Diffusion Models Demand Contrastive Guidance for Adversarial Purification to Advance

Mingyuan Bai [*1] **Wei Huang** [*2] **Tenghui Li** [*3 1 4] **Andong Wang** [1] **Junbin Gao** [5] **Cesar F Caiafa** [1 6] **Qibin Zhao** [1]

Abstract

In adversarial defense, adversarial purification can be viewed as a special generation task with the purpose to remove adversarial attacks and diffusion models excel in adversarial purification for their strong generative power. With different predetermined generation requirements, various types of guidance have been proposed, but few of them focuses on adversarial purification. In this work, we propose to guide diffusion models for adversarial purification using contrastive guidance. We theoretically derive the proper noise level added in the forward process diffusion models for adversarial purification from a feature learning perspective. For the reverse process, it is implied that the role of contrastive loss guidance is to facilitate the evolution towards the signal direction. From the theoretical findings and implications, we design the forward process with the proper amount of Gaussian noise added and the reverse process with the gradient of contrastive loss as the guidance of diffusion models for adversarial purification. Empirically, extensive experiments on CIFAR-10, CIFAR-100, the German Traffic Sign Recognition Benchmark and ImageNet datasets with ResNet and WideResNet classifiers show that our method outperforms most of current adversarial training and adversarial purification methods by a large improvement.

[*]Equal contribution [1]Tensor Learning Team, Center of Advanced Intelligence Project, RIKEN, Tokyo, 1030027, JAPAN [2]Deep Learning Theory Team, Center of Advanced Intelligence Project, RIKEN, Tokyo, 1030027, JAPAN [3]School of Automation, Guangdong University of Technology, Guangzhou, 510006, CHINA [4]Key Laboratory of Intelligent Detection and the Internet of Things in Manufacturing, Ministry of Education, Guangzhou, 510006, CHINA [5]Discipline of Business Analytics, The University of Sydney Business School, The University of Sydney, Darlington, NSW, 2006, AUSTRALIA [6]Instituto Argentino de Radioastronomìa, CONICET CCT La Plata/CIC-PBA/UNLP, V. Elisa, 1900, ARGENTINA. Correspondence to: Qibin Zhao <qibin.zhao@riken.jp>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Adversarial defense has been an important method to resist adversarial attack to deep learning models. These adversarial attacks are imperceptible by human sense but can cause deep neural networks (DNNs) to misclassify adversarially attacked data. To address this vulnerability of DNNs, many adversarial defense methods were designed, such as adversarial training (Madry et al., 2018), certified methods for training (Gowal et al., 2019; Zhang et al., 2018; 2020a; Frosio & Kautz, 2023), and adversarial purification (Shi et al., 2021; Yoon et al., 2021; Nie et al., 2022; Wang et al., 2022; Xiao et al., 2023). Adversarial training and certified methods for training rely on generated attacks, as a result, they can only defend against such attacks. For unseen threats, there are adversarial training methods designed to defend against them, however, at the cost of significant clean accuracy decay (Laidlaw et al., 2021; Dolatabadi et al., 2022).

Instead, adversarial purification methods do not assume the form of threat models and do not require retraining the classifier. For example, given adversarial examples, they use generative models to generate clean examples such that they are correctly classified (Samangouei et al., 2018; Hill et al., 2021; Shi et al., 2021; Yoon et al., 2021). In recent years, diffusion models have been attracting increasing attention for their strong generative power and achieve the state-of-the-art performance for adversarial purification (Nie et al., 2022; Xiao et al., 2023). Besides, it is still a relatively new topic for leveraging diffusion models to conduct adversarial purification. DiffPure (Nie et al., 2022) firstly diffuses the adversarial example with a small amount of Gaussian noise followed by a reverse diffusion process, and hence removes the adversarial attack, meanwhile preserving the semantic information. However, theoretical studies in DiffPure show that its upper bound of difference between purified data and clean data is larger than the adversarial attacks. Also, its purified data is still relatively indistinguishable among different classes under common pretrained classifiers. A guided diffusion model for adversarial purification uses the difference between adversarial example and purified example as a guidance (Wang et al., 2022) during purifying the adver-

sarial examples. Hence they provided a balance between retaining image semantics and destroying adversarial perturbations, achieving high standard and robust accuracy for adversarial purification. Other diffusion models undertake purification against backdoor attacks instead of adversarial attacks and also obtain impressive purification results (Shi et al., 2023).

Many diffusion models for image generation are designed with different guidance to accomplish different goals rather than adversarial purification. Classifier guided diffusion models (Dhariwal & Nichol, 2021) synthesize data by maximizing the joint probability of the synthesized data and the predetermined labels. Diffusion models with classifier-free guidance (Ho & Salimans, 2022) generate data conditioned on the predetermined labels or the text conditions. However, for adversarial purification, label and text condition information is unavailable. Lu et al. (2023) find that energy guidance for diffusion models can lead the data to fit a desired data distribution Also, they proved that using a contrastive loss to learn the energy provides the guarantee to converge to the exact guidance under unlimited model capacity and data samples.

In our work, inspired by Lu et al. (2023)'s theoretical findings, we aim to answer the question: *if we consider the adversarial examples distribution as the given data distribution, can contrastive guidance guide this distribution to the clean data distribution for diffusion models on adversarial purification?*

To this end, we theoretically study the diffusion models for adversarial purification and the effect of contrastive guidance on it. Accordingly, we propose contrastive guided diffusion models for adversarial purification. In the forward diffusion process, we gradually add Gaussian noises to adversarial examples, until adversarial attacks are diffused towards Gaussian noises, but label semantics are not destroyed. Experiments on CIFAR-10, CIFAR-100, German Traffic Sign Recognition Benchmark (GTSRB) and ImageNet datasets demonstrate that our proposed guided diffusion model for adversarial purification outperforms most baseline models against various adversarial attacks. Our contributions are as follows.

1. We theoretically derive the proper Gaussian noise level in the forward process and find the possible role of contrastive loss guidance: to facilitate the evolution towards the signal direction in the reverse process of diffusion models for adversarial purification.

2. We propose contrastive guided diffusion models for adversarial purification accordingly.

3. We conduct extensive experiments which show diffusion models demand contrastive guidance for adversarial purification to advance in three common benchmark datasets against various adversarial attacks.

2. Related Work

To resist adversarial attacks, there have been a large number of past attempts. Adversarial training (Madry et al., 2018) synthesize adversarial attacks and train DNNs to produce correct output. However, they are only effective to adversarial attacks which the DNNs are trained with. Unseen threats can be defended by some adversarial training methods, but incurring significant drop in performance (Laidlaw et al., 2021; Dolatabadi et al., 2022). Certified methods for training provide estimates of correct output bounds used for training (Gowal et al., 2019; Zhang et al., 2018; 2020a; Frosio & Kautz, 2023), but these methods also cannot defend against unseen threats. Unlike those aforementioned adversarial defense methods, adversarial purification methods remove adversarial attacks from data without assumptions on the threat models or downstream classifiers (Samangouei et al., 2018; Hill et al., 2021; Shi et al., 2021; Yoon et al., 2021; Nie et al., 2022; Xiao et al., 2023). They purify adversarial examples as data preprocessing in a plug-n-play manner before classification so as to make classifiers produce correct outputs. Generative adversarial networks (GANs) (Samangouei et al., 2018) were trained to generate clean data given adversarial examples. Energy-based models (EBMs) are trained with Markov-Chain Monte-Carlo (Hill et al., 2021) or denoising score matching (Yoon et al., 2021) to generate clean data, given adversarial examples. However, the performance of GAN-based or EBM-based adversarial purification models is constrained by their generative power. Due to the powerfulness of diffusion models on data generation, they came onto the stage for adversarial purification and showed significant improvement on performance. DiffPure (Nie et al., 2022) imposes a time condition to diffusion models for adversarial purification, where the number of steps to add Gaussian noises is selected to both diffuse adversarial attacks and preserve semantics in data. DensePure (Xiao et al., 2023) did one step more by improving certified robustness of classifiers. Other than adversarial attacks, diffusion models also succeed in purifying backdoor attacked examples (Shi et al., 2023).

In order to generate data with different purposes, a number of guidance for diffusion models are proposed. Classifier guidance for diffusion models (Dhariwal & Nichol, 2021) guides the whole generation process using the classification result from a classifier, i.e., maximizing the joint probability of the predicted semantic label and the generated image. Classifier-free guided diffusion models (Ho & Salimans, 2022) simply generate data using a combination of conditional diffusion model and a jointly trained unconditional diffusion model, given the conditions such as labels or text conditions which are not available for adversarial purifi-

cation tasks. Loss guidance (Song et al., 2023) samples different losses using Monte Carlo methods to reduce the approximation bias. Universal guidance (Bansal et al., 2024) controls diffusion models by arbitrary guidance modalities without retraining use-specific components. Lu et al. (2023) proved that energy guidance can guide a distribution to the desired distribution. Furthermore, energy learned using the contrastive loss is guaranteed to converge to the exact guidance, assuming unlimited model capacity and data samples. For adversarial training, Ouyang et al. (2023) theoretically and empirically demonstrated that data synthesized by contrastive guided diffusion models can enhance adversarial training.

3. Theoretical Study

In this section, we delve into theoretical findings from a feature learning perspective, focusing on a well-regarded distribution model celebrated for its simplicity and effectiveness in theoretically understanding intricate learning behaviors (Schmidt et al., 2018; Augustin et al., 2020). Initially, we illustrate how DiffPure (Nie et al., 2022) effectively purifies adversarial perturbations to enhance robust classification, followed by a demonstration of how a properly trained contrastive guidance can further boost robust accuracy.

For theoretical analysis, we adopt a high dimension setting and take the Gaussian mixture model as the data generation model. Lemma 3.1 first provides several concentration bounds for the inner product between noise and signal vectors. The main idea is that, the noise vector and signal vector is almost orthogonal to each other. Furthermore, we examine that is well-trained linear network can tell the true label of data example generated from Gaussian mixture model in Lemma 3.2. Later, we introduce an adversarial attack, which is verified to be effective by Lemma 3.3. Finally, the effectiveness of DiffPure method is demonstrated by Theorem 3.4. In the theorem, we find that reverse diffusion has a potential to recover the predicted label against attack.

3.1. Data and Adversarial Model

Consider a binary classification task where the input-label pair $(\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, 1\}$ follows the Gaussian mixture model (Schmidt et al., 2018):

$$\mathbf{x} = y\boldsymbol{\mu} + \boldsymbol{\xi}, \tag{1}$$

where the label y is uniformly distributed on the set $\{-1, 1\}$, while $\boldsymbol{\mu} \in \mathbb{R}^d$ represents a fixed 'feature' component. Additionally, $\boldsymbol{\xi}$ follows a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and represents the random 'noise' component, which is independent of the label y.

Suppose we are given n *i.i.d.* training examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ drawn from model (1), i.e., $\mathbf{x}_i = y_i\boldsymbol{\mu} + \boldsymbol{\xi}_i$, $\forall i \in [n]$. Following Schmidt et al. (2018), we adopt a linear classifier $\hat{y}(\mathbf{x}) := \text{sign}\left(\langle \boldsymbol{\theta}^*, \mathbf{x}\rangle\right)$ parameterized by

$$\boldsymbol{\theta}^* = \frac{1}{n}\sum_{i=1}^n y_i\mathbf{x}_i \in \mathbb{R}^d. \tag{2}$$

In the high-dimensional setting, it can be verified that each training data $(\mathbf{x}_i, y_i)$ can be correctly classified with high probability due to the concentration of measure.

For self-contained purpose, we introduce the following Lemma.

Lemma 3.1 (Lemma B.2 in Cao et al. (2022), and Lemma B.3 in Kou et al. (2023))**.** *Suppose that $\delta > 0$, $d = \Omega(\log(4n/\delta))$ and $\{\boldsymbol{\xi}_i\}_{i=1}^n \subset \mathbb{R}^d$ are the random 'noise' components corresponding to the input training data $\{\mathbf{x}_i\}_{i=1}^n$. Then it holds with probability at least $1-\delta$ that*

$$\begin{aligned} &d/2 \le \|\boldsymbol{\xi}_i\|_2^2 \le 3d/2, \\ &|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle| \le 2\sqrt{d\log(4n^2/\delta)}, \\ &|\langle \boldsymbol{\xi}_i, \boldsymbol{\mu}\rangle| \le \|\boldsymbol{\mu}\|_2\sqrt{2\log(8n/\delta)}, \end{aligned}$$

for all $i, i' \in [n]$.

Based on the above Lemma, we can have our first claim,

Lemma 3.2. *Suppose that $\delta > 0$ and $d > 16(n-1)^2\log(4n^2/\delta)$ and $\|\boldsymbol{\mu}\|_2 > 2\sqrt{2\log(8n/\delta)}$, then with probability at least $1-\delta$, it follows that*

$$\text{sign}(\langle \mathbf{x}_i, \boldsymbol{\theta}^*\rangle) = \text{sign}(y_i).$$

The proof of Lemma 3.2 can be found in Appendix A.

Now we consider an adversarial example $\mathbf{x}_i^a = \mathbf{x}_i + \boldsymbol{\delta}_i$. By Lemma 3.1, we find that the ℓ_2 norm of noise vector $\boldsymbol{\xi}_i$ is about $\sqrt{d}$. Thus we define a critical quantity named signal-to-noise ration as $\text{SNR} = \frac{\|\boldsymbol{\mu}\|_2}{\sqrt{d}}$. Then we claim that, the following

$$\boldsymbol{\delta}_i = -32n\text{SNR}^2 y_i \sum_{i'=1}^n y_{i'}\boldsymbol{\xi}_{i'} \tag{3}$$

is a successful attack, as summarized in the following Lemma.

Lemma 3.3. *Assuming the conditions specified in Lemma 3.2 are satisfied, and if $n\text{SNR} > 1$ and $d > \max\{8n\|\boldsymbol{\mu}\|_2\sqrt{2\log(8n/\delta)}, 64n^2\log(4n^2/\delta)\}$, then it can be established with probability at least $1-\delta$ that*

$$\text{sign}(\langle \mathbf{x}_i^a, \boldsymbol{\theta}^*\rangle) = -\text{sign}(y_i).$$

The proof of Lemma 3.3 can be found in Appendix A. In Lemma 3.3, we adopt the high dimension setting, which ensures that attack on noise vector can efficiently reverse the

prediction sign of trained network. On the other hand, the condition $n\text{SNR} > 1$ implies that the number of training sample and signal-to-noise are large enough such that the diffusion method has the potential to recover the prediction against attack examples.

3.2. A Feature Learning Justification for DiffPure

Given a data point $\mathbf{x}$ randomly generated from the Gaussian mixture model (1), we consider a forward diffusion process starting with $\mathbf{x}(0) = \mathbf{x}$. Suppose the probability density of $\mathbf{x}(t)$ conditioned on $\mathbf{x}(0)$ at time t is given as follows:

$$q_t(\mathbf{x}(t)|\mathbf{x}(0)) = \mathcal{N}(\mathbf{x}(t)|\alpha_t\mathbf{x}(0), \sigma_t^2\mathbf{I}). \tag{4}$$

Then, the corresponding diffusion process in reverse-time can be given as (Nie et al., 2022):

$$d\mathbf{x}(t) = \big[f(t)\mathbf{x}(t) - \frac{1}{2}g^2(t)\nabla_{\mathbf{x}(t)}\log q_t(\mathbf{x}(t))\big]dt + g(t)d\boldsymbol{\omega},$$

where functions $f(t) = \frac{d\log\alpha_t}{dt}$, $g^2(t) = \frac{d\sigma_t^2}{dt} - 2\frac{d\log\alpha_t}{dt}\sigma_t^2$, $\nabla_{\mathbf{x}(t)}\log q_t(\mathbf{x}(t))$ is the score function (Song et al., 2021), and $\boldsymbol{\omega}(t)$ is a standard d-dimensional reverse-time Wiener process. Given the forward process (4), we can deduce that the marginal distribution q_t is also a mixture of Gaussians:

$$\begin{aligned}
q_t(\mathbf{x}(t)) &= \int q_t(\mathbf{x}(t)|\mathbf{x}(0))q(\mathbf{x}(0))\mathrm{d}\mathbf{x}(0) \\
&= \frac{1}{2}\int (2\pi\sigma_t^2)^{-\frac{d}{2}}\exp\left(-\frac{(\mathbf{x}(t)-\alpha_t\mathbf{x}(0))^2}{2\sigma_t^2}\right)(2\pi)^{-\frac{d}{2}} \\
&\exp\left(-\frac{(\mathbf{x}(0)-\boldsymbol{\mu})^2}{2}\right)d\mathbf{x}(0) + \frac{1}{2}\int d\mathbf{x}(0)(2\pi\sigma_t^2)^{-\frac{d}{2}} \\
&\exp\left(-\frac{(\mathbf{x}(t)-\alpha_t\mathbf{x}(0))^2}{2\sigma_t^2}\right)(2\pi)^{-\frac{d}{2}}\exp\left(-\frac{(\mathbf{x}(0)+\boldsymbol{\mu})^2}{2}\right) \\
&= \sum_{y=\pm 1}\frac{1}{2}\mathcal{N}(\mathbf{x}(t)|\alpha_t y\boldsymbol{\mu}, (\sigma_t^2+\alpha_t^2)\mathbf{I}).
\end{aligned}$$

The score function can be expressed as follows,

$$\begin{aligned}
&\nabla_{\mathbf{x}(t)}\log q_t(\mathbf{x}(t)) \\
&= \frac{\sum_{y=\pm 1}\frac{1}{2}\mathcal{N}(\mathbf{x}(t)|\alpha_t y\boldsymbol{\mu}, (\sigma_t^2+\alpha_t^2)\mathbf{I})(\alpha_t y\boldsymbol{\mu} - \mathbf{x}(t))}{(\sigma_t^2+\alpha_t)\sum_{y=\pm 1}\frac{1}{2}\mathcal{N}(\mathbf{x}(t)|\alpha_t y\boldsymbol{\mu}, (\sigma_t^2+\alpha_t^2)\mathbf{I})} \\
&= (\tanh(\langle\mathbf{x}(t), \alpha_t^2\boldsymbol{\mu}\rangle)\alpha_t\boldsymbol{\mu} - \mathbf{x}(t))/(\sigma_t^2+\alpha_t^2).
\end{aligned}$$

In practice, obtaining an exact expression of the score function is not feasible, and researchers typically train a proxy model $\mathbf{s}(\cdot)$ to estimate the score function (Song et al., 2021), which is then used in the backward process. To thoroughly simplify the theoretical derivation and focus our efforts on uncovering the working mechanism of DiffPure (Nie et al., 2022), we have made a mild assumption that the score function is exactly estimated via a pre-trained model $\mathbf{s}(\cdot)$, *i.e.*, $\mathbf{s}(\mathbf{x}(t)) = \nabla_{\mathbf{x}(t)}\log q_t(\mathbf{x}(t))$.

Furthermore, we consider $\alpha_t = \exp(-t)$, and $\sigma_t^2 = 1-\exp(-2t)$. In this case, $\sigma_t^2 + \alpha_t^2 = 1$. Similarly a setting has been adopted in (Shah et al., 2023). Then we can conclude that $f(t) = -1$ and $g^2(t) = 2$. Then we find that the reverse diffusion process with an exactly estimated score function can be simplified as:

$$\begin{aligned}
d\mathbf{x}(t) &= [f(t)\mathbf{x}(t) - \frac{1}{2}g^2(t)\mathbf{s}(\mathbf{x}(t))]dt + g(t)d\boldsymbol{\omega}(t) \\
&= -\tanh(\langle\mathbf{x}(t), \alpha_t\boldsymbol{\mu}\rangle)\alpha_t\boldsymbol{\mu}dt + \sqrt{2}d\boldsymbol{\omega}(t).
\end{aligned}$$

The reverse process starts from t^*. By a choosing a proper t^*, the following theorem states that DiffPure is able to recover the sign of encoder on the attacked examples.

Theorem 3.4. *Assuming that* $t^* \le \log(\frac{\|\boldsymbol{\mu}\|_2}{20\sqrt{2\log(8n/\delta)}})$, $\|\boldsymbol{\mu}\|_2 \ge 4\sqrt{2n\log(8n/\delta)}$ *and* $n\text{SNR}^2 \le \frac{1}{16\sqrt{2\log(8n/\delta)}}$, *then with probability at least* $1-\delta$, *the recovered* $\overline{\mathbf{x}}(0)$ *example stratifies*

$$\begin{aligned}
y_i\langle\overline{\mathbf{x}}(0), \boldsymbol{\mu}\rangle \le \exp(-\frac{1}{2}) + \frac{1}{2}\exp(-t^*)\|\boldsymbol{\mu}\|_2^2 \\
- \exp(-\frac{1}{2}\exp(-2\|\boldsymbol{\mu}\|_2^2 t^*)).
\end{aligned}$$

The proof of Theorem 3.4 can be found in Appendix A.

3.3. Enhanced Purification Via Contrastive Guidance

Theorem 3.4 states that with a proper choice of t^*, the example recovered by reverse diffusion has a potential to be fixed. However, it requires carefully choice of the diffusion time t^*. A natural question is that can we design a guidance that can enhance reverse diffusion such that the learning direction in the signal vector ($\boldsymbol{\mu}$) can be boosted? Inspired by (Ouyang et al., 2023; Lu et al., 2023), we introduce the contrastive guidance. Following we provide an analysis for the effect of contrastive guidance. Suppose that $F(\mathbf{x}(t)) = \langle\boldsymbol{\mu}, \mathbf{x}(t)\rangle$. Here we assume a well-learned encoder $F(\mathbf{x}(t))$. Then the contrastive guidance follows:

$$\begin{aligned}
&\nabla_{\mathbf{x}(t)}\log\left(\frac{e^{F(\mathbf{x}(t))^\top F(\mathbf{x}(t))}}{\sum_{j=1}^M e^{F(\mathbf{x}(t))^\top F(\mathbf{x}_j(t))}}\right) \\
&= \nabla_{\mathbf{x}(t)}\log\left(\frac{e^{\langle\boldsymbol{\mu},\mathbf{x}(t)\rangle^2}}{\sum_{j=1}^M e^{\langle\boldsymbol{\mu},\mathbf{x}(t)\rangle\langle\boldsymbol{\mu},\mathbf{x}_j(t)\rangle}}\right) \\
&= 2\langle\mathbf{x}(t), \boldsymbol{\mu}\rangle\boldsymbol{\mu} - \frac{\sum_j\langle\mathbf{x}_j(t), \boldsymbol{\mu}\rangle e^{\langle\boldsymbol{\mu},\mathbf{x}(t)\rangle\langle\boldsymbol{\mu},\mathbf{x}_j(t)\rangle}}{\sum_j e^{\langle\boldsymbol{\mu},\mathbf{x}(t)\rangle\langle\boldsymbol{\mu},\mathbf{x}_j(t)\rangle}}\boldsymbol{\mu}.
\end{aligned} \tag{5}$$

It is found that constrative loss guidance (5), we observe that can further boost the learning in the signal direction. First, the term induced from positive pair $2\langle\mathbf{x}(t), \boldsymbol{\mu}\rangle\boldsymbol{\mu}$ is the direction of $y_i\boldsymbol{\mu}$. Second, the negative pair term provides

Algorithm 1 Adversarial Purification in Contrastive Guided Diffusion Models

Require: each minibatch of m adversarial examples $\mathcal{X}^a = \{\mathbf{x}^{a,(i)}\}_{i=1}^m$, temperature τ, time $t = t^*$, guidance strength hyperparameter λ, time step Δt
for $i = 1$ **to** m **do**
  $\mathbf{x}(t^*)^{(i)} = \sqrt{\alpha_{t^*}}\mathbf{x}^{a,(i)} + \sqrt{1-\alpha_{t^*}}\boldsymbol{\epsilon}$
end for
while $t \geq 0$ **do**
  for $i = 1$ **to** m **do**
    Choose $\mathbf{x}(t)_p^{(i)}$ as the positive pair of $\mathbf{x}(t)^{(i)}$
    Choose $\mathbf{x}(t)_n^{(i)}$ [1] as the negative pair of $\mathbf{x}(t)^{(i)}$
    $\tilde{\epsilon}_\theta(\mathbf{x}(t)^{(i)}) = \epsilon_\theta(\mathbf{x}(t)^{(i)}) + \lambda\nabla_{\mathbf{x}(t)}\ell(\mathbf{x}(t)_a, \mathbf{x}(t)_p; \tau)$
    $\mathbf{x}(t) = \text{SDEINT}(\mathbf{x}(t^*), f_{\text{rev}}, g_{\text{rev}}, \bar{\boldsymbol{\omega}}, t^*, t)$
    $t = t - \Delta t$
  end for
end while
Return purified data $\mathcal{X}_0 = \{\hat{\mathbf{x}}(0)^{(i)} = \mathbf{x}(0)^{(i)}\}_{i=1}^m$

guidance in the direction of $-y_j\boldsymbol{\mu}$. The negative pair implies that $y_j = -y_i$, leading the same guidance direction as the positive pair. Together, the constrastive loss guidance boosts the learning in direction of signal during reverse diffusion process.

4. Contrastive Guided Diffusion Models for Adversarial Purification

The theoretical studies in Section 3 unveil that diffusion models require proper noise levels (t^*) for forward processes and contrastive guidance for reverse processes to successfully purify adversarial examples. Hence, we propose Contrastive Guided Diffusion Models for Adversarial Purification.

4.1. Forward Diffusion Process for Adversarial Purification

Here we consider the continuous-time diffusion models where $t \in [0, 1]$ (Song et al., 2021). The forward diffusion process for adversarial purification is a forward SDE. It gradually adds Gaussian noises to adversarial examples $\mathbf{x}^a \in \mathbb{R}^d$ from $t = 0$ to $t = 1$, i.e., $\mathbf{x}(0) = \mathbf{x}^a$, and diffuses adversarial attacks into Gaussian noises. During the forward process, the clean data distribution $p(\mathbf{x})$ and the adversarial sample distribution $q(\mathbf{x})$ become closer, i.e., $\frac{\partial D_{\text{KL}}(p_t \| q_t)}{\partial t} \leq 0$, where the equality occurs only at $p_t = q_t$ as Gaussian noises, in other words, $t = 1$. Following DiffPure (Nie et al., 2022) and the findings in theoretical studies in Section 3.2, we aim to stop the forward process at $t^* \in (0, 1)$ to obtain a balance between removing local adversarial attacks and preserving global label semantics, as in Eq. (6). Then we stochastically solve for the purified data $\hat{\mathbf{x}}(0)$ in the reverse process starting from $\mathbf{x}(t^*)$ which can be expressed as

$$\mathbf{x}(t^*) = \sqrt{\alpha_{t^*}}\mathbf{x}^a + \sqrt{1-\alpha_{t^*}}\boldsymbol{\epsilon}, \tag{6}$$

where $\alpha_{t^*} = \prod_0^{t^*} \alpha_t$, $\alpha_t \in (0, 1)$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Hence, the classification results are more likely to be correct, compared to the case where the reverse process starts at $\mathbf{x}(1) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

4.2. Contrastive Guided Diffusion Model for Adversarial Purification

The theoretical studies in Section 3.3 prove that contrastive guidance is able to enhance the diffusion models for adversarial purification. Hence we propose to guide the diffusion model for adversarial purification using contrastive guidance.

As aforementioned in Section 4.1, the reverse process starts from $\mathbf{x}(t^*)$. Hence the whole reverse process is formulated as follows.

$$\hat{\mathbf{x}}(0) = \text{SDEINT}(\mathbf{x}(t^*), f_{\text{rev}}, g_{\text{rev}}, \bar{\boldsymbol{\omega}}, t^*, 0), \tag{7}$$

where SDEINT is a reverse SDE solver (Nie et al., 2022) with the initial value $\mathbf{x}(t^*)$, the drift coefficient f_{rev}, the diffusion coefficient g_{rev}, the Wiener process $\bar{\boldsymbol{\omega}}$, the initial time t^* and the end time 0. Here the drift coefficient and the diffusion coefficient are

$$f_{\text{rev}}(\mathbf{x}, t) := -\frac{1}{2}\sigma(t)[\mathbf{x} + 2\tilde{\epsilon}_\theta(\mathbf{x}, t)], \quad g_{\text{rev}}(t) = \sqrt{\sigma(t)},$$

where $\tilde{\epsilon}_\theta(\mathbf{x}, t)$ is the approximated score function $\epsilon_\theta(\mathbf{x}, t)$ plus the contrastive guidance. In specific, $\tilde{\epsilon}_\theta(\mathbf{x}(t, \bar{\boldsymbol{\omega}}), t)$ is

$$\tilde{\epsilon}_\theta(\mathbf{x}(t)) = \epsilon_\theta(\mathbf{x}(t)) + \lambda\nabla_{\mathbf{x}(t)}\ell(\mathbf{x}(t)_a, \mathbf{x}(t)_p; \tau), \tag{8}$$

where λ is the hyperparameter representing the strength of guidance. $\nabla_{\mathbf{x}(t)}\ell(\mathbf{x}(t)_a, \mathbf{x}(t)_p; \tau)$ is the contrastive guidance, where $\ell(\mathbf{x}(t)_a, \mathbf{x}(t)_p; \tau)$ is the contrastive loss. $\mathbf{x}(t)_a$ is the anchor at each time t. $\mathbf{x}(t)_p$ is its positive pair. There are a number of methods to select positive pairs. For example, for a minibatch $\mathcal{X}_t = \{\mathbf{x}(t)^i\}_{i=1}^m$ at the time t, the positive pair of an anchor $\mathbf{x}(t)_a^{(i)}, i = 1, 2, \cdots, m$ at the time t is $\mathbf{x}(t + \Delta t)^{(i)}$ from last time $t + \Delta t$ in the reverse process. We refer the audience to Appendix C.1 in Ouyang et al. (2023)'s work for the selection strategies. τ is the temperature which is a hyperparameter. From the theoretical findings in Section 3.3 and the theoretical results by Lu et al. (2023), we first adopt the InfoNCE loss for $\ell(\mathbf{x}(t), \mathbf{x}(t)_p; \tau)$, because it is derived to be the theoretical guaranteed loss to enhance the learning direction of signal. We also adopt the hard negative mining loss for $\ell(\mathbf{x}(t)_a, \mathbf{x}(t)_p; \tau)$ because of its empirical powerfulness (Ouyang et al., 2023).

In specific, at time t, the InfoNCE loss is

$$\ell_{\text{InfoNCE}}(\mathbf{x}(t)_a, \mathbf{x}(t)_p; \tau) = -\log\left(\frac{g_\tau\left(\mathbf{x}(t)_a, \mathbf{x}(t)_p\right)}{\sum_{k=1}^{m} \mathbf{1}_{k\neq a} g_\tau\left(\mathbf{x}(t)_a, \mathbf{x}(t)_k\right)}\right), \tag{9}$$

where m is the batch size. $g_\tau(\mathbf{x}(t)_a, \mathbf{x}(t)_p; \tau)$ is to measure the similarity between the anchor $\mathbf{x}(t)_a$ and its positive pair $\mathbf{x}(t)_p$, where $g_\tau(\mathbf{x}, \mathbf{x}'; \tau) = \exp(F(\mathbf{x})^\top F(\mathbf{x}')/\tau)$. $F(\cdot)$ is a feature extractor. Similarly, $g_\tau(\mathbf{x}(t)_a, \mathbf{x}(t)_k)$ measures the similarity between the anchor $\mathbf{x}(t)_a$ and its negative pair $\mathbf{x}(t)_k, k \neq a$. Here, we consider the negative pairs as all the samples in the same minibatch and at the time step t as the anchor $\mathbf{x}(t)_a$, which are not $\mathbf{x}(t)_a$. Overall, the InfoNCE loss aims to pull the positive pairs close and push the negative pairs away.

However, not all the samples except $\mathbf{x}(t)_a$ are true negatives. In order to resolve this issue, the hard negative mining (HNM) criterion (Chuang et al., 2020; Robinson et al., 2021) can be applied to construct the contrastive guidance as

$$\ell_{\text{HNM}}(\mathbf{x}(t)_a, \mathbf{x}(t)_p; \tau) = -\log\left(\frac{g_\tau\left(\mathbf{x}(t)_a, \mathbf{x}(t)_p\right)}{g_\tau\left(\mathbf{x}(t)_a, \mathbf{x}(t)_p\right) + \frac{m}{\tau^-} h_\tau\left(\mathbf{x}_a\right)}\right), \tag{10}$$

where τ^- is the probability of observing other classes than that of $\mathbf{x}(t)_a$. $h_\tau(\mathbf{x}_a)$ measures the similarity between different the anchor and the negative pair as

$$h_\tau(\mathbf{x}(t)_a) = \mathbb{E}_{\mathbf{x}(t)_n \sim q_\beta}\left[g_\tau\left(\mathbf{x}(t)_a, \mathbf{x}(t)_n\right)\right] - \tau^+ \mathbb{E}_{\mathbf{v}\sim q_\beta^+}[g_\tau(\mathbf{x}(t)_a, \mathbf{v})],$$

where q_β is an unnormalized von Mises-Fisher distribution with the mean direction $F(\mathbf{x})$. The concentration parameter β controls the hardness of negative mining. Approximation of q_β and q_β^+ can be achieved by Monte-Carlo importance sampling. For more details of the HNM loss, we kindly refer to work by Chuang et al. (2020) and Robinson et al. (2021). Note that HNM loss is still an unsupervised contrastive learning loss function. When we conduct guidance with it for adversarial purification, there is no label input. The HNM loss emphasizes on negative pairs whose representations are currently very similar. Hence, it can enhance the dissimilarity between the samples from different classes in the feature space.

It should be noted that the feature extractor $F(\cdot)$ requires training on each minibatch $\{\mathbf{x}^{(i)}\}_{i=1}^m$ in the training set. Also, Contrastive Guided Diffusion Model for Adversarial Purification does not require any change of the original training process of diffusion models.

5. Experiments

We investigate the empirical performance of Contrastive Guided Diffusion Model for Adversarial Purification on four benchmark datasets: CIFAR-10, CIFAR-100, the German Traffic Sign Recognition Benchmark (GTSRB) (Houben et al., 2013) and ImageNet datasets. We conduct our experiments against various adversarial attacks, compared with the adversarial defense methods with the state-of-the-art performance listed on RobustBench (Croce et al., 2020) and the other adversarial purification methods (Nie et al., 2022). The results demonstrate that the performance of Contrastive Guided Diffusion Model for Adversarial Purification excels the current state-of-the-art performance on the CIFAR-10 dataset. It also outperforms the baseline method DiffPure with the current state-of-the-art performance on the GTSRB dataset. On the CIFAR-100 dataset, we run ablation studies to scrutinize the effect of different sampling methods and the effect of contrastive guidance on diffusion models for adversarial purification. The code is available at `https://github.com/tenghuilee/ContrastDiffPurification`. The results show significant improvement on adversarial purification using contrastive guidance for diffusion models. The details are as follows.

5.1. Experimental Settings

Datasets As aforementioned, we apply four benchmark datasets: CIFAR-10, CIFAR-100, GTSRB and ImageNet datasets, for experiments. For the CIFAR-10 dataset (Krizhevsky et al., 2009), we follow Nie et al. (2022)'s settings to select data for evaluation. Note that because of the computational power constraint, we present the results from the randomly selected subsets of the dataset in Section 5.2. We also use the same setting on the CIFAR-100 (Krizhevsky et al., 2009) dataset for evaluation. The GTSRB dataset contains 39,252 training images in 43 classes and 12,629 images for testing, and the image sizes vary between 15×15 to 250×250. For the results from the full datasets, we kindly refer the audience to Appendices. In the experiment, all images from the first three datasets are reshaped to 32×32.

Classifiers, evaluation metrics and attacks We use different classifiers to evaluate the performance of our method against baselines on different datasets. For the CIFAR-10 dataset, we test the performance of all the baselines and Contrastive Guided Diffusion Model for Purification on WideResNet-28-10, WideResNet-70-16 and ResNet-50 as in RobustBench and Nie et al. (2022)'s work. The performance on the GTSRB dataset is examined on ResNet-18. In terms of the CIFAR-100 dataset, the ablation studies are conducted on WideResNet-28-10. These classifiers produce the standard accuracy and the robust accuracy. The standard accuracy measures the performance of these adversarial defense methods on test sets for all three datasets, whereas the robust accuracy measures their performance on these test sets adversarial attacked by AutoAttack and BPDA+EOT

attacks. In specific, we present our experimental results on strong adaptive attacks in this section, where AutoAttack ℓ_∞ and ℓ_2 threat models are applied. In order to address the stochasticity in diffusion models and denoising processes, we also apply expectation of time (EOT) to AutoAttack. Also, we apply the BPDA+EOT attack (Hill et al., 2021) for a fair comparison with other adversarial purification methods, including DiffPure (Nie et al., 2022).

5.2. Experimental Results

The overall experimental results on CIFAR-10 and GTSRB datasets demonstrate the improved adversarial defense using Contrastive Guided Diffusion Model for Adversarial Purification. On the CIFAR-10 dataset, we evaluate the performance of our method against adversarial training methods with the state-of-the-art performance listed on RobustBench and DiffPure in Tables 1 and 2 on AutoAttack $\ell_\infty(\epsilon = 8/255)$ and $\ell_2(\epsilon = 0.5)$ threat models, respectively. The results in Tables 1 indicate that our method outperforms adversarial training methods and DiffPure with the state-of-the-art performance defending against the AutoAttack ℓ_∞ threat model $(\epsilon = 8/255)$ evaluated by classifiers WideResNet-28-10 and WideResNet-70-16 on the robust accuracy by 12.17% and 11.52%, respectively. The standard accuracy of our method is also comparable with the state-of-the-art performance. Table 2 shows that our method can still achieve better performance than adversarial training methods and DiffPure with the state-of-the-art performance against ℓ_2 threat model $(\epsilon = 0.5)$ evaluated by WideResNet-28-10 by 0.7% in terms of the robust accuracy, with comparable performance to the state-of-the-art on standard accuracy. It indicates that the trade-off between the standard accuracy and the robust accuracy still is still unsolved by our methods and remains an interesting topic to study in the future work. Besides, for the same threat model, our method obtains higher robust accuracy and comparable standard accuracy to DiffPure, but lower standard accuracy and comparable robust accuracy to the state-of-the-art performance by adversarial training methods. It seems that our method tends to prefer stronger attacks and weaker classifiers. As adversarial training methods have seen adversarial attacks during training, it is impressive that our method outperform them without knowing the adversarial attacks beforehand. This is also validated in Table 3.

In Table 4, comparing with other adversarial purification methods evaluated against BPDA+EOT attack on WideResNet-28-10, the performance of our method is higher than the state-of-the-art performance by DiffPure in terms of the robust accuracy, at a little cost on the standard accuracy. It is consistent with our previous findings on the standard-accuracy-robust-accuracy trade-off and the preference of our method from results in Tables 1 and 2.

Table 1. Standard accuracy and robust accuracy against AutoAttack ℓ_∞ $(\epsilon = 8/255)$ on CIFAR-10, obtained by WideResNet-28-10 and WideResNet-70-16. ($t^* = 0.1$ for diffusion models)

Method	Extra Data	Standard Acc	Robust Acc
WideResNet-28-10			
(Zhang et al., 2020b)	✓	89.36	59.96
(Wu et al., 2020)	✓	88.25	62.11
(Gowal et al., 2020)	✓	89.48	62.70
(Cui et al., 2023)	✓	92.16	67.73
(Wang et al., 2023)	✗	92.44	67.31
(Xu et al., 2023)	✗	**93.69**	63.89
(Wu et al., 2020)	✗	85.36	59.18
(Rebuffi et al., 2021)	✗	87.33	61.72
(Gowal et al., 2021)	✗	87.50	65.24
(Nie et al., 2022)	✗	89.02	70.64
Ours	✗	91.41	**82.81**
WideResNet-70-16			
(Gowal et al., 2020)	✓	91.10	66.02
(Rebuffi et al., 2021)	✓	92.23	68.56
(Gowal et al., 2020)	✗	85.29	59.57
(Rebuffi et al., 2021)	✗	88.54	64.46
(Gowal et al., 2021)	✗	88.74	66.60
(Wang et al., 2023)	✗	**93.25**	70.69
(Nie et al., 2022)	✗	90.07	71.29
Ours	✗	92.97	**82.81**

Table 2. Standard accuracy and robust accuracy against AutoAttack ℓ_2 $(\epsilon = 0.5)$ on CIFAR-10, obtained by different classifier architectures. ($t^* = 0.075$ for diffusion models, and*methods use WideResNet-34-10, with the same width but more layers than the default one.)

Method	Extra Data	Standard Acc	Robust Acc
WideResNet-28-10			
(Augustin et al., 2020)*	✓	92.23	77.93
(Rony et al., 2019)	✗	89.05	66.41
(Ding et al., 2020)	✗	88.02	67.77
(Wu et al., 2020)*	✗	88.51	72.85
(Sehwag et al., 2021)*	✗	90.31	75.39
(Rebuffi et al., 2021)	✗	91.79	78.32
(Wang et al., 2023)	✗	**95.16**	83.68
(Nie et al., 2022)	✗	91.03	78.58
Ours	✗	93.75	**84.38**
WideResNet-70-16			
(Gowal et al., 2020)	✓	94.74	79.88
(Rebuffi et al., 2021)	✓	95.74	81.44
(Gowal et al., 2020)	✗	90.90	74.03
(Rebuffi et al., 2021)	✗	92.41	80.86
(Wang et al., 2023)	✗	**95.54**	**84.97**
(Nie et al., 2022)	✗	92.68	80.60
Ours	✗	90.63	82.82

Furthermore, we also scrutinize the performance on the GTSRB dataset against $\ell_\infty(\epsilon = 8/255)$ and $\ell_\infty(\epsilon = 0.5)$ threat models with AutoAttack, and the $\ell_\infty(\epsilon = 8/255)$ threat model with BPDA+EOT attack. According to results in Table 5, our method outperforms DiffPure (Nie et al., 2022) for both standard accuracy and robust accuracy by 1.56% and 27.34%, respectively against AutoAttacks with ℓ_∞ perturbations, $\epsilon = 8/255$. For AutoAttack with ℓ_2 perturbations, $\epsilon = 0.5$, our method achieves higher robust accuracy than DiffPure by 11.97% with 3.13% lower stan-

Table 3. Standard accuracy and robust accuracies against unseen threat models on ResNet-50 for CIFAR-10. We keep the same evaluation settings with (Laidlaw et al., 2021), where the attack bounds are $\epsilon = 8/255$ for AutoAttack ℓ_∞, and $\epsilon = 1$ for AutoAttack ℓ_2. The baseline results are reported from the respective papers. ($t^* = 0.125$ for diffusion models)

Method	Standard Acc	Robust Acc	
		ℓ_∞	ℓ_2
Adv. Training with ℓ_∞ (Laidlaw et al., 2021)	86.8	49.00	19.20
Adv. Training with ℓ_2 (Laidlaw et al., 2021)	85.0	39.5	47.80
PAT-self (Laidlaw et al., 2021)	82.40	30.20	34.90
ADV. CRAIG (Dolatabadi et al., 2022)	83.20	40.00	33.90
ADV. GRADMATCH (Dolatabadi et al., 2022)	83.10	39.20	34.10
DIFFPURE (Nie et al., 2022)	88.20	70.00	70.90
Ours	**96.36**	**73.44**	**79.12**

Table 4. Comparison with different adversarial purification methods using BPDA+EOT with ℓ_∞ perturbations. We evaluate on WideResNet-28-10 for CIFAR-10, and keep the experimental settings the same with (Hill et al., 2021), where $\epsilon = 8/255$. (*The purification is actually a variant of the LD sampling.)

Method	Purification	Standard Acc	Robust Acc
(Song et al., 2018)	Gibbs Update	**95.00**	9.00
(Yang et al., 2019)	Mask+Recon.	94.00	15.00
(Hill et al., 2021)	EBM+LD	84.12	54.90
(Yoon et al., 2021)	DSM+LD*	86.14	70.01
(Nie et al., 2022) ($t^* = 0.075$)	Diffusion	91.03	77.43
(Nie et al., 2022) ($t^* = 0.1$)	Diffusion	89.02	81.40
Ours ($t^* = 0.1$)	Diffusion	92.61	**81.94**

Table 5. Standard accuracy and robust accuracy on the GTSRB dataset evaluated on ResNet-18. (a) We evaluate using AutoAttack with ℓ_∞ perturbations, where $\epsilon = 8/255$. (b) We evaluate using AutoAttack with ℓ_2 perturbations with $\epsilon = 0.5$. (c) We evaluate using BPDA+EOT attack with ℓ_∞ perturbations where $\epsilon = 8/255$.

(a) AutoAttack with ℓ_∞ perturbations, $\epsilon = 8/255$

Method	Standard Acc	Robust Acc
(Nie et al., 2022) ($t^* = 0.0.075$)	77.35	42.19
Ours ($t^* = 0.075$)	**78.91**	**69.53**

(b) AutoAttack with ℓ_2 perturbations, $\epsilon = 0.5$

Method	Standard Acc	Robust Acc
(Nie et al., 2022) ($t^* = 0.075$)	**87.50**	59.38
Ours ($t^* = 0.075$)	84.37	**71.35**

(c) BPDA+EOT with ℓ_∞ perturbations, $\epsilon = 8/255$

Method	Standard Acc	Robust Acc
(Nie et al., 2022) ($t^* = 0.075$)	80.00	**61.43**
Ours ($t^* = 0.075$)	80.00	61.25

dard accuracy. Against BPDA+EOT with ℓ_∞ perturbations with $\epsilon = 8/255$, our method has the same standard accuracy with DiffPure, with 0.18% lower robust accuracy than DiffPure. This minor inferiority may be due to randomness. These results show that our method can significantly resist strong adaptive adversarial attacks. For weaker attacks, it can still at least achieve similar performance without the guidance. The ablation studies provide a deeper analysis on the effect of contrastive guidance of diffusion models for adversarial purification.

5.3. Ablation Studies

We embark on an exploration through ablation studies of our proposed method, examining its performance variations under distinct diffusion types. Additionally, we extend our analysis to another dataset, evaluating the model's robustness against ℓ_∞ perturbations ($\epsilon = 8/255$) using the AutoAttack technique. The experiments are conducted with the WideResNet-28-10 architecture, employing ($t^* = 0.1$).

Table 6. Standard accuracy and robust accuracy against AutoAttack ℓ_∞ ($\epsilon = 8/255$), obtained by WideResNet-28-10. ($t^* = 0.1$)

(a) Comparison of diffusion type under CIFAR-10

Method	Standard Acc	Robust Acc
VP-SDE	91.67	82.81
VP-ODE	93.75	69.79

(b) Comparison of diffusion under CIFAR-100

Method	Standard Acc	Robust Acc
Diffusion	62.50	8.60
Diffusion + Contrastive	57.82	24.22

Table 6a assesses the influence of various diffusion types, VP-SDE and VP-ODE, on our method in CIFAR-10. VP-SDE exhibits higher standard accuracy, while VP-ODE excels in robust accuracy. Comparisons on CIFAR-100, depicted in Table 6b, reveal the baseline method's superior standard accuracy, yet the proposed method outperforms in robust accuracy. For the results on the ImageNet dataset, we kindly refer the audience to Appendices.

6. Conclusion and Discussion

This research addresses the critical challenges in adversarial purification. We introduce Contrastive Guided Diffusion Model for Adversarial Purification. This framework effectively neutralizes adversarial attacks while preserving image semantics. Supported by the rigorous theory and extensive experiments, our approach excels, especially against strong adaptive adversarial threats. This work not only overcomes

existing limitations but also paves the way for innovative and practical advancements in adversarial defense.

Limitations While our approach excels against various adversarial threats, it may require further optimization to handle high-dimensional data efficiently. Additionally, the computational cost of our method, particularly in resource-constrained environments, remains a potential limitation.

Impact Statement

This paper presents work whose goal is to advance the field of adversarial defense. There are many potential societal and ethical consequences of our work, such as enhancing technological guarantees for social risk management on citizens' asset safety.

Acknowledgements

This research was supported by JSPS Invitational Fellowship 2023 Number S23142, and in part by National Natural Science Foundation of China under Grants 62103110. MB, WH and AW were partially supported by RIKEN Incentive Research Project 100847-202301062011. WH was partially supported by JSPS KAKENHI Grant Number 24K20848. TL was supported by RIKEN's IPA Program. We also want to thank Dr. Chao Li for his valuable advice and proofreading this work.

References

Augustin, M., Meinke, A., and Hein, M. Adversarial robustness on in- and out-distribution improves explainability. In *European Conference on Computer Vision*, pp. 228–245. Springer, 2020.

Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. Universal guidance for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Cao, Y., Chen, Z., Belkin, M., and Gu, Q. Benign overfitting in two-layer convolutional neural networks. *Advances in Neural Information Processing Systems*, 35:25237–25250, 2022.

Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A., and Jegelka, S. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775, 2020.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. RobustBench: A standardized adversarial robustness benchmark. *arXiv:2010.09670*, 2020.

Cui, J., Tian, Z., Zhong, Z., Qi, X., Yu, B., and Zhang, H. Decoupled Kullback-Leibler divergence loss. *arXiv*, 2305.13948, 2023.

Dhariwal, P. and Nichol, A. Q. Diffusion models beat GANs on image synthesis. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Ding, G. W., Sharma, Y., Lui, K. Y. C., and Huang, R. MMA training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020.

Dolatabadi, H. M., Erfani, S., and Leckie, C. ℓ_∞-robustness and beyond: Unleashing efficient adversarial training. In *European Conference on Computer Vision*, 2022.

Frosio, I. and Kautz, J. The best defense is a good offense: Adversarial agumentation against adversarial attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv:1810.12715*, 2019.

Gowal, S., Qin, C., Uesato, J., Mann, T., and Kohli, P. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv:2010.03593*, 2020.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D. A., and Mann, T. A. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34, 2021.

Hill, M., Mitchell, J. C., and Zhu, S.-C. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. In *International Conference on Learning Representations*, 2021.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv*, 2207.12598, July 2022.

Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., and Igel, C. Detection of traffic signs in real-world images: The German traffic sign detection benchmark. In *International Joint Conference on Neural Networks*, pp. 1–8, 2013.

Kou, Y., Chen, Z., Chen, Y., and Gu, Q. Benign overfitting in two-layer relu convolutional neural networks. In *International Conference on Machine Learning*, pp. 17615–17659. PMLR, 2023.

Krizhevsky, A. et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Laidlaw, C., Singla, S., and Feizi, S. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2021.

Lee, M. and Kim, D. Robust evaluation of diffusion-based adversarial purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 134–144, October 2023.

Lu, C., Chen, H., Chen, J., Su, H., Li, C., and Zhu, J. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning. In *International Conference on Machine Learning*, volume 202, pp. 22825–22855, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., and Anandkumar, A. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, volume 162, pp. 16805–16827, 2022.

Ouyang, Y., Xie, L., and Cheng, G. Improving adversarial robustness through the contrastive-guided diffusion process. In *International Conference on Machine Learning*, volume 202, pp. 26699–26723, 2023.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. A. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29935–29948, 2021.

Robinson, J. D., Chuang, C.-Y., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

Rony, J., Hafemann, L. G., Oliveira, L. S., Ayed, I. B., Sabourin, R., and Granger, E. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4322–4330, 2019.

Samangouei, P., Kabkab, M., and Chellappa, R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in Neural Information Processing Systems*, 31, 2018.

Sehwag, V., Mahloujifar, S., Handina, T., Dai, S., Xiang, C., Chiang, M., and Mittal, P. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2021.

Shah, K., Chen, S., and Klivans, A. Learning mixtures of Gaussians using the DDPM objective. *arXiv:2307.01178*, 2023.

Shi, C., Holtz, C., and Mishne, G. Online adversarial purification based on self-supervised learning. In *International Conference on Learning Representations*, 2021.

Shi, Y., Du, M., Wu, X., Guan, Z., Sun, J., and Liu, N. Black-box backdoor defense via zero-shot image purification. In *Advances in Neural Information Processing Systems*, 2023.

Song, J., Zhang, Q., Yin, H., Mardani, M., Liu, M.-Y., Kautz, J., Chen, Y., and Vahdat, A. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, volume 202, pp. 32483–32498, 2023.

Song, Y., Kim, T., Nowozin, S., Ermon, S., and Kushman, N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Wang, J., Lyu, Z., Lin, D., Dai, B., and Fu, H. Guided diffusion model for adversarial purification. *arXiv*, 2205.14969, June 2022.

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, volume 202, pp. 36246–36263, 2023.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2958–2969, 2020.

Xiao, C., Chen, Z., Jin, K., Wang, J., Nie, W., Liu, M., Anandkumar, A., Li, B., and Song, D. DensePure: Understanding diffusion models for adversarial robustness. In *International Conference on Learning Representations*, 2023.

Xu, Y., Sun, Y., Goldblum, M., Goldstein, T., and Huang, F. Exploring and exploiting decision boundary dynamics for adversarial robustness, 2023.

Yang, Y., Zhang, G., Katabi, D., and Xu, Z. ME-Net: Towards effective adversarial robustness with matrix estimation. In *International Conference on Machine Learning*, 2019.

Yoon, J., Hwang, S. J., and Lee, J. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, volume 139, pp. 12062–12072, 2021.

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhang, H., Chen, H., Xiao, C., Gowal, S., Stanforth, R., Li, B., Boning, D., and Hsieh, C.-J. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020a.

Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., and Kankanhalli, M. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2020b.

A. Proofs

Proof of Lemma 3.2. The Lemma can be proved by showing $y_i\langle \mathbf{x}_i, \boldsymbol{\theta}^*\rangle > 0$. By the definition of the model $\boldsymbol{\theta}^*$, we have

$$\begin{aligned} y_i\langle \mathbf{x}_i, \boldsymbol{\theta}^*\rangle =& \langle \boldsymbol{\mu} + y_i\boldsymbol{\xi}_i, \boldsymbol{\mu} + \frac{1}{n}\sum_{i'=1}^{n} y_{i'}\boldsymbol{\xi}_{i'}\rangle \\ =& \langle \boldsymbol{\mu}, \boldsymbol{\mu}\rangle + y_i\langle \boldsymbol{\xi}_i, \boldsymbol{\mu}\rangle + \frac{1}{n}\sum_{i'=1}^{n} y_{i'}\langle \boldsymbol{\mu}, \boldsymbol{\xi}_{i'}\rangle + \frac{1}{n}\langle y_i\boldsymbol{\xi}_i, y_i\boldsymbol{\xi}_i\rangle + \frac{1}{n}\sum_{i'\neq i}^{n} y_i y_{i'}\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle \\ \geq& \|\boldsymbol{\mu}\|_2^2 - |\langle \boldsymbol{\xi}_i, \boldsymbol{\mu}\rangle| - \frac{1}{n}\sum_{i'=1}^{n} |\langle \boldsymbol{\mu}, \boldsymbol{\xi}_{i'}\rangle| + \frac{1}{n}\|\boldsymbol{\xi}_i\|_2^2 - \frac{1}{n}\sum_{i'\neq i}^{n} |\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle|. \end{aligned} \tag{11}$$

By Lemma 3.1, we will have

$$y_i\langle \mathbf{x}_i, \boldsymbol{\theta}^*\rangle \geq \|\boldsymbol{\mu}\|_2^2 - 2\|\boldsymbol{\mu}\|_2\sqrt{2\log(8n/\delta)} + \frac{d}{2n} - 2\frac{n-1}{n}\sqrt{d\log(4n^2/\delta)} > 0,$$

according to the given conditions. □

Proof of Lemma 3.3. We prove the Lemma by showing $y_i\langle \boldsymbol{\theta}^*, \mathbf{x}_i^a\rangle < 0$.

$$\begin{aligned} y_i\langle \boldsymbol{\theta}^*, \mathbf{x}_i^a\rangle &= y_i\left\langle \frac{1}{n}\sum_{i'=1}^{n} y_{i'}\mathbf{x}_{i'}, \mathbf{x}_i + \boldsymbol{\delta}_i\right\rangle \\ &= \frac{1}{n}\sum_{i'=1}^{n} y_i y_{i'}\langle \mathbf{x}_{i'}, \mathbf{x}_i\rangle + \frac{1}{n}\sum_{i'=1}^{n} y_i y_{i'}\langle \mathbf{x}_{i'}, \boldsymbol{\delta}_i\rangle \\ &= \|\boldsymbol{\mu}\|_2^2 + \frac{1}{n}\|\boldsymbol{\xi}_i\|_2^2 + y_i\langle \boldsymbol{\xi}_i, \boldsymbol{\mu}\rangle + \frac{1}{n}\sum_{i'\neq i}^{n} y_i y_{i'}\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle + \frac{1}{n}\sum_{i'=1}^{n} y_{i'}\langle \boldsymbol{\mu}, \boldsymbol{\xi}_{i'}\rangle + y_i\langle \boldsymbol{\mu}, \boldsymbol{\delta}_i\rangle + \frac{1}{n}\sum_{i'=1}^{n} y_i y_{i'}\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\delta}_i\rangle \\ &\leq 2\left(\|\boldsymbol{\mu}\|_2^2 + \frac{1}{n}\|\boldsymbol{\xi}_i\|_2^2\right) + \left\langle \boldsymbol{\mu}, -32n\mathrm{SNR}^2\sum_{i'=1}^{n} y_{i'}\boldsymbol{\xi}_{i'}\right\rangle + \frac{1}{n}\sum_{i'=1}^{n} y_{i'}\left\langle \boldsymbol{\xi}_{i'}, -32n\mathrm{SNR}^2\sum_{i'=1}^{n} y_{i'}\boldsymbol{\xi}_{i'}\right\rangle \\ &\leq 2\|\boldsymbol{\mu}\|_2^2 + \frac{3d}{n} + 32n^2\mathrm{SNR}^2\|\boldsymbol{\mu}\|_2\sqrt{2\log(8n/\delta)} - 32n\mathrm{SNR}^2\left(\frac{d}{2} - 2n\sqrt{d\log(4n^2/\delta)}\right) \\ &< -\frac{1}{2}\|\boldsymbol{\mu}\|_2^2 \\ &< 0, \end{aligned}$$

where the first inequality holds due to the last line of Eq. (11), the second inequality is by Lemma 3.1 and the third inequality is by $n\mathrm{SNR} > 1$ and $d > \max\{8n\|\boldsymbol{\mu}\|_2\sqrt{2\log(8n/\delta)}, 64n^2\log(4n^2/\delta)\}$. □

Proof of Theorem 3.4. Define $\widetilde{\mathbf{x}}(t) = \mathbf{x}(t) - \sqrt{2}\boldsymbol{\omega}(t)$, then we find that

$$d\widetilde{\mathbf{x}}(t) = -\tanh(\langle \widetilde{\mathbf{x}}(t) + \sqrt{2}\boldsymbol{\omega}(t), \alpha_t\boldsymbol{\mu}\rangle)\alpha_t\boldsymbol{\mu}dt.$$

By concentration property stated in Lemma 3.1, we know that with probability at least $1-\delta$,

$$|\langle \boldsymbol{\omega}(t), \boldsymbol{\mu}\rangle| < \|\boldsymbol{\mu}\|_2\sqrt{2\log(8n/\delta)}.$$

Define $\gamma(t) \triangleq \langle \widetilde{\mathbf{x}}(t), \boldsymbol{\mu}\rangle$, and $\rho(t) \triangleq \sqrt{2}\langle \boldsymbol{\omega}(t), \boldsymbol{\mu}\rangle$, then we obtain,

$$d\gamma(t) = -\tanh(\alpha_t(\gamma(t) + \rho(t)))\alpha_t\|\boldsymbol{\mu}\|_2^2 dt.$$

According to the forward diffusion process, the distribution of attacked example at time t can be calculated

$$\begin{aligned}
q_t(\mathbf{x}^a(t)) &= \int q_t(\mathbf{x}^a(t)|\mathbf{x}^a(0))q(\mathbf{x}^a(0))\mathrm{d}\mathbf{x}^a(0) \\
&= \frac{1}{2}\int (2\pi\sigma_t^2)^{-\frac{d}{2}}\exp(-\frac{(\mathbf{x}^a(t)-\alpha_t\mathbf{x}^a(0))^2}{2\sigma_t^2})(2\pi\sigma_a^2)^{-\frac{d}{2}}\exp(-\frac{(\mathbf{x}^a(0)-\boldsymbol{\mu})^2}{2\sigma_a^2})d\mathbf{x}^a(0) \\
&\quad + \frac{1}{2}\int (2\pi\sigma_t^2)^{-\frac{d}{2}}\exp(-\frac{(\mathbf{x}^a(t)-\alpha_t\mathbf{x}^a(0))^2}{2\sigma_t^2})(2\pi\sigma_a^2)^{-\frac{d}{2}}\exp(-\frac{(\mathbf{x}^a(0)+\boldsymbol{\mu})^2}{2\sigma_a^2})d\mathbf{x}^a(0) \\
&= \sum_{y=\pm 1}\frac{1}{2}\mathcal{N}(\mathbf{x}^a(t)|y_i\alpha_t\boldsymbol{\mu},(\sigma_t^2+\alpha_t^2\sigma_a^2)\mathbf{I})
\end{aligned}$$

with a variance constant $\sigma_a^2 = 1 - 64n\|\boldsymbol{\mu}\|_2^2/d + 1024n^3\|\boldsymbol{\mu}\|_2^4/d^2$. Thus, the adversarial example diffused in the forward process at time t can be written as

$$\mathbf{x}^a(t) = y\alpha_t\boldsymbol{\mu} + \tilde{\boldsymbol{\xi}}^a(t),$$

where $\tilde{\boldsymbol{\xi}}^a(t)$ is the perturbation component satisfying $\tilde{\boldsymbol{\xi}}^a(t) \sim \mathcal{N}(\mathbf{0},(\sigma_t^2+\alpha_t^2\sigma_a^2)\mathbf{I}))$.

Now we choose a t^*, such that

$$y\gamma(t^*) > |\rho(t^*)|. \tag{12}$$

To achieve the above requirement, we have

$$\begin{aligned}
y\gamma(t^*) &= \langle \alpha_{t^*}\boldsymbol{\mu} + y\tilde{\boldsymbol{\xi}}^a(t),\boldsymbol{\mu}\rangle \\
&= \alpha_{t^*}\|\boldsymbol{\mu}\|_2^2 + y\langle\tilde{\boldsymbol{\xi}}^a(t),\boldsymbol{\mu}\rangle \\
&\geq \exp(-t^*)\|\boldsymbol{\mu}\|_2^2 - \sqrt{\sigma_t^2+\alpha_t^2\sigma_a^2}\|\boldsymbol{\mu}\|_2\sqrt{2\log(8n/\delta)} \\
&\geq \exp(-t^*)\|\boldsymbol{\mu}\|_2^2 - [1-\exp(-2t^*)+\exp(-t^*)(1+32n^{\frac{3}{2}}\|\boldsymbol{\mu}\|_2^2/d)]\|\boldsymbol{\mu}\|_2\sqrt{2\log(8n/\delta)} \\
&\geq \exp(-t^*)\|\boldsymbol{\mu}\|_2^2 - (\|\boldsymbol{\mu}\|_2\sqrt{2\log(8n/\delta)} + 2\sqrt{n}\exp(-t^*)\|\boldsymbol{\mu}\|_2) \\
&\geq \frac{1}{2}\exp(-t^*)\|\boldsymbol{\mu}\|_2^2.
\end{aligned}$$

Th first inequality is by Lemma 3.1, the second inequality is due to $\sqrt{x^2+y^2} \leq |x|+|y|$, and the third inequality is by $n\mathrm{SNR}^2 \leq \frac{1}{16\sqrt{2\log(8n/\delta)}}$. The last inequality is by $t^* \leq \log(\frac{\|\boldsymbol{\mu}\|_2}{4\sqrt{2\log(8n/\delta)}})$ and $\|\boldsymbol{\mu}\|_2 \geq 4\sqrt{2n\log(8n/\delta)}$.

On the other hand, by Lemma 3.1, we know

$$|\rho(t^*)| \leq \|\boldsymbol{\mu}\|_2\sqrt{4\log(8n/\delta)}.$$

As a result, with the condition that $t^* \leq \log(\frac{\|\boldsymbol{\mu}\|_2}{20\sqrt{2\log(8n/\delta)}})$, we conclude that

$$y\gamma(t^*) \geq 5|\rho(t^*)|.$$

Furthermore, we consider the discrete reverse process, which leads to

$$\gamma(t-1)-\gamma(t) = \tanh(\exp(-t)(\gamma(t)+\rho(t)))\exp(-t)\|\boldsymbol{\mu}\|_2^2.$$

By inequality (12), we know that, the sign of $\gamma(t^*)+\rho(t^*)$ is dominated by $\gamma(t^*)$. Without loss of generality, we consider

$y_i = 1$. Taking telescoping sum over $t = t^*, t^* - 1, \cdots, 0$ then gives

$$\begin{aligned}\gamma(0) &\leq \sum_{t=0}^{t^*} \tanh(\exp(-t)(\gamma(t) + \rho(t))) \exp(-t)\|\boldsymbol{\mu}\|_2^2 \\ &\leq \sum_{t=0}^{t^*} \exp(-2t)(\gamma(t) + \rho(t))\|\boldsymbol{\mu}\|_2^2 \\ &\leq \exp(-\frac{1}{2}) + \frac{1}{2}\exp(-t^*)\|\boldsymbol{\mu}\|_2^2 - \exp(-\frac{1}{2}\exp(-2\|\boldsymbol{\mu}\|_2^2 t^*)).\end{aligned}$$

□

B. More Experimental Results

B.1. A Discussion on Gradient of Diffusion Models Involved in Adversarial Purification

As aforementioned in Section 5.2, diffusion-based adversarial purification methods, i.e., DiffPure (Nie et al., 2022) and our proposed model, utilize the settings on the gradient as Nie et al. (2022) for the experimental design. In other words, the gradients of diffusion models and the gradients of classifiers are accessible to the attackers. In this section, we also demonstrate the results from the settings where attackers cannot access gradients of the diffusion-based purifiers: DiffPure, guided diffusion models for adversarial purification (GDMP) (Wang et al., 2022) and our proposed model in Table 7.

Table 7. Standard accuracy and robust accuracy against different attackers: AutoAttack with $\ell_\infty(\epsilon = 8/255)$, AutoAttack $\ell_2(\epsilon = 1)$, AutoAttack $\ell_2(\epsilon = 0.5)$, and the PGD+EOT attack on WideResNet-28-10, WideResNet-70-16 or ResNet-50 for the CIFAR-10 dataset, without accesses to the gradients of diffusion-based purifiers.

t^*	ϵ	Solver	Attacker	Classifier	Purifier	Standard Accuracy	Robust Accuracy
100	$8/255$	SDE	AutoAttack ℓ_∞	WideResNet-28-10	DiffPure	89.32 ± 0.79	77.86 ± 0.75
					GDMP	90.23 ± 1.94	77.80 ± 0.88
					Ours	89.71 ± 0.82	77.08 ± 1.06
100	$8/255$	ODE	AutoAttack ℓ_∞	WideResNet-28-10	DiffPure	90.95 ± 0.56	67.38 ± 0.32
					Ours	91.80 ± 1.05	68.36 ± 1.52
100	$8/255$	SDE	AutoAttack ℓ_∞	WideResNet-70-16	DiffPure	90.23 ± 0.28	80.21 ± 0.49
					Ours	91.08 ± 0.75	79.10 ± 0.42
125	$8/255$	SDE	AutoAttack ℓ_∞	ResNet-50	DiffPure	87.96 ± 0.92	76.69 ± 0.91
					Ours	88.35 ± 0.64	75.91 ± 0.33
125	1	SDE	AutoAttack ℓ_2	ResNet-50	DiffPure	88.02 ± 0.72	76.30 ± 2.08
					Ours	87.04 ± 0.96	74.80 ± 0.28
75	0.5	SDE	AutoAttack ℓ_2	WideResNet-28-10	DiffPure	91.15 ± 0.33	81.84 ± 0.84
					GDMP	91.02 ± 0.97	80.60 ± 1.44
					Ours	91.21 ± 0.89	80.92 ± 1.09
75	0.5	SDE	AutoAttack ℓ_2	WideResNet-70-16	DiffPure	92.97 ± 0.28	84.05 ± 1.12
					Ours	93.03 ± 0.51	83.01 ± 0.28
100	$8/255$	SDE	PGD+EOT	WideResNet-28-10	DiffPure	89.78 ± 1.30	84.44 ± 0.92
					GDMP	89.91 ± 1.34	86.00 ± 0.51
					Ours	89.78 ± 1.30	84.83 ± 0.97

In specific, we compare the performance of DiffPure in the cases where the attacker can and cannot obtain the gradients of the purifier against the PGD+EOT attack on WideResNet-28-10 for the CIFAR-10 dataset in Table 8. The results demonstrate that turning on the gradients of purifiers, i.e., accesses to the gradient of purifiers, can allow the attackers to attack the purifiers as well and hence reduce the effect of the purifiers, where the standard accuracy and the robust accuracy are both lower than the cases of no access to the gradients of purifiers.

Table 8. Standard accuracy and robust accuracy of DiffPure with attackers having accesses or no access to the gradient of the purifier, against the PGD+EOT attack with $t^* = 100$, $\epsilon = 8/255$, the solver as SDE, evaluated by the classifier WideResNet-28-10

Gradient On/Off	Standard Accuracy	Robust Accuracy
DiffPure Grad On	89.52 ± 1.21	81.70 ± 0.24
DiffPure Grad Off	89.78 ± 1.30	84.44 ± 0.92

B.2. Experimental Results on the PGD+EOT Attack

The PGD+EOT attack is recently scrutinized and identified as an effective method for measuring the robustness of purification methods against adversarial attacks (Lee & Kim, 2023). We further investigate our proposed method against the PGD+EOT attack, comparing with other two diffusion models for adversarial purification: DiffPure (Nie et al., 2022) and GDMP (Wang et al., 2022) on the CIFAR-100 dataset and the ImageNet dataset in Tables 9 and 10, with the attackers having no access to the gradients of the purifiers. For the CIFAR-10 dataset, the performance of DiffPure, GDMP and our method against the PGD+EOT attack is presented in the last row of Table 7.

Table 9. Standard accuracy and robust accuracy of DiffPure, GDMP and our proposed method with $t^* = 100$, $\epsilon = 8/255$ and the SDE solver, against the PGD+EOT attack without the access to the gradient of purifiers on the CIFAR-100 dataset, evaluated by the classifier WideResNet-28-10

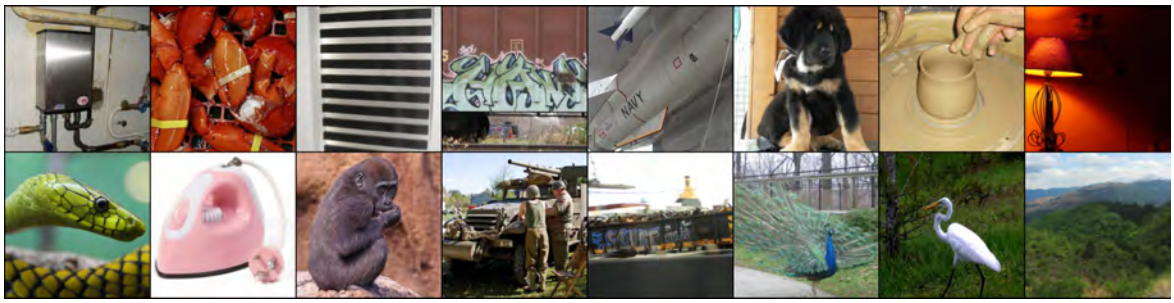
Method	Standard Accuracy	Robust Accuracy
DiffPure	50.20 ± 1.27	34.64 ± 0.09
GDMP	50.13 ± 1.21	34.90 ± 1.04
Ours	50.20 ± 1.27	34.70 ± 1.13

Table 10. Standard accuracy and robust accuracy of DiffPure, GDMP and our proposed method with $t^* = 100$, $\epsilon = 4/255$ and the SDE solver, against the PGD+EOT attack without the access to the gradient of purifiers on the ImageNet dataset, evaluated by the classifiers ResNet-50 and the transformer classifier xcit-small-24-p16-224 [3]

Classifier	Method	Standard Accuracy	Robust Accuracy
ResNet-50	DiffPure	70.41 ± 0.29	42.58 ± 0.20
ResNet-50	GDMP	70.41 ± 0.29	42.58 ± 0.00
ResNet-50	Ours	70.41 ± 0.29	41.70 ± 0.10
xcit-small-24-p16-224	DiffPure	76.56 ± 0.59	55.57 ± 0.10
xcit-small-24-p16-224	GDMP	76.56 ± 0.39	55.27 ± 0.00
xcit-small-24-p16-224	Ours	76.56 ± 0.59	55.47 ± 0.59

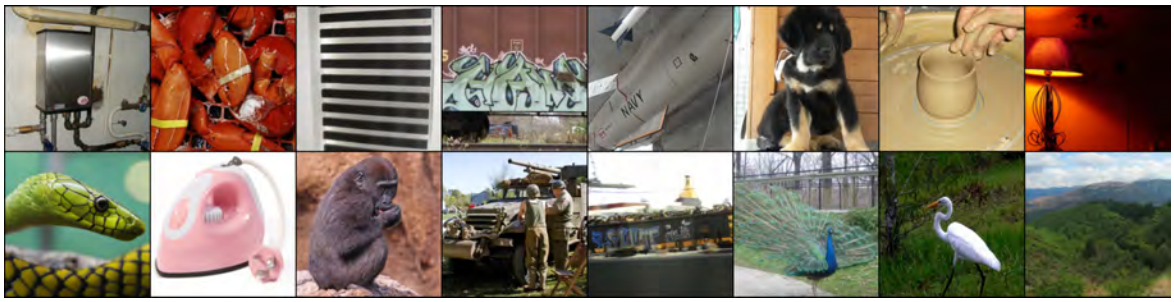
B.3. Qualitative Results

Furthermore, we also qualitatively examine the performance of our proposed method to evaluate the purification result from the human perception. We demonstrate the purification results on the ImageNet dataset in Figure 1 and the CIFAR-10 dataset in Figure 2, against the PGD+EOT and the AutoAttack, respectively. Figures 1 and 2 indicate that our proposed method is able to purify the attacked images successfully. The contours and the colors are well preserved in the purification results. From human perception we can easily recognize the saliency and the label. Also, the details of the purification results are vivid and very similar to the clean images, even for the difficult cases such as fingers, numbers and alphabetical letters.

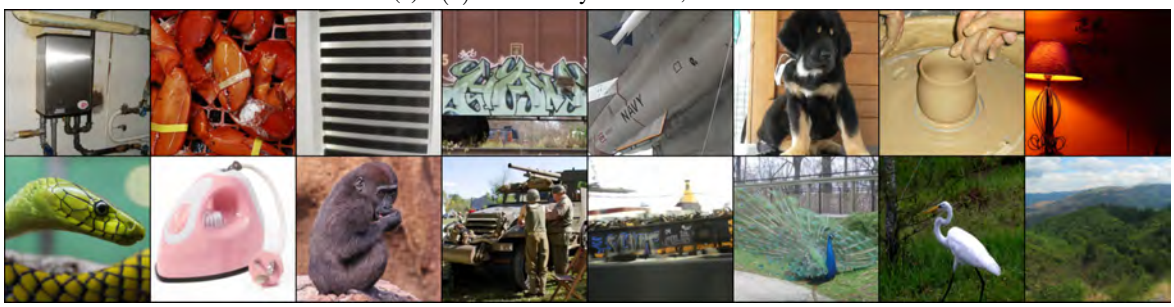


(a) Clean images

(b) $\mathbf{x}(t^*), t^* = 100$: Diffused adversarial examples by the PGD+EOT attack with $\epsilon = 4/255$ for 100 steps



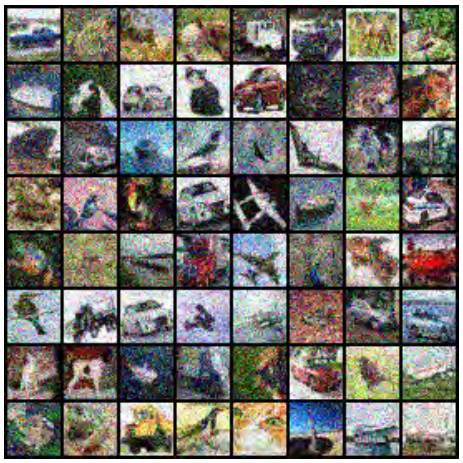
(c) $\hat{\mathbf{x}}(0)$: Purified by DiffPure, $t^* = 100$



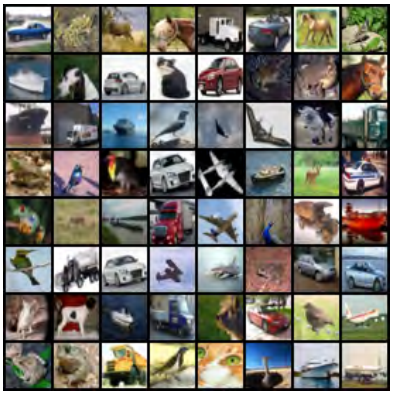
(d) $\hat{\mathbf{x}}(0)$: Purified by GDMP, $t^* = 100$

(e) $\hat{\mathbf{x}}(0)$: Purified by our method, $t^* = 100$

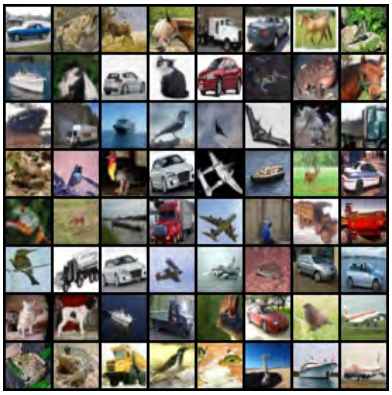
Figure 1. Qualitative results evaluated on the ImageNet dataset



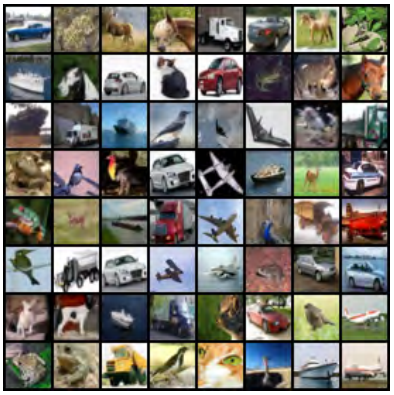
(a) $\mathbf{x}(t^*)$: Diffused adversarial examples by AutoAttack ℓ_∞ with $\epsilon = 8/255$ for 100 steps



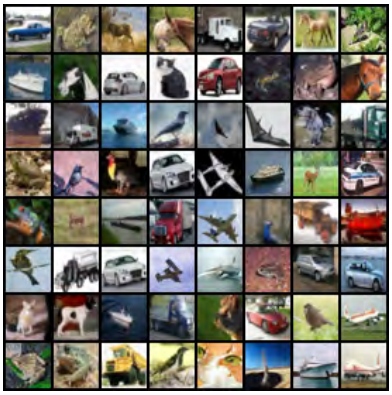
(b) Clean images



(c) $\hat{\mathbf{x}}(0)$: Purified by DiffPure, $t^* = 100$



(d) $\hat{\mathbf{x}}(0)$: Purified by GDMP, $t^* = 100$



(e) $\hat{\mathbf{x}}(0)$: Purified by our method, $t^* = 100$

Figure 2. Qualitative results on the CIFAR-10 dataset