21
Debenedetti *et al.* [7]*	XCiT-L12	70.76	35.08
Gowal *et al.* [12]*	WideResNet-70-16	69.15	36.88
Pang *et al.* [22]	WideResNet-28-10	63.66	31.08
Rebuffi *et al.* [23]	WideResNet-28-10	62.41	32.06
Wang *et al.* [34]	WideResNet-28-10	72.58	38.83
Pang *et al.* [22]	WideResNet-70-16	65.56	33.05
Rebuffi *et al.* [23]	WideResNet-70-16	63.56	34.64
Wang *et al.* [34]	WideResNet-70-16	**75.22**	42.67
MimicDiffusion (Our)	WideResNet-28-10	63.53 ± 6.17	$\mathbf{61.35 \pm 5.45}$
MimicDiffusion (Our)	WideResNet-70-16	64.32 ± 5.77	$\mathbf{62.26 \pm 4.31}$

Table 4. Standard accuracy and robust accuracy against BPDA+EOT ($\ell_\infty, \epsilon = 8/255$) on WideResNet-28-10 for CIFAR-10

Method	Purification	Standard Accuracy(%)	Robust Accuracy(%)
Song *et al.* [29]	Gibbs Update	95.00	9.00
Yang *et al.* [36]*	Mask+Recon.	94.00	15.00
Hill *et al.* [14]*	EBM+LD	70.76	35.08
Yong *et al.* [37]*	DSM+LD	86.14	70.01
Nie *et al.* [21]($t^* = 0.0075$)	Diffusion	**91.38**	77.62
NIe *et al.* [21]($t^* = 0.1$)	Diffusion	89.23	81.56
GDPM [33]	Diffusion	90.36	77.31
MimicDiffusion (Our)	Diffusion	92.5 ± 5.12	$\mathbf{92.00 \pm 6.1}$

g^l	g^s	d	Sampling	Standard(%)	Robust(%)
√		ℓ_2		86.42	16.7
	√	ℓ_2		10.89	10.00
√		ℓ_1		88.39	22.78
	√	ℓ_1		10.80	7.10
√	√	ℓ_2		91.03	57.07
√	√	ℓ_1		91.30	89.10
√	√	ℓ_1	√	93.30	92.60

Table 5. Ablation study for MimicDiffusion based on CIFAR-10 against AutoAttack($\ell_\infty, \epsilon = 8/255$) with WideResNet-28-10, where Sampling is sampling strategy, standard is standard accuracy, and robust is robust accuracy.

6. Conclusion

We proposed a new defense method called MimicDiffusion to achieve adversarial purification by mimicking the trajectory of the diffusion model using clean images as the input without serious settings. Specifically, using the two proposed guidance methods with the Manhattan distance can mitigate the negative impact caused by adversarial perturbation, as mentioned in Lemma 1. To show the robust performance of our method, we conducted thorough experiments on CIFAR-10, CIFAR-100, and ImageNet with three classifier backbones: WideResNet-70-16, WideResNet-28-10, and ResNet-50. The experimental results showed that our method performed best in defense of various strong adaptive attacks such as AutoAttack, PGD attack, C&W attack, and BPDA+EOT. These results show that MimicDiffusion could mimic the trajectory of the diffusion model using the clean image as the input.

Despite the large improvement, the proposed two guidance require calculating the gradients and will increase the computation cost. We will explore finding a gradient-free guided method in further work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (U2001211, U22B2060), Guangdong Basic and Applied Basic Research Foundation (2023A1515011400), Research Foundation of Science and Technology Plan Project of Guangzhou City (2023B01J0001, 2024B01W0004).

References

[1] Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in- and out-distribution improves explainability. In *Proceedings of the European Conference on Computer Vision*, pages 228–245, 2020. 7

[2] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4312–4321, 2021. 1

[3] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *CoRR*, abs/1608.04644, 2016. 6

[4] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *Proceedings of the International Conference on Learning Representations*, 2023. 1, 2, 3, 5

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the International Conference on Machine Learning*, pages 2206–2216, 2020. 2, 6

[6] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Proceedings of the Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 5, 6

[7] Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. abs/2209.07399, 2022. 8

[8] 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. 7

[9] Hadi M. Dolatabadi, Sarah M. Erfani, and Christopher Leckie. $l\ \infty$-robustness and beyond: Unleashing efficient adversarial training. In *Proceedings of the European Conference on Computer Vision*, pages 467–483. Springer, 2022. 1

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*, 2015. 1

[11] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A. Mann. Improving robustness using generated data. In *Proceedings of the Advances in Neural Information Processing Systems*. 2

[12] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy A. Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *CoRR*, abs/2010.03593, 2020. 7, 8

[13] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A. Mann. Improving robustness using generated data. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 4218–4233, 2021. 1, 7

[14] Mitch Hill, Jonathan Craig Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. In *Proceedings of the International Conference on Learning Representations*, 2021. 6, 8

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 1, 3

[16] Qiyu Kang, Yang Song, Qinxu Ding, and Wee Peng Tay. Stable neural ODE with lyapunov-stable equilibrium points for defending against adversarial attacks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 14925–14937, 2021. 2

[17] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *CoRR*, abs/2206.00364, 2022. 3, 6

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 5

[19] Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *ICCV*, 2023. 3, 6

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*, 2018. 2, 6

[21] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, pages 16805–16827. PMLR, 2022. 1, 2, 3, 6, 7, 8

[22] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *Proceedings of the International Conference on Machine Learning*, pages 17258–17277, 2022. 8

[23] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. Fixing data augmentation to improve adversarial robustness. *CoRR*, abs/2103.01946, 2021. 7, 8

[24] 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*, pages 4322–4330, 2019. 7

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 5, 6

[26] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *Proceedings of the International Conference on Learning Representations*, 2018. 2

[27] 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 *Proceedings of the International Conference on Learning Representations*, 2022. 7

[28] Changhao Shi, Chester Holtz, and Gal Mishne. Online adversarial purification based on self-supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2021. 1

[29] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *Proceedings of the International Conference on Learning Representations*, 2018. 8

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations*, 2021. 2, 3

[31] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion, 2023. 1

[32] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Proceedings of the Advances in Neural Information Processing Systems*, 2020. 1

[33] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification, 2022. 1, 3, 7, 8

[34] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. abs/2302.04638, 2023. 1, 8

[35] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Proceedings of the Advances in Neural Information Processing Systems*, 2020. 7

[36] Yuzhe Yang, Guo Zhang, Zhi Xu, and Dina Katabi. Me-net: Towards effective adversarial robustness with matrix estimation. In *Proceedings of the International Conference on Machine Learning*, pages 7025–7034, 2019. 8

[37] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *Proceedings of the International Conference on Machine Learning*, pages 12062–12072, 2021. 1, 2, 8

[38] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. *arXiv:2303.09833*, 2023. 5

[39] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016. 6

[40] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan S. Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *Proceedings of the International Conference on Learning Representations*, 2021. 7