Efficient Image-to-Image Diffusion Classifier for Adversarial Robustness

Hefei Mei[1], **Minjing Dong**[*1], **Chang Xu**[2]

[1]City University of Hong Kong, China
[2]University of Sydney, Australia
hefeimei2-c@my.cityu.edu.hk, minjdong@cityu.edu.hk, c.xu@sydney.edu.au

Abstract

Diffusion models (DMs) have demonstrated great potential in the field of adversarial robustness, where DM-based defense methods can achieve superior defense capability without adversarial training. However, they all require huge computational costs due to the usage of large-scale pre-trained DMs, making it difficult to conduct full evaluation under strong attacks and compare with traditional CNN-based methods. Simply reducing the network size and timesteps in DMs could significantly harm the image generation quality, which invalidates previous frameworks. To alleviate this issue, we redesign the diffusion framework from generating high-quality images to predicting distinguishable image labels. Specifically, we employ an image translation framework to learn many-to-one mapping from input samples to designed orthogonal image labels. Based on this framework, we introduce an efficient Image-to-Image diffusion classifier with a pruned U-Net structure and reduced diffusion timesteps. Besides the framework, we redesign the optimization objective of DMs to fit the target of image classification, where a new classification loss is incorporated in the DM-based image translation framework to distinguish the generated label from those of other classes. We conduct sufficient evaluations of the proposed classifier under various attacks on popular benchmarks. Extensive experiments show that our method achieves better adversarial robustness with fewer computational costs than DM-based and CNN-based methods.

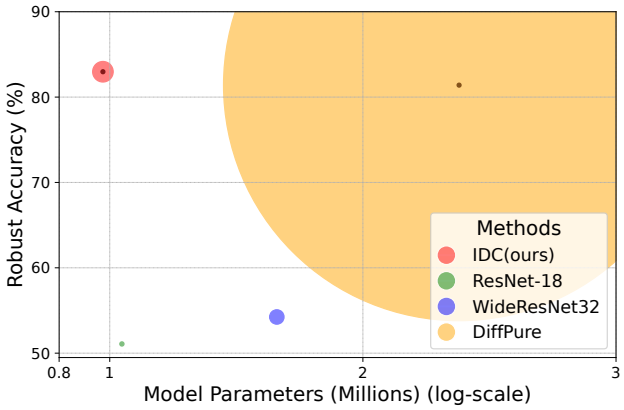


Figure 1: Comparison of robust accuracy with both CNN-based and DM-based benchmarks against BPDA+EOT attack. The area of a circle demonstrates the inference FLOPs where the FLOPs of our introduced IDC are 27.09G and those of DiffPure are 14.26T. The dark center of the circle is located at the accurate value.

Introduction

Diffusion models (DMs) have achieved excellent performance in high-quality generative tasks (Ho, Jain, and Abbeel 2020; Song et al. 2020; Song, Meng, and Ermon 2020; Dhariwal and Nichol 2021; Rombach et al. 2022; Ma et al. 2024a,b,c), leading to a range of relative applications, such as in-painting (Lugmayr et al. 2022; Zhang et al. 2023), audio or video synthesis (Ho et al. 2022; Singer et al. 2022; Wang et al. 2024; Chen et al. 2024), image super-resolution (Saharia et al. 2022b; Li et al. 2022) and image-to-image translation (Saharia et al. 2022a; Sasaki, Willcocks, and Breckon 2021; Li et al. 2023b).

Adversarial robustness refers to the capability of a model to maintain its accuracy when faced with deliberately crafted adversarial inputs intended to deceive it (Rice, Wong, and Kolter 2020; Liu et al. 2018; Dong and Xu 2023; Kang, Song, and Li 2024; Dong et al. 2022). Previous defense methods mainly focus on adversarial training (Madry et al. 2018) which involves generated adversarial examples during optimization. On the contrary, DM-based adversarial defense methods mainly utilize the inherent defense ability of DMs, such as adversarial purification that applies DMs to remove perturbations (Nie et al. 2022) and robust diffusion classifier (RDC) that treats DMs as naturally robust generative classifiers (Chen et al. 2023a). These DM-based methods have demonstrated superior performance in achieving adversarial robustness compared with CNNs.

Despite their advanced defense capability, the computational expense of these methods could be rather high, which makes it expensive to deploy them in real-world applications or even conduct a full evaluation on the entire test dataset compared with traditional CNN-based methods. For example, RDC (Chen et al. 2023a) adopts a Zero-shot Diffusion Classifier (ZDC) (Li et al. 2023a) as the prediction paradigm, which relies on a large-scale pre-trained DM

*Corresponding Author

and needs to iteratively perform a diffusion process for all classes to derive class distribution. DiffPure (Nie et al. 2022) connects pre-trained DM with traditional classifier (He et al. 2016; Zagoruyko and Komodakis 2016), where DM serves as a purification transformation of potential adversarial examples. However, its computation during the inference stage is boosted due to the complexity of large U-Net (Ronneberger, Fischer, and Brox 2015), diffusion timesteps, and an additional classifier. As illustrated in Figure 1, the FLOPs of ResNet-18 (He et al. 2016) are 1.12G, while those of DiffPure are 14.26T. Thus, we are motivated to explore a more efficient manner to incorporate diffusion models in adversarial robustness. A naive method is to reduce the size of the U-Net and decrease the number of diffusion timesteps, however, such simplification conflicts with the objective of high-quality generation in DMs which could play an important role in DM-based methods. We mainly attribute this issue to the gap between the generation task in DMs and the classification task in adversarial robustness.

In this paper, we aim to redesign the diffusion framework for adversarial robustness to eliminate this gap, relieving the pressure on DMs to generate high-quality images, which enables an efficient incorporation of DMs in adversarial robustness. Specifically, we regard classification tasks as image translation tasks achieved by DMs, where the image is translated to the pre-defined labels that are presented by different images. Since the objective is to align images with their corresponding image labels instead of high-quality image generation, the required complexity of DMs in adversarial robustness can be significantly reduced. With this simple framework, we introduce an Image-to-Image Diffusion Classifier (IDC) to effectively and efficiently apply DMs to adversarially robust image classification. In order to guarantee the diversity of pre-defined image labels for classification, we propose to construct orthogonal image labels in pixel space. Then IDC employs an image-to-image translation framework to learn a many-to-one mapping from input images to these orthogonal image labels, and the classification can be achieved by measuring the distances between translated results and all pre-defined image labels. Due to the switch of objectives, we are able to alleviate the computation of DMs in different aspects, including pruning the U-Net structure and reducing the diffusion timesteps. Furthermore, the vanilla optimization of translation frameworks only minimizes the distance between the translated result and its corresponding image labels without considering other classes. Thus, we incorporate a classification loss in the optimization of DM-based image translation to better fit the classification objective. Our introduced IDC achieves superior performance in the trade-offs between model efficiency and adversarial robustness. As shown in Figure 1, we achieve more competitive robustness with fewer model parameters than CNN-based adversarial training methods and much lower FLOPs cost than the DM-based method.

We summarize the main contributions as follows:

- We propose a novel Image-to-Image diffusion classifier with pre-defined orthogonal image labels, which naturally converts image generation tasks to classification tasks.
- Through task conversion, we propose to reduce the complexity of diffusion models without harming the performance, including network size and diffusion steps.
- We propose a classification loss during the optimization of IDC, which improves the adaptability of the image translation framework for classification tasks.
- We perform extensive experiments to empirically demonstrate the superiority of IDC on various benchmarks.

Related Work

Diffusion Models

As one of the advanced probabilistic generative models, Diffusion Models (DMs) (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Song and Ermon 2020; Dhariwal and Nichol 2021) employ a parameterized Markovian chain to estimate the target distribution, which has been used in many relative tasks, such as diffusion classification (DC) (Li et al. 2023a; Chen et al. 2023a), object detection (Chen et al. 2023b) and image-to-image translation (Saharia et al. 2022a; Li et al. 2023b). In this paper, we are concerned about the high potential of DMs in adversarial robustness.

CNN-based Adversarial Defense

The core strategies for CNN-based defense methods mainly focus on adversarial training (AT) (Madry et al. 2018; Zhang et al. 2019; Dong et al. 2022; Lin et al. 2024a) and adversarial purification (AP) (Samangouei, Kabkab, and Chellappa 2018; Grathwohl et al. 2020; Hill, Mitchell, and Zhu 2021). The former enhances model robustness by integrating adversarial examples into the training process while the latter focuses on removing adversarial noise from input samples.

DM-based Adversarial Defense

The existing diffusion classifier (Li et al. 2023a) calculates probabilities by applying Bayes' theorem to the outputs of a conditional diffusion model while RDC (Chen et al. 2023a) enhances its adversarial robustness. Diffusion purification methods (Nie et al. 2022; Zhang, Luo, and Zhang 2023; Wang et al. 2022; Wu, Ye, and Gu 2022; Lin et al. 2024b) purify adversarial samples through DMs before passing them to a cascaded general classifier. Distinct from prior research, we design an efficient DM-based robust classifier for the entire dataset evaluation.

Methodology

In this section, we explore the application of diffusion models in adversarial robustness. Specifically, we start with the preliminary of the diffusion model. Then we discuss the limitations of existing diffusion classifiers as well as diffusion models for adversarial robustness, which motivates us to introduce an efficient image-to-image diffusion classifier (IDC) to tackle the aforementioned problems. To further eliminate the gap between the diffusion-based optimization objective in IDC and the classification objective, a classification loss is thus incorporated into the loss function.

Preliminary

Diffusion Model Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) consider the generative task as a Markovian chain process. In the forward process, the DM adds Gaussian noise to real data $\boldsymbol{y}_0$ by T timesteps, which converts $\boldsymbol{y}_0$ to Gaussian noise $\boldsymbol{y}_T \sim \mathcal{N}(0,1)$. The diffusion forward process at each step can be defined as $\boldsymbol{y}_t = \sqrt{1-\beta_t}\boldsymbol{y}_0 + \sqrt{\beta_t}\boldsymbol{\epsilon}_t$, where β_t is a linearly increased scale factor, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0,1)$. To simplify the calculation, the forward process can be formulated as:

$$\boldsymbol{y}_t = \sqrt{\overline{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, \tag{1}$$

where $\alpha_t = 1-\beta_t, \overline{\alpha}_t = \prod_{s=1}^{t}\alpha_s$. In the reverse process, the DM is trained to denoise the sampled Gaussian noise $\boldsymbol{y}_t$ to $\boldsymbol{y}_{t-1}$, with which the generation of data $\hat{\boldsymbol{y}}_0$ follows another Markovian chain process. The reverse process of DM in timestep t can be formulated as:

$$\begin{aligned} p_\theta(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t) &= \mathcal{N}(\boldsymbol{y}_{t-1}; \mu_\theta(\boldsymbol{y}_t, t), \delta_t \boldsymbol{I}), \\ \mu(\boldsymbol{y}_t, t) &= \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{y}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\boldsymbol{y}_t, t)), \end{aligned} \tag{2}$$

where ϵ_θ is a noise predictor parameterized by θ. It is a common practice for DMs to utilize a U-Net (Ronneberger, Fischer, and Brox 2015) as the noise predictor, where θ can be optimized by maximizing the lower variation limit. The training loss of DMs can be formulated as:

$$\min_\theta \mathbb{E}_{t\sim U[0,1], \boldsymbol{y}_0, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left\|\epsilon_\theta\left(\boldsymbol{y}_t, t\right) - \boldsymbol{\epsilon}\right\|_2^2. \tag{3}$$

Diffusion Classifier Zero-shot Diffusion Classifier (Li et al. 2023a) performs training-free classification using the inherent capabilities of Diffusion model U_L. As shown in Figure 2(a), the core of ZDC is the application of Bayes' theorem to the outputs of a conditional generative model, combining the prediction likelihoods with prior knowledge of class distributions to calculate the posterior probability. For the input image $\boldsymbol{x}$ and a class label c_i, the ZDC utilizes the Evidence Lower Bound (ELBO) in place of $\log p_\theta(\boldsymbol{x} \mid c)$ then reformulated the posterior distribution as:

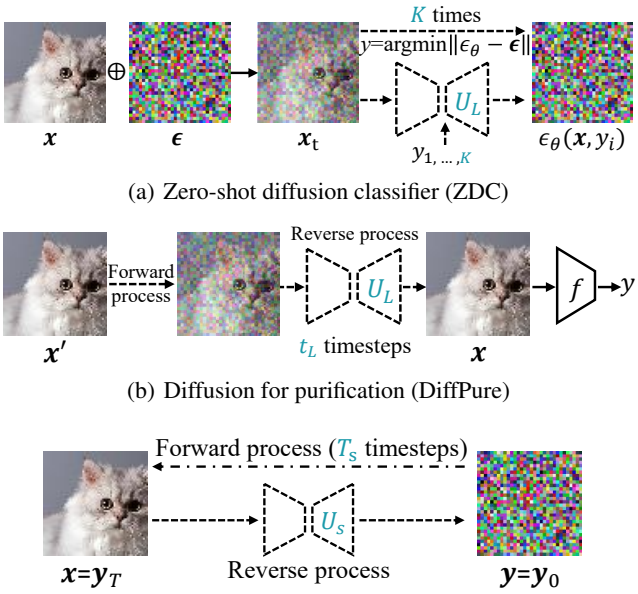
$$p_\theta(c_i \mid \boldsymbol{x}) = \frac{\exp\{-\mathbb{E}_{t,\epsilon}[\|\epsilon_\theta(\boldsymbol{x}_t, c_i) - \boldsymbol{\epsilon}\|^2]\}}{\sum_j \exp\{-\mathbb{E}_{t,\epsilon}[\|\epsilon_\theta(\boldsymbol{x}_t, c_j) - \boldsymbol{\epsilon}\|^2]\}}, \tag{4}$$

where the iteration $j \in [1, 2, 3, ..., K]$, and K is the number of the classes.

Diffusion Purification As shown in Figure 2(b), by combining the diffusion model with the classifier, Diffusion Purification (DiffPure) (Nie et al. 2022) achieves considerable adversarial robustness. During purification, the adversarial sample $\boldsymbol{x}'$ is diffused to the timestep t_L, which can be sampled as $\boldsymbol{x}(t_L) = \sqrt{\alpha(t_L)}\boldsymbol{x}' + \sqrt{1-\alpha(t_L)}\boldsymbol{\epsilon}$. Then the sample $\boldsymbol{x}'$ is purified through the reverse process by U-Net U_L and subsequently input into a CNN-based classifier to obtain the results, which can be formulated as:

$$p_{\theta_U,\theta_f}(c \mid \boldsymbol{x}) = f\left(\underset{t=1\ldots t_L}{\text{reverse}}\, U_L(\boldsymbol{x}(t_L), t)\right), \tag{5}$$

where the adversarial sample $\boldsymbol{x}'$ need to be purified by the t_L-timestep diffusion model and classified by classifier f.



(a) Zero-shot diffusion classifier (ZDC)

(b) Diffusion for purification (DiffPure)

(c) Image-to-Image diffusion classifier (IDC)

Figure 2: Comparison of different diffusion classifier paradigms. The number of iterations $K > T_s$ in our IDC while the timesteps $t_L \gg T_s$. The parameters of U_L in both ZDC and DiffPure are also larger than U_s in IDC.

Image-to-Image Diffusion Classifier

Applying a pre-trained diffusion model to classification tasks requires additional components or computations. For example, diffusion classifiers (Chen et al. 2023a; Li et al. 2023a) perform inference for each class separately, which leads to a computational overhead that scales proportionally with the number of classes K, as shown in Figure 2(a). DiffPure (Nie et al. 2022) requires an additional pre-trained classifier and t_L steps of the diffusion process, as shown in Figure 2(b). All the iterative steps in these methods heavily rely on the large-scale U-Net (Ronneberger, Fischer, and Brox 2015), which makes their inference phases much slower than traditional CNN-based classifiers.

Given these limitations, we are motivated to alleviate huge computational costs in DMs for adversarial robustness, such as the size of U-Net and the number of diffusion steps. However, reducing computation in diffusion models could significantly influence the image generation quality, which invalidates previous frameworks, such as purification in (Nie et al. 2022) and noise predictor in (Chen et al. 2023a). Thus, we propose to redesign the diffusion framework for adversarial robustness, which relieves the demand for high-quality image generation.

Orthogonal Image Labels Generation Different from the pre-trained DM classifier of ZDC (Li et al. 2023a), we employ the image translation task of DMs to the classification task, which converts the high-quality image generation into image label alignment. To achieve this, we first need to construct image labels for classification and then apply the

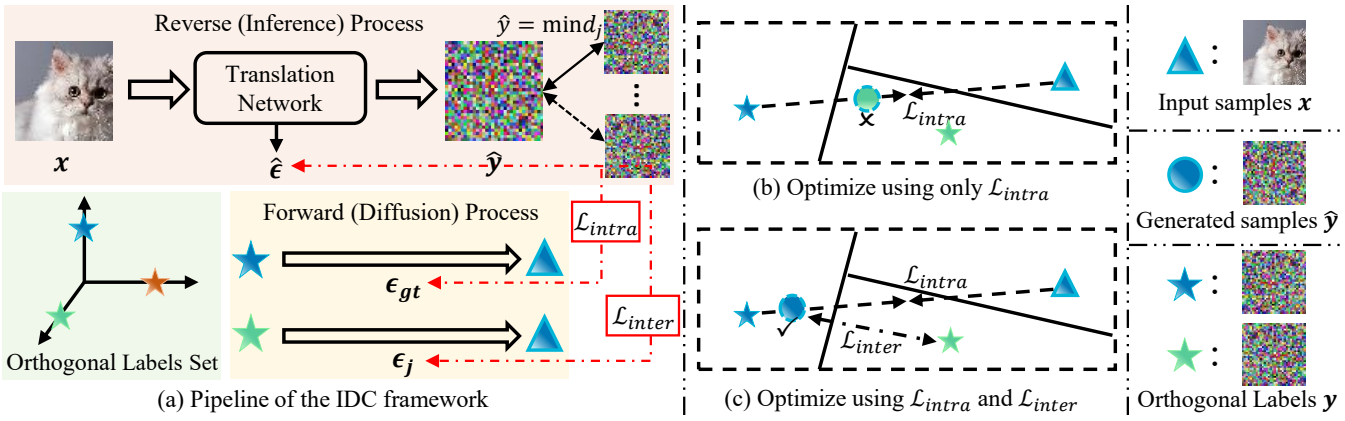


Figure 3: The illustration of our framework and optimization loss. Triangles represent input samples, while circles represent generated samples of the network. In Figure (a), we represent orthogonal labels in a high-dimensional pixel space using a three-dimensional schematic. The differently colored pentagrams each correspond to image labels of distinct categories.

translation framework to the classification task. Considering the labels $y \in \mathbb{N}^b$ for traditional classifiers (Simonyan and Zisserman 2021; He et al. 2016; Dosovitskiy et al. 2021), which maintain the same data distance and orthogonality, we construct orthogonal image labels $\boldsymbol{y} \in \mathbb{R}^{[b,c,h,w]}$ in the pixel space. To this end, we first generate a random noise vector $V \in \mathbb{R}^{[K,c*h*w]}$ in the space formed by classes and pixels where K is the number of classes. The orthogonal vectors can be obtained by performing QR decomposition (Francis 1961) on random vectors V, represented as $V = QR$, where $Q \in \mathbb{R}^{[K,c*h*w]}$ is an orthogonal matrix. By dividing and resizing the matrix Q, we can obtain the image label $\boldsymbol{y}^i \in \mathbb{R}^{[c,h,w]}$ for each class. Next, we normalize the images to ensure that the pixel ranges of the image labels and the inputs are consistent. During training, $\boldsymbol{y}^i$ can form a batch of image labels $\boldsymbol{y} \in \mathbb{R}^{[b,c,h,w]}$ based on original labels $y \in \mathbb{N}^b$.

Image Label Translation After constructing the image labels, we adopt the image translation framework BBDM (Li et al. 2023b) to realize the image translation from the inputs to labels as shown in Figure 2(c). Compared with ZDC (Li et al. 2023a) in Figure 2(a), which has a computational complexity of $O(K)$ for the number of classes, our framework achieves $O(1)$ complexity for the classes during reference. In the forward process, the image labels $\boldsymbol{y} = \boldsymbol{y}_0$ progressively introduce Gaussian noise, generating a series of increasing noise samples $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T$, which can be expressed as (specified in Appendix A.1):

$$\boldsymbol{y}_t = (1 - m_t)\boldsymbol{y}_0 + m_t\boldsymbol{x} + \sqrt{\delta_t}\boldsymbol{\epsilon}_t, \tag{6}$$

$$q(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}, \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}_t; \gamma\boldsymbol{y}_{t-1} + (m_t - \gamma m_{t-1})\boldsymbol{x}, (\delta_t - \gamma^2\delta_{t-1})\boldsymbol{I}), \tag{7}$$

where $\gamma = \frac{1-m_t}{1-m_{t-1}}$, Eq. (6) represents the forward process from $\boldsymbol{y}_0$ to $\boldsymbol{y}_t$, δ_t denotes the variance of diffusion process, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. In the reverse process, the DM learns the denoising process from input image $\boldsymbol{x} = \boldsymbol{y}_T$ to image label $\boldsymbol{y} = \boldsymbol{y}_0$. For timestep t, the reverse process of transition can be expressed as $q(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t, \boldsymbol{y}_0, \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}_{t-1}; \mu(\boldsymbol{y}_t, \boldsymbol{x}), \delta\boldsymbol{I})$ where the mean value μ is the noise predictor, and the prediction of the DM can be expressed as $p_\theta(\hat{\boldsymbol{y}}_{t-1} \mid \boldsymbol{y}_t, \boldsymbol{x}) = \mathcal{N}(\hat{\boldsymbol{y}}_{t-1}; \hat{\mu}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t), \delta_t\boldsymbol{I})$ where θ is the parameter of the network optimized by μ_θ.

Image-to-Image Classification Given the timestep T_s and the input image $\boldsymbol{x} = \boldsymbol{y}_T$, the distribution $p_\theta(\hat{\boldsymbol{y}}_0)$ can be generated following the Markovian chain of reverse process $p_\theta(\hat{\boldsymbol{y}}_{t-1} \mid \boldsymbol{y}_t, \boldsymbol{x})$. After sampling the translated image label $\hat{\boldsymbol{y}}_0$, the predicted class c can be derived by the distance between $\hat{\boldsymbol{y}}_0$ and all pre-defined image labels, which is formulated as (specified in Appendix A.3):

$$c = \arg\min_i \|\hat{\boldsymbol{y}}_0 - \boldsymbol{y}_0^i\|_1, \tag{8}$$

where $\boldsymbol{y}_0^i$ denotes the pre-defined image label of class i.

Diffusion Complexity Reduction

Although we reduce the computational complexity from $O(K)$ to $O(1)$ compared with ZDC (Li et al. 2023a), the large U-Net(Ronneberger, Fischer, and Brox 2015) structure and long timesteps still cause a computational burden during robustness evaluation. Fortunately, our framework transforms the task from generating high-quality images to aligning image labels. This simplification of the objective allows for a certain decrease in image generation quality, providing an opportunity to simplify the diffusion process by pruning the U-Net structure and reducing the diffusion timesteps.

Network Structure Pruning At the beginning of U-Net architecture (Ronneberger, Fischer, and Brox 2015), there is a convolutional layer to convert the channel of the input image to the model channel C_m. As the model channel directly affects all channels of the network, we reduce the original U-Net model channel to $c_m = C_m/k$, where k is used to adjust the degree of network pruning. Given the contracting path of U-Net architecture, it consists of several unpadded convolutions for downsampling and upscales the feature channels by a factor of u_1 while the overall upscale list can be $u = [u_1, u_2, ..., u_n]$. By adjusting it to lower multiplication

factors, parameters in the corresponding unpadded convolutions can be effectively reduced. Finally, we cut the number of ResNet blocks n_R between two unpadded convolutions, which achieves a further decrease in parameters. The structure of the expansive path in the U-Net is adapted to match the adjusted contracting path. Through the pruning methods, we reduce the parameters from the initial BBDM model of 237.09M to 9.39M (specified in Appendix C.1).

Diffusion Timestep Reduction To achieve high-quality image generation, hundreds or even thousands of timesteps are always required for diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Dhariwal and Nichol 2021), which results in a significant computational cost for the adversarial robustness evaluation. In our framework, the image generation process is greatly simplified, allowing the network to learn the diffusion process with only a few timesteps. As shown in Figure 2, we reduce the general diffusion timesteps t_L to T_s where $t_L \gg T_s$, greatly reducing the FLOPs of the inference process.

Classification Optimization of Diffusion Classifier

Thanks to the switch from high-quality image generation to image label translation, the complexity of the diffusion process can be significantly alleviated. However, unlike traditional CNN-based classifiers, the proposed image-to-image diffusion classifier cannot be optimized by the popular classification loss, such as cross-entropy, due to the diffusion-based framework. Thus, we need to reformulate the classification optimization of diffusion classifiers. We mainly divide the classification optimization into two parts, including the intra-class loss which minimizes the distance between the translated image label $\hat{\boldsymbol{y}}_0^i$ and the ground truth $\boldsymbol{y}_0^i$, and the inter-class loss which maximizes the distance between the translated image label $\hat{\boldsymbol{y}}_0^i$ and those of other classes $\boldsymbol{y}_0^j$ where $j \neq i$. First, we discuss the intra-class loss in the context of the diffusion classifier. It is obvious that the optimization of intra-class loss is naturally achieved by the training objective ELBO in the image-to-image translation framework, which is formulated as

$$\begin{aligned} -\mathbb{E}_q \Big[\sum_{t=2}^{T_s} D_{KL}(q(\boldsymbol{y}_{t-1}^i \mid \boldsymbol{y}_t^i, \boldsymbol{y}_0^i, \boldsymbol{x}) \| p_\theta(\hat{\boldsymbol{y}}_{t-1}^i \mid \boldsymbol{y}_t^i, \boldsymbol{x})) \\ - \log p_\theta(\hat{\boldsymbol{y}}_0^i \mid \boldsymbol{y}_1^i, \boldsymbol{x}) + D_{KL}(q(\boldsymbol{y}_T^i \mid \boldsymbol{y}_0^i, \boldsymbol{x}) \| p(\boldsymbol{y}_T^i \mid \boldsymbol{x}) \Big], \end{aligned} \tag{9}$$

where the last term can be ignored as $\boldsymbol{y}_T^i$ is equal to $\boldsymbol{x}$. Based on the reparametrization method (Ho, Jain, and Abbeel 2020; Li et al. 2023b), the above training objective can be simplified as the following loss (specified in Appendix A.2):

$$\mathcal{L}_{intra} = \| m_t(\boldsymbol{x} - \boldsymbol{y}_0^i) + \sqrt{\delta_t}\boldsymbol{\epsilon}_t - \epsilon_\theta(\boldsymbol{y}_t^i, t) \|_1. \tag{10}$$

Besides the intra-class in Eq. (10), the inter-class loss of the diffusion classifier remains unexplored. Thus, we propose to design a loss $\mathcal{L}_{inter}$ to push the translated image label $\hat{\boldsymbol{y}}_0^i$ away from other classes, as shown in Figure 3(c). Formally, we consider the most confusing class j of image $\boldsymbol{x}$ where the image label $\boldsymbol{y}_0^j$ is the most closest one to the translated $\hat{\boldsymbol{y}}_0^i$, as $j = \text{argmin}_j \| \hat{\boldsymbol{y}}_0^i - \boldsymbol{y}_0^j \|_1$ and $j \neq i$. In the framework of translation, the distance between $\hat{\boldsymbol{y}}_0^i$ and the most confusing image label $\boldsymbol{y}_0^j$ is expected to be maximized, which can be formulated in a simplified reverse format of ELBO as:

$$-\mathbb{E}_q \Big[- \textstyle\sum_{t=2}^{T_s} D_{KL}(q(\boldsymbol{y}_{t-1}^j \mid \boldsymbol{y}_t^j, \boldsymbol{y}_0^j, \boldsymbol{x}) \| p_\theta(\hat{\boldsymbol{y}}_{t-1}^i \mid \boldsymbol{y}_t^i, \boldsymbol{x})) \Big], \tag{11}$$

where the constant is ignored. Similar to the derivation in DDPM (Ho, Jain, and Abbeel 2020), the term $q(\boldsymbol{y}_{t-1}^j | \boldsymbol{y}_t^j, \boldsymbol{y}_0^j, \boldsymbol{x})$ can be derived through Bayes' theorem and formulated as $\mathcal{N}(\boldsymbol{y}_{t-1}^j; \mu(\boldsymbol{y}_t^j, \boldsymbol{y}_0^j, \boldsymbol{x}), \delta \boldsymbol{I})$. Through training a network to estimate the noise ϵ_θ in the mean value term μ, a simplified version of the ELBO objective in Eq. (11) can be derived, as shown in the following proposition.

Proposition 1 *For the objective in the Eq. (11), the training loss could be simplified as:*

$$\begin{aligned} \mathcal{L}_{inter} = -\| m_t(\boldsymbol{x} - \boldsymbol{y}_0^j) + m_{t-1}(\boldsymbol{y}_0^j - \boldsymbol{y}_0^i) \\ + \sqrt{\delta_t}\boldsymbol{\epsilon}_t - \epsilon_\theta(\boldsymbol{y}_t^i, t) \|_1. \end{aligned} \tag{12}$$

The detailed proof of Proposition 1 is provided in Appendix B.1. Finally, the overall objective loss of the IDC can be formulated as $\mathcal{L}_{cls} = \mathcal{L}_{intra} + \alpha \mathcal{L}_{inter}$, where α is a hyper-parameter to balance the influence of $\mathcal{L}_{inter}$.

Experiments

Experimental Settings

Datasets and Metrics. We conduct experiments on the CIFAR-10, CIFAR-100 datasets (Krizhevsky, Hinton et al. 2009) and Tiny-ImageNet (Deng et al. 2009). We compare our classifier with CNN-based and DM-based methods, using two metrics: standard accuracy and robust accuracy. Different from the setting of most DM-based methods, which randomly select a test subset for robust evaluation, we evaluate the performance of our IDC on the entire test dataset.

Adversarial Attacks. We evaluate our methods with CNN-based methods under five attacks: PGD (Madry et al. 2018), FGSM (Szegedy et al. 2014), MIFGSM (Dong et al. 2017), CW (Carlini and Wagner 2017) and AutoAttack (Croce and Hein 2020). For each attack, ϵ is set to 8/255, the attack steps of PGD and MIFGSM are set to 20 and 5, and the steps and learning rate of CW are 1000 and 0.01. Following (Lee and Kim 2023; Nie et al. 2022), we evaluate DM-based method under BPDA+EOT attack (Hill, Mitchell, and Zhu 2021) with ℓ_∞ perturbations and PGD+EOT attack (Athalye, Carlini, and Wagner 2018) while $\epsilon = 8/255$, steps are 200 and EOT is set to 20.

Diffusion Classifier Setup. We train our classifier using the Adam optimizer with 256 batch size cross 4 Tesla V100-32GB GPUs, CUDA V10.2 in PyTorch V1.7.1 (Paszke et al. 2019). The diffusion timesteps are set to $T_s = 4$, and the learning rate is set to 0.0001. For the CIFAR-10 dataset, we train 400 epochs in total while we train 600 epochs for the CIFAR-100 dataset. The hyper-parameter α is set to 0.2. The model channels of the U-Net are set to $c_m = 64$, and the number of ResNet blocks is set to $n_R = 1$. For the CIFAR-10 dataset, the upscale list of channels is set to $u = [1, 4]$ while that for CIFAR-100 is set to $u = [1, 4, 8]$.

Dataset	Method	Params	AT	Standard	PGD^{20}	FGSM	MIFGSM	CW	AutoAttack
CIFAR-10	ResNet-18	11.17M	✗	95.0	0.0	43.8	2.1	66.8	0.0
			✓	82.7	51.5	57.3	55.0	78.9	48.5
	WideResNet32	46.16M	✗	96.1	0.0	51.5	3.9	69.7	0.0
			✓	86.6	54.7	61.4	58.8	**83.4**	52.5
	IDC(Ours)	9.39M	✗	85.9	**84.8**	**83.6**	**84.4**	81.9	**59.9**
CIFAR-100	ResNet-18	11.17M	✗	76.4	0.0	8.3	0.2	37.3	0.0
			✓	54.2	27.8	30.7	29.7	49.9	24.0
	WideResNet32	46.16M	✗	80.2	0.0	17.1	1.2	41.6	0.0
			✓	59.6	30.9	34.1	32.7	55.1	**27.2**
	IDC(Ours)	42.84M	✗	59.4	**33.4**	**42.0**	**38.8**	**56.0**	17.7

Table 1: The adversarial robustness evaluation of CNN-based models and our method on both CIFAR-10 and CIFAR-100. 'AT' denotes the adversarial training for the networks.

Method	AT	Natural	PGD^{20}	FGSM	CW	AutoAttack
ResNet-18	✓	38.06	20.21	21.44	32.52	16.91
	✗	44.40	0.26	5.14	12.23	0.00
WideResNet32	✓	44.63	**24.21**	**25.88**	**38.41**	**20.93**
	✗	46.80	0.01	0.84	12.95	0.0
IDC(Ours)	✗	37.65	16.03	22.84	34.33	13.48†

Table 2: The adversarial robustness comparison with CNN-based defense methods on Tiny-ImageNet.

Method	Params	Standard	Robust Acc	
			BPDA+EOT	PGD+EOT
ResNet-18	11.17M	82.65	51.08	51.17
WideResNet32	46.16M	86.59	54.24	54.25
RDC†	111.48M	93.16	73.24	-
MimicDiffusion†	140.4M	92.50	**92.00**	-
ContrastDiffPure†	278M	92.61	81.94	-
DiffPure†	244.30M	89.02	81.40	46.84
IDC(Ours)	9.39M	86.10	82.97	**84.70**

Table 3: The adversarial robustness evaluation of both CNN-based and DM-based methods against BPDA+EOT and PGD+EOT attack on CIFAR-10. The symbol † denotes the evaluation is on a subset of the dataset.

Comparison with CNN-based Methods

CIFAR-10. The upper part of Table 1 shows the performance of the CNN-based defense models and our IDC on CIFAR-10 dataset against PGD^{20}, FGSM, MIFGSM, CW and AutoAttack ℓ_∞ ($\epsilon = 8/255$) where PGD^{20} denotes the attack steps of PGD is 20. IDC maintains an accuracy exceeding 80% on PGD^{20}, MIFGSM, and 59.9% on AutoAttack while CNN-based methods without AT perform near zero accuracy. Even compared with AT methods, IDC improves an average robust accuracy by 20.9% over ResNet-18 and 17.18% over WideResNet32. Among them, our IDC improves the robust accuracy by 32.6% and 29.7% against PGD^{20} compared with ResNet-18 and WideResNet32.

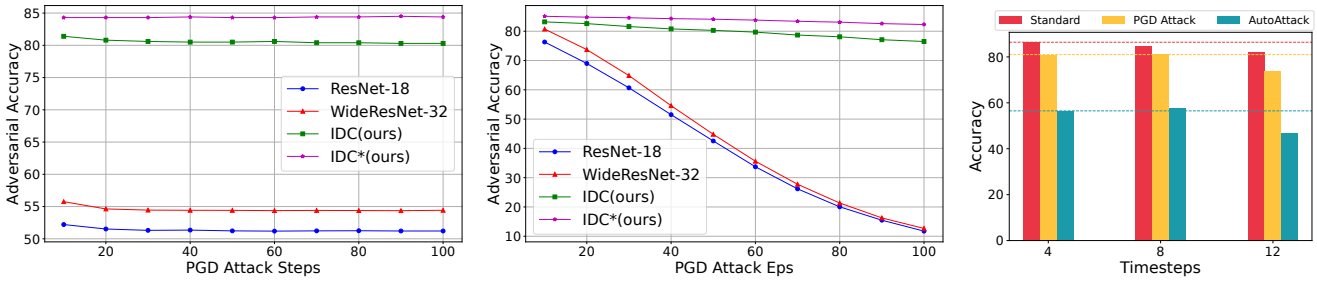
CIFAR-100. We compare the robust performance on CIFAR-100 against PGD^{20}, FGSM, MIFGSM, CW and AutoAttack ℓ_∞ ($\epsilon = 8/255$) on the lower part of Table 1. Our IDC maintains competitive standard accuracy with CNN-based classifiers with AT. As for the robust accuracy, the IDC can improve by 4.7%, 10.7%, 8.5% over ResNet-18 and 1.7%, 6.6%, 4.7% over WideResNet32 against PGD^{20}, FGSM, MIFGSM attack. It also shows competitive performance against CW attack and AutoAttack.

Tiny-ImageNet. Compared to AT methods, IDC performs better than ResNet-18 and slightly less than WideResNet32. However, compared to methods without AT, IDC has significant performance advantages. We will consider improving the robustness of IDC in complex datasets in the future.

Defense Against Adaptive Attacks

CIFAR-10. We compare our method against BPDA+EOT and PGD+EOT attacks in Table 3. Different from other DM-based defense methods, we conduct a full evaluation of CIFAR-10. Compared with ResNet-18 and WideResNet32, IDC improves the robust accuracy by 31.89%, 28.73% on BPDA+EOT attack and 33.53%, 30.45% on PGD+EOT attack with lower parameters. As for DM-based methods, IDC achieves more competitive performance with significantly reduced computations. Concretely, compared with DiffPure (Nie et al. 2022), we improve the accuracy by 1.57% against BPDA+EOT attacks and 37.86% against PGD+EOT attacks, with only 3.84% of the parameters. Compared with novel approaches RDC, MimicDiffusion (Song et al. 2024) and ContrastDiffPure (Bai et al. 2024), IDC also shows competitive performance with exceptionally lower parameters.

CIFAR-100. Table 4 shows the robustness performance against BPDA+EOT attack on CIFAR-100 while our method also achieves the competitive standard accuracy with fewer parameters. The robust accuracy of IDC is also advanced



(a) Robustness under different PGD attack steps (b) Robustness under different PGD attack eps (c) Ablation of diffusion timesteps

Figure 4: The evaluation of robustness under different PGD attack settings and different diffusion timesteps. (a) denotes larger attack iterations. (b) denotes larger perturbation sizes. The symbol * denotes the adversarial robustness of IDC without inter-class loss. (c) Standard and robust accuracy under PGD attack and AutoAttack ℓ_∞ ($\epsilon = 8/255$) with different timesteps.

Method	Params	AT	Standard Acc	Robust Acc
ResNet-18	11.17M	✓	54.23	27.53
WideResNet32	46.16M	✓	59.57	**30.36**
IDC(Ours)	42.84M	✗	59.97	28.02

Table 4: Comparison with CNN-based defense methods using the BPDA+EOT attack on CIFAR-100.

Inter-C	OL	Standard Acc	Robust Acc	
			PGD^{20}	AutoAttack
✓	✗	87.5	63.7	34.6
✗	✓	86.4	81.0	56.5
✓	✓	86.0	**84.8**	**59.9**

Table 5: Effectiveness of orthogonal image labels in our methods with PGD attack and AutoAttack ℓ_∞ ($\epsilon = 8/255$) on CIFAR-10. 'Inter-C' denotes incorporating the inter-class loss, while 'OL' denotes the orthogonality of labels setting.

with the improvement over the ResNet-18 by 0.49%.

Ablation Study

Effectiveness of Proposed Modules. We ablate the proposed orthogonal image labels and inter-class loss in Table 5. The inter-class loss can improve the robust accuracy by 3.8% against PGD attack and by 3.4% against AutoAttack. Incorporating inter-class loss can also consistently enhance the adversarial robustness of the PGD attack in Figure 4. The above results demonstrate that using the inter-class loss can effectively enhance the ability to generate images that are distinguishable among different class labels, thereby strengthening the resilience against attacks. For the orthogonal image labels, to maintain similar distances between image labels like traditional one-hot labels, we enhance the orthogonality between images. Table 5 confirms the efficacy of this configuration, while orthogonal image labels enhance the network's robustness performance by 21.1% under the PGD attack and by 25.3% under AutoAttack.

Different PGD Attacks Settings. We adjust the steps and ϵ of the PGD attack, and Figure 4 shows the robustness performance of our IDC and CNN-based methods. As shown in Figure 4(a), the IDC can maintain robustness performance as the attack steps increase, achieving a performance improvement of 32.99% and 29.8% compared to ResNet-18 and WideResNet, respectively. In Figure 4(b), as the attack ϵ increases, CNN-based methods experience a significant performance decline, whereas our method manages to maintain relatively stable performance. Compared to WideResNet32, our method shows an improvement of 4.36% at $2/255$, and this improvement escalates to 69.64% when $\epsilon = 20/255$.

Different Diffusion Timesteps. For a better illustration of the reduction of diffusion timesteps, we compile statistics on the standard accuracy and robustness accuracy under PGD and AutoAttack of the IDC in Figure 4(c), with timesteps L_s of 4, 8, and 12, respectively. The performance at timesteps $L_s = 8$ is comparable to that at $L_s = 4$, but it incurs a substantial increase in computational cost during inference. When $L_s = 12$, the standard performance decreases by 4.5%, and the adversarial robustness declines by 7.1% and 9.6%, respectively. A possible reason is that when there are excessive timesteps, the noise addition process in each timestep is not sufficiently pronounced, inadvertently increasing the learning difficulty for the network.

Conclusion

In this paper, we propose a new framework called IDC for DM-based adversarial robustness, which transforms the task of DMs from generating high-quality images to predicting distinguishable image labels. Based on our framework, we streamline the diffusion network, evaluate performance across the entire dataset, and compare it with both CNN-based and DM-based methods. A classification loss is also incorporated to enhance the adaptability of the DM-based framework to adversarial robustness. We select multiple representative attacks on the CIFAR-10 and CIFAR-100 datasets to demonstrate that our IDC achieves competitive results in terms of parameters and adversarial robustness.

Acknowledgments

This work was supported in part by the Start-up Grant (No. 9610680) of the City University of Hong Kong, Young Scientist Fund (No. 62406265) of NSFC, and the Australian Research Council under Projects DP240101848 and FT230100549.

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283. PMLR.

Bai, M.; Huang, W.; Li, T.; Wang, A.; Gao, J.; Caiafa, C. F.; and Zhao, Q. 2024. Diffusion Models Demand Contrastive Guidance for Adversarial Purification to Advance. In *International Conference on Machine Learning*.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. Ieee.

Chen, H.; Dong, Y.; Wang, Z.; Yang, X.; Duan, C.; Su, H.; and Zhu, J. 2023a. Robust classification via a single diffusion model. arXiv:2305.15241.

Chen, Q.; Ma, Y.; Wang, H.; Yuan, J.; Zhao, W.; Tian, Q.; Wang, H.; Min, S.; Chen, Q.; and Liu, W. 2024. Follow-Your-Canvas: Higher-Resolution Video Outpainting with Extensive Content Generation. *arXiv preprint arXiv:2409.01055*.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023b. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19830–19843.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *Advances in neural information processing systems*, volume 34, 8780–8794.

Dong, M.; Chen, X.; Wang, Y.; and Xu, C. 2022. Random normalization aggregation for adversarial defense. *Advances in Neural Information Processing Systems*, 35: 33676–33688.

Dong, M.; and Xu, C. 2023. Adversarial robustness via random projection filters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4077–4086.

Dong, Y.; Liao, F.; Pang, T.; Hu, X.; and Zhu, J. 2017. Discovering adversarial examples with momentum. arXiv:1710.06081.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Francis, J. G. 1961. The QR transformation a unitary analogue to the LR transformation—Part 1. *The Computer Journal*, 4(3): 265–271.

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2020. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

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

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. arXiv:2210.02303.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, volume 33, 6840–6851.

Kang, M.; Song, D.; and Li, B. 2024. DiffAttack: Evasion Attacks Against Diffusion-Based Adversarial Purification. *Advances in Neural Information Processing Systems*, 36.

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

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

Li, A. C.; Prabhudesai, M.; Duggal, S.; Brown, E.; and Pathak, D. 2023a. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2206–2217.

Li, B.; Xue, K.; Liu, B.; and Lai, Y.-K. 2023b. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, 1952–1961.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.

Lin, G.; Li, C.; Zhang, J.; Tanaka, T.; and Zhao, Q. 2024a. Adversarial Training on Purification (AToP): Advancing Both Robustness and Generalization. In *International Conference on Learning Representations*.

Lin, G.; Tao, Z.; Zhang, J.; Tanaka, T.; and Zhao, Q. 2024b. Robust Diffusion Models for Adversarial Purification. arXiv:2403.16067.

Liu, X.; Cheng, M.; Zhang, H.; and Hsieh, C.-J. 2018. Towards robust neural networks via random self-ensemble. In

Proceedings of the european conference on computer vision (ECCV), 369–385.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024a. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4117–4125.

Ma, Y.; He, Y.; Wang, H.; Wang, A.; Qi, C.; Cai, C.; Li, X.; Li, Z.; Shum, H.-Y.; Liu, W.; et al. 2024b. Follow-Your-Click: Open-domain Regional Image Animation via Short Prompts. *arXiv preprint arXiv:2403.08268*.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024c. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint arXiv:2406.01900*.

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

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion models for adversarial purification. In *International conference on machine learning*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Rice, L.; Wong, E.; and Kolter, Z. 2020. Overfitting in adversarially robust deep learning. In *International conference on machine learning*, 8093–8104. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM Special interest group on computer graphics and interactive techniques conference proceedings*, 1–10.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4713–4726.

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

Sasaki, H.; Willcocks, C. G.; and Breckon, T. P. 2021. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv:2104.05358.

Simonyan, K.; and Zisserman, A. 2021. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. arXiv:2209.14792.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Song, K.; Lai, H.; Pan, Y.; and Yin, J. 2024. MimicDiffusion: Purifying Adversarial Perturbation via Mimicking Clean Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24665–24674.

Song, Y.; and Ermon, S. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33: 12438–12448.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. In *International conference on learning representations*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

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

Wang, J.; Ma, Y.; Guo, J.; Xiao, Y.; Huang, G.; and Li, X. 2024. COVE: Unleashing the Diffusion Feature Correspondence for Consistent Video Editing. *arXiv preprint arXiv:2406.08850*.

Wu, Q.; Ye, H.; and Gu, Y. 2022. Guided diffusion model for adversarial purification from random noise. arXiv:2206.10875.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. In *BMVC*.

Zhang, B.; Luo, W.; and Zhang, Z. 2023. Purify++: Improving Diffusion-Purification with Advanced Diffusion Models and Control of Randomness. arXiv:2310.18762.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhang, Z.; Han, L.; Ghosh, A.; Metaxas, D. N.; and Ren, J. 2023. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6027–6037.