Transferable 3D Adversarial Shape Completion using Diffusion Models

Xuelong Dai[1] and Bin Xiao[1]

The Hong Kong Polytechnic University
xuelong.dai@connect.polyu.hk, b.xiao@polyu.edu.hk

Abstract. Recent studies that incorporate geometric features and transformers into 3D point cloud feature learning have significantly improved the performance of 3D deep-learning models. However, their robustness against adversarial attacks has not been thoroughly explored. Existing attack methods primarily focus on white-box scenarios and struggle to transfer to recently proposed 3D deep-learning models. Even worse, these attacks introduce perturbations to 3D coordinates, generating unrealistic adversarial examples and resulting in poor performance against 3D adversarial defenses. In this paper, we generate high-quality adversarial point clouds using diffusion models. By using partial points as prior knowledge, we generate realistic adversarial examples through shape completion with adversarial guidance. The proposed adversarial shape completion allows for a more reliable generation of adversarial point clouds. To enhance attack transferability, we delve into the characteristics of 3D point clouds and employ model uncertainty for better inference of model classification through random down-sampling of point clouds. We adopt ensemble adversarial guidance for improved transferability across different network architectures. To maintain the generation quality, we limit our adversarial guidance solely 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.

Keywords: 3D Black-box Adversarial Attacks · Diffusion Models · Model Uncertainty

1 Introduction

Deep-learning models have demonstrated their overwhelming performance on 2D [13, 23] and 3D computer vision [10, 28, 36] 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 [2,5,19,21,25], 3D point cloud deep learning is also susceptible to adversarial attacks [22, 35, 40]. These 3D adversarial attacks generate adversarial examples by introducing perturbations to the xyz coordinates.

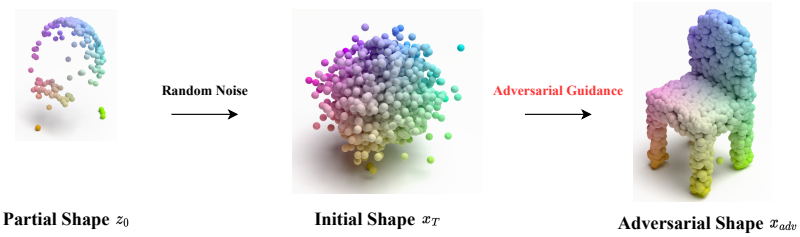

Fig. 1: The adversarial shape completion. Starting from the partial shape z_0, we construct our adversarial shape x_{adv} by utilizing diffusion models with proposed adversarial guidance.

However, such perturbations often lead to a significant degradation in visual quality, which can be easily detected by humans. Subsequent studies [15, 32, 39] 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 [16]. Moreover, most existing attacks primarily focus on white-box settings, limiting their practicality in real-world scenarios. Existing black-box attacks [11, 12] mainly target early 3D point cloud deep-learning models, leaving a substantial gap in the learning between adversarial and benign models.

In this paper, our objective is to execute high-quality black-box 3D adversarial 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 [7, 29] and 3D point clouds [38, 43]. It has been demonstrated that 2D diffusion models can generate adversarial examples [4, 6] 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 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. [18] 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.

2 Background

2.1 3D Point Cloud Classification

The field of 3D point cloud classification poses unique challenges compared to 2D image classification, primarily due to the disorder and discrete nature of 3D point cloud data. Traditional 2D deep-learning models find it challenging to process such data efficiently. PointNet [26] stood out as the pioneering approach to address the challenge of 3D feature learning. PointNet largely enhanced the performance of 3D classification tasks by employing a symmetric function that effectively extracts features from the inherently disorderly input of 3D point cloud data. The success of PointNet [26] has sparked a surge in research focused on 3D deep learning. In an effort to enhance the performance of 3D feature learning, researchers have integrated graph convolutional operations to extract features from both local neighbors and the global shape of the point cloud. Two notable state-of-the-art 3D deep learning networks, PointNet++ [27] and DGCNN [31], had successfully adopted graph convolutional layers. Recent approaches have further improved 3D point cloud classification by incorporating geometry features and transformers [10, 28, 36]. These advancements contribute to achieving satisfying performance in the challenging task of 3D point cloud classification.

2.2 3D Point Cloud Adversarial Attack

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. [22] and Xiang et al. [35], involve adding points generated from 2D FGSM, PGD, and C&W attack methods. Zheng et al. [40] 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 [39], GeoA3 [32], SI-Adv [15], and PF-Attack [12] achieved imperceptible shifting by leveraging geometric and shape information from clean point clouds. LG-GAN [41] and AdvPC [11] 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.

3 Preliminary

3.1 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:

$$\min D(x, x_{adv}), \qquad \text{s.t. } f(x_{adv}) = y_{adv} \tag{1}$$

Equation 1 is designed to generate an imperceptible adversarial example x_{adv} from the original point cloud x. This paper primarily concentrates on untargeted attacks, where y_{adv} can be any label distinct from the ground truth label y.

3.2 3D Point Cloud Generation and Completion

Recent advancements in diffusion models [7,14,17,29] applied to 2D image generation have showcased remarkable performance in terms of both generation quality and diversity. Likewise, recent studies on 3D diffusion models [24,38,43] 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:

$$\begin{aligned} q(x_{0:T}) &= q(x_0) \prod_{t=1}^{T} q(x_t|x_{t-1}) \\ p_\theta(x_{0:T}) &= p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \end{aligned} \tag{2}$$

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 $\beta_1, \ldots, \beta_T$:

$$\begin{aligned} q(x_t|x_{t-1}) &:= \mathcal{N}(x_t : \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}) \\ p_\theta(x_{t-1}|x_t) &:= \mathcal{N}(x_{t-1} : \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}) \end{aligned} \tag{3}$$

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 $z_0 \in \mathcal{P}^{K_p \times 3}$ [43]. The *forward diffusion process* and *reverse generative process* are formulated as:

$$\begin{aligned} q(\tilde{x}_t|\tilde{x}_{t-1}, z_0) &:= \mathcal{N}(\tilde{x}_t : \sqrt{1-\beta_t} \tilde{x}_{t-1}, \beta_t \mathbf{I}) \\ p_\theta(\tilde{x}_{t-1}|\tilde{x}_t, z_0) &:= \mathcal{N}(\tilde{x}_{t-1} : \mu_\theta(x_t, z_0, t), \sigma_t^2 \mathbf{I}) \end{aligned} \tag{4}$$

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 paper, 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.

4 Methodology

4.1 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 [4, 6], 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:

$$p_\theta(\tilde{x}_{t-1}|\tilde{x}_t, z_0) := \mathcal{N}(\tilde{x}_{t-1} : \mu_\theta(x_t, z_0, t), \beta_t \mathbf{I}) - a\beta_t \nabla_{x_t} \mathcal{L}(f(x_t), y) \tag{5}$$

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 [4].

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 [43]:

$$\tilde{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\tilde{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \tilde{\alpha}_t}} \epsilon_\theta(\tilde{x}_t, z_0, t) \right) + \sqrt{\beta_t}\varepsilon, \tag{6}$$

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

4.2 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 [1,20] 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 [18] demonstrates improved adversarial transferability. However, the application 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 [42]. Therefore, we are able to straightforwardly adopt model uncertainty to 3D deep-learning models with the *MC dropout*-like [8] 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:

$$\nabla_{x_t}\mathcal{L}_{\mathrm{MU}}(f(x_t),y)) = \frac{1}{M}\sum_{s=1}^{M}\nabla_{x_s}\mathcal{L}(f(x_s),y) \tag{7}$$

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

$$P_i(\tilde{x}_t) = \{1_x | x \in \tilde{x}_t, 1_x \sim Ber(0.5)\} \tag{8}$$

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. Formally, with n_{ens} substitute models, the ensemble adversarial objective function is defined as:

$$\mathcal{L}(f_{ens}(x_t), y) = -\log(\mathrm{softmax}\sum_{n=1}^{n_{\mathrm{ens}}} w_n p_{f_n}(y|x_t)) \tag{9}$$

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 [40] has shown that individual 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:

$$\mathrm{score}_x = \sum_{3}\frac{\partial\mathcal{L}(f(x_t),y)}{\partial x} \tag{10}$$

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.

4.3 Transferable 3D Adversarial Shape Completion Attack

We summarize the proposed black-box 3D adversarial attack in Algorithm 1. 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 [9] function to the ℓ_{inf} norm to limit the perturbation in adversarial guidance.

Algorithm 1 Transferable 3D Adversarial Shape Completion Attack Algorithm

Require: f_{ens}: substitute models
Require: z_0: partial shape for shape completion
Require: y: class label for shape completion
Require: T: reverse generation process timestep for LDM
Require: T_{adv}: timestep for adversarial guidance
Require: N: number of perturbed points at each diffusion step
Require: 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 4
    for $m = 1, \ldots, M$ **do**
      Simple random sampling with Equation 8
      Obtain the ensemble adversarial loss with Equation 9
    end for
    Monte Carlo estimate with Equation 7
    Calculate the saliency score of $\tilde{x}_{t-1}$ with Equation 10
    Update top-N points from step 11 of $\tilde{x}_{t-1}$ with Equation 5
    $\tilde{x}_{t-1} = \text{Clip}(\tilde{x}_{t-1})$
  else
    Sample $\tilde{x}_{t-1}$ with Equation 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}

4.4 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 2, 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. 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 of various models in Figure 2. 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.

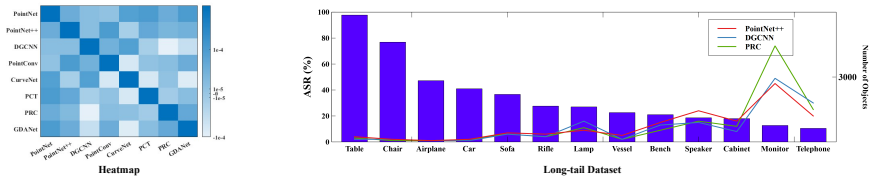

Fig. 2: The challenging 3D black-box adversarial attacks. The value in the Heatmap is re-scaled for better visualization. We use the top 13 classes from the ShapeNet dataset to demonstrate the long-tailed dataset problem. We use PGD with $\ell_{\text{inf}} = 0.16$ on PointNet to evaluate the black-box attack success rate (ASR).

Table 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 ($3\text{DAdvDiff}_{\text{ens}}$ 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
	$3\text{DAdvDiff}_{\text{ens}}$	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
	$3\text{DAdvDiff}_{\text{ens}}$	99.9	**97.0**	***99.9***	**94.5**	**93.5**	**80.5**	***99.9***	**85.2**	**90.1**

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 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 Experiments

5.1 Experimental Setup

Dataset. Due to ModelNet40 being insufficient to train the diffusion model, we use the ShapeNet [3] dataset for major evaluations. The ShapeNetCore split is adopted, which contains 55 categories with 42003 data, of which 31535 samples are used for training and 10468 samples are used for testing. We select PVD [43] for the diffusion model in this paper. The proposed attack does not require additional training in the diffusion model, we follow settings as in the original PVD paper for selecting shape completion's partial shapes. Public checkpoints [43] from Airplane, Chair, and Car are selected for repeatability. Experiments on ModelNet40 are discussed in the Appendix.
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 [26], Pointnet++ (SSG) [27], DGCNN [31], PointConv (SSG) [33], CurveNet [36], PCT [10], PRC [28], and GDANet [37].
Comparisons. We have chosen four white-box 3D adversarial attacks as our baseline for comparison, namely: PGD [22], KNN [30], GeoA3 [32], and SI-Adv [15]. We also employ existing black-box 3D adversarial attacks, specifically: AdvPC [11] and PF-Attack [12]. 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 [42], SOR [42], DUP-Net [42], IF-Defense [34], and Adversarial Hybrid Training [16] for evaluation under defenses. All the defense settings are followed according to [16].
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 Attack Success Rate (ASR). The experiment results are averaged over 10 attacks.

5.2 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 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 developed 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,

Table 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$_{\text{ens}}$	99.9	**85.9**	**49.1**	**36.9**	22.5	**96.1**

3DAdvDiff$_{\text{ens}}$ 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 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$_{\text{ens}}$ is effective against outlier removal defenses. However, the performance of 3DAdvDiff$_{\text{ens}}$ 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.

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 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 3 for a more comprehensive demonstration. The point clouds generated by 3DAdvDiff$_{\text{ens}}$ is smoother than existing attacks.

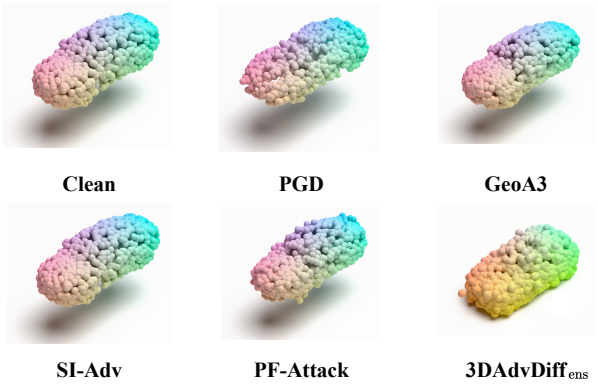

Fig. 3: The visual quality of adversarial examples. The black-box adversarial examples are relatively unnatural compared to white-box adversarial examples.

Table 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$_{ens}$
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 overwhelmingly performance on black-box adversarial attacks. The generation speed of diffusion models is a critical problem to influence its development. As shown in Table 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. Detailed discussion is given in the Appendix.

Table 4: The average running time to generate one adversarial example.

Method	PGD	KNN	GeoA3	SI-Adv	AdvPC	PF-Attack	3DAdvDiff$_{ens}$
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, 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: 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**

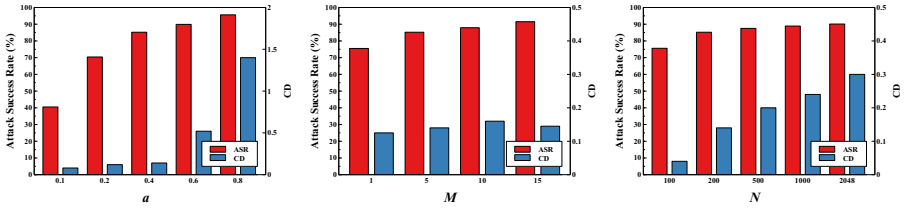

Fig. 4: The ablation study of proposed 3DAdvDiff$_{\text{ens}}$. The results are evaluated on the Chair class of the ShapeNet dataset. We use average ASR to test the black-box attack performance.

5.3 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 4. However, 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 4 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 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 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 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.

Table 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 Augmentation. Current 3D distance measurements take into account the difference between the entire point set. Therefore, to improve the 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 4.

6 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.

7 Conclusion

In this paper, 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 improvement 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.

Acknowledgements

This work was supported in part by HK RGC GRF under Grant PolyU 15201323.

References

1. Carbone, G., Wicker, M., Laurenti, L., Patane, A., Bortolussi, L., Sanguinetti, G.: Robustness of bayesian neural networks to gradient-based attacks. Advances in Neural Information Processing Systems **33**, 15602–15613 (2020)
2. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. IEEE (2017)
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
4. Chen, X., Gao, X., Zhao, J., Ye, K., Xu, C.Z.: Advdiffuser: Natural adversarial example synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4562–4572 (2023)
5. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)
6. Dai, X., Liang, K., Xiao, B.: Advdiff: Generating unrestricted adversarial examples using diffusion models. arXiv preprint arXiv:2307.12499 (2023)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
8. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
10. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: Pct: Point cloud transformer. Computational Visual Media **7**, 187–199 (2021)
11. Hamdi, A., Rojas, S., Thabet, A., Ghanem, B.: Advpc: Transferable adversarial perturbations on 3d point clouds. In: Proceedings of the European Conference on Computer Vision. pp. 241–257 (2020)
12. He, B., Liu, J., Li, Y., Liang, S., Li, J., Jia, X., Cao, X.: Generating transferable 3d adversarial point cloud via random perturbation factorization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 764–772 (2023)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
15. Huang, Q., Dong, X., Chen, D., Zhou, H., Zhang, W., Yu, N.: Shape-invariant 3d adversarial point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15335–15344 (June 2022)
16. Ji, Q., Wang, L., Shi, C., Hu, S., Chen, Y., Sun, L.: 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. pp. 4295–4304 (2023)

17. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2426–2435 (2022)
18. Li, Q., Guo, Y., Zuo, W., Chen, H.: Making substitute models more bayesian can enhance transferability of adversarial examples. In: The Eleventh International Conference on Learning Representations (2022)
19. Li, Y., Li, Y., Dai, X., Guo, S., Xiao, B.: Physical-world optical adversarial attacks on 3d face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24699–24708 (2023)
20. Li, Y., Bradshaw, J., Sharma, Y.: Are generative classifiers more robust to adversarial attacks? In: International Conference on Machine Learning. pp. 3804–3814. PMLR (2019)
21. Liang, K., Xiao, B.: Styless: boosting the transferability of adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8163–8172 (2023)
22. Liu, D., Yu, R., Su, H.: Extending adversarial attacks and defenses to deep 3d point cloud classifiers. In: Proceedings of the International Conference on Image Processing. pp. 2279–2283. IEEE (2019)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
24. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2837–2845 (2021)
25. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018)
26. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
27. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 5099–5108 (2017)
28. Ren, J., Pan, L., Liu, Z.: Benchmarking and analyzing point cloud classification under corruptions. In: International Conference on Machine Learning. pp. 18559–18575. PMLR (2022)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
30. Tsai, T., Yang, K., Ho, T.Y., Jin, Y.: Robust adversarial objects against deep learning models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 954–962 (2020)
31. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. Acm Transactions On Graphics **38**(5), 1–12 (2019)
32. Wen, Y., Lin, J., Chen, K., Chen, C.P., Jia, K.: Geometry-aware generation of adversarial point clouds. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(6), 2984–2999 (2020)

33. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 9621–9630 (2019)
34. Wu, Z., Duan, Y., Wang, H., Fan, Q., Guibas, L.J.: If-defense: 3d adversarial point cloud defense via implicit function based restoration. arXiv preprint arXiv:2010.05272 (2020)
35. Xiang, C., Qi, C.R., Li, B.: Generating 3d adversarial point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9136–9144 (2019)
36. Xiang, T., Zhang, C., Song, Y., Yu, J., Cai, W.: Walk in the cloud: Learning curves for point clouds shape analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 915–924 (October 2021)
37. Xu, M., Zhang, J., Zhou, Z., Xu, M., Qi, X., Qiao, Y.: Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 3056–3064 (2021)
38. Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: Lion: latent point diffusion models for 3d shape generation. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. pp. 10021–10039 (2022)
39. Zhao, Y., Wu, Y., Chen, C., Lim, A.: 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. pp. 1201–1210 (2020)
40. Zheng, T., Chen, C., Yuan, J., Li, B., Ren, K.: Pointcloud saliency maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1598–1606 (2019)
41. Zhou, H., Chen, D., Liao, J., Chen, K., Dong, X., Liu, K., Zhang, W., Hua, G., Yu, N.: 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. pp. 10356–10365 (2020)
42. Zhou, H., Chen, K., Zhang, W., Fang, H., Zhou, W., Yu, N.: Dup-net: Denoiser and upsampler network for 3d adversarial point clouds defense. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1961–1970 (2019)
43. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5826–5835 (2021)