ADVERSARIAL ROBUSTNESS WITH DIFFUSION MODELS

XUELONG DAI

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Computing

Adversarial Robustness with Diffusion Models

Xuelong Dai

A thesis
submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

May 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Xuelong Dai (Name of student)

Abstract

Artificial Intelligence (AI) and Deep Learning (DL) have experienced rapid development and widespread industry deployment in recent years. Among the various deep learning models, Computer Vision (CV) stands out as one of the most advanced fields. DL models have achieved performance comparable to human experts across a range of 2D and 3D tasks. However, adversarial attacks pose a significant threat to the further application of DL-based CV techniques. These attacks involve adding small perturbations to input data, which do not affect human classification but lead to high-confidence misclassification by the target deep learning network. This challenge highlights the urgent need to evaluate and enhance the adversarial robustness of deep learning models.

Diffusion models, a recently proposed generative model known for its outstanding performance, have made a significant impact due to their impressive data generation capabilities and user-friendly interface. In addition to their excellent generative performance, these models have demonstrated the ability to conduct high-quality adversarial attacks by generating adversarial data, posing a new threat to the security of deep learning models. Consequently, it is important to investigate the attack capabilities of diffusion models under various threat scenarios and to explore strategies for enhancing adversarial robustness against attacks driven by these models.

Firstly, we observe that current adversarial attacks utilizing diffusion models typically employ PGD-like gradients to guide the creation of adversarial examples. However, the generation process of diffusion models should adhere strictly to the learned diffusion process. As a result, these current attacks often produce low-quality adversarial examples with limited effectiveness. To address these issues, we introduce AdvDiff, a theoretically provable adversarial attack method that leverages diffusion models. We have developed two novel adversarial guidance techniques to sample adversarial examples by following the trained reverse generation process of diffusion models. These guidance techniques are effective and stable, enabling the generation of high-quality, realistic adversarial examples by integrating the gradients of the tar-

get classifier in an interpretable manner. Experimental results on the MNIST and ImageNet datasets demonstrate that AdvDiff excels in generating unrestricted adversarial examples, surpassing state-of-the-art unrestricted adversarial attack methods in both attack performance and generation quality.

Secondly, we note that in no-box adversarial scenarios, where the attacker lacks access to both the training dataset and the target model, the performance of existing attack methods is significantly hindered by limited data access and poor inference from the substitute model. To overcome these challenges, we propose a no-box adversarial attack method that leverages the generative and adversarial capabilities of diffusion models. Specifically, our approach involves generating a synthetic dataset using diffusion models to train a substitute model. We then fine-tune this substitute model using a classification diffusion model, taking into account model uncertainty and incorporating noise augmentation. Finally, we generate adversarial examples from the diffusion models by averaging approximations over the diffusion substitute model with multiple inferences. Extensive experiments on the ImageNet dataset demonstrate that our proposed attack method achieves state-of-the-art performance in both no-box and black-box attack scenarios.

Thirdly, we find that existing adversarial research on 3D point cloud models predominantly focuses on white-box scenarios and struggles to achieve successful transfer attacks on recently developed 3D deep-learning models. Moreover, the adversarial perturbations in current 3D attacks often cause noticeable shifts in point coordinates, resulting in unrealistic adversarial examples. To address these challenges, we propose a high-quality adversarial point cloud shape completion method that leverages the generative capabilities of 3D diffusion models. By using partial points as prior knowledge, we generate realistic adversarial examples through shape completion with adversarial guidance. To enhance attack transferability, we explore the characteristics of 3D point clouds and utilize model uncertainty for improved model classification inference through random down-sampling of point clouds. We employ ensemble adversarial guidance to improve transferability across different network architectures. To maintain generation quality, we restrict our adversarial guidance to

the critical points of the point clouds by calculating saliency scores. Extensive experiments demonstrate that our proposed attacks outperform state-of-the-art adversarial attack methods against both black-box models and defenses. Our black-box attack establishes a new baseline for evaluating the robustness of various 3D point cloud classification models.

Fourthly, we notice that while current diffusion-based adversarial purification methods offer effective and practical defense against adversarial attacks, they suffer from low time efficiency and limited performance against recently developed unrestricted adversarial attacks. To address these issues, we propose an effective and efficient diffusion-based adversarial purification method that counters both perturbation-based and unrestricted adversarial attacks. Our defense is inspired by the observation that adversarial attacks typically occur near the decision boundary and are sensitive to pixel changes. To tackle this, we introduce adversarial anti-aliasing to mitigate adversarial modifications. Additionally, we propose adversarial super-resolution, which uses prior knowledge from clean datasets to benignly recover images. These approaches do not require additional training and are computationally efficient, as they do not involve gradient calculations. Extensive experiments against both perturbation-based and unrestricted adversarial methods demonstrate that our defense method outperforms state-of-the-art adversarial purification techniques.

Acknowledgments

I would like to express my deepest gratitude to my Ph.D. supervisor, Prof. Bin Xiao. Throughout my Ph.D. studies, Prof. Xiao has provided invaluable guidance on research directions, experimental support, and paper writing and reviewing. His dedication and effort in helping to revise my thesis have been instrumental, and I am immensely grateful for his supervision and assistance.

I also wish to thank my colleagues and collaborators in the research group. Their valuable insights and support over the years have been crucial to my research. We have shared many memorable moments during our time in Hong Kong, making my experience at PolyU truly precious.

I am grateful to PolyU and its staff for their comprehensive support during my studies in Hong Kong. They have provided assistance in nearly all aspects of living, studying, and working.

I would also like to thank the thesis reviewers for their time and effort in reviewing my work. Additionally, I am grateful to the researchers in related fields for their pioneering contributions, which have greatly informed my research.

Finally, I extend my heartfelt thanks to my parents. Their encouragement and wise advice have been essential in completing my Ph.D. studies and thesis. Their support has been invaluable during both challenging and joyful times throughout my years of study.

Publications Arising from the Thesis

Published Papers:

1. **Xuelong Dai**, Yanjie Li, Mingxing Duan, Bin Xiao, "Diffusion Models as Strong Adversaries", IEEE Transactions on Image Processing, 33:6734 - 6747, 2024.

2. **Xuelong Dai**, Kaisheng Liang, Bin Xiao, "AdvDiff: Generating Unrestricted Adversarial Examples using Diffusion Models", In *Proceedings of the European Conference on Computer Vision*, 2024.

3. **Xuelong Dai**, Bin Xiao, "Transferable 3D Adversarial Shape Completion using Diffusion Models", In *Proceedings of the European Conference on Computer Vision*, 2024.

Under Review:

1. **Xuelong Dai**, Dong Wang, Mingxing Duan, Bin Xiao, "Gradient-Free Adversarial Purification with Diffusion Models", In *Proceedings of the International Conference on Computer Vision*, 2025.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

AI and DL have achieved significant breakthroughs in both efficiency and accuracy across numerous challenging tasks. These advanced technologies have been widely adopted in industries such as medical care, security identification, autonomous driving, and smart cities. Their impressive performance in various fields has garnered increasing attention in research. Major directions in DL research include computer vision, natural language processing, and more. There are virtually no other algorithms that can surpass deep learning models in the field of computer vision when it comes to both usability and accuracy. The convolutional layer has significantly enhanced the performance of deep learning models across various challenging computer vision tasks. Due to their reliable performance in tasks such as image classification and object detection, an increasing number of real-world applications, such as face recognition and smart driving, have been developed based on deep learning models.

While deep learning models have shown significant improvements in various computer vision tasks, researchers have discovered that these models are highly vulnerable to adversarial attacks. An adversarial attack involves adding small perturbations to input data that are imperceptible to humans but can easily alter the classification results of a deep learning classifier with high confidence. These modified inputs are known as adversarial examples. To further deploy DL models in security-critical applications, there has been considerable interest among researchers in both adversarial attacks and defenses. Based on the knowledge accessible to the adversary, adversarial

attacks are categorized into two types: white-box adversarial attacks and black-box adversarial attacks. White-box attacks assume that the adversary has full knowledge of the target model, allowing adversarial examples to be crafted directly using the gradient of the target model's loss function. On the other hand, black-box adversarial attacks do not permit direct access to the parameters and architecture of the target model. Instead, the adversary conducts attacks by querying the target model or exploiting the transferability of adversarial examples for effective black-box attacks.

Deep learning models commonly process 2D data, such as images and videos. However, in practice, people also encounter 3D data, like 3D point clouds or 3D grids. Learning from 3D data is fundamentally different from 2D data, and 3D deep learning require more computational resources. Since Qi et al. introduced PointNet, a deep learning model that uses a specialized layer to extract global features from 3D point clouds, there has been a surge in 3D deep learning research. PointNet++ and DGCNN are two widely recognized models in this field. Additionally, 3D adversarial attacks have been shown to be effective against 3D deep learning models. Recent works in this area can be categorized into three types: 2D-attack-based methods, such as IFGM and C&W attacks on 3D point clouds; point-modification-based methods, like isometry transformation attacks and point occlusion attacks; and generative-based methods, such as LG-GAN and AdvPC. These various attack algorithms employ different strategies to target 3D deep learning models, and most achieve a high success rate when attacking state-of-the-art 3D DL models.

However, the perturbations generated by most attack algorithms are easily detectable by humans, as they often produce noisy patterns on images. Consequently, these attacks can be countered by various defense methods and are challenging to implement in the physical world. Therefore, it is valuable to develop a natural and realistic adversarial example generation algorithm to enhance the effectiveness of adversarial attacks.

1.1 Unrestricted Adversarial Attacks with Diffusion Models

With the development of generative models, these models bring new threats to the robustness of the deep learning models. The adversary adopts the generation ability of the Generative Adversarial Networks (GAN) models to craft adversarial examples by generating perturbations. However, these existing methods require re-training of the GAN models and harm the original generation performance of the benign GAN models. Therefore, their performances are limited by their perturbation-based attack algorithms. Unrestricted adversarial attacks, on the other hand, craft adversarial examples from scratch. These adversarial examples are visually indistinguishable from the benign samples while deceiving the deep-learning models with high confidence. Followed by Song et al. pioneering work, more and more GAN-based unrestricted adversarial attacks are proposed.

Diffusion models have demonstrated superior performance in image generation compared to their competitor GANs. With the development of diffusion models, recent works demonstrated that diffusion models can be used to generate unrestricted adversarial examples, although these studies have been limited to black-box scenarios and have not thoroughly explored the capabilities of diffusion models as adversaries.

Therefore, our work will exploit the remarkable generation ability of diffusion models for discussing attack performance under no-box scenarios with a comprehensive and end-to-end discussion from the generation of the training dataset and adversarial examples.

1.2 Thesis Contribution

In our thesis, we conduct a comprehensive investigation into the adversarial robustness of diffusion models across a wide range of topics, including 2D and 3D scenarios, white-box and black-box settings, and both attack and defense strategies. The general framework of our work is given in Figure. 1.1. Leveraging the strong generative

capabilities of diffusion models, we design effective adversarial guidance to direct the diffusion model in generating high-quality, unrestricted adversarial examples by adhering to the benign generation diffusion process. Our adversarial guidance does not interfere with the trained sampling process of diffusion models, thereby producing adversarial examples with superior generation quality and attack performance. Moreover, we explore the efficacy of diffusion models in more challenging attack scenarios, namely black-box and no-box environments. We synthesize datasets using diffusion models and enhance attack performance through a diffusion classification substitute model. With advancements in diffusion models, they have demonstrated impressive performance in 3D point cloud generation. To provide a comprehensive discussion on the adversarial capabilities of diffusion models, we introduce a transferable adversarial shape completion method utilizing diffusion models. We begin by evaluating the robustness of recently proposed 3D point cloud classifiers, achieving state-of-the-art performance in black-box attacks.

The denoising-like generation process of diffusion models facilitates diffusion-based adversarial purification for defensive purposes. However, current diffusion-based defenses often suffer from low time efficiency and limited effectiveness against unrestricted adversarial attacks. To address these issues, we propose a gradient-free adversarial defense method based on diffusion models. Our approach offers a more effective defense against unrestricted adversarial attacks.

1.2.1 Generating Unrestricted Adversarial Examples

Recent research has demonstrated that diffusion models are capable of executing unrestricted adversarial attacks. However, existing attack methods frequently incorporate the gradient from traditional perturbation-based adversarial attacks into the generation process of diffusion models. This practice can substantially diminish the generation quality, rendering the attacks easily detectable by humans and current defense mechanisms. Consequently, it is essential to develop an unrestricted adversarial attack that aligns with the benign generation process of diffusion models.

We introduce AdvDiff, an interpretable method for executing unrestricted adver-

sarial attacks using diffusion models. Our approach involves adding two effective adversarial guidance techniques to the reverse generation process of the diffusion model. Notably, our attack does not require retraining the diffusion model; instead, we utilize a pre-trained conditional diffusion model. The two adversarial guidance techniques we propose are: 1) Incrementally incorporating *adversarial guidance* throughout the reverse generation process by increasing the likelihood of the target attack label, and 2) Repeatedly executing the reverse generation process while infusing adversarial prior knowledge into the initial noise through *noise sampling guidance*. We provide a theoretical analysis of our attack method to demonstrate that the adversarial guidance does not alter the original sampling patterns of benign diffusion models. To further validate the effectiveness of AdvDiff, we conduct extensive experiments on two datasets and evaluate using four metrics: attack success rate, generation quality, transfer attack performance, and attack performance against defenses. The experimental results show that our attack achieves state-of-the-art performance compared to perturbation-based attacks and previous diffusion-based unrestricted adversarial attacks.

1.2.2 Strong No-Box Unrestricted Adversarial Attack

In a black-box threat model, access to the model parameters is restricted, while in a no-box threat model, access to the training data is not permitted. These two threat models are more practical than white-box attacks and are better suited for evaluating model robustness. However, current no-box adversarial attacks still require access to a limited amount of data from the training set. Conducting a no-box attack without any access to the target model's data remains a significant challenge.

Leveraging the generative capabilities of diffusion models, we can construct a training dataset exclusively using these models. Our approach offers a solution for training a substitute model for no-box adversarial attacks using a synthetic dataset generated by the diffusion model. To further enhance the transferability of our adversarial examples, we employ a classification diffusion model as the substitute model. This model utilizes the probability distribution of labels to infer input data, which

can be combined with uncertainty estimation techniques to improve attack transferability. Additionally, we incorporate scheduled noise during the training phase of the substitute model. Once the substitute model is trained, we use the same diffusion model to generate the dataset for executing no-box unrestricted adversarial attacks. We adopt an ensemble-like strategy by applying the Monte Carlo sampling method across multiple conditional distribution predictions from the diffusion substitute model. We conduct experiments in both black-box and no-box scenarios to demonstrate the effectiveness of our proposed attack. Compared to existing attack methods, our no-box unrestricted adversarial attack achieves superior performance in terms of attack success rate and generation quality.

1.2.3 Transferable Adversarial 3D Shape Completion

3D point cloud data store xyz coordinates, and perturbation-based adversarial attacks can lead to the generation of outlier points by adding perturbations to this data. Even more concerning, existing attacks struggle to overcome the defenses of recently proposed 3D point cloud classifiers. Consequently, generating natural and realistic adversarial point clouds against state-of-the-art target models has become an important research topic.

Diffusion models have demonstrated strong generative capabilities for 3D point clouds. However, their ability to generate 3D adversarial point clouds has not been thoroughly explored. We propose a 3D point cloud adversarial attack method using diffusion models, leveraging a shape completion task to enhance generation quality. To conduct effective black-box adversarial attacks, we first use a Monte Carlo estimate over the inference of multiple down-sampled point clouds to account for model uncertainty, thereby improving attack transferability. Secondly, we employ ensemble logits to incorporate adversarial guidance into the 3D diffusion model. To further enhance generation quality, we restrict the application of adversarial guidance to selected critical points identified by our proposed saliency scores. Experimental results demonstrate that the proposed transferable adversarial 3D shape completion method achieves state-of-the-art black-box performance across a wide range of 3D

target models, including recently proposed 3D point cloud classifiers.

1.2.4 Gradient-Free Diffusion-Based Adversarial Purification

The diffusion model's generation process gradually removes noise from the latent space, making it suitable for eliminating adversarial perturbations from adversarial examples. However, the sampling speed of the diffusion generation process is slow. Existing diffusion-based adversarial purification methods are less time-efficient compared to previous approaches. Additionally, their performance is limited when defending against unrestricted adversarial examples. Developing a time-efficient diffusion-based adversarial purification method that effectively counters both perturbation-based and unrestricted adversarial attacks remains a significant challenge.

We identify common characteristics between perturbation-based and unrestricted adversarial examples, noting that these examples are generated near the decision boundary with minimal alterations, which makes them sensitive to pixel changes. To address this, our defense first applies a preprocessing step of adversarial anti-aliasing, which extracts the semantic shape from adversarial examples by blurring the adversarial perturbations. Next, we employ diffusion models to achieve adversarial super-resolution by upscaling the anti-aliased adversarial examples, utilizing prior knowledge of clean data from pre-trained diffusion models. To demonstrate the effectiveness of our proposed defense, we further evaluate its performance by using upscaled adversarial examples as input for adversarial purification. Experiments conducted across various datasets show that our defense outperforms state-of-the-art adversarial defenses in terms of adversarial purification.

1.3 Thesis Outline

The remainder of this thesis is organized as follows: Chapter 2 introduces the background knowledge for this thesis. Chapter 3 showcases our work on AdvDiff. Chapter 4 presents the no-box adversarial attack method utilizing diffusion models. Chapter 5 proposes our work on transferable 3D adversarial shape completion. Chapter 6

presents a gradient-free adversarial purification approach using diffusion models. Finally, we provide the conclusion and outline directions for future work in Chapter 7.

The primary research outputs are selected from the following references:

- *AdvDiff: Generating Unrestricted Adversarial Examples using Diffusion Models*, **Xuelong Dai**, Kaisheng Liang, Bin Xiao. ECCV 2024.
- *Diffusion Models as Strong Adversaries*, **Xuelong Dai**, Yanjie Li, Mingxing Duan, Bin Xiao. IEEE Transactions on Image Processing.
- *Transferable 3D Adversarial Shape Completion using Diffusion Models*, **Xuelong Dai**, Bin Xiao. ECCV 2024.
- *Gradient-Free Adversarial Purification with Diffusion Models*, **Xuelong Dai**, Dong Wang, Mingxing Duan, Bin Xiao. Under Review.

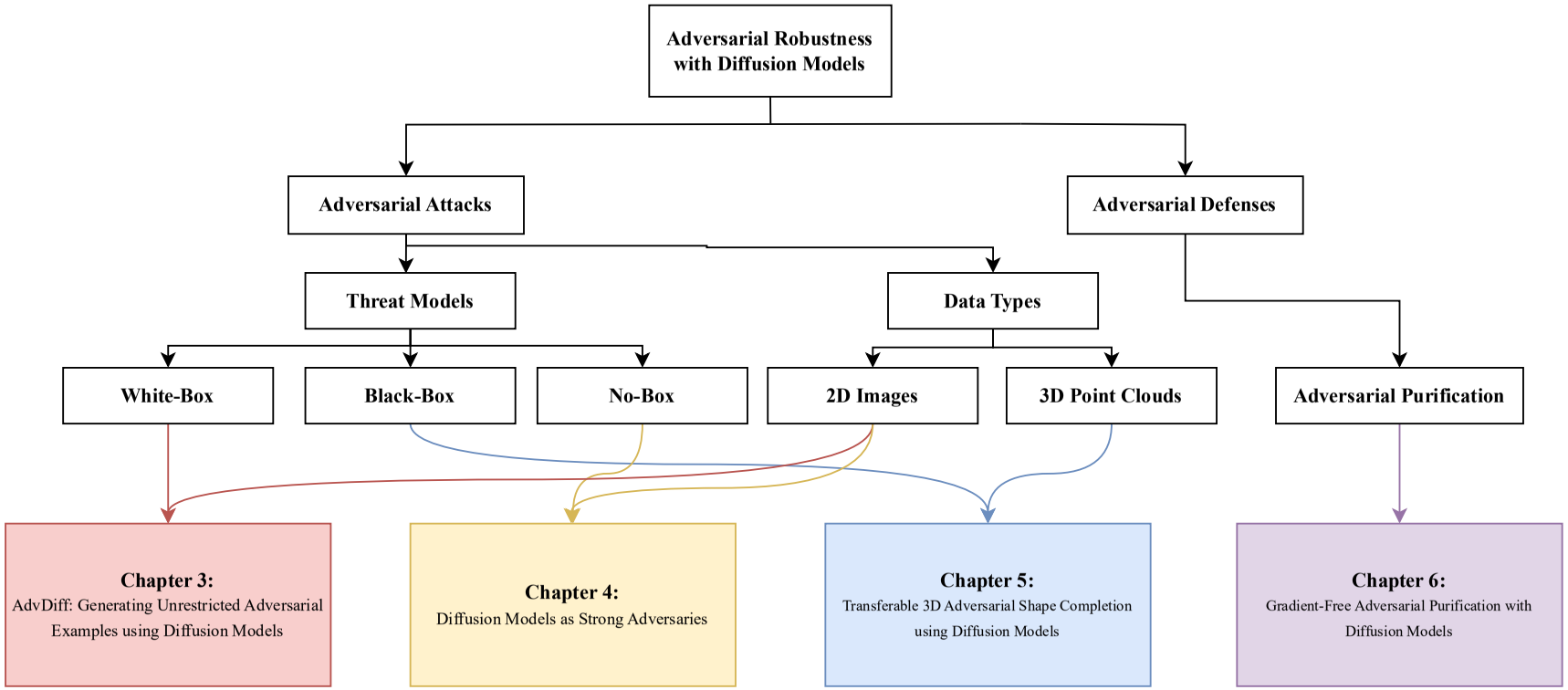


Figure 1.1: The structure of our work.

Chapter 2

Preliminary

In this chapter, we provide the foundational knowledge necessary for the thesis. The chapter is organized as follows: First, we introduce the background of 2D and 3D deep learning. Then, we explore diffusion models in the context of image synectics tasks. Next, we discuss related work on adversarial attacks, covering both 2D and 3D approaches, perturbation-based and unrestricted attacks, as well as white-box and black-box threat models. Finally, we discuss various strategies for adversarial defense.

2.1 Deep Learning

Neural Network A neural network is a machine learning algorithm that consists of multiple neurons connected to each other like the human brain neurons. Each neuron in a neural network processes the input from the given data or a previously activated neuron and produces its activation to the next node. An example of a neuron is given in Figure 2.1. A neuron node is made up of inputs, weights, activation function, and the output.

In a standard neural network, there are multiple layers of connected neurons. The layers are categorized into three types: the input layer, the hidden layer, and the output layer. Figure 2.2 shows the layers of a neural network. The three layers are connected to each other by taking the input from the previous layer's output. The neurons in the same layer are not connected to each other. By training the neural

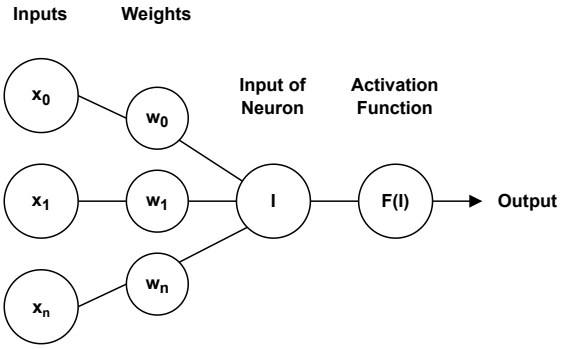


Figure 2.1: An example of a neuron

network with sufficient data, the neural networks have wide applications in many industries: computer vision [29], medical care [91], and speech processing [86].

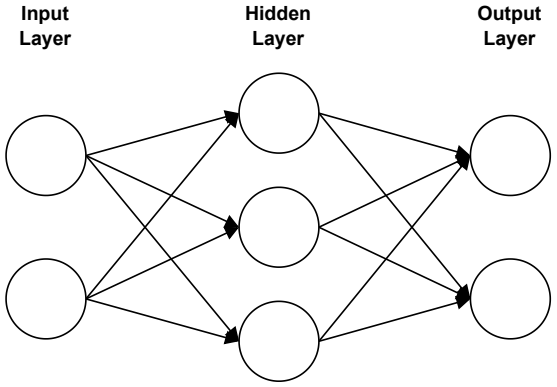


Figure 2.2: Layers of a neural network

Deep Learning Deep learning has been extensively studied as it largely improved the performance of neural networks in various tasks. The deep learning network's structure is similar to the standard neural network with multiple hidden layers. With more hidden layers, the deep learning network has more capability to simulate more complicated functions. The deep learning networks can be further classified according to their basic neuron: Convolutional Neural Networks (CNNs), recurrent neural networks (RNNs), deep belief networks, and so on. The Convolutional Neural Networks (CNNs) will be the focus of this paper because it has been proved that high effective at computer vision tasks.

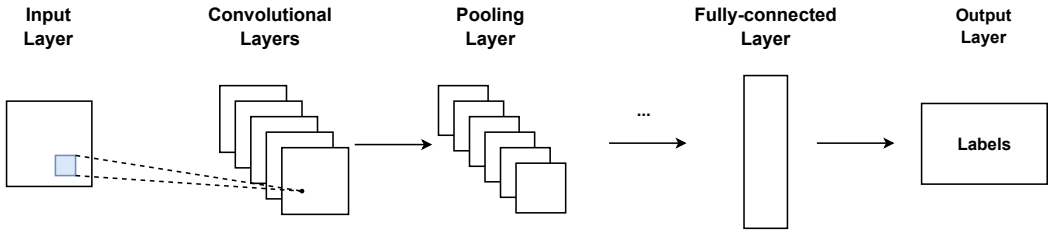


Figure 2.3: The CNN pipeline

Figure 2.3 shows the pipeline of a convolutional neural network [59]. A CNN is consists of the input layer, the output layer, the convolutional layers, the pooling layers, and the fully connected layers. The convolutional layer adopts a convolutional operation to process the input image or feature maps. The convolutional operations share the same weight (convolutional kernel) to process the same feature map. The convolutional kernel learns the local information from the data, and it is invariant to the location. The pooling layer is adopted to reduce the dimension of the network parameters and the feature maps. Finally, the fully-connected layer converts the 2D feature map into a 1D vector, which outputs the final classification labels.

Deep learning networks are achieving leading performance in various major directions of machine learning tasks like computer vision [59], natural language processing [119], etc. Also, deep learning has been widely applied in many scenarios like healthcare [30], automotive [55, 82], smart city [14, 57], and so on.

2.1.1 3D deep learning

Since the development of LiDAR (light detection and ranging) and 3D scanner, 3D data has become easier to access by the consumer. A variety of large 3D datasets have emerged. However, feature learning in the 3D dataset was a difficult task as it contains richer information than 2D data. Also, 3D data have different types, like point cloud, mesh, voxel, etc. Deploying 3D data in the real-life scenario can achieve more accuracy and robustness than only using 2D data. Since the emergence of the deep learning technique, 3D feature learning has been received rapid development. PointNet [95]

was the first approach to solve the 3D feature learning problem, and remarkably improved the performance of the 3D classification task. It learns 3D features by adopting a symmetric function to extract features with the disorder input. Since the success of PointNet, 3D deep learning has received a surge of related research. To further improve the performance of 3D feature learning, the researchers adopt graph convolutional operations to learn features from both local neighbors and global shapes. PointNet++ [96] and DGCNN [122] are two state-of-art 3D deep learning networks that adopted graph convolutional layers. Therefore, We select PointNet, PointNet++, and DGCNN as the targeted network in the adversarial settings for their state-of-art performance on the current 3D dataset.

2.2 Diffusion Models

Diffusion models have shown great generation quality and diversity in the image synthesis task since Ho et al. [46] proposed a probabilistic diffusion model for image generation that greatly improved the performance of diffusion models. Diffusion models for conditional image generation are extensively developed for more usable and flexible image synthesis. Dhariwal & Nichol [25] proposed a conditional diffusion model that adopted classifier-guidance for incorporating label information into the diffusion model. They trained the classifier separately and used its gradient for conditional image generation. Jonathan Ho & Tim Salimans [48], on the other hand, performed conditional guidance without an extra classifier. They trained a conditional diffusion model alongside a standard diffusion model and used a combination of the two models during sampling. Their idea is motivated by an implicit classifier with the Bayes rule. Followed by [25, 48]'s works, many research [31, 81, 100] have been proposed to achieve state-of-the-art performance on image generation, image inpainting, and text-to-image generation tasks. Latent Diffusion Model (LDM) [100] and its text-to-image variant, Stable Diffusion, are capable of generating photo-realistic images. They are able to generate data that is highly related to the dataset of the target model with certain prompts or conditional labels, especially on open-source high-quality datasets

like ImageNet [23]. In this paper, we adopt the Denoising Diffusion Implicit Models (DDIM) for image generation. The DDIM consists of two main processes: the forward diffusion process and the reverse generation process, as shwon in Figure 2.4. The forward diffusion process gradually adds Gaussian noise to the sampled data x_0 with the predefined scheduling parameter α and pre-defined T time steps:

$$q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\mathbf{I}) \tag{2.1}$$

where $q_\sigma(x_T|x_0) = \mathcal{N}(\sqrt{\alpha_T}x_0, (1-\alpha_T)\mathbf{I})$ and σ is the magnitude of the Gaussian noise.

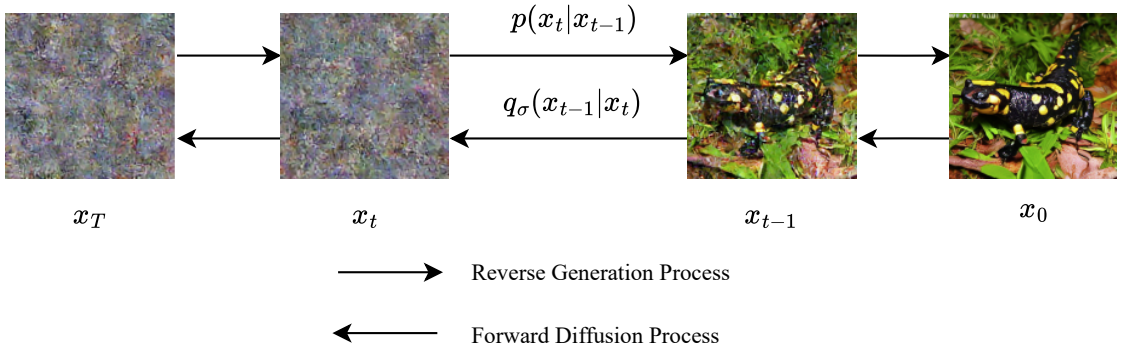


Figure 2.4: The diffusion pipeline

The reverse generation process aims to recover the data x_0 by a denoising-like process starting with a random noise. With T time steps, we generate a sample x_{t-1} from a sample x_t:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \boldsymbol{\epsilon}_\theta^{(t)}(x_t) + \sigma_t\boldsymbol{\epsilon}_t \tag{2.2}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$ is an independent Gaussian noise, and $\boldsymbol{\epsilon}_\theta$ is the trainable model to predict the added Gaussian noise in the forward diffusion process. After training

the $\boldsymbol{\epsilon}_\theta$, we will be able to sample high-quality data with a random initial noise.

Besides the data synthesis task, diffusion models achieve satisfying performance on various tasks like classification [41], segmentation [6], and representation learning [94].

2.3 Adversarial Attacks

White-box attacks. Szegedy et al. [114] demonstrated that these models can be vulnerable to imperceptible perturbations, denoted as $x_{\text{adv}} = x + \delta$, which maximize the network's prediction error. The objective of the white-box attack is to find the perturbation that satisfies the constraint $||\delta||_p < \text{dist}$, where δ represents the perturbation bounded by the l_p norm. In this scenario, the attacker has full knowledge of the target model, including its parameters and network architecture. The perturbations are typically guided by the gradient of the target model's loss function. Adversarial methods such as FGSM [34], I-FGSM [60], and PGD [84] are commonly used to perform white-box attacks. Simultaneously, effective defense methods [67,76,80,84,87,89,143] are proposed to defend against the adversary.

Black-box attacks. In a black-box attack scenario, the attacker does not have access to the parameters of the target model and can only make limited queries to the model. Existing black-box attack methods achieve adversarial attacks by leveraging the transferability of a substitute model or estimating the gradient of the target model through multiple queries. However, query-based attacks [7,12] typically require a large number of queries to successfully execute a single attack, which may not be feasible in many cases. Recent research efforts have focused on enhancing the adversarial transferability by modifying the backpropagation computation, as seen in approaches like LinBP [38], ILA++ [39], TAIG [52], and LGV [36]. Another direction is to increase the input diversity to improve the success rate of black-box attacks. Techniques such as TIM [27], SIM [75], Admix [121], and MBA [66] have been proposed to achieve this goal. These methods aim to find adversarial examples by exploring different input variations and perturbations. With the capability of generative models, researchers find new effective attacks [105, 144] against the data-

free black-box threat model, where the adversary uses the synthetic data by generative models to perform the black-box attack by querying the target models. Nonetheless, despite the data-free threat model being more practical than the traditional black-box threat model, it still requires querying the target model multiple times and can be inapplicable to security-concerned applications that are able to detect aggressive queries.

2.3.1 3D Point Cloud Adversarial Attacks

3D deep-learning models exhibit vulnerability to adversarial attacks, even when using 2D adversarial approaches. However, the perturbations applied to 3D point cloud data are more perceptible to humans due to the specific data structure of point clouds. Adversarial perturbations that shift coordinates lead to noticeable changes in the original shape of 3D objects, presenting a challenge in devising stronger and more realistic adversarial attack methods. Early adversarial attack methods, such as those proposed by Liu et al. [78] and Xiang et al. [129], involve adding points generated from 2D FGSM, PGD, and C&W attack methods. Zheng et al. [146] demonstrated high attack performance on the PointNet network by dropping points with the lowest salience scores based on the saliency map. However, these attacks are easily detectable as they alter the number of points in the clean point cloud.

Subsequent works aim to create imperceptible perturbations by shifting point coordinates within the clean point clouds. Approaches like ISO [145], GeoA3 [123], SI-Adv [51], and PF-Attack [42] achieved imperceptible shifting by leveraging geometric and shape information from clean point clouds. LG-GAN [147] and AdvPC [40] utilized generative models to generate camouflaged perturbations effectively. However, only AdvPC and PF-Attack achieved effective black-box attacks against 3D point cloud classifiers. Nonetheless, these methods face challenges in being effective against recently proposed state-of-the-art 3D deep-learning models, resulting in a huge gap in the development between adversarial attacks and benign models.

2.4 Unrestricted Adversarial Attacks

With the development of generative models, these models bring new threats to the robustness of the deep learning models. The adversary adopts the generation ability of the GAN models to craft adversarial examples by generating perturbations [4, 93]. However, these existing methods require re-training of the GAN models and harm the original generation performance of the benign GAN models. Therefore, their performances are limited by their perturbation-based attack algorithms.

Unrestricted adversarial attacks, on the other hand, craft adversarial examples from scratch. These adversarial examples are visually indistinguishable from the benign samples while deceiving the deep-learning models with high confidence. Followed by Song et al. [110] pioneering work, more and more GAN-based unrestricted adversarial attacks [64, 92] are proposed. With the development of diffusion models, recent works [13, 16, 20, 21] demonstrated that diffusion models can be used to generate unrestricted adversarial examples, although these studies have been limited to black-box scenarios and have not thoroughly explored the capabilities of diffusion models as adversaries. Therefore, our work will exploit the remarkable generation ability of diffusion models for discussing attack performance under no-box scenarios with a comprehensive and end-to-end discussion from the generation of the training dataset and adversarial examples.

Particularly, our unrestricted adversarial examples are defined as:

$$x_{\text{UAE}} = \mathcal{G}(z_{adv}, y), s.t.\ y \neq f(x_{\text{UAE}}) \tag{2.3}$$

where z_{adv} and y are the input adversarial latent of the generate model and class label, respectively, $\mathcal{G}$ is the generator, and $f(\cdot)$ is the target classifier. The z_{adv} is commonly sampled from random Gaussian noise.

The unrestricted adversarial examples (UAE) are generated by generative models from scratch. Because these examples are not crafted by adding gradient perturbation to clean images, UAEs are hard to detect and defend by current perturbation-based adversarial defense methods.

2.5 Adversarial Defenses

Adversarial Training Adversarial training (AT) is one of the most practical methods for enhancing a model's robustness against adversarial attacks. It involves training the model with both benign and adversarial data simultaneously during the training phase. However, robustness against unseen attacks remains a significant challenge that affects the defense performance of traditional adversarial training [84]. To address this, Gowal et al. [35] and Rebuffi et al. [98] have incorporated generated and augmented data to improve generalization by increasing data diversity. In addition to leveraging diverse data, refining the objective formulation of AT has also proven effective. By considering model weights, a wide range of adversarial training methods [54, 126] have been proposed.

Adversarial Purification Adversarial purification aims to eliminate adversarial perturbations in adversarial examples without requiring the re-training of deep learning models. These methods leverage the generative capabilities of generative models. Previous works utilizing GANs [103] and score-based matching models [111, 136] have demonstrated state-of-the-art performance compared to adversarial training. With the advent of diffusion models, Nie et al. [89] discovered that diffusion-based adversarial purification methods outperform previous approaches in recovering clean images. However, finding the optimal generation steps for diffusion-based adversarial purification remains challenging. Additionally, adversarial images can negatively impact the reverse generation process of diffusion models. To address these issues, several works [63, 109, 120] have proposed various solutions to enhance the performance of adversarial purification. Increasing the number of purification steps improves defense performance [89, 120]. However, they cannot utilize the full diffusion process for purification because they need to preserve image consistency and the clean data prior. Recent works ([63, 109]) show that gradient-based guidance is an effective method to advance adversarial purification, although it is not time-efficient. Moreover, Lin et al. [74] present an alternative involving supervised additional training on the diffusion model, which tends to suffer in terms of usability and transferability.

Chapter 3

AdvDiff: Generating Unrestricted Adversarial Examples using Diffusion Models

3.1 Introduction

While the DL community continues to explore the wide range of applications of DL models, researchers [114] have demonstrated that these models are highly susceptible to deception by adversarial examples. Adversarial examples are generated by adding perturbations to clean data. The perturbed examples can deceive DL classifiers with high confidence while remaining imperceptible to humans. Many strong attack methods [10, 19, 26, 68, 71, 84] are proposed and investigated to improve the robustness of DL models.

In contrast to existing perturbation-based adversarial attacks, Song et al. [110] found that using a well-trained generative adversarial network with an auxiliary classifier (AC-GAN) [90] can directly generate new adversarial examples without perturbing the clean data. These newly generated examples are considered **unrestricted** as they are obtained by optimizing input noise vectors without any norm restrictions. Compared to traditional adversarial examples, unrestricted adversarial

examples [22, 97] are more aggressive against current adversarial defenses. A malicious adversary can also generate an unlimited number of unrestricted adversarial examples using a trained GAN.

Diffusion models [46] are likelihood-based generative models proposed recently, which emerged as a strong competitor to GANs. Diffusion models have outperformed GANs for image synthesis tasks [25, 56, 100]. Compared with GAN models, diffusion models are more stable during training and provide better distribution coverage. Diffusion models contain two processes: a forward diffusion process and a reverse generation process. The forward diffusion process gradually adds Gaussian noise to the data and eventually transforms it into noise. The reverse generation process aims to recover the data from the noise by a denoising-like technique. A well-trained diffusion model is capable of generating images with random noise input. Similar to GAN models, diffusion models can achieve adversarial attacks by incorporating adversarial objectives [13, 15, 16].

GAN-based unrestricted adversarial attacks often exhibit poor performance on high-quality datasets, particularly in terms of visual quality, because they directly add the PGD perturbations to the GAN latents without theoretic supports. These attacks tend to generate low-quality adversarial examples compared to benign GAN examples [110]. Therefore, these attacks are not imperceptible among GAN synthetic data. Diffusion models, however, offer state-of-the-art generation performance [25] on challenging datasets like LSUN [137] and ImageNet [23]. The conditional diffusion models can generate images based on specific conditions by sampling from a perturbed conditional Gaussian noise, which can be carefully modified with adversarial objectives. These properties make diffusion models more suitable for conducting unrestricted adversarial attacks. Nevertheless, existing adversarial attack methods using diffusion models [13, 15, 16] adopt similar PGD perturbations to the sample in each reverse generation process, making them generate relatively low-quality adversarial examples.

In this chapter, we propose a novel and interpretable unrestricted adversarial attack method called AdvDiff that utilizes diffusion models for adversarial examples

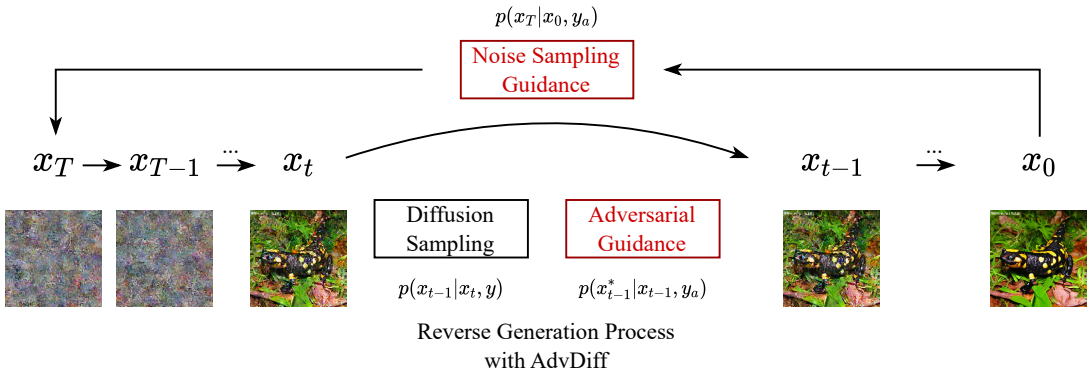


Figure 3.1: **The two new guidance techniques in our AdvDiff to generate unrestricted adversarial examples.** During the reverse generation process, the *adversarial guidance* is added at timestep x_t, which injects the adversarial objective y_a into the diffusion process. The *noise sampling guidance* modifies the original noise by increasing the conditional likelihood of y_a.

generation, as shown in Figure 3.1. Specifically, AdvDiff uses a trained conditional diffusion model to conduct adversarial attacks with two new adversarial guidance techniques. 1) During the reverse generation process, we gradually add *adversarial guidance* by increasing the likelihood of the target attack label. 2) We perform the reverse generation process multiple times, adding adversarial prior knowledge to the initial noise with the *noise sampling guidance*.

Our theoretical analysis indicates that these adversarial guidance techniques can effectively craft adversarial examples by the reverse generation process with adversarial conditional sampling. Furthermore, the sampling of AdvDiff benefits from stable and high sample quality of the diffusion models sampling, which leads to the generation of realistic unrestricted adversarial examples. Through extensive experiments conducted on two datasets, i.e., the high-quality dataset ImageNet, and the small, robust dataset MNIST, we have observed a significant improvement in the attack performance using AdvDiff with diffusion models. These results prove that our proposed AdvDiff is more effective than previous unrestricted adversarial attack methods in conducting unrestricted adversarial attacks to generate high-fidelity and diverse examples without decreasing the generation quality.

Our contributions can be summarized as follows:

- We propose AdvDiff, the new form unrestricted adversarial attack method that utilizes the reverse generation process of diffusion models to generate realistic adversarial examples.
- We design two new effective adversarial guidance techniques to the sampling process that incorporate adversarial objectives to the diffusion model without re-training the model. Theoretical analysis reveals that AdvDiff can generate unrestricted adversarial examples while preserving the high-quality and stable sampling of the conditional diffusion models.
- We perform extensive experiments to demonstrate that AdvDiff achieves an overwhelmingly better performance than GAN models on unrestricted adversarial example generation.

3.2 Preliminaries

In this section, we introduce the diffusion model and the classifier guidance for constructing our adversarial diffusion model.

3.2.1 Classifier-Guided Guidance

Dhariwal et al. [25] achieved conditional diffusion sampling by adopting a trained classifier. The conditional information is injected into the diffusion model by modifying the mean value $\mu_\theta(x_t, t)$ of the samples according to the gradient of the prediction of the target class y by the trained classifier. They adopted log probability to calculate the gradient, and the mean value is given by:

$$\hat{\mu}_\theta(x_t, t) = \mu_\theta(x_t, t) + s \cdot \nabla_{x_t} \log p_\phi(y|x_t) \tag{3.1}$$

where s is the guidance scale.

3.2.2 Classifier-Free Guidance

Ho et al. [47] recently proposed a new conditional diffusion model using classifier-free guidance that injects class information without adopting an additional classifier. The classifier-free guidance utilizes a conditional diffusion model $p_\theta(x|y)$ for image synthesis with given labels. For effective training, they jointly train the unconditional diffusion model $p_\theta(x|\emptyset)$ and the conditional diffusion model $p_\theta(x|y)$, where the unconditional diffusion model is simply replacing the label information with $\emptyset$. Sampling is performed by pushing the model towards the latent space of $p_\theta(x|y)$ and away from $p_\theta(x|\emptyset)$:

$$\hat{\epsilon}_\theta(x_t|y) = \epsilon_\theta(x_t|\emptyset) + w \cdot (\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset)) \tag{3.2}$$

where w is the weight parameter for class guidance and $\emptyset$ is the empty set.

The idea of classifier-free guidance is inspired by the gradient of an implicit classifier $p^i(y|x) \propto p(x|y)/p(x)$, the gradient of the classifier would be:

$$\begin{aligned} \nabla_x log p^i(y|x) &\propto \nabla_x log p(x|y) - \nabla_x log p(x) \\ &\propto \epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset) \end{aligned} \tag{3.3}$$

The classifier-free guidance has a good capability of generating high-quality conditional images, which is critical for performing adversarial attacks. The generation of these images does not rely on a classification model and thus can better fit the conditional distribution of the data.

3.3 Adversarial Diffusion Sampling

3.3.1 Rethinking Unrestricted Adversarial Examples

Song et al. [110] presented a new form of adversarial examples called UAEs. These adversarial examples are not generated by adding perturbations over the clean data but are directly generated by any generative model. UAEs can be viewed as false negative errors in the classification tasks, and they can also bring severe security

problems to deep learning models. These generative-based UAEs can be formulated as:

$$A_{\text{UAE}} \triangleq \{x \in \mathcal{G}(z_{\text{adv}}, y) | y \neq f(x)\} \tag{3.4}$$

where $f(\cdot)$ is the target model for unrestricted adversarial attacks. The unrestricted adversarial attacks aim to generate UAEs that fool the target model while still can be visually perceived as the image from ground truth label y.

Previous UAE works adopt GAN models for the generation of UAEs, and these works perturb the GAN latents by maximizing the cross-entropy loss of the target model, i.e., $\max_{z_{\text{adv}}} \mathcal{L}(f(\mathcal{G}(z_{\text{adv}}, y)), y)$. Ideally, the generated UAEs should guarantee similar generation quality to the samples crafted by standard z because successful adversarial examples should be imperceptible to humans. In other words, UAEs should not be identified among the samples with adversarial latents and standard latents.

However, due to GAN's poor interpretability, there's no theoretical support on z_{adv} that can craft UAEs with normally trained GANs. The generator of GAN is not trained with $z_{adv} = z + \nabla\mathcal{L}$ but only $z \sim \mathcal{N}(0, \mathbf{I})$. Therefore, GAN-based UAEs encounter a significant decrease in generation quality because samples with z_{adv} are not well-trained compared with samples with $z \sim \mathcal{N}(0, \mathbf{I})$. Moreover, the GAN latents are sampled from low dimensional latent spaces. Therefore, GANs are extremely sensitive to the latent z [72, 106]. If we inject gradients of the classification results into GAN latents, GAN-based methods are more likely to generate flipped-label UAEs (images corresponding to the targeted attack label y_a instead of the conditional generation label y) and distorted UAEs. However, these generation issues are hard to address only by attack success rate (ASR). In other words, even with a high ASR, some of the successful UAEs with GAN-based methods should be identified as failure cases for poor visual quality. However, such cases can not be reflected by ASR but can be evaluated by generation quality. All these problems may indicate that GAN models are not suitable for generative-based adversarial attacks.

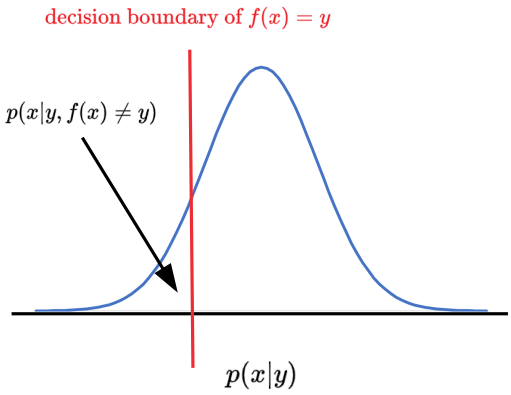


Figure 3.2: **Unrestricted adversarial examples generated by the diffusion model.** The generated adversarial examples should be visually indistinguishable from clean data with label y but wrongly classified by the target classifier f.

Diffusion models have shown better performance on image generation than GAN models [25]. They are log-likelihood models with interpretable generation processes. In this chapter, we aim to generate UAEs by injecting the adversarial loss with theoretical proof and without sabotaging the benign generation process, where we increase the conditional likelihood on the target attack label by following the diffusion process. The perturbations are gradually injected with the backward generation process of the diffusion model by the same sample procedure. As shown in Figure 3.2, the diffusion model can sample images from the conditional distribution $p(x|y)$. The samples from $p(x|y, f(x) \neq y)$ are the adversarial examples that are misclassified by $f(\cdot)$. These examples also follow the data distribution $p(x|y)$ but on the other side of the label y 's decision boundary of $f(\cdot)$. Moreover, the diffusion model's generation process takes multiple sampling steps. Thus, we don't need one strong perturbation to the latent like GAN-based methods. The AdvDiff perturbations at each step are unnoticeable, and perturbations are added to the high dimensional sampled data rather than low dimensional latents. Therefore, AdvDiff with diffusion models can preserve the generation quality and barely generates flipped-label or distorted UAEs.

3.3.2 Adversarial Diffusion Sampling with Theoretical Support

There are several existing adversarial attack methods [13, 15, 16] that adopt diffusion models to generate adversarial examples. However, these methods still adopt PGD or I-FGSM gradients to perturb the diffusion process for constructing adversarial examples. As discussed earlier, the generation process of diffusion models is a specially designed sampling process from given distributions. Such adversarial gradients change the original generation process and can harm the generation quality of the diffusion model. Additionally, these methods fail to give a comprehensive discussion of the adversarial guidance with theoretical analysis. Therefore, we aim to design a **general** and **interpretable** method to generate adversarial examples using diffusion models **without** affecting the benign diffusion process.

3.3.3 Adversarial Guidance

Inspired by Dhariwal's work [25] that achieves the conditional image generation by classifier gradient guidance $\nabla_{x_t} \log p_\phi(y|x_t)$, we generate our UAEs with adversarial gradient guidance over the reverse generation process. Our attack aims at utilizing a conditional diffusion model $\epsilon_\theta(x_t, y)$ to generate x_0 that fits the ground truth label y while deceiving the target classifier with $p_f(x_0) \neq y$. These generated samples are the false negative results in p_f's classification results.

Normally, we will obtain the images with label y by following the standard reverse generation process with classifier-free guidance:

$$x_{t-1} = \mu(x_t, y) + \sigma_t \varepsilon \tag{3.5}$$

where $\mu(x_t, y)$ is the conditional mean value and ε is sampled from $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$.

Sampling by Equation 3.5, we obtain the samples with the generation process $p(x_{t-1}|x_t, y)$. Following the above-mentioned definition of UAEs, we can get our adversarial examples by adding adversarial guidance to the standard reverse pro-

cess, which is performing another sampling with the adversarial generation process $p(x^*_{t-1}|x_{t-1}, f(x) \neq y)$. We find that specifying a target label for the adversarial generation process is more effective during experiments. Suggest the target label y_a is the target for the adversarial attacks, the adversarial example is sampled by the following steps:

$$x^*_{t-1} = x_{t-1} + \sigma_t^2 s \nabla_{x_{t-1}} \log p_f(y_a|x_{t-1}) \tag{3.6}$$

where s is the adversarial guidance scale. The derivation of Equation 3.6 is given in the Appendix. Intuitively, the adversarial guidance encourages the generation of samples with a higher likelihood of the target label.

In practice, we utilize the classifier-free guidance to train a conditional diffusion model $\epsilon_\theta(\cdot)$ as our basic generation model.

3.3.4 Noise Sampling Guidance

We can improve the reverse process by adding an adversarial label prior to the noise data x_T. The UAEs are a subset of the dataset labeled with y. They can be viewed as the conditional probability distribution with $p(x|y, f(x) = y_a)$ during sampling, and y_a is the target label for the adversarial attack. Therefore, we can add the adversarial label prior to x_T with Bayes' theorem:

$$\begin{aligned} p(x_T|y_a) =& \frac{p(y_a|x_T)p(x_T)}{p(y_a)} = \frac{p(y_a|x_T, x_0)p(x_T|x_0)}{p(y_a|x_0)} \\ =& p(x_T|x_0)e^{\log p(y_a|x_T) - \log p(y_a|x_0)} \end{aligned} \tag{3.7}$$

We can infer the x_T with the adversarial prior by Equation 3.16, i.e.,

$$x_T = (\mu(x_0, y) + \sigma_t \varepsilon) + \bar{\sigma}_T^2 a \nabla_{x_0} \log p_f(y_a|x_0) \tag{3.8}$$

where a is the noise sampling guidance scale. See the Appendix for detailed proof.

Equation 3.8 is similar to Equation 3.6 as they both add adversarial guidance to the reverse generation process. However, the noise sampling guidance is added to x_T

Algorithm 1 DDPM Adversarial Diffusion Sampling

Require: y_a: target label for adversarial attack, y: ground truth class label, s, a: adversarial guidance scale, w: classification guidance scale, N: noise sampling guidance steps, T: reverse generation process timestep

$x_T \sim \mathcal{N}(0, \mathbf{I})$
$x_{adv} = \varnothing$
for $i = 1 \dots N$ **do**
  for $t = T, \dots, 1$ **do**
    $\tilde{\epsilon}_t = (1 + w)\epsilon_\theta(x_t, y) - w\epsilon_\theta(x_t)$
    Classifier-free sampling x_{t-1} with $\tilde{\epsilon}_t$.
    Input x_{t-1} to target model and get the gradient $\log p_f(y_a|x_{t-1}))$
    $x^*_{t-1} = x_{t-1} + \sigma_t^2 s \nabla_{x_{t-1}} \log p_f(y_a|x_{t-1})$
  end for
  Obtain classification result from $f(x_0)$
  Compute the gradient with $\log p_f(y_a|x_0)$
  Update x_T by $x_T = x_T + \bar{\sigma}_T^2 a \nabla_{x_0} \log p_f(y_a|x_0)$
  $x_{adv} \leftarrow x_0$ if $f(x_0) = y_a$
end for
return x_{adv}

according to the final classification gradient $\nabla_{x_0} \log p_f(y_a|x_0)$, which provides a strong adversarial guidance signal directly to the initial input of the generative model. The gradient of Equation 3.8 is effective as it reflects the eventual classification result of the target classifier.

3.3.5 Training-Free Adversarial Attack

The proposed adversarial attack does not require additional modification on the training of the diffusion model. The adversarial examples are sampled by using Algorithm 1 over the trained classifier-free diffusion model $\epsilon_\theta(\cdot)$.

3.4 Experiments

Datasets and Target Models. We use two datasets for major evaluation: MNIST [24] and ImageNet [23]. MNIST is a 10-classes dataset consisting of handwritten numbers from 0 to 9. We adopt the MNIST dataset to evaluate our method for low-quality robust image generation. ImageNet is a large visual database with 1000

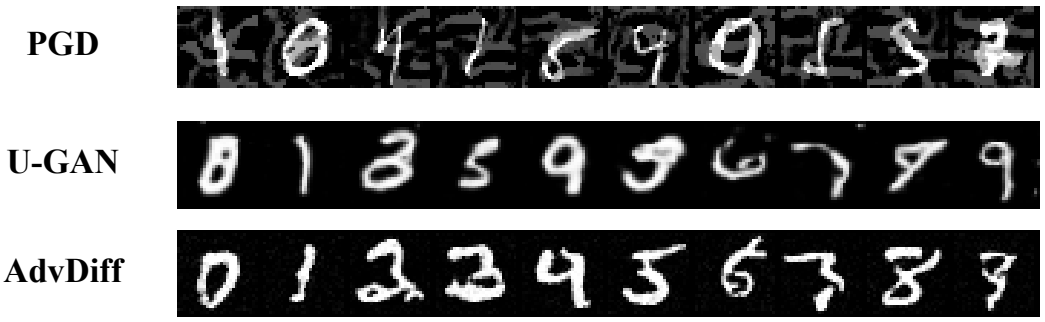


Figure 3.3: **Adversarial examples on the MNIST dataset.** Perturbation-based attack methods generate noise patterns to conduct attacks, while unrestricted adversarial attacks (U-GAN and AdvDiff) are imperceptible to the clean data.

object classes and is used for the high-quality generation task. For target classifiers, we adopt simple LeNet5 [61], and ResNet18 [43] for the MNIST dataset, and the widely-used ResNet50 [43] and WideResNet50-2 [140] for the ImageNet dataset.

Comparisons. It is not applicable to give a clear comparison between perturbation attacks and unrestricted attacks because perturbation attacks have the corresponding ground truth while unrestricted attacks do not. We mainly compare our method with the unrestricted adversarial attack U-GAN [110] and give the discussion with the AutoAttack [19], PGD [84], BIM [26], and C&W [10] perturbation-based attacks under norm $\ell_{\text{inf}} = 8/255$. For U-GAN, We adopt AC-GAN [90] for the MNIST dataset, and SAGAN [142] and BigGAN [8] for the ImageNet dataset, as AC-GAN has shown poor performance on ImageNet. We use the official code from DiffAttack [13] and implement AdvDiffuser by ourselves [15] for comparisons. We do not compare with Chen et al. [16], because they use a similar method as DiffAttack and without official code. Because existing diffusion model attacks are all untargeted attacks, we include the untargeted version of AdvDiff for a clear comparison, which is represented by "AdvDiff-Untargeted".

Implementation Details. Because our adversarial diffusion sampling does not require additional training to the original diffusion model, we use the pre-trained diffusion model in our experiment. We adopt DDPM [46] with classifier-free guidance

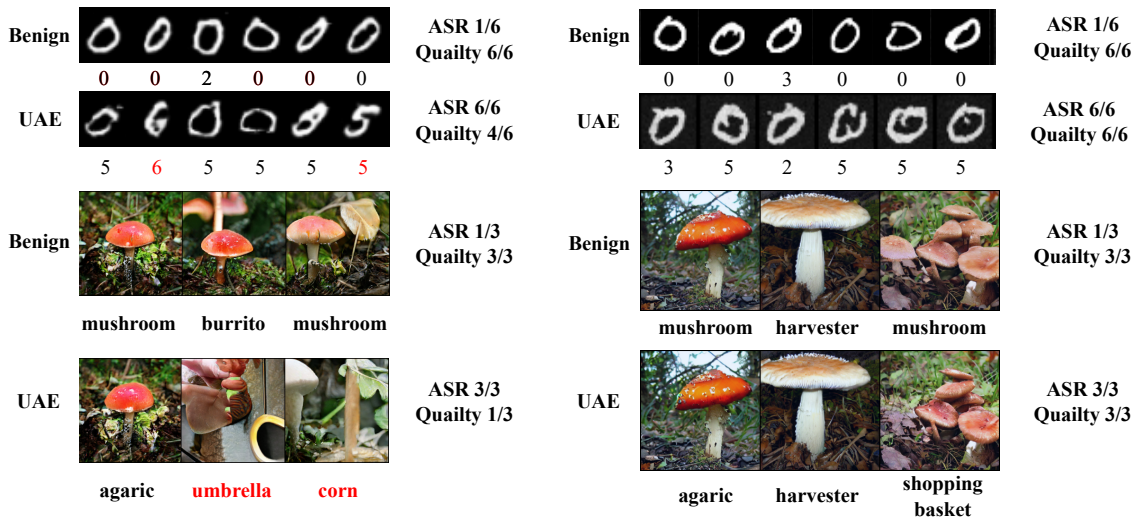


Figure 3.4: **Comparisons of unrestricted adversarial attacks between GANs and diffusion models on two datasets.**. Left: generated samples from U-GAN (BigGAN for ImageNet dataset). Right: generated samples from AdvDiff. We generate unrestricted adversarial examples on the MNIST "0" label and ImageNet "mushroom" label. U-GAN is more likely to generate adversarial examples with the target label, i.e., examples with red font. However, AdvDiff tends to generate the "false negative" samples by the target classifier by combing features from the target label.

for the MNIST dataset and LDM [100] with DDIM sampler for the ImageNet dataset. For MNIST dataset, we use $N = 10$, $s = 0.5$, and $a = 1.0$, And $N = 5$, $s = 0.7$, and $a = 0.5$ for ImageNet dataset.

Evaluation Metrics. We utilize the top-1 classification result to evaluate the ASR on different attack methods under untargeted attack settings. As discussed earlier, GAN-based UAEs often encounter severe generation quality drops compared to benign GAN samples. Therefore, we give comparisons of generation performance on ImageNet to evaluate the attack performance of different UAEs in imperceptibly. The results are averaged with five runs. We use ResNet50 as the target model for default settings.

3.4.1 Attack Performance

MNIST We show the attack success rate against the normally trained model and adversarially trained model [84] in the MNIST dataset. All the selected adversarial

Table 3.1: **The attack success rate on MNIST dataset.**

Method	ASR(%)			
	LeNet5		ResNet18	
	Clean	PGD-AT	Clean	PGD-AT
PGD	99.8	25.6	99.3	20.8
BIM	99.6	34.6	**100**	31.5
C&W	**100**	68.6	**100**	64.5
U-GAN	88.5	*79.4*	85.6	*75.1*
AdvDiff	94.2	***88.6***	92.1	***86.5***

Table 3.2: **The attack success rate on ImageNet dataset.** U-SAGAN and U-BigGAN represent the base GAN models for U-GAN are SAGAN and BigGAN, respectively.

Method	ASR(%)						Time (s)
	ResNet50			WideResNet50-2			
	Clean	DiffPure	PGD-AT	Clean	DiffPure	PGD-AT	
AutoAttack	95.1	22.2	56.2	94.9	20.6	55.4	0.5
U-SAGAN	99.3	30.5	80.6	98.9	28.6	70.1	10.4
U-BigGAN	96.8	40.1	81.5	96.5	35.5	78.4	11.2
AdvDiffuser	95.4	28.9	90.6	94.6	26.5	88.9	38.6
DiffAttack	92.8	30.6	88.4	90.6	27.6	85.3	28.2
AdvDiff	**99.8**	41.6	92.4	**99.9**	38.5	90.6	**9.2**
AdvDiff-Untargeted	99.5	**75.2**	**94.5**	99.4	**70.5**	**92.6**	9.6

attacks achieve over 90% attack success rate against the normally trained model. The adversarially trained model can effectively defend against perturbation-based adversarial attacks for their noise-like perturbation generation patterns, as reported in Table 3.1. However, the UAEs obviously perform better with their non-perturbed image generation. Despite the fact that the unrestricted attack can break through the adversarial defenses, the crafted adversarial examples should also be imperceptible to humans for a reasonably successful attack. The visualized adversarial examples in Figure 3.3 show that the perturbation-based adversarial attacks tend to blur the original images while U-GAN can generate mislabeled adversarial examples.

ImageNet It is reported that deep learning models on ImageNet are extremely vulnerable to adversarial attacks. However, the state-of-the-art adversarial defense DiffPure [89] and adversarial training [84] can still defend against the perturbation-based attacks, as reported in Table 3.2. More UAEs evade the current defenses, but the

generation quality of U-GAN is relatively poor compared to our adversarial examples. This phenomenon also shows that the performance of UAEs is heavily affected by the generation quality of the generation model. The adversarial examples generated by AdvDiff are more aggressive and stealthy than U-GAN's. Meanwhile, the generation speed of AdvDiff is the best among all the unrestricted adversarial attack methods. Note that we adopt the clean images generated by LDM to achieve DiffAttack and AutoAttack for a fair comparison.

3.4.2 Generation Quality: True ASR for UAEs

We witness similar ASR with U-GAN and AdvDiff. However, imperceptibility is also critical for a successful unrestricted adversarial attack, so we adopt the evaluation metrics in [25] to compare the generation quality with and without performing unrestricted attacks. Table 3.3 shows that the AdvDiff achieves an overwhelming better IS score and similar FID score on the large-scale ImageNet dataset, where FID [44] and IS [101] scores are commonly adopted for evaluating the quality of a generative model. Because the generation of UAEs does not modify the data distribution of the generated images, the Precision score can be inferred as generation quality, while the Recall score indicates the flipped-label problems. We witness the frequent generation of flipped-label UAEs and low-quality UAEs from GAN-based methods, which is reflected by the decrease in the Precision score and the increase in the Recall score. Figure 3.4 illustrates this problem with some examples. It can be further proved that U-BigGAN achieves much higher image quality on non-reference metrics than reference metrics, as shown in Table 3.4.

We find the IS score is heavily affected by the transferability of adversarial examples due to the calculation method. Therefore, we further compare the image quality of adversarial examples by commonly used metrics in Table 3.4. The results show that AdvDiff (average 5 out of 5) and AdvDiff-Untargeted (average 4 out of 5) outperform existing adversarial attack methods using diffusion models. The perturbation-based adversarial attacks, i.e., AutoAttack, achieve much worse image quality compared with UAEs.

Table 3.3: **The generation performance on the ImagetNet dataset.**

Method	FID (↓)	sFID (↓)	IS (↑)	Precision (↑)	Recall (↑)
SAGAN	41.9	50.2	26.7	0.50	0.51
BigGAN	19.3	45.7	250.3	0.95	0.21
LDM	12.3	25.4	385.5	0.94	0.73
U-SAGAN	52.8/+26%	52.2/+4%	12.5/-53%	0.58	0.57
U-BigGAN	25.4/+31%	52.1/+14%	129.4/-48%	0.81	0.35
AdvDiffuser	26.8/+117%	38.6/+51%	206.8/-46%	0.70	0.75
DiffAttack	20.5/+66%	40.2/+58%	264.3/-31%	0.83	0.73
AdvDiff	**16.2**/+31%	**30.4**/+20%	**343.8**/-10%	**0.90**	0.75
AdvDiff-Untargeted	22.8/+85%	33.4/+28%	220.8/-45%	0.85	**0.76**

Table 3.4: **The image quality on the ImagetNet dataset.**

Method	FID (↓)	LPIPS (↓)	SSIM (↑)	BRISQUE [85] (↓)	TRES (↑)
AutoAttack	26.5	0.72	0.21	34.4	69.8
U-BigGAN	25.4	0.50	0.32	19.4	80.3
AdvDiffuser	26.8	0.21	0.84	18.9	75.6
DiffAttack	20.5	0.15	0.75	22.6	67.8
AdvDiff	**16.2**	**0.03**	**0.96**	**18.1**	**82.1**
AdvDiff-Untargeted	22.8	0.14	0.85	23.4	76.8

3.4.3 UAEs against Defenses and Black-box Models

Current defenses assume the adversarial examples are based on perturbations over data from the training dataset, i.e., $x_{adv} = x + \nabla\mathcal{L}, x \in D$. However, UAEs are synthetic data generated by the generative model. Because of different data sources, current defenses are hard to defend UAEs, which brings severe security concerns to deep learning applications. The proposed AdvDiff achieves an average of 36.8% ASR against various defenses, while AutoAttack only achieves 30.7% ASR with significantly lower image quality. We also test the attack transferability of AdvDiff and the results show that the untargeted version of AdvDiff achieves the best performance against black-box models. Experiment results are given in Table 3.5.

Table 3.5: **The attack success rates (%) of ResNet50 examples for transfer attack and attack against defenses on the ImagetNet dataset.**

Method	ResNet-152 [43]	Inception v3 [113]	ViT-B [28]	BEiT [5]
AutoAttack	32.5	38.6	9.3	45.3
U-BigGAN	30.8	35.3	30.1	69.4
AdvDiffuser	18.3	20.0	18.5	79.4
DiffAttack	21.1	**43.9**	17.4	78.0
AdvDiff	20.5	14.9	17.8	78.8
AdvDiff-Untargeted	**52.0**	42.7	**36.0**	**81.5**
Method	Adv-Inception [84]	AdvProp [131]	DiffPure [89]	HGD [73]
AutoAttack	14.6	69.6	22.2	20.5
U-BigGAN	40.6	75.2	40.1	22.6
AdvDiffuser	24.4	84.0	30.5	10.8
DiffAttack	30.9	85.1	30.6	20.5
AdvDiff	19.4	89.7	41.6	17.8
AdvDiff-Untargeted	**60.1**	**95.3**	**75.2**	**53.8**
Method	R&P [132]	RS [17]	NRP [88]	Bit-Red [135]
AutoAttack	20.6	38.9	39.4	19.8
U-BigGAN	14.2	34.5	30.9	13.1
AdvDiffuser	15.4	38.4	40.5	11.4
DiffAttack	23.7	40.8	38.5	20.1
AdvDiff	17.4	47.6	45.2	15.8
AdvDiff-Untargeted	**56.8**	**82.8**	**74.2**	**52.6**

3.4.4 Better Adversarial Diffusion Sampling

We present detailed comparisons with DiffAttack and AdvDiffuser. The results show that the proposed adversarial guidance achieves significantly higher generation quality than PGD-based adversarial guidance. With PGD gradient guidance, the diffusion model generates images with a similar Recall score but a much lower Precision score, which indicates that the PGD gradient influences the benign generation process and causes the generation of low-quality images. The result proves that the adversarial guidance of diffusion models should be carefully designed without affecting the benign sampling process. Meanwhile, the generation speed of AdvDiff is the best among the existing diffusion attack methods. Note that AdvDiff (36.8%) sightly outperforms AdvDiffuser (32.0%) and DiffAttack (36.2%) against defenses. However, previous attacks achieve slightly better transfer attack performance than the original AdvDiff. The reason could be the gradient of the cross-entropy loss is shared among nearly

all the deep learning models and is better at attack transferability against these models. Nevertheless, the untargeted version of AdvDiff achieves overwhelmingly better performance, which further demonstrates the effectiveness of the proposed adversarial sampling. But the generation quality is affected, we leave a better design in the future work.

3.4.5 Ablation Study

We discuss the impact of the parameters of AdvDiff in the subsection. Note that our proposed method does not require re-training the conditional diffusion models. The ablation study is performed only on the sampling process.

Adversarial Guidance Scale s and a. The magnitudes of s and a greatly affect the ASR of AdvDiff, as shown in Figure 3.5. Noted that we witness the generation of unrealistic images when setting the adversarial guidance extremely large.

Noise Sampling Guidance Steps N. Like the iteration times of GAN-based unrestricted adversarial attacks, larger steps N can effectively increase the attack performance against an accurate classifier, as shown in Figure 3.5. However, it can affect the initial noise distribution and hence decreases the generation quality. During experiments, we observe that adversarial guidance is already capable of generating adversarial examples with high ASR. Thus, we can set a small noise sampling guidance step N for better sample quality.

Adversarial Guidance Timestep t^*. The reverse diffusion process gradually denoises the input noise. Therefore we generally get noisy images at most timesteps. Because the target classifier is not able to classify the noisy input, the adversarial guidance is not effective in the early reverse diffusion process. Figure 3.5 shows our results, and we can improve the performance of adversarial guidance by training a separate classifier, which we leave for future work.

3.5 Conclusion

In this work, we propose a new method called AdvDiff, which can conduct unrestricted adversarial attacks using any pre-trained conditional diffusion model. We propose two novel adversarial guidance techniques in AdvDiff that lead diffusion models to obtain high-quality, realistic adversarial examples without disrupting the diffusion process. Experiments show that our AdvDiff vastly outperforms GAN-based and diffusion-based attacks in terms of attack success rate and image generation quality, especially in the ImageNet dataset. AdvDiff indicates that diffusion models have demonstrated effectiveness in adversarial attacks, and highlights the need for further research to enhance AI model robustness against unrestricted attacks.

3.6 Appendix

3.6.1 Detailed Proof of Equation 3.6

We can obtain the sample x_{t-1} with condition label y, according to the sampling with the classifier-free guidance. To get the unrestricted adversarial example x^*_{t-1}, we add adversarial guidance to the conditional sampling process with Equation 8. With Bayes' theorem, we want to deduce the adversarial sampling with adversarial guidance at timestep t by:

$$p(x^*_{t-1}|y_a) = \frac{p(y_a|x^*_{t-1})p(x^*_{t-1})}{p(y_a)} \tag{3.9}$$

with Equation 3.9, we want to sample the adversarial examples with the target label y_a. Starting from x_t, the sampling of the reverse generation process with AdvDiff is:

$$p(x^*_{t-1}|x_t, y_a) = \frac{p(y_a|x^*_{t-1}, x_t)p(x^*_{t-1}|x_t)}{p(y_a|x_t)} \tag{3.10}$$

Noted that Equation 3.10 is the same as the deviation of classifier-guidance in [25]'s Section 4.1, where they treated $p(y_a|x_t)$ as a constant. Because $p(x^*_{t-1}|x_t)$ is the known

sampling process by our conditional diffusion sampling, we evaluate $\frac{p(y_a|x^*_{t-1},x_t)}{p(y_a|x_t)}$ by:

$$\log p_f(y_a|x^*_{t-1}) - \log p_f(y_a|x_t) \tag{3.11}$$

We can approximate Equation 3.11 using a Taylor expansion around $x^*_{t-1} = \mu(x_t)$ as:

$$\begin{aligned}\log p_f(y_a|x^*_{t-1}) - \log p_f(y_a|x_t) &\approx \log p_f(y_a|\mu(x_t)) \\ &\quad + (x^*_{t-1} - \mu(x_t))\nabla_{\mu(x_t)} \log p_f(y_a|\mu(x_t)) \\ &\quad - \log p_f(y_a|x_t) + C \\ &= (x^*_{t-1} - \mu(x_t))\nabla_{\mu(x_t)} \log p_f(y_a|\mu(x_t)) + C \end{aligned} \tag{3.12}$$

Assume $p(x^*_{t-1}|x_t) = \mathcal{N}(x^*_{t-1}; \mu(x_t), \sigma_t^2\mathbf{I}) \propto e^{-(x^*_{t-1}-\mu(x_t))^2/2\sigma_t^2}$, we have:

$$\begin{aligned} p(x^*_{t-1}|x_t, y_a) &\propto e^{-(x^*_{t-1}-\mu(x_t))^2/2\sigma_t^2 + (x^*_{t-1}-\mu(x_t))\nabla_{\mu(x_t)} \log p_f(y_a|\mu(x_t))} \\ &\propto e^{-(x^*_{t-1}-\mu(x_t)-\sigma_t^2\nabla_{\mu(x_t)} \log p_f(y_a|\mu(x_t)))^2/2\sigma_t^2 + (\nabla_{\mu(x_t)} \log p_f(y_a|\mu(x_t)))^2/2\sigma_t^2} \\ &\propto e^{-(x^*_{t-1}-\mu(x_t)-\sigma_t^2\nabla_{\mu(x_t)} \log p_f(y_a|\mu(x_t)))^2/2\sigma_t^2 + C} \\ &\approx \mathcal{N}(x^*_{t-1}; \mu(x_t) + \sigma_t^2\nabla_{\mu(x_t)} \log p_f(y_a|\mu(x_t)), \sigma_t^2\mathbf{I}) \end{aligned} \tag{3.13}$$

Sampling with Equation 3.13 should be:

$$x^*_{t-1} = \mu(x_t, y) + \sigma_t\varepsilon + \sigma_t^2 s\nabla_{\mu(x_t)} \log p_f(y_a|\mu(x_t)) \tag{3.14}$$

where $\mu(x_t, y)$ is the conditional mean value and ε is sampled from $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. Note that $\mu(x_t, y) + \sigma_t\varepsilon$ is the normal sampling process that we will get x_{t-1}. In practice, in each diffusion step, the difference between x_{t-1} and $\mu(x_t)$ should be small enough [25, 46] for a reasonable and stable diffusion sampling. Therefore, we adopt x_{t-1} to calculate the adversarial gradient after the sampling with the conditional diffusion model, and we have:

$$x^*_{t-1} = \mu(x_t, y) + \sigma_t\varepsilon + \sigma_t^2 s\nabla_{\mu(x_t)} \log p_f(y_a|\mu(x_t)) \approx x_{t-1} + \sigma_t^2 s\nabla_{x_{t-1}} \log p_f(y_a|x_{t-1}) \tag{3.15}$$

where s is the adversarial guidance scale. $\square$

3.6.2 Detailed Proof of Equation 3.8

The deviation of Equation 10 is similar to Equation 8, where the noise sampling guidance is added with the forward diffusion process. Similarly, we have Equation 9:

$$p(x_T|y_a) = \frac{p(y_a|x_T)p(x_T)}{p(y_a)} = \frac{p(y_a|x_T, x_0)p(x_T|x_0)}{p(y_a|x_0)} \tag{3.16}$$

And Taylor expansion around $x_T = x_0$ to evaluate $\frac{p(y_a|x_T,x_0)}{p(y_a|x_0)}$.

$$\log p_f(y_a|x_T) - \log p_f(y_a|x_0) = (x_T - x_0)\nabla_{x_0} \log p_f(y_a|x_0) + C \tag{3.17}$$

From x_0 to x_T, we gradually add the Gaussian noise with the predefined schedule [46]:

$$p(x_T|x_0) = \mathcal{N}(x_T; \sqrt{\bar{\alpha}_T}x_0, (1 - \bar{\alpha}_T)\mathbf{I}) \tag{3.18}$$

The noise sampling guidance is as follows:

$$\begin{aligned} x_T &\approx (\bar{\mu}(x_0, y) + \bar{\sigma}_T\varepsilon) + \bar{\sigma}_T^2 a\nabla_{x_0} \log p_f(y_a|x_0) \\ &= x_T + \bar{\sigma}_T^2 a\nabla_{x_0} \log p_f(y_a|x_0) \end{aligned} \tag{3.19}$$

where $\bar{\mu}(x_0, y) + \bar{\sigma}_T\varepsilon$ is the forward diffusion process to get x_T with x_0 and a is the noise sampling guidance scale. $\square$

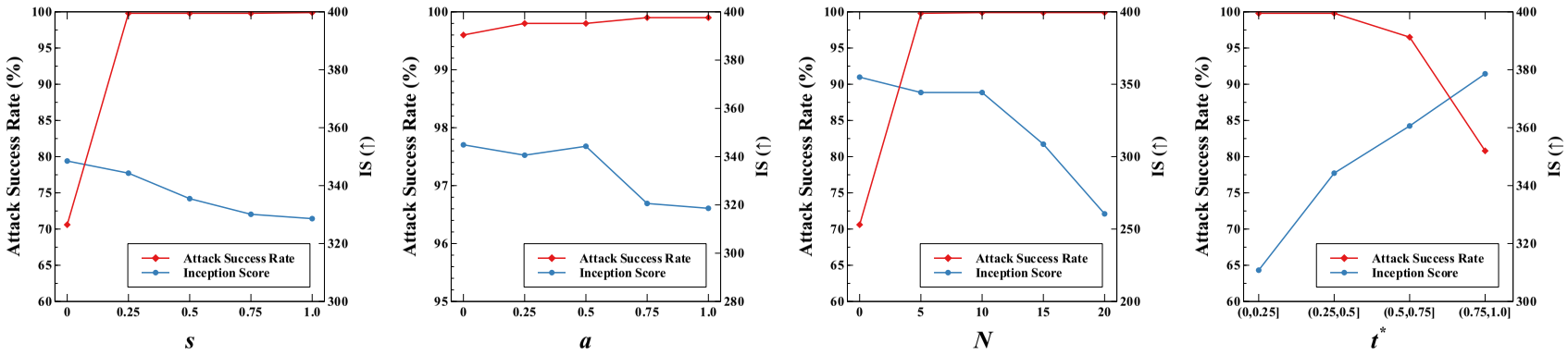


Figure 3.5: **Ablation study of the impact of parameters in AdvDiff.** The results are generated from the ImageNet dataset against the ResNet50 model. We adopt the ASR and IS scores to show the impact of attack performance and generation quality.

Chapter 4

Diffusion Models as Strong Adversaries

4.1 Introduction

The adversarial vulnerability [114] of deep learning models is a severe security issue that threatens the deployment of AI applications. Adversarial attacks aim to deceive deep learning models by introducing small perturbations to the input data [1, 34, 60, 114]. Based on the knowledge the attacker possesses to generate these perturbations, adversarial attacks can be categorized as white-box attacks or black-box attacks. White-box attacks assume that the attacker has access to the target model's parameters or network structure, allowing them to craft effective adversarial examples. On the other hand, black-box attacks assume that the adversary has no such access and can only interact with the model through input-output queries. Despite this limitation, black-box attacks have shown the ability to achieve high attack success rates against state-of-the-art models in practical scenarios. As a result, deep learning model applications face threats from potential adversaries.

In previous black-box attacks, the transferability of adversarial examples was exploited to deceive the target model. These attacks involved generating adversarial examples against a substitute model that was trained on the same dataset as the target model. However, a more practical scenario is that the adversary may not have

Figure 4.1: **The generated clean images (left) and UAE from the diffusion model.** Our proposed attack is capable of producing high-quality UAEs under the no-box threat model. It's worth noting that these UAEs are generated without any conspicuous noisy patterns, unlike perturbation-based attacks.

access to the training dataset of the target model, where we call this type of attack as a no-box attack. The no-box threat model, introduced by Li et al. [65], imposes more practical constraints on the adversary. In this scenario, the attacker is not allowed to access the training data or the outputs of the black-box target model. Only a few correctly labeled data are leaked to the adversary, which limits their knowledge about the target model. Existing works on no-box adversarial attacks leverage the transferability of adversarial examples from a substitute model [65, 112]. However, these works still rely on using data from the validation set of the target model, which may not be available or permissible in many security-concerned applications. Additionally, these attacks require a relatively larger norm perturbation than black-box attacks for successful adversarial examples generation. Therefore, it is still a challenge to conduct effective adversarial attacks under a no-box scenario.

With the advancements in generative models, there is a growing concern regarding their potential threats to humans and deep learning applications. Diffusion models [46, 108] are particularly powerful generative models that have gained attention from both the research community and the general public. Large-scale public text-to-image diffusion models, such as Stable Diffusion [100], have demonstrated their ability to generate AI-manipulated images that can deceive humans with false in-

formation. This raises important security issues that require the attention of the research community to address and mitigate the risks involved. Given the impressive generative capabilities of diffusion models across various tasks, it is worth exploring whether diffusion models can serve as strong adversaries by self-generating training data for adversarial attacks. However, only a few works [13, 15, 16, 20] have discussed the ability of diffusion models for adversarial attacks. And none of them perform adversarial attacks under the no-box scenario.

In this chapter, we investigate and demonstrate the effectiveness of diffusion models as powerful adversaries under the no-box and black-box threat models. Specifically, the training data of the proposed attack is **only** consisting of generated data by the diffusion model for no-box attack. We leverage a technique called classifier-free guidance [48] to conditionally generate data using label information from the target model's training dataset, which we utilized the generated data as the training dataset. To provide a comprehensive discussion of the diffusion model, we utilize a classification diffusion model as the substitute model in our attack. This substitute model estimates the distribution of labels based on the input data, employing uncertainty estimation techniques. To improve the transferability of adversarial examples, we introduce scheduled noise during the training of the substitute model. Once the substitute model is trained, we utilize the same diffusion model to generate a dataset for performing no-box unrestricted adversarial attacks as shown in Figure 4.1. We adopt an ensemble-like approach using the Monte Carlo sampling method over multiple conditional distribution predictions from the diffusion substitute model. The generation pipeline of our proposed attack is given in Figure 4.2. We conduct experiments on the ImageNet [23] dataset to demonstrate the effectiveness of diffusion models as strong adversaries against deep learning models even in a no-box attack threat model. Our work emphasizes the need for the community to focus on developing more robust defenses against adversarial attacks involving diffusion models. Besides the no-box attack, we also test the attack performance of diffusion models under the standard black-box attack scenario.

Our contributions are summarized as follows:

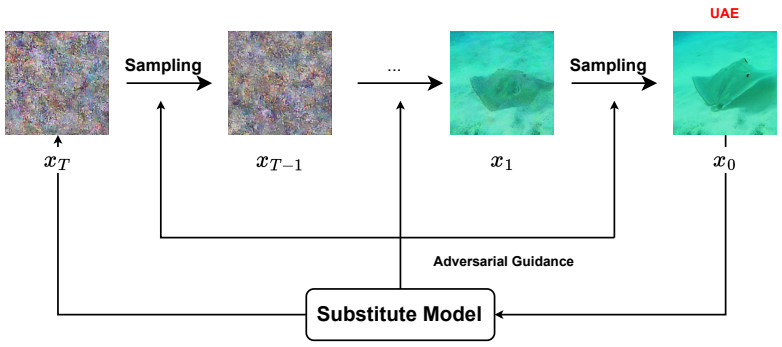


Figure 4.2: **The attack pipeline of the proposed adversarial attack.** The generation of our unrestricted adversarial examples follows the normal reverse generation pattern of the diffusion models, where we incorporate adversarial guidance from the substitute model to adversarially sample the UAEs.

- We propose an effective no-box adversarial attack method using diffusion models against existing deep learning models. Our proposed attack does not rely on any training data or queries from the target model, making it a practical approach for no-box attacks. Additionally, the proposed method demonstrates significant effectiveness in black-box adversarial attacks.

- We design effective approaches to generate no-box adversarial examples with diffusion models under the no-box threat model, including the generation method for constructing the dataset, a special fine-tuning method that incorporates model uncertainty and noise augmentation to enhance the model transferability, and a novel ensemble-like no-box unrestricted adversarial attack method that leverages the average prediction from the diffusion substitute model for the generation of strong adversarial examples.

- We conduct extensive experiments to validate the effectiveness of our approach. Our results show that the proposed attack can generate effective no-box and black-box adversarial examples, achieving a state-of-the-art attack success rate compared to existing methods.

4.2 Background

4.2.1 Adversarial Attacks

No-box attacks.

We give the definition of the no-box threat model in our chapter following Li et al. [65] that we assume the attacker can access neither the whole training dataset nor any pre-trained target model. Accessing the validation data or testing data is also prohibited. The attacker can only have some basic information about the dataset and the target model following the label-only data-free setting [144], such as label encoding, label information, data structure, model input and output structure, and any other auxiliary information.

In this chapter, we utilize the generative capabilities of diffusion models to construct the training dataset for the substitute model. The selection of diffusion models for this purpose needs to meet two requirements: (1) the diffusion models should be open-source and publicly available for practical reasons, and (2) the diffusion models should be capable of generating data that is similar to the training data of the target classifier. To generate the training dataset, we employ conditional labels for DDIM models with classifier-free guidance and prompts with label text for text-to-image diffusion models. By using these techniques, we create a dataset that closely approximates the target classifier's training data. Once the training dataset is obtained, we can train the substitute model using this data to perform the no-box adversarial attack. This allows us to craft adversarial examples that can successfully fool the target classifier, even without direct access to its training data or the ability to query it.

4.2.2 Adversarial Attacks with Generative Models

Inspired by Dai's work [20], diffusion models are a powerful model to generate human imperceptible UAEs. To sample UAEs with the diffusion model, adversarial guidance is added in the reverse generation process:

$$
\begin{aligned}
x_{t-1} = \sqrt{\alpha_{t-1}} & \left(\frac{x_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \right) \\
& + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \boldsymbol{\epsilon}_\theta^{(t)}(x_t) \\
& + \sigma_t \boldsymbol{\epsilon}_t + a_1 \cdot \sqrt{1-\alpha_t} \nabla_{x_t} \log f(y_a | x_t)
\end{aligned} \tag{4.1}
$$

where a_1 is the scale of the adversarial guidance and y_a is the target label for the adversarial attack.

The noise sampling guidance is added to the initial noise to better sample the UAEs with prior knowledge:

$$
x_T = x_T + a_2 \cdot \sqrt{1-\alpha_T} \nabla_{x_0} \log p_f(y_a | x_0) \tag{4.2}
$$

where a_2 is the noise sampling guidance scale.

4.3 Methodology

The proposed attack achieves no-box adversarial attacks by UAEs generated by the diffusion models. The whole attack pipeline is given in Figure 4.3. The substitute model is trained by the generated dataset with the same diffusion models for attack. We will introduce our training mechanisms for the substitute model with the generative ability of diffusion models in Section III.A, and the fine-tuning method with model uncertainty in Section III.B. The no-box adversarial attack algorithms will be illustrated in Section III.C with detailed discussions.

4.3.1 Training Mechanisms with Diffusion Models

With the development of diffusion models like the LDM [100] and its successor Stable Diffusion, these models have shown remarkable capabilities in generating high-quality and high-resolution images. Previous works have demonstrated that utilizing generative models as an additional source of training data can enhance the performance of

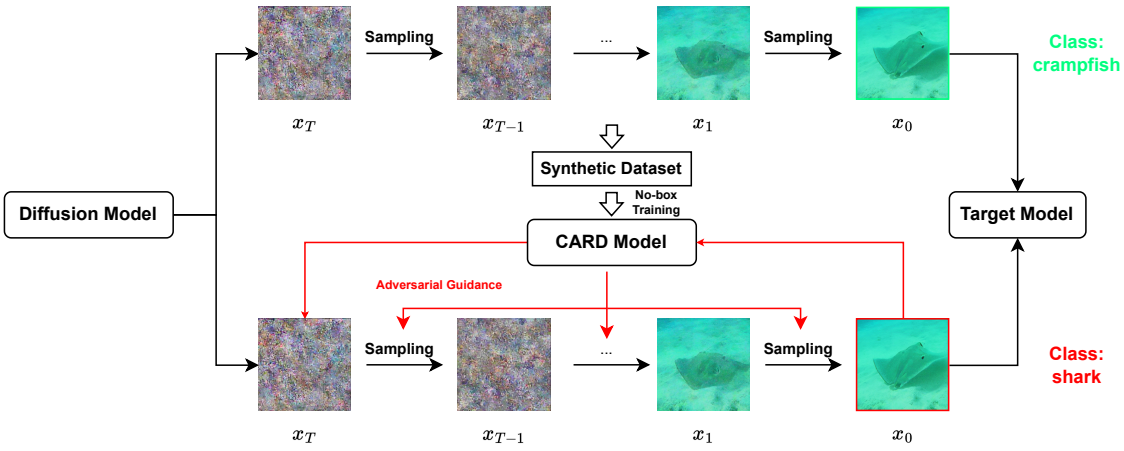


Figure 4.3: **The attack pipeline of our proposed no-box adversarial attacks.** Firstly, we employ the diffusion model to generate the training dataset. This generation is guided by conditional sampling with class information from the original training dataset of the no-box models (Section III.A). Secondly, we train the substitute Classification And Regression Diffusion (CARD) model using the synthetic dataset. A unique fine-tuning mechanism is implemented to enhance the performance of the proposed attack (Section III.B). Finally, we execute the unrestricted adversarial attack against the substitute CARD model using the diffusion model. We leverage adversarial guidance from multiple inferences of the CARD model to sample the image adversarially (Section III.C). Ideally, images from the synthetic training dataset should be accurately classified by the target model, while images with adversarial guidance should mislead the target model, resulting in incorrect classification.

classifiers. However, a crucial question arises: Can we solely rely on generative models for training a classifier? In white-box settings, it is unlikely that the classifier trained solely using generative models will be compatible or optimal. Generative models excel at producing realistic samples, but they may not capture all the complexities and nuances of the real training data that the target classifier has been trained on. However, in real-world scenarios, such as applications concerned with privacy, we cannot always access the training details of the target classifier. This practical limitation inspires our approach of exclusively using data generated from diffusion models to train the substitute model. While we recognize that relying solely on diffusion models for training may not produce a classifier that perfectly mirrors the target classifier in white-box settings, our goal is to devise effective adversarial attacks within the constraints of real-world settings. In these scenarios, accessing the training details of the target classifier is often impractical or even prohibited.

Our work considers two training scenarios based on how the diffusion model is trained for a comprehensive discussion on no-box attacks. For standard no-box setting, we adopt pre-trained class-conditional LDM with public checkpoints. The generation of the training dataset is formulated as follows:

$$D \triangleq \{x \sim p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y)\} \tag{4.3}$$

where y is the label encoding of the generated data.

A strict no-box scenario is that we assume the diffusion model is trained on multiple datasets without any fine-tuning on the training dataset of the target model. In our work, we use Stable Diffusion 2.0 [100], a text-to-image diffusion model available to the public. To construct the training dataset, we utilize the label text from the target model as prompts for text-to-image generation, which is formulated as:

$$\hat{\boldsymbol{\epsilon}}_\theta^{(t)}(x_t|y) = \boldsymbol{\epsilon}_\theta^{(t)}(x_t|\emptyset) + w \cdot (\boldsymbol{\epsilon}_\theta^{(t)}(x_t|\tau_\theta(y)) - \boldsymbol{\epsilon}_\theta^{(t)}(x_t|\emptyset)) \tag{4.4}$$

where the conditional guidance is incorporated with classifier-free guidance [48], and

$\tau_\theta(y)$ is the text prompt.

After obtaining the training dataset generated using diffusion models, we proceed with the standard training of the substitute model. Besides, we also include standard geometric transformations to enhance the performance of the substitute model in the initial training.

4.3.2 Fine-Tuning with Model Uncertainty

Recent works demonstrate that uncertainty learning is beneficial for the decision-making capabilities of deep learning models. Li et al. [66] also found that adopting an approximate Bayesian inference technique to the substitute model can enhance the performance of black-box attacks by a large margin. In the case of no-box attacks, it is crucial to avoid overconfident predictions from the substitute model, which may arise due to its under-fitted training on the synthetic dataset generated by diffusion models. To address this, we propose a fine-tuning method that leverages model uncertainty to enhance the transferability of the substitute model.

Diffusion probabilistic models, such as the Classification And Regression Diffusion (CARD) model proposed by Han et al. [41], provide an effective way to capture model uncertainty through variational inference. The inference for the classification task is formulated as follows:

$$y \sim p_{\text{CARD}}(y_T) \prod_{t=1}^{T} p_{\text{CARD}\theta}(y_{t-1}|y_t, x) \tag{4.5}$$

where $y_{t-1} = \gamma_0 \hat{y}_0 + \gamma_1 y_t + \gamma_2 f_\phi(x) + \sqrt{\tilde{\beta}_t}\boldsymbol{\epsilon}_t$, $y_T \sim \mathcal{N}(f_\phi(x), \mathbf{I})$, γ is the pre-defined hyper-parameter, $\boldsymbol{\epsilon}_t$ is the forward diffusion noise, and f_ϕ is the pre-trained substitute model.

Mathematically, the CARD model adopts the diffusion process to finally predict the probability of k^{th} class by:

$$\Pr(y = k) = \frac{\exp(-(y_0 - 1)_k^2)}{\sum_j \exp(-(y_0 - 1)_j^2)} \tag{4.6}$$

where y_0 is the output of the CARD model, and $(y_0 - 1)_k$ represent the k-th dimension of the y_0 vector.

CARD model performs classification tasks through the classification likelihoods through its probability predictions. Therefore, it is more suitable to calculate the adversarial guidance which also uses log-likelihoods to guide the diffusion model to sample UAEs. The unrestricted adversarial attacks aiming at target label y_a with diffusion models and the CARD model are performed by replacing the original adversarial guidance with:

$$\nabla_{x_t} \log p_f(y_a|x_t) = \nabla_{x_t} \log \frac{\exp(-((y_0|x_t) - 1)_a^2)}{\sum_j \exp(-((y_0|x_t) - 1)_j^2)} \tag{4.7}$$

The original CARD model did not consider the input of noisy images. Therefore, the performance of the CARD model with the input of images from the internal sampling steps of the diffusion model is limited. Data augmentations like noise injection [150] are effective methods to reduce over-fitting and improve the robustness of a deep learning model. As our no-box attack follows the reverse diffusion process to generate adversarial examples, utilizing noise augmentation would further improve the attack performance which makes the substitute model able to classify noisy inputs. Hence, different from simply adding Gaussian noise, our proposed noise augmentation method injects noises from the forward diffusion process. More specifically, the fine-tuning training algorithm is given in Algorithm 2. Noted that we first train a ResNet-50 [43] model as f_ϕ.

The proposed noise augmentation method assists the substitute model in classifying samples from the reverse diffusion process. We only select the last 20% of T steps for sampling the noisy images, as the early sampling steps merely generate noise patterns that are barely recognizable. After fine-tuning the CARD model, we execute our no-box adversarial attacks using the CARD model as the substitute model.

Algorithm 2 Fine-Tuning Training Algorithm

Require: f_ϕ: pre-trained substitute model, x_0: original sampled image without noise, $\bar{\alpha}$: linear noise schedules for CARD, $\hat{\alpha}$: linear noise schedules for LDM, T: reverse generation process timestep for CARD, T_{LDM}: reverse generation process timestep for LDM

1: **repeat**

2: $t_{\text{ft}} \sim \text{Uniform}(\{1 \dots T_{\text{LDM}}\})$

3: Sample x_{noise} with forward process of the LDM model

$$q(x_{\text{noise}}|x_0) = \mathcal{N}(x_{\text{noise}}; \sqrt{\hat{\alpha}_{t_{\text{ft}}}}x_0, (1 - \hat{\alpha}_{t_{\text{ft}}})\mathbf{I})$$

4: $y_0 \sim q(y_0|x_{\text{noise}})$

5: $t \sim \text{Uniform}(\{1 \dots T\})$

6: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

7: Compute noise estimation loss

$$\begin{aligned}\mathcal{L}_\epsilon = \big|\big|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\big(&x_{\text{noise}}, \sqrt{\bar{\alpha}_t}y_0 \\ &+ \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon} + (1 - \sqrt{\bar{\alpha}_t})f_\phi(x_{\text{noise}}), f_\phi(x_{\text{noise}}), t\big)\big|\big|^2\end{aligned}$$

8: Optimization over $\nabla_\theta \mathcal{L}_\epsilon$

9: **until** Convergence

4.3.3 No-box Adversarial Attacks with Diffusion Models

When conducting adversarial attacks against a classification diffusion model, the goal is to find perturbations that can deceive the model's softmax output, resulting in misclassification. This process is similar to standard adversarial attacks, where the objective is to find small perturbations that can fool the model's decision-making process. Under the no-box attack scenario, it is more practical that we utilize the generative diffusion model that constructs the training data to conduct unrestricted adversarial attacks against the substitute diffusion model. The no-box adversarial attack with diffusion models samples the adversarial examples with the guidance of the gradient from the substitute diffusion model, which is formulated as Equation 4.1 and 4.2.

As the classification diffusion model can also be viewed as an approach to model $p(y|x)$, we can approximate the exact inference by adopting the Monte Carlo sampling method $p(y|x) = \frac{1}{M}\sum_{i=1}^{M} p(y_i|x)$, where $p(y_i|x)$ is obtained by multiple sampling. We select the ground truth class for the no-box adversarial attack in this chapter.

Algorithm 3 No-Box Adversarial Attack Algorithm

Require: f_{CARD}: pre-trained CARD model, y_{gt}: ground truth class label, N: noise sampling guidance steps, T_{LDM}: reverse generation process timestep for LDM, T_{adv}: timestep for adversarial guidance

$x_{T_{\text{LDM}}} \sim \mathcal{N}(0, \mathbf{I})$
$x_{adv} = \varnothing$
$y_0 = \text{OneHotEnc}(y_{\text{gt}})$
for $i = 1 \dots N$ **do**
    for $t = T_{\text{LDM}}, \dots, 1$ **do**
        if t is in T_{adv} **then**
            Obtain adversarial guidance with Equation 4.8
            Sample x_{t-1} with Equation 4.1
        else
            Sample x_{t-1} with Equation 3.9
        end if
    end for
    Obtain adversarial guidance with Equation 4.8.
    Update $x_{T_{\text{LDM}}}$ with Equation 4.2
    $x_{adv} \leftarrow x_0$ if $f_{\text{CARD}}(x_0) \neq y_{\text{gt}}$
end for
return x_{adv}

The proposed no-box unrestricted adversarial attack is achieved with the ensemble of multiple inferences. Detailed attack algorithm is given in Algorithm 3. We use DDIM with classifier-free guidance from the ground truth label y_{gt} for diffusion sampling.

$$\log p_{\hat{f}}(y_a|x_t) = -\frac{1}{M}\sum_{i=1}^{M} \log \frac{\exp(-((y_0|x_t, \epsilon_i) - 1)^2_{\text{gt}})}{\sum_j \exp(-((y_0|x_t, \epsilon_i) - 1)^2_j)} \tag{4.8}$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \mathbf{I})$.

The multiple inferences are accomplished by M independent classification results from the trained CARD model, with M instances of random initial diffusion noise $\boldsymbol{\epsilon}$. The proposed attack achieves an ensemble-like adversarial attack, leveraging the characteristics of diffusion models without the need to train multiple substitute models.

4.4 Experiments

Datasets and Substitute Models. We use the ImageNet [23] dataset for major evaluation. Under no-box settings, the dataset is generated by diffusion models with 224×224 pixels. The substitute model is ResNet-50 [43]. Under black-box settings, the dataset is from the ImageNet 2012 validation dataset. The base model for generating the adversarial examples is also ResNet-50 [43]. Following existing previous work [38,66], we adopt the common settings for transfer-based adversarial attacks. We randomly sample 5 data from each class of the training dataset (no-box) or validation set (black-box) to conduct the adversarial attack for baselines.

Parameter Settings. LDM and Stable Diffusion v2.0 [100] are selected as source model in our work. While sampling data, the timestep for the diffusion process is set to 200 for LDM and 50 for Stable Diffusion. Both diffusion models' η are set to 0 for deterministic sampling. The classifier-free guidance scale w is set to 3.0 for LDM and 9.0 for Stable Diffusion. We use public checkpoints from the official release for LDM and Stable Diffusion. In fine-tuning with CARD [41], the diffusion timestep for classification is set to 100. The linear noise schedules are set accordingly as the official implementation. In generating the adversarial examples, we set $N = 5$, $a_1 = 0.5$, $a_2 = 0.5$ for adversarial guidance, and $M = 10$ for multiple inferences. We use DMSA_{LDM} to denote the proposed attack with LDM implementation and DMSA_{SD} for Stable Diffusion implementation. The basic attack is performed by replacing the f_{CARD} with ResNet-50 trained by our synthetic dataset in Algorithm 3.

Combining with Perturbation-Based Attacks. Our method aims to generate high-quality non-perturbed adversarial examples with the benign diffusion process. In other words, the sampled adversarial examples can be treated as benign images. Therefore, it is possible to interrogate the perturbation-based adversarial attacks with our generated adversarial examples. Besides using the diffusion model for generating adversarial examples, we perform 200 steps I-FGSM over the generated examples to enhance their performance on no-box models.

Target Models. We select various widely adopted target models for ImageNet

Figure 4.4: **Comparisons of no-box adversarial examples with our method and Li et al.'s method.** Note that our method achieves a similar ASR with a significantly lower perturbation. The adversarial examples are generated by latent inversion.

to test the attack performance: ResNet-50 [43], VGG-19 [107], ResNet-152 [43], Inception v3 [113], DenseNet-121 [50], MobileNet v2 [104], SENet-154 [49], ResNeXt-101 [133], WRN-101 [139], PNASNet [77], and MNASNet [115]. Because these models take different resolution inputs, we adopt different re-scale functions before performing the adversarial attack according to their original implementation. We adopt the public checkpoints from the timm [124] library for these models.

Evaluation. We mainly use ASR to evaluate the performance of various adversarial attack methods. As our attack is aimed at attacking no-box and black-box models, the attack success rate is calculated by how many transfer-based adversarial examples can fool the target model. The adversarial examples are first generated by attacking the substitute model training by the adversary. Then, we adopt the generated adversarial examples to check if they can be misclassified by the no-box or black-box target model. We do not perform any query to the target model to fine-tune the adversarial examples. To ensure fair comparisons, we evaluate our attacks by using latent inversion [100] from images in the validation set to generate the adversarial examples when compared with previous attacks (The process of training and fine-tuning the substitute model is unchanged).

4.4.1 No-Box Threat Model

Under the no-box threat model, we compare our method with the state-of-the-art method [65], where they used supervised ResNets and unsupervised auto-encoders for no-box attacks (Naïve† and Prototypical). It's important to note that Li et al.'s method uses 20 images from the original training dataset to train the substitute model. As a result, their method only supports attacking a limited number of images and requires a relatively large perturbation under $\delta = 0.1$. Furthermore, their method does not support attacking all classes simultaneously, which further restricts the applicability of their method. For a fair comparison with their works, we use the latent inversion [100] from images in the validation set to generate our adversarial examples. The performance of the state-of-the-art methods and our proposed methods are displayed in Table 4.1 and Figure 4.4.

For Li et al.'s attack, they achieve around 68% ASR on different no-box target models with perturbations of $\delta = 0.1$. Their best performance is achieved with 20 decoders Prototypical*, 200 steps I-FGSM, and 100 steps ILA attack. Furthermore, using an auto-encoder over clean images brings visual quality problems as the reconstructed images can be identified by humans. Training of the Prototypical* requires random selections of the required training images, which limits the reproducibility of their methods. Therefore, their method is very limited in usability.

On the other hand, our method achieves the state-of-the-art attack success rate without adding I-FGSM gradients. This result indicates that diffusion models, coupled with our proposed attack methods, are more formidable adversaries with superior performance than traditional perturbation-based methods. Adopting the CARD model can notably improve the transfer ASR by around 15% without largely increasing the magnitude of the perturbation. This is more effective than adopting 20 decoders in Li et al.'s work with only a 2% increase in ASR. Moreover, Figure 4.4 further demonstrates that without using I-FGSM, the visual quality of the proposed method is noticeably better than Li et al.'s work. Note that both DMSA_{LDM} and DMSA_{SD} use synthetic dataset rather than validation set from the no-box target

model to train the substitute model. Therefore, the performance of proposed attacks can be further improved by using the substitute model from Li et al.'s work.

Furthermore, we test the performance of proposed attacks using the randomly sampled latents. Table 4.2 demonstrates that the proposed attack methods significantly outperform the previous attacks. The substitute model for the previous attacks is trained with the same dataset as DMSA_{LDM}. By only using the synthetic dataset generated by diffusion models, the state-of-the-art attack methods perform much worse than our methods. These findings further indicate the effectiveness of the proposed attacks. The proposed attack with Stable Diffusion performs slightly worse than the LDM settings. This could be due to the different training data of the original model. Stable Diffusion utilizes multiple large-scale datasets for training. Consequently, the data distribution of the no-box training dataset generated by Stable Diffusion is likely to be inconsistent with the original ImageNet dataset. As a result, the adversarial examples from the trained substitute model may struggle to transfer to the no-box target model.

Figure 4.4 demonstrates that the benign images generated by diffusion models attain a classification accuracy similar to the standard ImageNet validation dataset, which attests to the generation quality of our method. Moreover, the basic and CARD attacks achieve over 50% ASR with significantly less noticeable perturbations compared to Li et al.'s method.

4.4.2 Image Quality

We evaluate the image quality of no-box adversarial examples using the FID score [44]. As shown in Table 4.1, our proposed attacks achieve significantly better image quality compared to those by Li et al., demonstrating that diffusion models can serve as potent adversaries to deep learning models. However, it is important to note that adversarial guidance can negatively impact the original generation quality of the diffusion model. Since adversarial guidance is integrated based on the benign diffusion guidance of the model, it should be within the range of [0, 1.0]. For more stable and high-quality image generation, we recommend setting a_1 and a_2 to values smaller than 0.5 for no-

Table 4.1: **Attack success rates of transfer-based no-box attacks on ImageNet with ResNet-50 as the substitute model, the perturbation of baseline is ℓ_∞ with $\delta = 0.1$.** We use latent inversion from the data of the baseline to generate our adversarial examples.

Method	VGG-19	Inception v3	ResNet-152	DenseNet	SENet	WRN	PNASNet	MobileNet	Average	FID (↓)
Naïve†	23.80%	19.14%	16.24%	21.06%	13.00%	15.84%	13.04%	27.56%	18.71%	**10.2**
Prototypical	80.22%	63.54%	62.08%	70.84%	55.44%	62.72%	51.42%	82.22%	66.06%	77.8
Prototypical*	81.26%	66.32%	65.28%	73.94%	57.64%	66.86%	54.98%	83.66%	68.74%	85.4
DMSA_{LDM}	65.72%	53.15%	60.77%	71.44%	45.74%	63.25%	45.53%	75.27%	60.10%	15.6
+ CARD	**82.11%**	**68.62%**	**78.74%**	**81.81%**	**61.26%**	**77.29%**	**60.18%**	**89.54%**	**74.94%**	26.8
DMSA_{SD}	58.57%	49.62%	62.31%	64.68%	42.78%	46.96%	41.53%	66.13%	54.07%	13.1
+ CARD	74.31%	62.85%	78.26%	78.62%	59.47%	60.02%	57.66%	79.89%	68.89%	24.4

Table 4.2: **Attack success rates of transfer-based no-box attacks on ImageNet with ResNet-50 as the substitute model, the perturbation of baseline is ℓ_∞ with $\delta = 0.1$.** We use the generated images from the LDM model as the clean data for the previous attacks.

Method	VGG-19	Inception v3	ResNet-152	DenseNet	SENet	WRN	PNASNet	MobileNet	Average
I-FGSM	24.17%	20.87%	19.67%	21.37%	18.97%	20.47%	18.47%	25.73%	21.22%
ILA++ (2022)	40.51%	22.65%	31.51%	26.03%	26.81%	36.33%	29.05%	48.27%	32.65%
MBA (2023)	45.67%	32.93%	34.43%	41.63%	33.37%	38.03%	32.90%	53.80%	39.10%
DMSA_{LDM}	52.95%	30.41%	36.06%	39.67%	29.37%	40.37%	26.89%	50.22%	38.24%
+ CARD	59.60%	38.70%	55.30%	59.48%	37.75%	57.08%	36.49%	72.16%	52.07%
+ CARD I-FGSM	**93.98%**	**81.23%**	**87.54%**	**91.47%**	**83.15%**	**87.58%**	**79.18%**	**96.15%**	**87.53%**
DMSA_{SD}	35.80%	36.70%	34.29%	35.13%	37.21%	34.04%	32.21%	39.17%	35.61%
+ CARD	54.40%	45.78%	53.52%	50.21%	53.87%	46.32%	38.05%	56.32%	49.80%
+ CARD I-FGSM	80.53%	66.48%	68.98%	74.06%	68.32%	77.52%	58.01%	86.30%	72.53%

box adversarial attacks. Additionally, employing stronger diffusion models, such as Stable Diffusion, can enhance generation quality. Utilizing the CARD ensemble attack is also more effective in improving the ASR than merely increasing the adversarial guidance.

4.4.3 Black-Box Threat Model

For a comprehensive discussion on the adversarial ability of diffusion models, we perform standard black-box adversarial attacks with the proposed attack. A variety of state-of-the-art black-box adversarial attacks are selected as comparisons, including LinBP [38], ILA++ [39], TAIG [52] and LGV [36], TIM [27], SIM [75], Admix [121] and MBA [66] with the ℓ_∞ attack budget $\delta = 8/255$. We also include more black-box target networks for complete comparisons.

Prior to the year of 2022, previous works achieved relatively lower performance

than white-box attacks because black-box attacks could not access the gradient of the target model. As a result, existing methods tend to enhance attack transferability by better inferring the gradient of the black-box model with the substitute model. The MBA attack by Li et al. [66] significantly improves the ASR of black-box attacks by using an ensemble-like approach with Bayesian fine-tuning. However, their method necessitates re-training the standard substitute model, and the ensemble attack further restricts the efficiency of their attacks.

For our proposed attack, we utilize the LDM as the base model. We use the latent inversion [100] from images in the validation set to generate our adversarial examples. The standard pre-trained ResNet-50 is adopted as the substitute model. Table 4.3 shows that under standard settings, the proposed adversarial attack with the diffusion model already achieves an 89% ASR without any fine-tuning. This result demonstrates that diffusion models have the potential to execute stronger and more concealed black-box adversarial attacks than traditional perturbation-based attacks. When adopting the CARD model for black-box adversarial attack, the proposed attack outperforms the state-of-the-art attack methods without adding gradient-based perturbations. The attack performance of our proposed method can be further enhanced by combining it with simple perturbation-based attacks, which are similar to the no-box attack settings.

It's worth noting that most of the black-box target networks employ similar convolution blocks for feature learning, which results in low robust accuracy against transfer adversarial examples from the ResNet-50 networks. These networks also use the same image transformations before feeding the input to the networks. Consequently, we observe a significant drop in ASR on target networks with different network structures and image transformations, such as Inception v3 and PNASNet. This insight could contribute to the design of better defenses against both perturbation-based adversarial attacks and diffusion-model-based adversarial attacks.

Table 4.3: **Attack success rates of transfer-based black-box attacks on ImagetNet with ResNet-50 as the substitute model, the perturbation is ℓ_∞ with $\delta = 8/255$.** We use latent inversion from the data of the baseline to generate our adversarial examples.

Method	ResNet-50	VGG-19	ResNet-152	Inception v3	DenseNet	MobileNet
I-FGSM	**100.00%**	39.22%	29.18%	15.60%	35.58%	37.90%
TIM (2019)	**100.00%**	44.98%	35.14%	22.21%	46.19%	42.67%
SIM (2020)	**100.00%**	53.30%	46.80%	27.04%	54.16%	52.54%
LinBP (2020)	**100.00%**	72.00%	58.62%	29.98%	63.70%	64.08%
Admix (2021)	**100.00%**	57.95%	45.82%	23.59%	52.00%	55.36%
TAIG (2022)	**100.00%**	54.32%	45.32%	28.52%	53.34%	55.18%
ILA++ (2022)	99.96%	74.94%	69.64%	41.56%	71.28%	71.84%
LGV (2022)	**100.00%**	89.02%	80.38%	45.76%	88.20%	87.18%
MBA (2023)	**100.00%**	97.79%	97.13%	73.12%	98.02%	97.49%
DMSA_{LDM}	**100.00%**	93.95%	94.26%	77.04%	94.57%	97.91%
DMSA_{LDM} + CARD	**100.00%**	**98.02%**	**98.45%**	**84.21%**	**98.12%**	**99.33%**

Method	SENet	ResNeXt	WRN	PNASNet	MNASNet	Average
I-FGSM	17.66%	26.18%	27.18%	12.80%	35.58%	27.69%
TIM (2019)	22.47%	32.11%	33.26%	21.09%	39.85%	34.00%
SIM (2020)	27.04%	41.28%	42.66%	21.74%	50.36%	41.69%
LinBP (2020)	41.02%	51.02%	54.16%	29.72%	62.18%	52.65%
Admix (2021)	30.28%	41.94%	42.78%	21.91%	52.32%	42.40%
TAIG (2022)	24.82%	38.36%	42.16%	17.20%	54.90%	41.41%
ILA++ (2022)	53.12%	65.92%	65.64%	44.56%	70.40%	62.89%
LGV (2022)	54.82%	71.22%	75.14%	46.50%	84.58%	72.28%
MBA (2023)	85.41%	94.16%	95.39%	77.60%	**97.15%**	91.33%
DMSA_{LDM}	79.33%	89.77%	94.05%	78.18%	95.82%	89.49%
DMSA_{LDM} + CARD	**88.21%**	**95.25%**	**97.56%**	**85.71%**	96.73%	**94.16%**

4.4.4 Adversarial Robust Models and Vision Transformers

It has been reported that adversarial defense methods like adversarial training can effectively improve the adversarial robustness of deep learning models. It is practical to test the performance of adversarial attacks under defenses to test the performance on real-world scenarios. We test the performance of various attack methods against adversarial robust models using latent inversion, including adversarial-trained Inception v3, EfficientNet-B0, ResNet-50, a robust DeiT-S [116], and a diffusion-based adversarial purification method DiffPure [89]. Checkpoints from Inception v3, EfficientNet-B0, ResNet-50 follows [66]. Table 4.4 shows that adversarial training is effective at defend-

ing black-box adversarial attacks, especially for adversarial-trained ResNet-50 which successfully defends around 90% of the state-of-the-art black-box attack methods. Our proposed attack does not directly add the adversarial gradient to the generated adversarial examples. Therefore, our methods remarkably outperform perturbation-based black-box attacks on various adversarial-trained and denoising deep-learning models.

Vision transformers are recent transformer-based models with state-of-the-art performance but different network artifacts. They achieve relatively high robust accuracy under adversarial attacks for their special feature learning techniques. We also test the attack performance of adversarial examples with recent vision transformers using latent inversion, i.e., ViT-B [28], a DeiT-B [116], a Swin-B [79], and a BEiT [5]. Table 4.4 demonstrates that perturbation-based adversarial examples hardly transfer to vision transformers for adversarial attacks. As vision transformers adopt special patch embedding for feature learning, the adversarial perturbations are very likely to be sabotaged during patching. Therefore, vision transformers are robust to perturbation-based attacks even without any defenses. However, our proposed attacks seek adversarial examples by adversarial sampling, which generates UAEs with adversarial global features rather than special perturbation patterns. Hence, the proposed attacks achieve overwhelmingly better performance against vision transformers than previous methods.

The adversarial examples sampled by diffusion models are more effective at deceiving defense methods and vision transformers due to their adversarial sampling with the diffusion process, rather than simply adding noise patterns to the image. This presents significant challenges to current deep-learning applications and underscores the need for effective designs of adversarial defense methods.

We also compare our attacks with the state-of-the-art diffusion model based black-box attacks, DiffAttack [13]. The results are given in Table 4.5. Our attack significantly outperforms DiffAttack on our Basic attack in both ASR and FID score [44] for generation quality on most black-box target models (6 out of 8).

Table 4.4: **Attack success rates of transfer-based black-box attacks on ImagetNet against robust models and vision transformers with ResNet-50 as the substitute model, the perturbation is ℓ_∞ with $\delta = 8/255$.** We use latent inversion from the data of the baseline to generate our adversarial examples.

Method	Vision transformers				Robust models				
	ViT-B	DeiT-B	Swin-B	BEiT	Inception v3	EfficientNet	ResNet-50	DeiT-S	DiffPure
I-FGSM	4.70%	5.92%	5.18%	3.64%	11.94%	9.48%	9.26%	10.68%	6.65%
ILA++ (2022)	9.48%	21.34%	14.88%	11.76%	15.54%	30.90%	10.08%	11.08%	7.25%
LGV (2022)	7.18%	20.02%	12.14%	11.66%	18.00%	39.06%	10.56%	11.50%	8.10%
MBA (2023)	21.66%	43.53%	21.84%	29.78%	25.89%	67.05%	11.02%	12.02%	9.45%
DMSA_{LDM}	49.33%	56.21%	52.85%	83.24%	53.31%	88.55%	75.91%	63.21%	74.22%
DMSA_{LDM} + CARD	**61.68%**	**63.36%**	**59.74%**	**88.52%**	**67.67%**	**93.20%**	**82.25%**	**71.08%**	**78.30%**

Table 4.5: **Attack success rates of transfer-based black-box attacks on ImagetNet comparing with DiffAttack, the perturbation is ℓ_∞ with $\delta = 8/255$.** FID is evaluated on our selected ImageNet validation data. We use latent inversion from the data of the baseline to generate our adversarial examples.

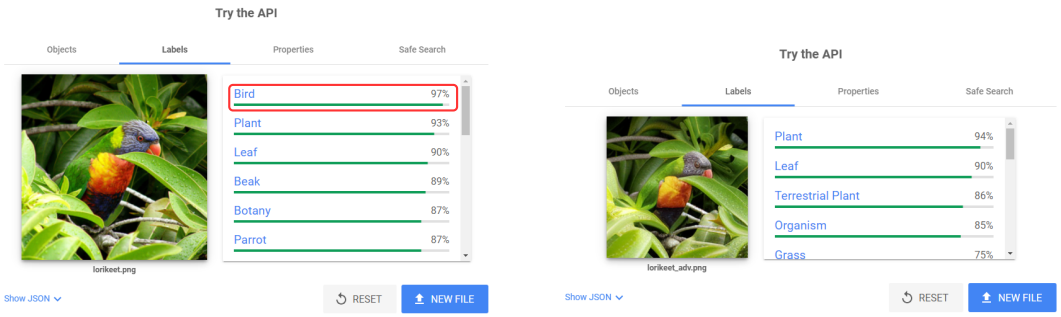
Method	CNNs				Vision transformers				FID (↓)
	ResNet-50	VGG-19	MobileNet	Inception v3	ViT-B	Swin-B	DeiT-B	DeiT-S	
DiffAttack (2023)	96.3%	75.6%	77.1%	69.0%	51.2%	56.2%	50.5%	55.0%	*25.2*
DMSA_{LDM}	100.00%	93.95%	97.91%	77.04%	49.33%	52.85%	56.21%	64.52%	**16.4**
DMSA_{LDM} + CARD	**100.00%**	**98.02%**	**97.97%**	**84.21%**	**61.68%**	**59.74%**	**63.36%**	**69.91%**	25.9

4.4.5 Time Efficiency

The proposed attack model is trained and evaluated on a Nvidia GeForce RTX 3090 GPU. We demonstrate the time efficiency of each component to perform our attacks in Table 4.6. Since the CARD model requires only a single training session, the training cost of the proposed methods can be considered negligible. However, the time efficiency in generating the no-box adversarial image is relatively low due to the use of the CARD model and multiple inferences. Nonetheless, adopting the CARD model significantly increases the attack success rate, presenting a tradeoff between ASR and time efficiency. We can reduce the time cost by adopting time-efficient diffusion substitute models and diffusion models, which we plan to explore in future work.

Table 4.6: **Time cost (s) of proposed DMSA attacks in training and attacking process.**

Method	Training Dataset (per image)	Training CARD	Fine-tuning CARD	Adversarial Attack (per image)	Adversarial Attack with CARD (per image)
$DMSA_{LDM}$	6.4	9251.5	5428.8	11.3	41.5
$DMSA_{SD}$	4.9	9214.2	4412.1	9.6	29.7



Clean Image | **$DMSA_{LDM}$ CARD Attack**

Figure 4.5: **A successful attack against Google Vision.** The confidence level for "bird" is reduced, causing it to drop out of the top three labels.

4.4.6 Attacking Commercial CNNs

A practical scenario for our proposed no-box adversarial attacks involves targeting commercial CNNs, such as Google Vision. To further validate the effectiveness of our methods, we randomly selected 100 images from the no-box adversarial examples from $DMSA_{LDM}$ to test the attack success rate against Google Vision. An attack is considered successful if it reduces the confidence of the correct label out of the top three labels. An example of a successful attack is shown in Figure 4.5. Out of the 100 images, 86 successfully deceived Google Vision, demonstrating that no-box adversarial attacks pose a significant threat to deep learning models.

Google Vision is a multi-label classification model capable of detecting over 1,000 ImageNet classes. The results suggest that the proposed attack can successfully deceive the no-box target model with only practical knowledge of the label information, which we plan to explore further in future work.

4.5 Ablation Studies

The proposed attacks contain three complicated processes to perform the no-box attack. We give comprehensive ablation studies on each important process that contributes to the attack performance in this section. We select the LDM model with the CARD substitute model for major experiments. We use the no-box threat model to conduct the adversarial attacks using random latents. For clarity, we only cover a part of the target model to test the attack performance. By default, the attack success rate is the average ASR over the no-box 8 target models.

4.5.1 Training Dataset

In this section, we delve into the crucial role that the quality of the substitute model plays in the performance of transfer-based adversarial attacks, particularly in the context of no-box adversarial attacks. The impact of the scale of the training dataset on the proposed attacks is explored by adopting four different quantities of images per class to construct the training dataset, with the results summarized in Table 4.7. The outcomes clearly illustrate that a larger training dataset significantly enhances both clean accuracy and attack transferability. Notably, when utilizing a dataset with only 100 images per class, the substitute model tends to be under-fitted, resulting in the poorest performance compared to models trained on larger datasets. However, the ASR did not largely increase when the training dataset was set to 2000 images per class. The reason may be the over-fitting of the substitute model. Remarkably, even in the absence of real data from the original training dataset, our proposed substitute model achieves an impressive approximately 80% top-5 classification accuracy on the validation dataset of ImageNet. This outcome underscores the efficacy of our novel training method.

4.5.2 CARD Model

The CARD model is a diffusion model, which may cause large computation overheads to the attack algorithm. In this section, we investigate the time efficiency and attack

Table 4.7: **Attack success rates of transfer-based no-box attacks on ImageNet with ResNet-50 as the substitute model in terms of the scale of the training dataset.** n represents the scale of images per class. Substitute model classification accuracy on the ImageNet validation set is further evaluated.

Method	VGG-19	Inception v3	ResNet-152	DenseNet	SENet	WRN	PNASNet	MobileNet	Average	Clean Top-5 Acc
DMSA_{LDM} $n = 1000$	59.60%	38.70%	55.30%	59.48%	37.75%	57.08%	36.49%	72.16%	52.07%	79.85%
DMSA_{LDM} $n = 100$	50.32%	28.41%	35.05%	40.98%	24.30%	37.87%	20.71%	55.30%	36.61%	58.32%
DMSA_{LDM} $n = 500$	57.14%	33.19%	43.31%	48.83%	32.91%	46.84%	29.64%	64.75%	44.57%	68.50%
DMSA_{LDM} $n = 2000$	**62.10%**	**39.21%**	**56.21%**	**60.32%**	**40.80%**	**58.45%**	**38.01%**	**75.50%**	**53.82%**	**80.65%**

Table 4.8: **Attack success rates of transfer-based no-box attacks on ImageNet with ResNet-50 as the substitute model in terms of the fine-tuning for the CARD model.**

Method	VGG-19	Inception v3	ResNet-152	DenseNet	SENet	WRN	PNASNet	MobileNet	Average
DMSA_{LDM} w/o fine-tuning	54.84%	32.83%	50.00%	53.96%	28.06%	50.38%	25.79%	60.85%	44.59%
DMSA_{LDM}	59.60%	38.70%	55.30%	59.48%	37.75%	57.08%	36.49%	72.16%	52.07%

performance of the CARD model under different numbers of diffusion timesteps. Figure 4.6 shows the results that larger timesteps for the CARD model will cause a significant increase in average time to generate one adversarial example. However, the ASR does not notably increase after the settings of $|T| = 100$. Furthermore, we test the power of fine-tuning in Table 4.8. With the proposed fine-tuning, the ASR of the proposed attack is improved by 8% on average. At the same time, the proposed fine-tuning does not require additional computation or lower the generation quality of the generated adversarial examples.

4.5.3 Model Uncertainty

The proposed adversarial attack method exhibits superior performance compared to state-of-the-art methods by leveraging model uncertainty. However, the utilization of multiple inferences introduces additional computational demands. In this section, we assess the transferability of generated adversarial examples and analyze the time complexity of the attack algorithm across varying numbers of inferences. The results, illustrated in Figure 4.7, indicate that a higher number of inferences can enhance the attack performance of the proposed method. Nonetheless, this comes at the cost of significantly slowing down the attack speed of the diffusion model. The figure depicts a clear trade-off: while an increased number of inferences improves attack

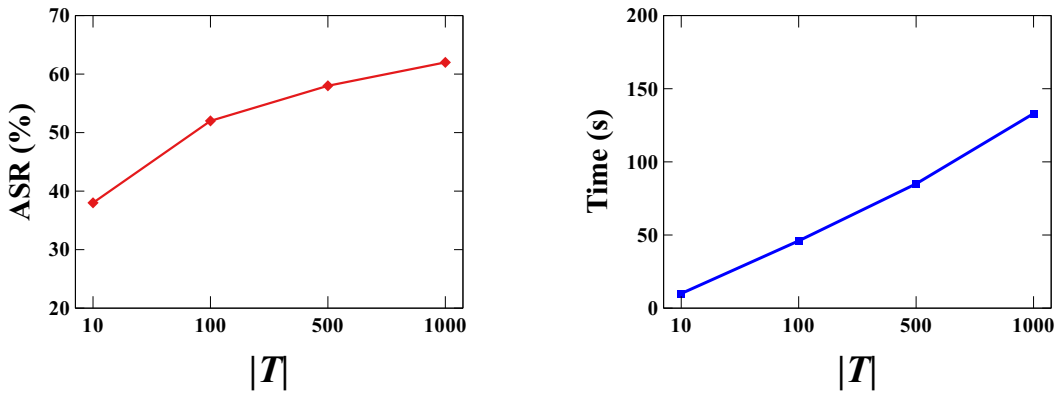


Figure 4.6: **The performance of our proposed attacks under different settings of diffusion timesteps for the CARD model.** Time represents the average time to generate one UAE.

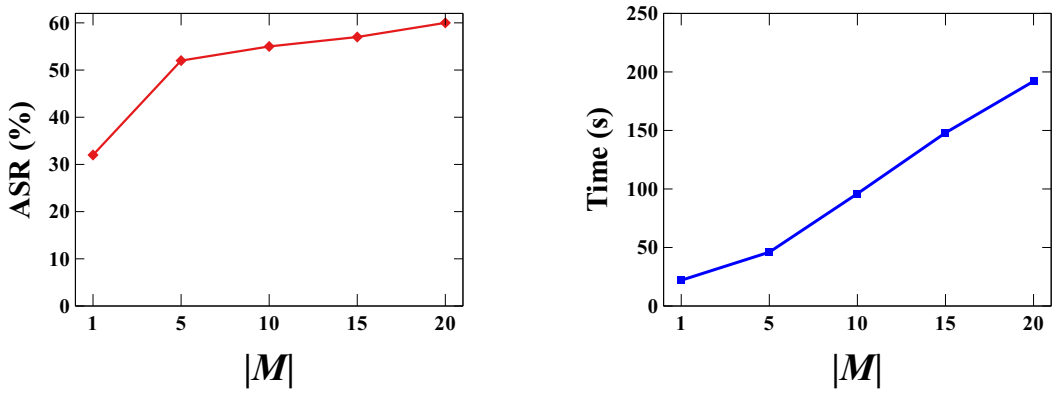


Figure 4.7: **The performance of our proposed attacks under different numbers of inferences from the substitute CARD model.**

performance, it concurrently imposes a notable delay on the execution speed of the diffusion model. Notably, the CARD model employed in this section utilizes 100 timesteps for classification. Importantly, the findings reveal that despite the computational overhead, employing multiple inferences substantially boosts the attack transferability compared to relying on a single deterministic substitute model. This trade-off underscores the importance of carefully considering the computational resources available and the desired balance between attack speed and transferability when implementing the proposed adversarial attack method.

4.5.4 Adversarial Guidance

In this section, we systematically evaluate the performance of our proposed attacks across different settings of a_1 and a_2. Intuitively, one would expect the attack success rate to increase as a_1 and a_2 are set to relatively large values. However, a critical trade-off exists, while higher values of a_1 and a_2 may enhance the attack success rate, they can simultaneously lead to a decrease in generation quality. To quantify this, we assess the generation quality using the Frechet Inception Distance (FID) score [44]. Figure 4.8 illustrates a significant decrease in the FID score as adversarial guidance, represented by a_1 and a_2, increases. Concurrently, the ASR surpasses 80%. Even with an 80% ASR, the FID score of our proposed attack still outperforms the PGD attack with $\delta = 8/255$. For a more visual understanding, we provide a comparison of adversarial examples generated by our attacks and the PGD attack in Figure 4.9. Notably, our attacks tend to produce distinctive textures to deceive the target network. It is noteworthy that the adversarial examples generated by our attack maintain a natural and realistic appearance, especially when a_1 and a_2 are set to relatively small values. The a_2 tends to generate unrealistic examples when set to a larger value. The reason could be modification to the original x_T disturbs the distribution of the initial latent and hence decreases the generation quality. This observation underscores the nuanced balance between maximizing attack success and preserving the visual coherence of the generated adversarial examples.

4.6 Discussion

Experiment results show that even without training data from the target model, our method can achieve state-of-the-art ASR under the no-box threat model. Note that the above 95% benign sampled images from the diffusion model can be correctly classified by the target model. However, the basic attack of our method's ASR is relatively lower than the baseline. The reason may be the different data of our proposed attack. Because the data generated by the diffusion model are not from the standard validation set of the training data, they may perform worse on the transferability.

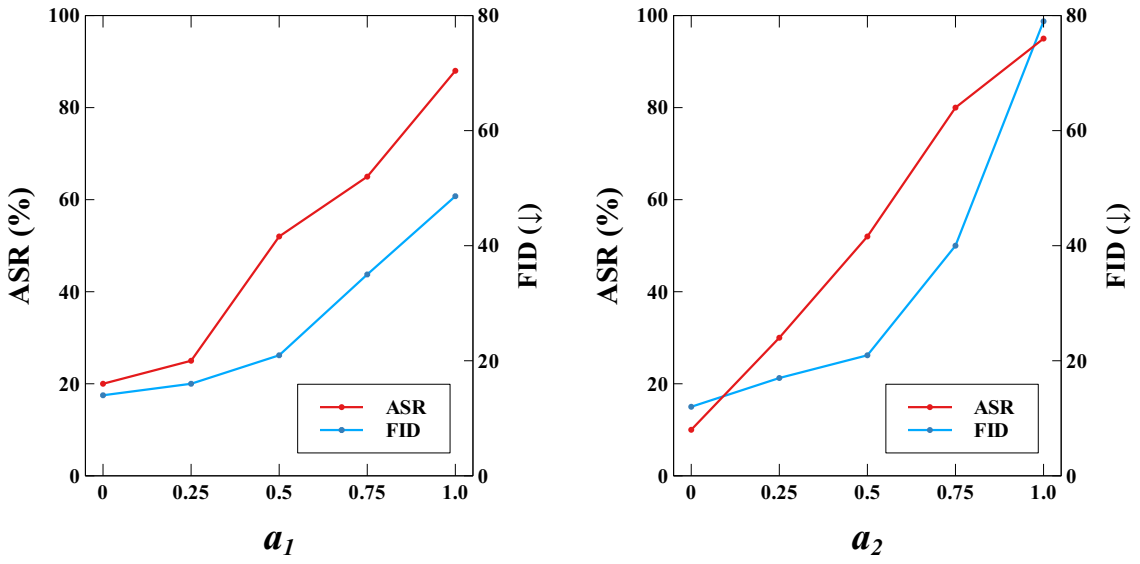


Figure 4.8: **The performance of our proposed attacks under different settings of a_1 and a_2 for adversarial guidance.**

Figure 4.9: **The visual comparison of different no-box adversarial examples.** Details are zoomed in for better comparisons. The perturbations of UAEs are more camouflaged.

Therefore, it is better we use some data from the original training dataset to enhance the attack performance. Moreover, our adversarial examples exhibit overwhelming performance against robust models and vision transformers. This emphasizes the imperative need for designing effective defense mechanisms. The discussion on defending against adversarial attacks by diffusion models is crucial, as these models introduce a potent and novel form of adversarial attack, posing new challenges to the enhancement of deep learning models' robustness.

4.7 Ethic Concerns

The chapter is driven by the comprehensive evaluation of the adversarial capabilities of diffusion models within the context of the no-box attack scenario. Capitalizing on the robust generation prowess of diffusion models, we demonstrate their capacity to generate adversarial examples without necessitating access to the training dataset of the target model. Notably, the motivation arises from the realization that current defense methods focus on fortifying defenses against perturbation-based attacks. Unfortunately, these defenses exhibit bad performance when defended with adversarial examples generated by diffusion models and fare even worse in the face of a combination of both perturbation-based and diffusion attacks. In light of these challenges, we advocate for the development of effective defense mechanisms specifically tailored to counter adversarial diffusion models. The overarching goal is to augment the usability and robustness of deep learning models, acknowledging the evolving threat landscape posed by advanced attack methodologies such as those involving diffusion models.

4.8 Weakness

While the proposed attack achieves state-of-the-art performance in the no-box adversarial attack setting, it does face a limitation: the generated images may appear visually unrealistic when compared to the adversarial examples produced under the black-box scenario. This discrepancy arises due to the potential impact of adversar-

ial guidance on the normal diffusion process, especially when setting $\nabla_{x_t} \log f(y_a|x_t)$ as $-\nabla x_t \log f(y_{\text{gt}}|x_t)$. Furthermore, it is acknowledged that the proposed method for generating the training dataset still exhibits gaps in comparison to the original dataset. As a consequence, there is room for improvement in the performance of the proposed attack. This improvement may be achieved through refining adversarial guidance and adopting more effective methods for generating datasets. Enhancing the time efficiency of the diffusion model can further improve the usability of the proposed attacks.

4.9 Conclusion

In this chapter, we investigate the attack ability of diffusion models as strong adversaries. Our attacks offer a novel solution to no-box adversarial attacks without requiring access to the entire dataset of the no-box target model. Additionally, our work is pioneering in incorporating diffusion models as substitute models for adversarial attacks. Specifically, we first train the substitute model with the data generated by the diffusion models with label priors from the original training dataset. To further fine-tune the performance of the substitute model, we adopt the classification diffusion probabilistic model to obtain the inference for the classification task. We introduce noise augmentation during the training of the substitute model. After training the substitute model, the adversarial examples are generated by the diffusion model with an ensemble-like attack over the multiple inferences from the classification diffusion substitute model. Extensive experiments on the ImageNet dataset have demonstrated the performance of the proposed attack. We show the strong adversarial ability of diffusion models even without any data or information from the target model. Our work urges effective defense mechanisms against adversarial examples generated by diffusion models.

Chapter 5

Transferable 3D Adversarial Shape Completion using Diffusion Models

5.1 Introduction

Deep-learning models have demonstrated their overwhelming performance on 2D [43, 79] and 3D computer vision [37, 99, 130] tasks. An increasing number of applications rely on deep-learning models to achieve efficient and accurate services. Therefore, the security of deep-learning models is crucial and significant.

Similar to the 2D scenario [10, 19, 68, 71, 84], 3D point cloud deep learning is also susceptible to adversarial attacks [78, 129, 146]. These 3D adversarial attacks generate adversarial examples by introducing perturbations to the xyz coordinates. However, such perturbations often lead to a significant degradation in visual quality, which can be easily detected by humans. Subsequent studies [51, 123, 145] have aimed to create less perceptible perturbations by taking into account geometric characteristics. Despite this, these attacks have been shown to perform poorly against defenses [53]. Moreover, most existing attacks primarily focus on white-box settings, limiting their practicality in real-world scenarios. Existing black-box attacks [40, 42] mainly target early 3D point cloud deep-learning models, leaving a substantial gap in the learning between adversarial and benign models.

In this chapter, our objective is to execute high-quality black-box 3D adversar-

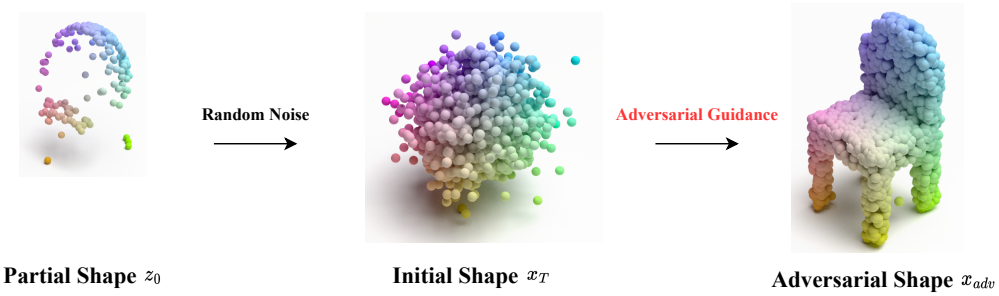


Figure 5.1: **The adversarial shape completion.** Starting from the partial shape z_0, we construct our adversarial shape x_{adv} by utilizing diffusion models with proposed adversarial guidance.

ial attacks using diffusion models. To generate natural adversarial point clouds, we employ diffusion models, which are state-of-the-art generative models known for creating high-quality 2D images [25, 100] and 3D point clouds [141, 149]. It has been demonstrated that 2D diffusion models can generate adversarial examples [15, 20] by altering the diffusion process. By extension, it is intuitive that 3D diffusion models, with their impressive generation performance, are capable of creating adversarial examples. Specifically, we craft adversarial examples by employing diffusion models for shape completion tasks, as shown in Figure 5.1. Using a partial shape as prior knowledge, our attack generates adversarial examples by completing shapes with the proposed adversarial guidance. Our approach to conducting adversarial attacks involves generating unseen data rather than introducing perturbations to clean data, effectively addressing the issue of unrealistic perturbations to xyz coordinates.

In order to enhance the transferability of our crafted adversarial examples against black-box 3D models, we initially incorporate model uncertainty into the gradient inference of the substitute models. Li et al. [66] demonstrated that the introduction of probability measures to the substitute models can significantly enhance the performance of black-box attacks. They execute adversarial attacks by training the substitute model in a Bayesian manner. In our attack, we leverage the characteristics of 3D point clouds and incorporate model uncertainty through a Monte Carlo estimate over the inference from multiple down-sampled point clouds. Additionally, to

improve the attack transferability against various network architectures, we employ ensemble logits to generate the adversarial guidance for the 3D diffusion model. To preserve the generation quality, we limit our adversarial guidance solely to the critical points that are selected based on the saliency scores. Our proposed black-box attack is capable of conducting black-box adversarial attacks against state-of-the-art 3D point cloud deep-learning models without the need to re-train the diffusion model.

Our contributions are summarized as follows:

- We generate adversarial examples through shape completion using diffusion models, offering a novel perspective on the creation of imperceptible adversarial examples. The proposed attack introduces diffusion models to the topic of 3D adversarial robustness.

- We propose a variety of strategies to enhance the transferability of the proposed attacks without compromising the quality of generation. These strategies include: employing model uncertainty for improved inference of predictions, ensemble adversarial guidance to boost attack performance against unseen models, and generation quality augmentation to identify critical points and maintain the quality of generation.

- We conduct a comprehensive evaluation against existing state-of-the-art black-box 3D deep-learning models. Our experiments demonstrate that our proposed attack achieves state-of-the-art performance against both black-box models and defenses.

5.2 Preliminary

5.2.1 Threat Model

Consider a point cloud $x \in \mathcal{P}^{K\times 3}$ consisting of K points, where each point $x_i \in \mathcal{P}^3$ is represented by 3D xyz coordinates. A classifier f is employed to classify the input point cloud and assign a label, denoted as $f(x) \rightarrow y$. In the context of adversarial

attacks, an adversary seeks to generate an adversarial example x_{adv} with the objective of causing the target classifier f to produce an incorrect classification result, represented as y_{adv}. Formally, the goal of the point cloud adversarial attack is defined as:

$$\min D(x, x_{adv}), \qquad \text{s.t. } f(x_{adv}) = y_{adv} \tag{5.1}$$

Equation 5.1 is designed to generate an imperceptible adversarial example x_{adv} from the original point cloud x. This chapter primarily concentrates on untargeted attacks, where y_{adv} can be any label distinct from the ground truth label y.

5.2.2 3D Point Cloud Generation and Completion

Recent advancements in diffusion models [25, 46, 56, 100] applied to 2D image generation have showcased remarkable performance in terms of both generation quality and diversity. Likewise, recent studies on 3D diffusion models [83, 141, 149] have demonstrated state-of-the-art performance in 3D point cloud generation tasks. The 3D denoising diffusion probabilistic model generates 3D point clouds with a denoising generation process. Starting from Gaussian noise x_T, the denoising process gradually produces the final output by a sequence of denoising-like steps, i.e., $x_T, x_{T-1}, \ldots, x_0$.

The generative diffusion model, denoted as $p_\theta(x_{0:T})$, aims to learn the Gaussian transitions from $p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$ by reconstructing x_0 from the diffusion data distribution $q(x_{0:T})$. This distribution introduces Gaussian noise to x_0 over the course of T steps. More specifically, these processes of adding noise and subsequent denoising can be formulated as a Markov transition:

$$\begin{aligned} q(x_{0:T}) &= q(x_0) \prod_{t=1}^{T} q(x_t | x_{t-1}) \\ p_\theta(x_{0:T}) &= p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} | x_t) \end{aligned} \tag{5.2}$$

where we name the $q(x_t|x_{t-1})$ as *forward diffusion process* and $p_\theta(x_{t-1}|x_t)$ as *reverse generative process*. Each detailed transition for each process is defined in accordance

with the scheduling function $\beta_1, \ldots, \beta_T$:

$$\begin{aligned} q(x_t|x_{t-1}) &:= \mathcal{N}(x_t : \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \\ p_\theta(x_{t-1}|x_t) &:= \mathcal{N}(x_{t-1} : \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}) \end{aligned} \tag{5.3}$$

where $\mu_\theta(x_t, t)$ is the inference of the diffusion model to predict the shape of the point cloud. We set $\sigma_t^2 = \beta_t$ based on empirical knowledge.

The 3D point cloud generation task can be easily modified to achieve shape completion with an fixed partial shape $z_0 \in \mathcal{P}^{K_p \times 3}$ [149]. The *forward diffusion process* and *reverse generative process* are formulated as:

$$\begin{aligned} q(\tilde{x}_t|\tilde{x}_{t-1}, z_0) &:= \mathcal{N}(\tilde{x}_t : \sqrt{1-\beta_t}\tilde{x}_{t-1}, \beta_t\mathbf{I}) \\ p_\theta(\tilde{x}_{t-1}|\tilde{x}_t, z_0) &:= \mathcal{N}(\tilde{x}_{t-1} : \mu_\theta(x_t, z_0, t), \sigma_t^2\mathbf{I}) \end{aligned} \tag{5.4}$$

While recent studies have extensively explored the generation capabilities of 3D diffusion models, their potential in crafting adversarial point clouds remains largely unexplored. In this chapter, we aim to generate high-quality adversarial point clouds with the reverse generative process of pre-trained 3D diffusion models. Note that we don't modify the training part of pre-trained models.

5.3 Methodology

5.3.1 Diffusion Model for 3D Adversarial Shape Completion

In crafting high-quality adversarial examples, our aim is to utilize diffusion models for their superior performance in 3D point cloud generation. Unlike previous generative models, the denoising generation process of diffusion models can naturally incorporate adversarial objectives [15,20], which can be viewed as a process of iterative adversarial attacks. Previous perturbation-based adversarial attacks perturb each point in the clean point cloud, commonly altering the shape of the original point cloud. In our work, we aim to minimize the impact of adversarial perturbations on the point cloud data and achieve adversarial attacks with our proposed method, the 3D adversarial

shape completion attack.

The proposed attack generates adversarial point clouds with a fixed partial shape $z_0 \in \mathcal{P}^{K_p \times 3}$. We utilize any pre-trained 3D shape completion diffusion model ϵ_θ to gradually generate the completed adversarial point cloud $x_0 = (z_0, \tilde{x}_0)$ through the reverse generative process $p_\theta(\tilde{x}_{t-1}|\tilde{x}_t, z_0)$, $t = T, \ldots, 1$. For any intermediate shape $x_t = (z_0, \tilde{x}_t)$, the adversarial generative process is defined as:

$$p_\theta(\tilde{x}_{t-1}|\tilde{x}_t, z_0) := \mathcal{N}(\tilde{x}_{t-1} : \mu_\theta(x_t, z_0, t), \beta_t \mathbf{I}) - a\beta_t \nabla_{x_t} \mathcal{L}(f(x_t), y) \tag{5.5}$$

where y represents the ground truth label of the original point cloud, $\mathcal{L}$ denotes the cross, and the scale of adversarial guidance $a \in (0, 1)$. We employ the untargeted I-FGSM-like gradient as the adversarial guidance for the adversarial generative process [15].

We sample benign $\tilde{x}_{t-1}$ from $\mathcal{N}(\tilde{x}_{t-1} : \mu_\theta(x_t, z_0, t), \beta_t \mathbf{I})$ by following PVD [149]:

$$\tilde{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\tilde{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \tilde{\alpha}_t}} \epsilon_\theta(\tilde{x}_t, z_0, t) \right) + \sqrt{\beta_t} \varepsilon, \tag{5.6}$$

where α and β are hyper-parameters from the pre-trained ϵ_θ, and $\varepsilon \sim N(0, \mathbf{I})$.

5.3.2 Diffusion Model with Boosting Transferbility

In order to improve the effectiveness of the proposed attack on a black-box target model, we have outlined several effective strategies to enhance the transferability of the generated 3D point clouds, all without increasing the magnitude of the adversarial guidance.

Employing Model Uncertainty. Previous works [9,69] have shown that leveraging model uncertainty for feature learning is proposed to be more robust to adversarial attacks compared to standard deep learning models. These Bayesian deep neural networks are probabilistic models that predict input by computing expectations from maximum likelihood estimation over model parameters. Furthermore, utilizing model uncertainty [66] demonstrates improved adversarial transferability. However, the ap-

plication of model uncertainty in 3D contexts is currently underexplored. Considering the characteristics of 3D point clouds, which comprise unordered 3D points, the removal of some points does not alter the classification outcome of the original point cloud [148]. Therefore, we are able to straightforwardly adopt model uncertainty to 3D deep-learning models with the *MC dropout*-like [32] approach over the input. In our attack, we adopt Simple Random Sampling over the 3D point clouds and use the Monte Carlo estimate over M re-sampled point clouds to obtain the estimated adversarial guidance:

$$\nabla_{x_t}\mathcal{L}_{\mathrm{MU}}(f(x_t),y)) = \frac{1}{M}\sum_{s=1}^{M}\nabla_{x_s}\mathcal{L}(f(x_s),y) \tag{5.7}$$

The x_s is obtained by simple random sampling from $x_t = (z_0, \tilde{x}_t)$:

$$P_i(\tilde{x}_t) = \{1_x | x \in \tilde{x}_t, 1_x \sim Ber(0.5)\} \tag{5.8}$$

where x is sampled from a $Bernoulli(0.5)$ distribution to indicate the existence of x in the $x_s = (z_0, \tilde{x}_s)$ point cloud re-sampled from i^{th} point of $\tilde{x}_t$, and z_0 is not re-sampled.

Ensemble Adversarial Guidance. In the 2D attack scenario, the ensemble attack is an effective way to enhance the attack transferability by utilizing multiple white-box models to calculate the average gradient of the objective loss. Ensemble gradient in 2D results in perturbation in the given pixel of the 2D image. In our attack, we ensemble the logits of selected substitute models according to the generative process in Equation 5.5. Formally, with n_{ens} substitute models, the ensemble adversarial objective function is defined as:

$$\mathcal{L}(f_{ens}(x_t), y) = -\log(\mathrm{softmax}\sum_{n=1}^{n_{\mathrm{ens}}} w_n p_{f_n}(y|x_t)) \tag{5.9}$$

where w_n is the weight parameters, and we use the proportion of correctly classified point clouds for an adaptive ensemble attack; p_f is the predictive distribution of f.

Generation Quality Augmentation. Previous work [146] has shown that individ-

ual points within a point cloud can have varying degrees of impact on the classification outcome of a 3D deep-learning model. This insight suggests that identifying critical points within the point cloud could achieve strong adversarial attacks. Due to the significant reduction in visual quality caused by perturbations to 3D coordinates, it is advisable to control these perturbations by constraining the ℓ_0 distance between the adversarial and benign point clouds. Thus, our objective is to create adversarial examples by altering only a subset of N points of the benign point cloud. The saliency score of given point x is calculated as:

$$\text{score}_x = \sum_3 \frac{\partial \mathcal{L}(f(x_t), y)}{\partial x} \tag{5.10}$$

where the saliency score is the sum of xyz channels of point x. Moreover, we further adopt ℓ_{inf} norm restriction to the perturbation at each diffusion step for a fair comparison with perturbation-based adversarial attacks.

5.3.3 Transferable 3D Adversarial Shape Completion Attack

We summarize the proposed black-box 3D adversarial attack in Algorithm 4. In the early generation process, the generated point clouds are disorganized. Therefore, we only perform adversarial guidance at given timestep T_{adv}. We apply the Clip [34] function to the ℓ_{inf} norm to limit the perturbation in adversarial guidance.

5.3.4 Revisiting 3D Black-Box Adversarial Attack

Black-box adversarial attacks present a significantly greater challenge than white-box adversarial attacks, with 3D black-box adversarial attacks proving even more difficult than their 2D counterparts. As illustrated in Figure 5.4, the data distribution of the existing ShapeNet 3D dataset is long-tailed. Consequently, existing adversarial attack methods tend to achieve a higher ASR on classes with less data (the top 5 classes contain 50% data but only contribute 14% success adversarial examples). This issue is similar in the ModelNet40 dataset, in which the top 5 classes contain 30% of data.

Algorithm 4 Transferable 3D Adversarial Shape Completion Attack Algorithm

Require: f_{ens}: substitute models, z_0: partial shape for shape completion, y: class label for shape completion, T: reverse generation process timestep for LDM, T_{adv}: timestep for adversarial guidance, N: number of perturbed points at each diffusion step, M: number of simple random sampling

$\tilde{x}_T \sim \mathcal{N}(0, \mathbf{I})$, $x_T = (z_0, \tilde{x}_T)$
$x_{adv} = \varnothing$
for $t = T, \ldots, 1$ **do**
    if t is in T_{adv} **then**
        Sample $\tilde{x}_{t-1}$ with Equation 5.4
        for $m = 1, \ldots, M$ **do**
            Simple random sampling with Equation 5.8
            Obtain the ensemble adversarial loss with Equation 5.9
        end for
        Monte Carlo estimate with Equation 5.7
        Calculate the saliency score of $\tilde{x}_{t-1}$ with Equation 5.10
        Update top-N points from step 11 of $\tilde{x}_{t-1}$ with Equation 5.5
        $\tilde{x}_{t-1} = \text{Clip}(\tilde{x}_{t-1})$
    else
        Sample $\tilde{x}_{t-1}$ with Equation 5.4
    end if
end for
$x_0 = (z_0, \tilde{x}_0)$
$x_{adv} \leftarrow x_0$ if $f_{\text{ens}}(x_0) \neq y$
return x_{adv}

Another significant challenge in 3D black-box adversarial attacks lies in the varying model architectures. To provide a comprehensive discussion on the transferability between different 3D models, we have demonstrated the cosine similarity from the logit outputs by the same input of various models in Figure 5.4. The results indicate that gradients from models with different architectures vary significantly, thus posing a considerable challenge for 3D black-box adversarial attacks. These challenging problems make existing 3D black-box adversarial attacks effective against only a few 3D models on the ModelNet40 dataset.

To execute an effective black-box 3D adversarial attack, we employ diffusion models to directly generate adversarial examples. The gradual diffusion generation process allows for the introduction of adversarial guidance with significantly less perturbation than existing adversarial attacks. Adversarial shape completion aids in identifying

Table 5.1: **The attack success rate (ASR %) of transfer attack on the ShapeNet dataset.** The adversarial examples of existing attack methods are generated from the PointNet model. The Average ASR is calculated among the seven black-box models ($3DAdvDiff_{ens}$ is calculated among the five black-box models).

Dataset	Method	PointNet	PointNet++	DGCNN	PointConv	CurveNet	PCT	PRC	GDANet	Average
Chair	PGD	99.7	1.0	0.9	1.2	0.7	1.4	0.9	2.1	1.2
	KNN	99.2	0.8	0.8	1.0	0.4	1.2	1.0	2.1	1.0
	GeoA3	99.6	0.9	0.8	1.2	0.7	0.8	1.0	0.9	0.9
	SI-Adv	82.4	1.2	1.2	1.5	1.5	1.4	2.3	2.2	1.6
	AdvPC	71.8	2.2	0.9	1.5	1.8	2.1	2.6	2.0	1.6
	PF-Attack	99.0	20.2	5.6	4.8	3.2	1.0	2.5	1.6	5.5
	3DAdvDiff	99.9	60.6	8.7	23.5	9.8	6.9	14.9	8.9	19.0
	$3DAdvDiff_{ens}$	99.9	**94.5**	***99.9***	**91.3**	**88.6**	**65.8**	***99.9***	**85.6**	**85.2**
Dataset	Method	PointNet	PointNet++	DGCNN	PointConv	CurveNet	PCT	PRC	GDANet	Average
All	PGD	99.9	2.1	0.7	0.8	0.5	0.4	0.7	1.6	0.9
	KNN	99.9	2.2	0.7	0.7	0.5	0.6	1.1	1.6	1.1
	GeoA3	99.8	2.0	1.5	1.4	0.9	0.6	0.9	1.1	1.2
	SI-Adv	92.5	2.0	1.7	1.5	1.2	1.0	1.3	1.0	1.4
	AdvPC	89.6	0.4	0.2	0.5	0.4	0.6	0.7	0.5	0.5
	PF-Attack	99.6	24.2	6.7	5.1	3.8	1.2	2.4	1.9	6.2
	3DAdvDiff	99.9	73.2	12.6	55.3	40.5	32.6	25.9	16.0	36.6
	$3DAdvDiff_{ens}$	99.9	**97.0**	***99.9***	**94.5**	**93.5**	**80.5**	***99.9***	**85.2**	**90.1**

the vulnerable rotation for more potent adversarial attacks and ensures the reliable generation of natural point clouds, surpassing shape generation tasks. In addition to utilizing an ensemble attack approach, we also employ random sampling to leverage model uncertainty and enhance performance against defenses. By taking into account the characteristics of 3D point clouds and the generation performance of diffusion models, we are able to achieve an effective and high-quality black-box 3D adversarial attack.

5.4 Experiments

5.4.1 Experimental Setup

Dataset. Due to ModelNet40 being insufficient to train the diffusion model, we use the ShapeNet [11] dataset for major evaluations. The ShapeNetCore split is adopted, which contains 42003 point clouds with 55 categories, of which 31535 samples are used for training and 10468 samples are used for testing. We select PVD [149] for the diffusion model in this chapter. The proposed attack does not require additional training in the diffusion model, we follow settings as in the original PVD chapter for

selecting shape completion's partial shapes. Public checkpoints [149] from Airplane, Chair, and Car are selected for repeatability.

Target Models. For a better evaluation of different network architectures, we select eight widely adopted 3D deep-learning models as the black-box models, including PointNet [95], Pointnet++ (SSG) [96], DGCNN [122], PointConv (SSG) [127], CurveNet [130], PCT [37], PRC [99], and GDANet [134].

Comparisons. We have chosen four white-box 3D adversarial attacks as our baseline for comparison, namely: PGD [78], KNN [117], GeoA3 [123], and SI-Adv [51]. We also employ existing black-box 3D adversarial attacks, specifically: AdvPC [40] and PF-Attack [42]. We use PointNet as the substitute model by default and the perturbations are constrained under the ℓ_{inf}-normal ball with a radius of 0.16. We use 3DAdvDiff to denote the white-box version of the proposed attack and $\text{3DAdvDiff}_{\text{ens}}$ for boosting transferability version.

Defenses. We select SRS [148], SOR [148], DUP-Net [148], IF-Defense [128], and Adversarial Hybrid Training [53] for evaluation under defenses. All the defense settings are followed according to [53].

Attack Settings. We select PointNet, DGCNN, and PRC for ensemble adversarial guidance on $\text{3DAdvDiff}_{\text{ens}}$. The hyper-parameters of the proposed attack are set to: $a = 0.4, T = 1000, T_{\text{adv}} = (0, 0.2T], N = 200, M = 5, K = 2048$. We also adopt $\ell_{\text{inf}} = 0.16$ restriction to the adversarial guidance. We set 200 points for partial shapes. For each partial shape, we generate 20 views and only save the views that successfully attack the substitute models. To evaluate the attack performance, we use the top-1 accuracy of the target model to evaluate the ASR. The experiment results are averaged over 10 attacks.

5.4.2 Attack Performance

Transfer Attack. We evaluate the transfer attack performance of current point cloud adversarial attack methods on selected robust classes. The results are given in Table 5.1. As we discussed in Section 4.4, the adversarial examples from state-of-the-art attacks merely transfer to different models, particularly those recently de-

Table 5.2: **The attack success rate (ASR %) of different adversarial attack methods against defenses.** All attacks are evaluated under white-box settings against the PointNet model.

Method	ASR	SRS	SOR	DUP-Net	IF-Defense	HybridTraining
PGD	99.9	5.9	1.0	0.7	13.8	1.9
KNN	99.9	4.0	0.9	0.4	13.0	1.3
GeoA3	99.8	4.9	1.6	0.8	13.6	2.2
SI-Adv	92.5	10.8	0.9	0.9	14.9	2.0
AdvPC	89.6	4.1	1.5	0.7	13.2	1.9
PF-Attack	99.6	8.5	3.6	2.8	13.9	2.0
3DAdvDiff	99.9	82.2	9.9	9.6	**30.0**	9.4
$3DAdvDiff_{ens}$	99.9	**85.9**	**49.1**	**36.9**	22.5	**96.1**

veloped 3D models. Models trained on long-tailed datasets typically exhibit limited generalization. However, our proposed white-box 3DAdvDiff achieves notably better performance even on the black-box adversarial attack. Furthermore, $3DAdvDiff_{ens}$ considerably boosts the attack performance of 3DAdvDiff without augmenting the magnitude of the adversarial guidance, thereby validating the effectiveness of our proposed methods.

Adversarial Defenses. We evaluate the adversarial examples against a variety of defenses under white-box settings, as shown in Table 5.2. The findings indicate that current defenses can effectively counter existing adversarial attacks, even with simple SRS (Simple Random Sampling). Defense methods that rely on outlier point removal exhibit the best performance among all defenses, suggesting that perturbation-based attack methods tend to displace points outside the original shape by adding perturbations to xyz coordinates. Our proposed 3DAdvDiff significantly outperforms state-of-the-art adversarial attacks. Due to its utilization of model uncertainty, 3DAdvDiff is particularly effective against random sampling. The proposed critical point selection of $3DAdvDiff_{ens}$ is effective against outlier removal defenses. However, the performance of $3DAdvDiff_{ens}$ against IF-Defense is not satisfying due to the selection of critical points. Balancing generation quality and defense performance remains a challenge. In future work, we aim to enhance attack performance against reconstruction-based defenses.

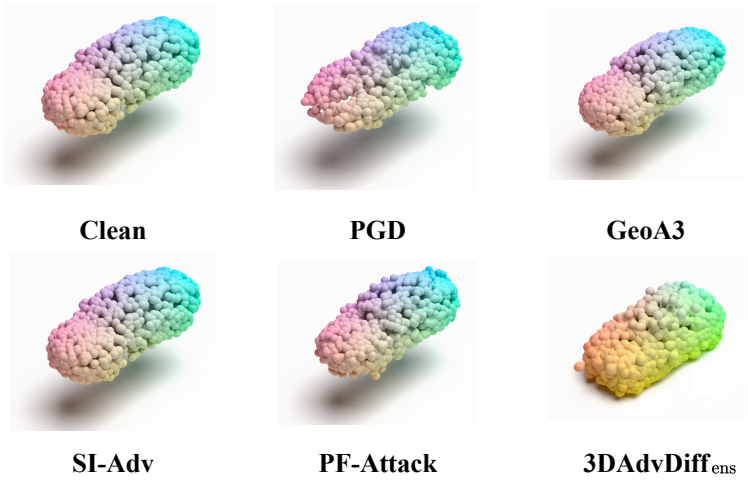


Figure 5.2: **The visual quality of adversarial examples.** The black-box adversarial examples are relatively unnatural compared to white-box adversarial examples.

Generation Quality. We further assess the distance between benign and adversarial examples to evaluate the visual quality of existing adversarial attack methods, as shown in Table 5.3. The Chamfer Distance (CD), Hausdorff Distance (HD), and Mean Square Error (MSE) are selected. Given that we apply the same $\ell_{\text{inf}} = 0.16$ norm to limit the perturbation for each attack, the visual quality across different attack methods is relatively similar. However, it is hard to give a fair comparison with 3DAdvDiff's adversarial examples, because the adversarial sampling of diffusion models can lead to the generation of new point clouds with completely different shapes. Therefore, the generation quality of 3DAdvDiff$_{\text{ens}}$ is evaluated by the difference between the benign samples and the adversarial examples with fixed sampling. A visual comparison is provided in Figure 5.2 for a more comprehensive demonstration. The point clouds generated by 3DAdvDiff$_{\text{ens}}$ is smoother than existing attacks.

Table 5.3: **The generation quality on the ShapeNet dataset.** The CD distance is multiplied by 10^{-2}.

Method	PGD	KNN	GeoA3	SI-Adv	AdvPC	PF-Attack	3DAdvDiff$_{\text{ens}}$
HD	0.136	0.105	0.039	0.071	**0.028**	0.046	*0.098*
CD	0.46	0.42	**0.10**	0.33	0.27	0.25	*0.14*
MSE	2.71	2.42	1.50	3.08	2.04	1.85	***1.18***

Time Efficiency. Despite the proposed 3DAdvDiff achieves overwhelming performance on black-box adversarial attacks. The generation speed of diffusion models is a critical problem that influences their development. As shown in Table 5.4, the running time of the proposed 3DAdvDiff is relatively slower than previous perturbation-based attack methods. However, we can improve the sampling speed by adopting DDIM sampling to PVD.

Table 5.4: **The average running time to generate one adversarial example.**

Method	PGD	KNN	GeoA3	SI-Adv	AdvPC	PF-Attack	3DAdvDiff$_{\text{ens}}$
Time (s)	1.1	17.3	81.6	7.0	2.5	38.6	60.8

Integration with Other Methods. To completely demonstrate the effectiveness of the proposed transferability boosting methods, we integrate the proposed improvement methods with existing attacks. As shown in Table 5.5, our proposed enhancement methods markedly improve the performance of PGD, SI-Adv, and AdvPC on black-box attacks. However, the performance increase of adversarial attacks is limited without the diffusion models.

Table 5.5: **The ensemble of proposed boosting transferability methods with existing attack methods.** The experiments are performed on the **whole** test dataset of the ShapeNet dataset.

Method	PointNet	PointNet++	DGCNN	PointConv	CurveNet	PCT	PRC	GDANet	Average
PGD	**99.8**	10.8	8.9	11.1	7.1	7.3	9.1	10.1	9.2
PGD + 3DAdvDiff	99.5	**48.9**	***93.6***	**21.7**	**25.6**	**14.2**	***96.1***	**14.5**	**25.0**
SI-Adv	**97.6**	12.2	10.2	11.9	7.5	8.8	12.8	8.3	10.2
SI-Adv + 3DAdvDiff	70.5	**42.8**	***45.9***	**19.2**	**24.9**	**20.4**	***38.6***	**21.7**	**25.8**
AdvPC	**96.9**	7.7	6.1	6.3	10.9	5.4	6.8	6.1	7.0
AdvPC + 3DAdvDiff	95.2	**57.5**	***75.8***	**38.1**	**35.4**	**21.8**	***63.0***	**16.1**	**33.8**

5.4.3 Ablation Study

We conduct a series of ablation studies to investigate the effectiveness of various approaches in 3DAdvDiff$_{\text{ens}}$ for enhancing transferability, including model uncertainty, ensemble adversarial guidance, and generation quality augmentation.

Adversarial Guidance. The parameter a of the adversarial guidance is critical to the attack success rate and the generation quality, as shown in Figure 5.3. However,

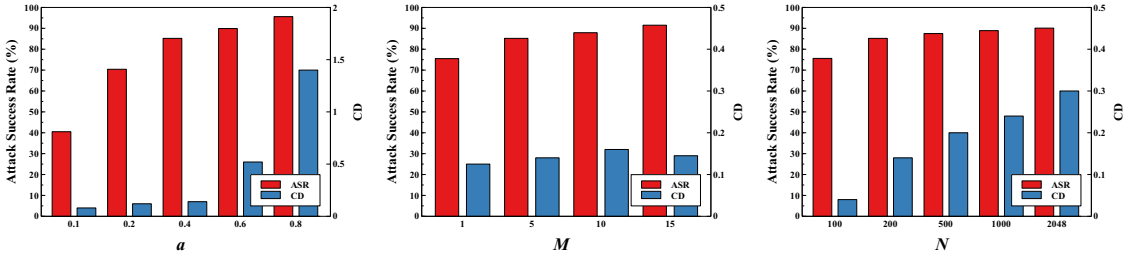


Figure 5.3: **The ablation study of proposed 3DAdvDiff$_{\text{ens}}$.** The results are evaluated on the Chair class of the ShapeNet dataset. We use average ASR to test the black-box attack performance.

our proposed 3DAdvDiff generates adversarial examples by finding the most vulnerable rotation from multiple views. Therefore, we can easily balance ASR and the generation quality without largely decreasing ASR.

Model Uncertainty. We evaluate the performance of model uncertainty with varying settings of M. Figure 5.3 indicates that attack transferability increases with a larger M. However, this significantly impacts the time efficiency required to generate adversarial examples. As shown in Table 5.6, incorporating model uncertainty significantly improves the transfer attack performance of 3DAdvDiff combined with the sampling of the diffusion model. These results further validate the effectiveness of our proposed model uncertainty approach.

Table 5.6: **The ensemble of model uncertainty with 3DAdvDiff.** The experiments are performed on the Chair class of the ShapeNet dataset.

Method	PointNet	PointNet++	DGCNN	PointConv	CurveNet	PCT	PRC	GDANet	Average
3DAdvDiff	99.9	60.6	8.7	23.5	9.8	6.9	14.9	8.9	19.0
3DAdvDiff + MU	99.9	**82.6**	**78.6**	**85.6**	**84.2**	**68.1**	**59.5**	**70.2**	**75.5**

Ensemble Adversarial Guidance. We test the performance of 3DAdvDiff with ensemble adversarial guidance. Table 5.7 shows that the proposed adversarial guidance can effectively improve the performance of transfer attacks against black-box models. Simultaneously, the use of ensemble adversarial guidance does not compromise the generation quality of the proposed attack.

Generation Quality Augmentation. Current 3D distance measurements take into account the difference between the entire point set. Therefore, to improve the

Table 5.7: **The performance of ensemble adversarial guidance.** The experiments are performed on the Chair class of the ShapeNet dataset.

Method	PointNet	PointNet++	DGCNN	PointConv	CurveNet	PCT	PRC	GDANet	Average
3DAdvDiff	99.9	60.6	8.7	23.5	9.8	6.9	14.9	8.9	19.0
3DAdvDiff + EAG	99.9	**70.8**	***99.9***	**79.5**	**75.9**	**45.3**	***99.9***	**54.3**	**65.2**

generation quality, we should limit the ℓ_0 distance between the adversarial and benign examples. The proposed augmentation notably enhances the quality of the generated point clouds without compromising the attack performance. The results are given in Figure 5.3.

5.5 Discussion

Experiments demonstrate that current attacks perform poorly against black-box models under the $\ell_{\text{inf}} = 0.16$ constraint, particularly in the Chair, Airplane, and Car categories. However, these black-box models are extremely vulnerable to the proposed 3DAdvDiff due to the long-tail training dataset. Consequently, we advocate for a more balanced training approach for 3D point cloud models and the creation of large-scale datasets with a similar scale to the 2D ImageNet. While 3DAdvDiff delivers satisfactory attack performance, its major weakness lies in the need for improved time efficiency to ensure better generalization.

5.6 Conclusion

In this chapter, we introduce the first-ever method designed to execute a black-box adversarial attack on recently developed 3D point cloud classification models. Our research is also a pioneering work in the use of diffusion models for 3D adversarial attacks. Specifically, we generate adversarial examples through 3D adversarial shape completion, ensuring reliable and high-quality point cloud generation. We propose several strategies to enhance the transferability of our proposed attack, including the use of model uncertainty for improved prediction inference, enhancing adversarial guidance through ensemble logits from various substitute models, and the improve-

ment of generation quality via critical points selection. Comprehensive experiments on the robust dataset validate the effectiveness of our proposed attacks. Our methods establish a solid baseline for future development in black-box 3D adversarial attacks.

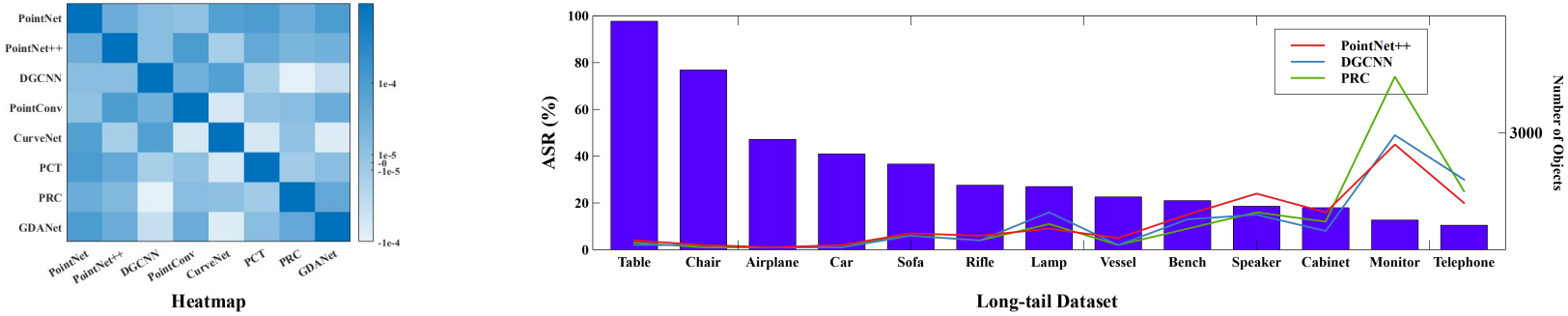


Figure 5.4: **The challenging 3D black-box adversarial attacks.** The value in the Heatmap is re-scaled for better visualization. We use the top 13 classes from the ShapeNet dataset to demonstrate the long-tailed dataset problem. We use PGD with $\ell_{\text{inf}} = 0.16$ on PointNet to evaluate the black-box ASR.

Chapter 6

Gradient-Free Adversarial Purification with Diffusion Models

6.1 Introduction

Deep learning models have demonstrated remarkable performance across various tasks [43,79,130]. With the rapid advancement and widespread deployment of these models, their security and robustness are garnering increasing attention.

It is widely recognized that deep learning models are highly vulnerable to adversarial attacks [10, 84]. These attacks are performed by adding imperceptible perturbations to clean images. The perturbed images, known as adversarial examples, can deceive trained deep learning classifiers with high confidence while appearing natural and realistic to human observers. To mitigate adversarial attacks and ensure the stability of deep learning models, adversarial training [35, 84] has been developed. This approach aims to defend against adversarial attacks by training the classifier with adversarial examples. However, adversarial training tends to perform poorly against unknown attacks.

Recently, with the development of diffusion models [25, 100], adversarial purification [89, 109] has shown promising defense performance by recovering the adversarial examples to clean images. These works adopt the diffusion model's reverse generation process to gradually remove the Gaussian noise from the forward process and the

Figure 6.1: **The proposed adversarial defense pipeline.** We give an adversarial example of "cock" class with AutoAttack $\ell_{\text{inf}} = 8/255$ on ImageNet dataset. Adversarial anti-aliasing aims to eliminate adversarial perturbations, while adversarial super-resolution seeks to restore benign images from blurred adversarial examples using prior knowledge from the clean dataset.

adversarial perturbations. Nevertheless, these methods require heavy computational resources during the purification, which may not be practical in real-time scenarios.

Diffusion models also facilitate stronger unrestricted adversarial attacks [15,16,20]. These UAEs are generated through the reverse generation process by incorporating adversarial guidance. Unlike traditional perturbation-based adversarial attacks, UAEs exhibit superior attack performance against current defenses due to their distinct threat models. These attacks pose a new threat to the development of deep learning models and urgently need to be addressed. Even worse, existing defenses have merely covered the discussion against UAEs.

In this chapter, we propose an effective adversarial defense method that detects both perturbation-based adversarial examples and unrestricted adversarial examples. To achieve the defense objective, we locate and utilize the common characteristic of these two types of attacks that both adversarial examples are generated close to the decision boundary for minimal perturbations, which makes these adversarial examples susceptible to changes in pixels.

Our defense employs zero-shot adversarial purification by extracting the "semantic shape" information from images without the image details, as illustrated in Figure 6.1. Specifically, we use adversarial anti-aliasing with specialized filters to blur the

detailed adversarial modifications in the adversarial examples. Following this, we apply adversarial super-resolution to the anti-aliased adversarial examples, upscaling the blurred images using details from pre-trained clean super-resolution diffusion models. These two methods are time-efficient and do not require any modifications to the original models. To demonstrate the effectiveness of our proposed defense, we further validate its performance by using the upscaled adversarial examples as input for adversarial purification. Experiments on various datasets show that our defense outperforms state-of-the-art adversarial defenses in adversarial purification.

Our contributions are summarized as follows:

- We propose a novel adversarial defense capable of countering both perturbation-based adversarial examples and unrestricted adversarial examples, addressing the current gap in effective defenses against unrestricted adversarial attacks.

- We introduce various zero-shot and gradient-free defense strategies that preserve the semantic information of adversarial examples while eliminating adversarial modifications. These strategies include adversarial anti-aliasing for "semantic" extraction and adversarial super-resolution for incorporating benign priors and recovering benign details from adversarial examples.

- We conduct extensive experiments on various datasets against adaptive adversarial attacks. The results demonstrate the effectiveness of our proposed defense method compared to state-of-the-art adversarial defenses. Moreover, anti-aliased and upscaled adversarial examples effectively integrate with existing diffusion-based adversarial purification, validating the usability and scalability of our approach.

6.2 Preliminary

6.2.1 Threat Model

Adversarial examples conduct attacks by deceiving the target model with wrong classification results. Considering the untargeted attack scenario, the perturbation-based adversarial examples are defined as:

$$A_{\text{AE}} \triangleq \{x_{\text{adv}} = x + \delta | y \neq f(x), x \in D, |\delta| \leq \epsilon\} \tag{6.1}$$

where δ is the adversarial perturbation, $f(\cdot)$ is the target model, D is the clean dataset, and ϵ is the perturbation norm constraint.

These adversarial examples are generated by adding the perturbations to the clean images. However, such perturbations can degenerate the image quality. By utilizing the generation models, Song et al. [110] presented unrestricted adversarial examples by directly generating adversarial examples with the generation tasks, which can be formulated as:

$$A_{\text{UAE}} \triangleq \{x_{\text{adv}} \in \mathcal{G}(z_{\text{adv}}, y) | y \neq f(x)\} \tag{6.2}$$

where $\mathcal{G}$ is the generation model, z_{adv} is the latent code for generation.

These two adversarial examples are generated with different threat models. However, they both can successfully conduct attacks against the given target model. A robust defense method should be able to defend against these attacks simultaneously.

6.2.2 Diffusion-Based Adversarial Purification

The diffusion model [46] learns to recover the image from the denoising-like process, i.e., *reverse generation process*. The reverse generation process takes T time steps to obtain a sequence of noisy data $\{x_{T-1}, \ldots, x_1\}$ and get the data x_0 at the last step. Specifically, it can be formulated as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1} : \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}) \tag{6.3}$$

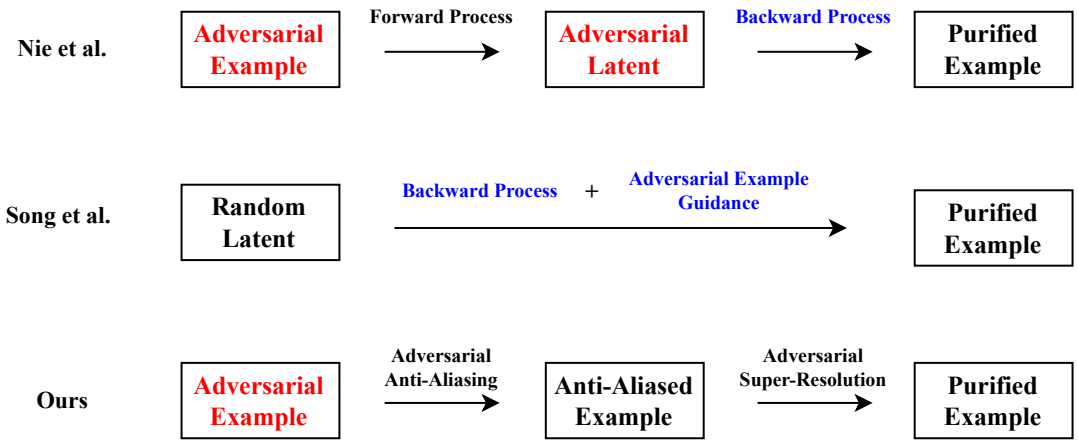


Figure 6.2: **The comparisons of state-of-the-art diffusion-based adversarial purification pipelines.** We mark the defense process in blue to represent time-consuming approaches. We use red font to indicate non-purified adversarial input.

The *forward diffusion process* is where we iteratively add Gaussian noise to the data for training the diffusion model to learn $p_\theta(x_{t-1}|x_t)$. It is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t : \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{6.4}$$

where σ is the noise schedule.

Nie et al. [89] attempted to find the optimal t^* where it satisfy that:

$$\begin{aligned} x_{t^*} =& \sqrt{\sigma_{t^*}}x_{\text{adv}} + \sqrt{1-\sigma_{t^*}}\varepsilon \\ =& \sqrt{\sigma_{t^*}}(x+\delta) + \sqrt{1-\sigma_{t^*}}\varepsilon \end{aligned} \tag{6.5}$$

where ε is the Gaussian noise $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. After we obtain the optimal t^*, we can utilize the reverse generation process over x_{adv} to recover the clean x.

Song et al. [109] utilized the whole reverse generation process from T diffusion timesteps; they used adversarial sample x_{adv} as guidance rather than an intermediate time step state. At each time step t, the guidance is added to the x_t after the original reverse generation process and can be formulated as:

$$\nabla_x \log p(x_{\text{adv}}|x_t; t) = -R_t \nabla_{x_t} d(\hat{x}_t, x_{\text{adv}}) \tag{6.6}$$

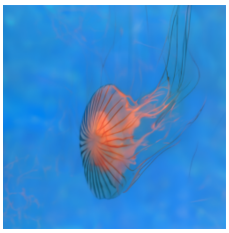
AutoAttack Example
Robust Acc: 0%

RGB conversion
Robust Acc: 38.25%

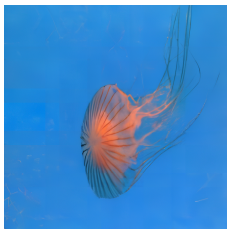
Adv. Anti-Aliasing
Robust Acc: 55.85%

Figure 6.3: **The vulnerability of adversarial examples to the changes in pixels.** The RGB conversion is performed by converting the images to RGB space after the ImageNet normalization and achieves 38% robust accuracy. The proposed adversarial anti-aliasing is more effective while preserving the image quality.

AutoAttack Example



MimicDiffusion



Adv.
Super-Resolution

Figure 6.4: **The example of proposed adversarial super-resolution.** Our method achieves similar adversarial purification without any gradient calculation of diffusion models.

where R_t is the scale factor at t time step, $d(\cdot)$ is the distance measurement, and $\hat{x}_t$ is the estimation for x_0 at t time step. The $\hat{x}_t$ is defined as:

$$\hat{x}_t = \frac{x_t - \sqrt{1-\sigma_t} s_\theta(x_t)}{\sqrt{\sigma_t}} \tag{6.7}$$

where the s_θ known score function is defined as [111].

6.3 Methodology

6.3.1 Motivation

With the advancement of diffusion models, diffusion-based adversarial purification has emerged as a leading approach for adversarial defenses. However, current methods still face significant challenges that impact their effectiveness. Figure 6.2 illustrates typical diffusion-based purification pipelines from state-of-the-art methods. Nie et al. [89] achieved purification by utilizing the adversarial latent generated by the forward process of adversarial examples. Unfortunately, this approach can introduce adversarial perturbations into the purified examples, as these perturbations persist in the adversarial latent. Song et al. [109] sought to mitigate the impact of adversarial perturbations by using random latents, employing adversarial examples solely as guidance. However, this method requires gradient calculations at each step of the reverse process, making it computationally intensive. Consequently, achieving both **time-efficient** and **perturbation-isolated** diffusion-based adversarial purification remains a challenge. Furthermore, existing defenses fail to defend against the recently proposed unrestricted adversarial attacks.

6.3.2 Perturbation-Isolated Adversarial Purification

Perturbation-based adversarial examples are precisely calculated based on the gradient of the loss function, whereas unrestricted adversarial examples are sampled near the decision boundary. Despite employing different threat models, both types of attacks produce adversarial examples that are sensitive to pixel changes. Since adversarial examples are designed to be imperceptible compared to clean images, the semantic shapes of objects within the images should correspond to their original labels. Therefore, our defense strategy focuses on extracting the semantic shapes from the adversarial examples and eliminating the adversarial pixel-level details.

Adversarial Anti-Aliasing

To achieve effective defenses against both unrestricted and perturbation-based adversarial attacks, it is essential to address their common characteristics. One critical factor is the value range of images: a valid RGB value is an integer between 0 and 255. However, the modifications introduced by various adversarial attacks are often performed using non-integer data types for gradient calculations. These modifications can become ineffective when transformed back to the RGB image format. Figure 6.3 supports our findings, showing that approximately 38% of adversarial examples from AutoAttack fail with the combinations of RGB conversions and image normalization for deep-learning models. The reasons for this phenomenon could be that adversarial examples are typically located near the decision boundary and are sensitive to pixel changes. However, simple RGB conversion can be effectively compromised by adaptive attacks [2]. Therefore, in this chapter, we aim to propose more effective transformations.

Anti-aliasing is a straightforward, zero-shot method for smoothing image details, including adversarial perturbations [70, 118]. Unlike previous works, we have found that anti-aliasing with non-square filters is particularly effective against adversarial attacks while preserving clean accuracy. Additionally, using the average value from neighboring pixels, excluding the original pixel, has also proven effective. This is because adversarial perturbations are calculated on a pixel-wise basis and are sensitive to pixel changes. These two approaches greatly enhance the effectiveness of anti-aliasing. Even with simple anti-aliasing, we achieve moderate defense performance, underscoring the effectiveness of our approach. Although adversarial anti-aliasing can produce blurred images, the semantic features are preserved because the adversarial perturbation should remain imperceptible. Therefore, it effectively reduces the magnitude of adversarial perturbations while maintaining the semantic information necessary for classification. To maintain the resolution of the output image, we use padding, which is calculated as follows:

$$R_{out} = \lfloor R_{in} + 2 \times \text{Padding} - \text{filter\_size} \rfloor \tag{6.8}$$

Table 6.1: **The standard and robust accuracy against left: AutoAttack ($\ell_{\text{inf}} = 8/255$), right: PGD-EOT ($\ell_{\text{inf}} = 8/255$) on CIFAR-10.**

Method	Target Model	Standard Acc(%)	AutoAttack Acc(%)	PGD-EOT Acc(%)
Wu *et al.* [126]	WideResNet-28-10	85.36	59.18	62.16
Gowal *et al.* [35]	WideResNet-28-10	87.33	61.72	64.68
Rebuffi *et al.* [98]	WideResNet-28-10	87.50	65.24	68.89
Wang *et al.* [120]	WideResNet-28-10	84.85	71.18	68.36
Nie *et al.* [89]	WideResNet-28-10	89.23	71.03	46.84
Lee *et al.* [63]	WideResNet-28-10	90.16	70.47	55.82
Song *et al.* [109]	WideResNet-28-10	92.10	75.45	68.20
Ours	WideResNet-28-10	**92.54 ± 1.66**	**82.02 ± 1.17**	**80.86 ± 1.33**
Rebuffi *et al.* [98]	WideResNet-70-16	88.54	64.46	68.23
Gowal *et al.* [35]	WideResNet-70-16	88.74	66.60	69.48
Nie *et al.* [89]	WideResNet-70-16	91.04	71.84	51.13
Lee *et al.* [63]	WideResNet-70-16	90.43	66.06	56.88
Song *et al.* [109]	WideResNet-70-16	93.25	76.60	69.55
Ours	WideResNet-70-16	**93.42 ± 1.51**	**83.65 ± 2.90**	**81.60 ± 1.75**

where R is the shape of the data. We use stride = 1.

Adversarial Super-Resolution

During the adversarial anti-aliasing phase, we significantly reduce adversarial perturbations by directly decreasing pixel-wise modifications of adversarial examples. However, this approach may not be effective against unrestricted adversarial examples, as they are not generated by adding explicit perturbations. Additionally, blurring the images can negatively impact the clean accuracy of the target model. Super-resolution offers an effective way to recover high-quality images from our adversarial anti-aliased images. Previous super-resolution methods [33, 62] typically modify the original pixels of the low-resolution image and use the residual features of the original low-resolution image. These methods can inadvertently transfer negative effects from the adversarial examples to the final high-resolution images, making them ineffective for adversarial super-resolution. Diffusion-model-based super-resolution [100,138] provides a more isolated approach for super-resolution. These models generate high-resolution images through a denoising-like process over randomly sampled noise, using the low-resolution image as conditions.

In this work, we adopt the ResShift method by Yue et al. [138] for our super-resolution process. This super-resolution model can also incorporate benign priors for defense, as it is trained with the clean dataset of the target model. Figure 6.4 demonstrates that the proposed super-resolution method achieves results comparable to diffusion-based adversarial purification [109], which do not require the calculation of gradient.

Adversarial Purification

The proposed adversarial purification is performed by the combination of adversarial anti-aliasing and adversarial super-resolution. We resize the purified images after the adversarial super-resolution for the target model. Additionally, our approach does not require any training of the target model or the defense model.

$$y = \{f(\mathrm{SR}(\mathrm{AA}(x_{\mathrm{adv}})))\} \tag{6.9}$$

6.3.3 Discussions on Improved Time Efficiency

As previously discussed, employing the entire reverse process with adversarial example guidance is computationally intensive, while using only a partial reverse process diminishes defense performance. In this chapter, we propose a two-fold solution to address this issue. First, we introduce an **effective preprocessing approach, specifically anti-aliasing, to mitigate the impact of adversarial perturbations**. Previous research has shown that diffusion-based adversarial purification should avoid introducing adversarial perturbations into the diffusion model. Therefore, a more effective strategy is to remove some of these perturbations before feeding adversarial examples into the diffusion models. Unlike previous methods that directly utilize adversarial examples for purification, our approach offers a preliminary filtering step.

Second, we employ **diffusion-based super-resolution instead of diffusion-based image generation**. It is well-known that the reverse process of diffusion models is time-consuming, as illustrated in Figure 6.2, and the gradient calculation

exacerbates this issue. However, we may not require the entire reverse process for purification, given that we already have a reference adversarial example, which is also discussed in Nie et al. [89]'s work. Since adversarial perturbations are pixel-wise, we opt for a relatively lightweight generation task, namely super-resolution, which also focuses on pixel modification. The diffusion-based super-resolution method used in this chapter requires only tens of steps, compared to the hundreds or thousands of diffusion steps needed in previous works. With these two approaches, we significantly enhance the time efficiency of diffusion-based adversarial purification without compromising defense performance.

6.4 Experiments

6.4.1 Experimental Setup

Dataset and Target Models. We consider CIFAR-10 [58] and ImageNet [23] for major evaluation. For target models, we adopt WideResNet-28-10 and WideResNet-70-16 [140] for CIFAR-10 dataset and ResNet50 [43] for ImageNet dataset. These are commonly adopted backbones for adversarial robustness evaluation.

Comparisons. We compared our defense methods with various state-of-the-art defenses by the standardized benchmark: RobustBench [18]. We compare four diffusion-based adversarial purification methods: Nie et al.'s DiffPure [89], Wang et al.'s [120], Lee et al.'s [63] and Song et al.'s MimicDiffusion [109]. We mainly compare our method with MimicDiffusion as it is the current state-of-the-art method. We use the Score SDE [111] implementation of MimicDiffusion on CIFAR-10 for fair comparisons. The defense methods that use extra data are not compared for fairness. We only evaluate the adversarial purification methods against unrestricted adversarial attacks as the adversarial training's different threat model.

Attack Settings. We evaluate our method with both perturbation-based attacks and diffusion-based unrestricted adversarial attacks. For perturbation-based attacks, we select AutoAttack [19], PGD [84]. For diffusion-based unrestricted adversarial

attacks, we use DiffAttack [13] and AdvDiff [20] for comparisons. DiffAttack is only evaluated on the ImageNet dataset according to the original chapter. To ensure a fair comparison with previous diffusion-based adversarial purification, we include the evaluation against the adaptive attack, i.e., reverse pass differentiable approximation (BPDA) [45]. We also evaluate the performance against PGD+EOT that is discussed in [63]. On CIFAR-10, the attack settings follow DiffPure [89]. On ImageNet, we randomly sample 5 images from each class and average over 10 runs. The PGD+EOT settings all follow Lee et al. [63].

Implementation Details. We adopt the mean filter with $[[1, 1], [1, 1]]$ for adversarial anti-aliasing on CIFAR-10, and $[[1, 1, 1, 1, 1], [1, 1, 0, 1, 1], [1, 1, 1, 1, 1]]$ in ImageNet. ResShift [138] is utilized for adversarial super-resolution. We use the official Score SDE [111] checkpoint for CIFAR-10 and LDM [100] checkpoint for ImageNet to generate UAEs.

Evaluation Metrics. Following Nie et al. [89], we use *standard accuracy* and *robust accuracy* as the evaluation metrics. Both are calculated according to the top-1 classification accuracy.

6.4.2 Attack Performance

CIFAR-10

Perturbation-based Adversarial Attack. Table 6.1 presents the defense performance against AutoAttack ($\ell_{\text{inf}} = 8/255$) on the CIFAR-10 dataset. The results demonstrate that our proposed method achieves better standard accuracy and robust accuracy than previous attack methods. Because images in the CIFAR-10 dataset are only with 32×32 resolution, we set our anti-aliasing filter to a relatively small size. Table 6.2 indicates that the robustness performance of the proposed method is on par with the state-of-the-art method [89]. This finding suggests that our method is more suitable for high-resolution images, as 32×32 may not be large enough to effectively extract the semantic shape for our approach. However, we can further enhance our performance by incorporating adversarial purification techniques from previous work.

Table 6.2: **The standard and robust accuracy against BPDA ($\ell_{\mathbf{inf}} = 8/255$) on the CIFAR-10 dataset with WideResNet-28-10 as the target model.**

Method	Purification	Standard Acc(%)	Robust Acc(%)
Nie *et al.* [89]($t^* = 0.0075$)	Diffusion	91.38	77.62
Nie *et al.* [89]($t^* = 0.1$)	Diffusion	89.23	**81.56**
Wang *et al.* [120]	Diffusion	90.36	77.31
Song *et al.* [109]	Diffusion	91.41	76.45
Ours	Diffusion	**91.52 ± 1.28**	81.24 ± 2.51

Table 6.3: **The standard and robust accuracy against AdvDiff on the CIFAR-10 dataset.**

Method	Target Model	Standard Acc(%)	Robust Acc(%)
Nie *et al.* [89]	WideResNet-28-10	95.42	21.56
Wang *et al.* [120]	WideResNet-28-10	95.86	26.68
Lee *et al.* [63]	WideResNet-28-10	95.29	24.94
Song *et al.* [109]	WideResNet-28-10	96.21	23.23
Ours	WideResNet-28-10	**96.80 ± 0.37**	**33.97 ± 0.77**

Our defense's performance against PGD-EOT showcases its ability to defend against adaptive attacks. This is because our approach focuses on extracting and recovering the semantic features from adversarial images, rather than inferring and denoising the adversarial perturbations. As a result, our defense maintains similar effectiveness against both adaptive and standard attacks.

Unrestricted Adversarial Attack. Unrestricted adversarial examples on the CIFAR-10 dataset are challenging to defend against, as shown in Table 6.3. Our purification method outperforms the previous adversarial purification approach [109] by an average of 10%, validating the effectiveness of our proposed defense.

ImageNet

Perturbation-based Adversarial Attack. Tables 6.4 and 6.5 demonstrate that the proposed defense method achieves significantly higher performance in both standard accuracy and robust accuracy. Our defense's standard accuracy notably sur-

Table 6.4: **The standard and robust accuracy against AutoAttack ($\ell_{\mathbf{inf}} = 8/255$) on the ImageNet dataset.**

Method	Target Model	Standard Acc(%)	Robust Acc(%)
Engstrom *et al.* [18]	ResNet50	62.56	31.06
Wong *et al.* [125]	ResNet50	55.62	26.95
Salman *et al.* [102]	ResNet50	64.02	37.89
Bai *et al.* [3]	ResNet50	67.38	35.51
Nie *et al.* [89]	ResNet50	68.22	43.89
Song *et al.* [109]	ResNet50	66.92	61.53
Ours	ResNet50	**75.28 ± 1.06**	**67.61 ± 1.95**

Table 6.5: **The standard and robust accuracy against left: PGD ($\ell_{\mathbf{inf}} = 4/255$), right: PGD+EOT ($\ell_{\mathbf{inf}} = 4/255$) on ImageNet dataset.**

Method	Target Model	Standard Acc(%)	PGD Acc(%)	PGD+EOT Acc(%)
Wong *et al.* [125]	ResNet50	55.62	26.24	30.51
Salman *et al.* [102]	ResNet50	64.02	34.96	38.62
Bai *et al.* [3]	ResNet50	67.38	40.27	43.42
Nie *et al.* [89]	ResNet50	68.22	42.88	38.71
Lee *et al.* [63]	ResNet50	70.74	46.31	42.15
Wang *et al.* [120]	ResNet50	70.17	68.78	40.22
Song *et al.* [109]	ResNet50	66.92	62.16	52.66
Ours	ResNet50	**75.28 ± 1.06**	**69.75 ± 2.61**	**66.87 ± 1.85**

passes previous work, further validating that adversarial super-resolution effectively leverages prior knowledge from the training dataset to achieve better classification accuracy. Adversarial anti-aliasing proves to be particularly effective on the ImageNet dataset, where the filter successfully blurs adversarial perturbations in the detailed pixels of adversarial examples. The performance against PGD-EOT further validates the effectiveness of our proposed defense pipeline.

Unrestricted Adversarial Attack. We present the defense performance of various methods against the unrestricted adversarial attack AdvDiff and DiffAttack in Table 6.6 and 6.7. The results indicate that current defenses are ineffective against the recently proposed unrestricted adversarial attacks. The high standard accuracy can

be attributed to the strong generative performance of benign diffusion models. Our defense method is capable of achieving significantly higher robust accuracy compared to previous defenses while preserving the standard accuracy.

Table 6.6: **The standard and robust accuracy against AdvDiff on the ImageNet dataset.**

Method	Target Model	Standard Acc(%)	Robust Acc(%)
Nie *et al.* [89]	ResNet50	91.48	24.82
Wang *et al.* [120]	ResNet50	92.31	26.74
Lee *et al.* [63]	ResNet50	91.80	25.34
Song *et al.* [109]	ResNet50	92.54	25.35
Ours	ResNet50	**97.83 ± 1.36**	**42.21 ± 3.41**

Table 6.7: **The standard and robust accuracy against DiffAttack on the ImageNet dataset.**

Method	Target Model	Standard Acc(%)	Robust Acc(%)
Nie *et al.* [89]	ResNet50	68.22	59.15
Wang *et al.* [120]	ResNet50	69.54	62.33
Lee *et al.* [63]	ResNet50	70.74	61.56
Song *et al.* [109]	ResNet50	66.92	60.17
Ours	ResNet50	**75.28 ± 1.06**	**65.51 ± 1.33**

6.4.3 Time efficiency

We evaluate the average time for defending against one adversarial example as shown in Table 6.8. The results indicate that our proposed method achieves better robust accuracy with significantly lower time costs, as it does not require any gradient calculations over the diffusion model. Notably, our adversarial anti-aliasing can defend against approximately 57% of adversarial examples in just $3e^{-3}$ seconds. Furthermore, we can enhance the defense performance of our method by combining it with previous purification methods, with only a minimal tradeoff in time cost.

Table 6.8: **The average time cost of defending one image against PGD ($\ell_{\mathbf{inf}} = 4/255$) on the ImageNet dataset.**

Method	Defend Method	Time Cost(s)	Robust Acc(%)
Nie *et al.* [89]	Diffusion	13.3	42.88
Wang *et al.* [120]	Diffusion	62.8	68.78
Lee *et al.* [63]	Diffusion	32.4	46.31
Song *et al.* [109]	Diffusion	146.1	62.16
Ours	Adversarial Anti-Aliasing	$3e^{-3}$	57.61
+	Adversarial Super-Resolution	**1.1**	**69.75**

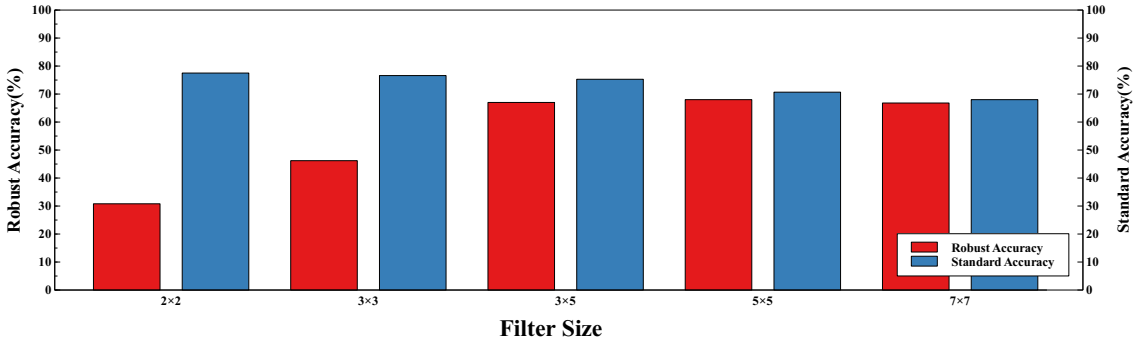


Figure 6.5: **The ablation study of filter size.** The weight of the filter at each position is set to 1 except for the center, which we set to 0.

6.4.4 Ablation Study

We perform ablation studies to validate the performance of the proposed methods. We evaluate the defense method against AutoAttack ($\ell_{\text{inf}} = 8/255$) on the ImageNet dataset by default.

Adversarial Anti-Aliasing. Despite the satisfactory robustness performance of the proposed adversarial anti-aliasing, the choice of filter settings is critical for optimal defense performance. We present the defense performance with different filters in Figure 6.5. The results indicate a tradeoff between robust accuracy and standard accuracy. Robust accuracy tends to stabilize when using a filter larger than 3×3 in size. Therefore, it is relatively straightforward to identify a suitable filter with a few attempts. Furthermore, the filter settings are generalized across different adversarial attacks within the same dataset, as demonstrated in Tables 6.4, 6.5, and 6.6.

Table 6.9: **The ablation study of proposed methods.**

(a) **The ablation study of proposed adversarial super-resolution.**

Method	Robust Acc(%)
Nie *et al.* [89]	43.89
Song *et al.* [109]	61.53
Adversarial AA	55.85
Adversarial SR	41.23
Adversarial AA+SR	**67.61**

(b) **The performance of integrating our method with previous adversarial purification.**

Method	Robust Acc(%)
Nie *et al.* [89]	43.89
+ Ours	69.44
Song *et al.* [109]	61.53
+ Ours	72.18

Adversarial Super-Resolution. The proposed adversarial super-resolution achieves a similar purification function to previous diffusion-based adversarial purification methods, but without the need for computationally expensive gradient calculations. Table 6.9a demonstrates that our method slightly outperforms traditional adversarial purification when using anti-aliased adversarial examples as input. However, it is crucial to use anti-aliased adversarial examples for optimal performance in adversarial super-resolution, as we do not account for the adversarial gradient during the super-resolution process.

Adversarial Purification. We can enhance diffusion-based adversarial purification methods from previous works by replacing the adversarial input with the adversarial examples after the proposed purification. The processed adversarial examples are more benign and closer to the clean images, thereby enabling better purification performance, as demonstrated in Table 6.9b.

6.5 Conclusion

In this chapter, we present an effective and efficient adversarial defense method against both perturbation-based and unrestricted adversarial attacks. The proposed techniques, adversarial anti-aliasing and adversarial super-resolution, effectively eliminate adversarial modifications and recover benign images with minimal computational overhead. Comprehensive experiments on the CIFAR-10 and ImageNet datasets

validate that our proposed defense outperforms state-of-the-art defense methods. Our work demonstrates that simple adversarial anti-aliasing can achieve moderate model robustness with almost no additional cost. Furthermore, the proposed super-resolution method can perform adversarial purification without requiring the calculation of the diffusion model's gradient. We hope our work will serve as a baseline for the further development of adversarial defenses.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

Generative models, particularly diffusion models, hold significant potential for advancing AI Generated Content (AIGC) research. Recent progress in Multi-model Large Language Models (MLLMs) with image and text input has further highlighted the importance of diffusion models. There is an urgent need to thoroughly investigate the adversarial capabilities of diffusion models to ensure the secure and robust deployment of AIGC models.

In the work of unrestricted adversarial attacks, we propose AdvDiff, which provides a interpretable unrestricted adversarial attack using diffusion models by following the benign diffusion generation process. Existing diffusion-based adversarial attacks typically use the gradients of the target model's loss function to guide the generation process. However, these methods can compromise the generation quality of diffusion models, resulting in low-quality adversarial examples. AdvDiff offers two effective forms of adversarial guidance: adversarial guidance and noise sampling guidance. These strategies follow the diffusion generation process and enhance the generation of adversarial examples by increasing the conditional likelihood of the target attack label. Our experiments demonstrate that AdvDiff significantly improves the generation quality of diffusion-based unrestricted adversarial attacks across various evaluation metrics. Additionally, we achieve a higher attack success rate in

both white-box and black-box scenarios, with improved time efficiency in generating adversarial examples.

In the work of no-box adversarial attacks, we propose a practical approach using a diffusion classification model and a diffusion model. The no-box attack threat model prohibits access to the target model's training dataset, whereas existing methods still rely on a small sub-dataset to train a substitute model. To perform no-box adversarial attacks without any dataset access, we harness the generative capabilities of diffusion models. Our training dataset is entirely generated by diffusion models using only label information. Subsequently, we employ the diffusion classification model as the substitute model, fine-tuning it with model uncertainty to enhance the transferability of adversarial examples. The no-box unrestricted adversarial examples are generated by the diffusion model using ensemble-like Monte Carlo sampling methods from the substitute model. Through extensive experiments, we demonstrate that our no-box adversarial attack achieves state-of-the-art performance in both no-box and black-box adversarial attack scenarios. Additionally, our approach improves generation quality and performance against various defenses.

In the work of 3D adversarial attacks, we propose a transferable adversarial 3D shape completion method using diffusion models. Creating natural 3D adversarial point clouds is more challenging than working with 2D images, as perturbations in 3D point clouds lead to shifts in 3D coordinates. Moreover, existing adversarial attacks often struggle to successfully execute black-box attacks on recently developed 3D classifiers. Our proposed 3D adversarial attack utilizes the strong generative capabilities of diffusion models to produce adversarial examples within the shape completion task. To enhance black-box adversarial attack performance, we employ a Monte Carlo estimate over multiple down-sampled point clouds to infer the model's gradient, and we aggregate logits from multiple substitute models. Our adversarial guidance is applied only to selected critical points, identified by proposed saliency scores, to preserve the quality of point cloud generation. Experimental results demonstrate that our adversarial 3D shape completion method achieves leading performance against a wide range of black-box 3D target classifiers.

In the work of diffusion-based adversarial purification, we introduce a gradient-free approach that incorporates adversarial anti-aliasing and adversarial super-resolution. Diffusion-based adversarial purification methods have shown promising defense capabilities due to their denoising-like generation process. However, these defenses often struggle against newly developed unrestricted adversarial attacks and suffer from poor time efficiency due to the iterative nature of diffusion timesteps. Our defense offers an effective and efficient preprocessing step with adversarial anti-aliasing to extract semantic shapes from both perturbation-based and unrestricted adversarial attacks. We then deploy super-resolution diffusion models to leverage the clean prior from benign data to purify adversarial examples. Experimental results demonstrate that our proposed purification method significantly improves the time efficiency of diffusion-based adversarial purification across various datasets. We achieve state-of-the-art performance in defending against both perturbation-based and unrestricted adversarial attacks. Our approach provides a new pipeline for diffusion-based adversarial defense with enhanced time efficiency.

7.2 Future Work

With the development of text-to-image diffusion models, such as Stable Diffusion, and MLLMs, it enables stronger adversarial attacks with diffusion models. Our research aims to further explore these techniques. In future work, we plan to focus on utilizing text-to-image diffusion models to generate more aggressive adversarial examples by incorporating both adversarial guidance and text prompts. Additionally, we intend to train adversarial LoRA models for efficient adversarial example generation. Simultaneously, we will explore attacking MLLMs using diffusion models. Enhancing diffusion-based purification methods to defend against unrestricted adversarial attacks is also a key objective to address security concerns.

7.2.1 Effective Adversarial Sampling with Prompt

Text-to-image diffusion models enable more precise and consistent image synthesis based on user prompts, offering the potential for creating more camouflaged and flexible adversarial examples compared to using adversarial guidance alone. In future work, we plan to design aggressive prompts that incorporate adversarial gradients. By using adversarial prompts, we aim to reduce the reliance on adversarial guidance during sampling and enhance the quality of the generated adversarial examples. Furthermore, we intend to design prompts that utilize ensemble logits and transferable loss objectives to improve the transferability of adversarial attacks.

7.2.2 Training Adversarial LoRA

Low-Rank Adaptation (LoRA) of Large Language Models facilitates effective and efficient fine-tuning for text-to-image diffusion models, particularly for generating content from specific domains. Current diffusion-based adversarial attacks suffer from low time efficiency when generating adversarial examples. Training a LoRA can help reduce the computational overhead during adversarial sampling. We plan to propose adversarial attack methods by training adversarial LoRAs for efficient sampling of adversarial examples. Our approach will involve training LoRAs with adversarial objectives against a target model. Once trained, we can directly generate adversarial examples without the need for adversarial guidance during diffusion sampling. With a trained LoRA, we will be able to generate adversarial examples with significantly lower time costs compared to existing diffusion-based adversarial attacks. Additionally, the generation quality is improved without relying on adversarial guidance.

7.2.3 Breaking through Multi-model Large Language Models

LLMs and MLLMs have demonstrated remarkable performance in content generation and autonomous task solving. The development of LLMs is becoming a key trend in the advancement of AI. However, LLMs have been shown to be vulnerable to adversarial attacks or exploitation by malicious users. These security risks significantly

impact the deployment of LLMs in security-related applications. To thoroughly investigate the security of LLMs and MLLMs, we plan to propose effective attacks against MLLMs using diffusion models. As MLLMs accept both text and image inputs from users, it is possible to use images generated by diffusion models to breach MLLM defenses and generate content prohibited by user policies. We will utilize gradients from MLLMs to create adversarial guidance for the adversarial sampling of diffusion models. Our attack aims to elicit sensitive responses from MLLMs using malicious prompts and adversarial examples from diffusion models.

Another adversarial attack strategy against MLLMs involves adversarial fine-tuning using data generated by diffusion models. MLLMs support user-driven fine-tuning to provide customized services based on a user's dataset or specified tasks. This fine-tuning capability also allows us to manipulate the functionality of MLLMs. In the future, we plan to generate malicious content from diffusion models through MLLM fine-tuning. We will create adversarial examples from diffusion models to induce MLLMs to output malicious content. These adversarial examples can be used as a fine-tuning dataset to compromise the defenses of MLLMs and provoke malicious responses.

7.2.4 Robust Adversarial Purification against Unrestricted Adversarial Attack

Unrestricted adversarial examples generated by diffusion models pose significant security concerns for adversarial defenses, as they employ different threat models compared to traditional perturbation-based attacks. In future work, we plan to propose a diffusion-based adversarial purification method specifically designed to counter unrestricted adversarial attacks. Our defense strategy will begin with preprocessing to extract the semantic shape from adversarial examples. We will then use latent inversion with a text-to-image diffusion model, employing safety-oriented prompts to guide the model in generating benign images. Our adversarial defense will focus on reconstructing the semantic objects from adversarial examples and will be more

generalized across different attack threat models.

References

[1] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Crafting adversarial perturbations via transformed image component swapping. *IEEE Transactions on Image Processing*, 31:7338–7349, 2022.

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

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

[4] Shumeet Baluja and Ian Fischer. Learning to attack: Adversarial transformation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[6] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4175–4186, 2022.

[7] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

[8] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[9] Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane, Luca Bortolussi, and Guido Sanguinetti. Robustness of bayesian neural networks to gradient-based attacks. *Advances in Neural Information Processing Systems*, 33:15602–15613, 2020.

[10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[11] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[12] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020.

[13] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *arXiv preprint arXiv:2305.08192*, 2023.

[14] Qi Chen, Wei Wang, Fangyu Wu, Suparna De, Ruili Wang, Bailing Zhang, and Xin Huang. A survey on an emerging area: Deep learning for smart city data. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(5):392–410, 2019.

[15] Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4562–4572, 2023.

[16] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *arXiv preprint arXiv:2305.10665*, 2023.

[17] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

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

[19] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[20] Xuelong Dai, Kaisheng Liang, and Bin Xiao. Advdiff: Generating unrestricted adversarial examples using diffusion models. In *European Conference on Computer Vision*, pages 93–109. Springer, 2025.

[21] Xuelong Dai and Bin Xiao. Transferable 3d adversarial shape completion using diffusion models. In *European Conference on Computer Vision*, pages 392–408. Springer, 2025.

[22] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[24] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[25] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[26] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[27] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[29] Michael Egmont-Petersen, Dick de Ridder, and Heinz Handels. Image processing with neural networks—a review. *Pattern recognition*, 35(10):2279–2301, 2002.

[30] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[31] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022.

[32] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[33] Shangqi Gao and Xiahai Zhuang. Multi-scale deep neural networks for real image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[34] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[35] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

[36] Martin Gubri, Maxime Cordy, Mike Papadakis, Yves Le Traon, and Koushik Sen. Lgv: Boosting adversarial example transferability from large geometric vicinity. *arXiv preprint arXiv:2207.13129*, 2022.

[37] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021.

[38] Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. In *NeurIPS*, 2020.

[39] Yiwen Guo, Qizhang Li, Wangmeng Zuo, and Hao Chen. An intermediate-level attack framework on the basis of linear regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[40] Abdullah Hamdi, Sara Rojas, Ali Thabet, and Bernard Ghanem. Advpc: Transferable adversarial perturbations on 3d point clouds. In *Proceedings of the European Conference on Computer Vision*, pages 241–257, 2020.

[41] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

[42] Bangyan He, Jian Liu, Yiming Li, Siyuan Liang, Jingzhi Li, Xiaojun Jia, and Xiaochun Cao. Generating transferable 3d adversarial point cloud via random perturbation factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 764–772, 2023.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[44] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

[46] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[47] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[48] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[49] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[50] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[51] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, and Nenghai Yu. Shape-invariant 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15335–15344, June 2022.

[52] Yi Huang and Adams Wai-Kin Kong. Transferable adversarial attack based on integrated gradients. *arXiv preprint arXiv:2205.13152*, 2022.

[53] Qiufan Ji, Lin Wang, Cong Shi, Shengshan Hu, Yingying Chen, and Lichao Sun. Benchmarking and analyzing robust point cloud recognition: Bag of tricks for defending adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4295–4304, 2023.

[54] Gaojie Jin, Xinping Yi, Dengyu Wu, Ronghui Mu, and Xiaowei Huang. Randomized adversarial training via taylor expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16447–16457, 2023.

[55] Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *Proceedings of the International Conference on Robotics and Automation*, pages 8248–8254. IEEE, 2019.

[56] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[57] İbrahim Kök, Mehmet Ulvi Şimşek, and Suat Özdemir. A deep learning model for air quality prediction in smart cities. In *Proceedings of the International Conference on Big Data*, pages 1983–1990. IEEE, 2017.

[58] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[59] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in neural information processing systems*, pages 1097–1105, 2012.

[60] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[61] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[62] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[63] Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 134–144, 2023.

[64] Dongze Li, Wei Wang, Hongxing Fan, and Jing Dong. Exploring adversarial fake images on face manifold. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5789–5798, 2021.

[65] Qizhang Li, Yiwen Guo, and Hao Chen. Practical no-box adversarial attacks against dnns. *Advances in Neural Information Processing Systems*, 33:12849–12860, 2020.

[66] Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Making substitute models more bayesian can enhance transferability of adversarial examples. *arXiv preprint arXiv:2302.05086*, 2023.

[67] Xiao Li, Ziqi Wang, Bo Zhang, Fuchun Sun, and Xiaolin Hu. Recognizing object by components with human prior knowledge enhances adversarial robustness of deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8861–8873, 2023.

[68] Yanjie Li, Yiquan Li, Xuelong Dai, Songtao Guo, and Bin Xiao. Physical-world optical adversarial attacks on 3d face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24699–24708, 2023.

[69] Yingzhen Li, John Bradshaw, and Yash Sharma. Are generative classifiers more robust to adversarial attacks? In *International Conference on Machine Learning*, pages 3804–3814. PMLR, 2019.

[70] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and Xiaofeng Wang. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Transactions on Dependable and Secure Computing*, 18(1):72–85, 2018.

[71] Kaisheng Liang and Bin Xiao. Styless: boosting the transferability of adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8163–8172, 2023.

[72] Yuanbang Liang, Jing Wu, Yu-Kun Lai, and Yipeng Qin. Exploring and exploiting hubness priors for high-quality gan latent sampling. In *International Conference on Machine Learning*, pages 13271–13284. PMLR, 2022.

[73] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1778–1787, 2018.

[74] Guang Lin, Chao Li, Jianhai Zhang, Toshihisa Tanaka, and Qibin Zhao. Adversarial training on purification (atop): Advancing both robustness and generalization. *arXiv preprint arXiv:2401.16352*, 2024.

[75] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

[76] Aishan Liu, Xianglong Liu, Hang Yu, Chongzhi Zhang, Qiang Liu, and Dacheng Tao. Training robust deep neural networks via adversarial noise propagation. *IEEE Transactions on Image Processing*, 30:5769–5781, 2021.

[77] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

[78] Daniel Liu, Ronald Yu, and Hao Su. Extending adversarial attacks and defenses to deep 3d point cloud classifiers. In *Proceedings of the International Conference on Image Processing*, pages 2279–2283. IEEE, 2019.

[79] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[80] Shao-Yuan Lo and Vishal M Patel. Defending against multiple and unforeseen adversarial videos. *IEEE Transactions on Image Processing*, 31:962–973, 2021.

[81] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[82] Chenxu Luo, Xiaodong Yang, and Alan Yuille. Self-supervised pillar motion learning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3183–3192, June 2021.

[83] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021.

[84] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[85] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, pages 723–727. IEEE, 2011.

[86] David P Morgan and Christopher L Scofield. Neural networks and speech processing. In *Neural Networks and Speech Processing*, pages 329–348. Springer, 1991.

[87] Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29:1711–1724, 2019.

[88] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 262–271, 2020.

[89] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

[90] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the International conference on machine learning*, pages 2642–2651, 2017.

[91] Will Penny and David Frost. Neural networks in clinical medicine. *Medical Decision Making*, 16(4):386–398, 1996.

[92] Omid Poursaeed, Tianxing Jiang, Harry Yang, Serge Belongie, and Ser-Nam Lim. Robustness and generalization via generative adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15711–15720, 2021.

[93] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4422–4431, 2018.

[94] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022.

[95] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[96] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5099–5108, 2017.

[97] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 19–37. Springer, 2020.

[98] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

[99] Jiawei Ren, Liang Pan, and Ziwei Liu. Benchmarking and analyzing point cloud classification under corruptions. In *International Conference on Machine Learning*, pages 18559–18575. PMLR, 2022.

[100] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[101] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[102] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *Proceedings of the Advances in Neural Information Processing Systems*, 2020.

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

[104] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[105] Mingwen Shao, Lingzhuang Meng, Yuanjian Qiao, Lixu Zhang, and Wangmeng Zuo. Data-free black-box attack based on diffusion model. *arXiv preprint arXiv:2307.12872*, 2023.

[106] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[107] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[108] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[109] Kaiyu Song, Hanjiang Lai, Yan Pan, and Jian Yin. Mimicdiffusion: Purifying adversarial perturbation via mimicking clean diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24665–24674, 2024.

[110] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 8322–8333, 2018.

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

[112] Chenghao Sun, Yonggang Zhang, Wan Chaoqun, Qizhou Wang, Ya Li, Tongliang Liu, Bo Han, and Xinmei Tian. Towards lightweight black-box attack against deep neural networks. *Advances in Neural Information Processing Systems*, 35:19319–19331, 2022.

[113] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[114] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[115] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2820–2828, 2019.

[116] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[117] Tzungyu Tsai, Kaichen Yang, Tsung-Yi Ho, and Yier Jin. Robust adversarial objects against deep learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 954–962, 2020.

[118] Cristina Vasconcelos, Hugo Larochelle, Vincent Dumoulin, Rob Romijnders, Nicolas Le Roux, and Ross Goroshin. Impact of aliasing on generalization in deep convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10529–10538, 2021.

[119] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[120] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022.

[121] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16158–16167, 2021.

[122] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics*, 38(5):1–12, 2019.

[123] Yuxin Wen, Jiehong Lin, Ke Chen, CL Philip Chen, and Kui Jia. Geometry-aware generation of adversarial point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2984–2999, 2020.

[124] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[125] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *Proceedings of the International Conference on Learning Representations*, 2020.

[126] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969, 2020.

[127] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 9621–9630, 2019.

[128] Ziyi Wu, Yueqi Duan, He Wang, Qingnan Fan, and Leonidas J Guibas. If-defense: 3d adversarial point cloud defense via implicit function based restoration. *arXiv preprint arXiv:2010.05272*, 2020.

[129] Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9136–9144, 2019.

[130] Tiange Xiang, Chaoyi Zhang, Yang Song, Jianhui Yu, and Weidong Cai. Walk in the cloud: Learning curves for point clouds shape analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 915–924, October 2021.

[131] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 819–828, 2020.

[132] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

[133] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[134] Mutian Xu, Junhao Zhang, Zhipeng Zhou, Mingye Xu, Xiaojuan Qi, and Yu Qiao. Learning geometry-disentangled representation for complementary

understanding of 3d object point cloud. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3056–3064, 2021.

[135] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[136] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, pages 12062–12072. PMLR, 2021.

[137] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[138] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024.

[139] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[140] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[141] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: latent point diffusion models for 3d shape generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 10021–10039, 2022.

[142] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[143] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[144] Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15115–15125, 2022.

[145] Yue Zhao, Yuwei Wu, Caihua Chen, and Andrew Lim. On isometry robustness of deep 3d point cloud models under adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1201–1210, 2020.

[146] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Point-cloud saliency maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1598–1606, 2019.

[147] Hang Zhou, Dongdong Chen, Jing Liao, Kejiang Chen, Xiaoyi Dong, Kunlin Liu, Weiming Zhang, Gang Hua, and Nenghai Yu. Lg-gan: Label guided adversarial network for flexible targeted attack of point cloud based deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10356–10365, 2020.

[148] Hang Zhou, Kejiang Chen, Weiming Zhang, Han Fang, Wenbo Zhou, and Nenghai Yu. Dup-net: Denoiser and upsampler network for 3d adversarial point clouds defense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1961–1970, 2019.

[149] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021.

[150] Mo Zhou, Tianyi Liu, Yan Li, Dachao Lin, Enlu Zhou, and Tuo Zhao. Toward understanding the importance of noise in training neural networks. In *International Conference on Machine Learning*, pages 7594–7602. PMLR, 2019.