PointDP: DIFFUSION-DRIVEN PURIFICATION AGAINST 3D ADVERSARIAL POINT CLOUDS

Anonymous authors
Paper under double-blind review

ABSTRACT

3D Point cloud is a critical data representation in many real-world applications, such as autonomous driving, robotics, and medical imaging. Although the success of deep learning further accelerates the adoption of 3D point clouds in the physical world, deep learning is notoriously vulnerable to adversarial attacks. Various defense solutions have been proposed to build robust models against adversarial attacks. In this work, we identify that the state-of-the-art empirical defense, adversarial training, has a major limitation in 3D point cloud models due to gradient obfuscation, resulting in significant degradation of robustness against strong attacks. To bridge the gap, we propose *PointDP*, a purification strategy that leverages diffusion models to defend against 3D adversarial attacks. Since *PointDP* does not rely on predefined adversarial examples for training, it can defend against diverse threats. We extensively evaluate *PointDP* on six representative 3D point cloud architectures and leverage **sixteen** strong and adaptive attacks to demonstrate its lower-bound robustness. Our evaluation shows that *PointDP* achieves significantly better (*i.e.,* 12.6%-40.3%) adversarial robustness than state-of-the-art methods under strong attacks bounded by different ℓ_p norms.

1 INTRODUCTION

Point cloud data is emerging as one of the most broadly used representations in 3D computer vision. It is a versatile data format available from various sensors like LiDAR and stereo cameras and computer-aided design (CAD) models, which depict physical objects by many coordinates in the 3D space. Many deep learning-based 3D perception models have been proposed [59, 34, 43, 60, 41, 9] and thus realized several safety-critical applications (*e.g.,* autonomous driving) [81, 46, 45]. Although deep learning models [41, 42] have exhibited performance boost on many challenging tasks, extensive studies show that they are notoriously vulnerable to adversarial attacks [5, 49, 68], where attackers manipulate the input in an imperceptible manner, which will lead to incorrect predictions of the target model. Because of the broad applications of 3D point clouds in safety-critical fields, it is imperative to study the adversarial robustness of point cloud recognition models.

The manipulation space for 2D adversarial attacks is to change pixel-level numeric values of the input images. Unlike adversarial examples in 2D applications, the flexible representation of 3D point clouds results in an arguably larger attack surface. For example, adversaries could shift and detach existing points [88], add new points into the pristine point cloud [50], or even generate totally new point clouds [89] to launch attacks. Different strategies, including limits on the number of altered points and constraints on the maximal magnitude of shifted points [50] were proposed to make attacks less perceptible. The flexibility of 3D point cloud data formats enables diverse attacks, thus hindering a practical and universal defense design.

Given the safety-critical property involved in 3D point cloud applications, various studies have been devoted to advancing the robustness of 3D point cloud recognition models. DUP-Net [90] and GvG-PointNet++ [14] pioneered to add statistical outlier removal (SOR) modules as pre-processing and in-network blocks, respectively, as mitigation strategies. More lately, Sun *et al.* [51] broke the robustness of DUP-Net and GvG-PointNet++ by specific adaptive attacks. Adversarial training has been acknowledged as the most potent defense to deliver strong empirical robustness on PointNet, DGCNN, and PCT [50]. Meanwhile, advanced purification strategies like IF-Defense [66] and LPC [25] leverage more complex modules to cleanse the adversarial point clouds. However, given that point cloud is a sparse and unstructured data format, it motivates us to re-think that *whether*

the current adversarial training and purification-based methods are robust enough against stronger adversarial attacks?

In this work, our journey starts with revisiting the prior arts and exploring their truly adversarial robustness. By designing various types of strong adaptive attacks, we, for the first time, demonstrate that standard adversarial training [33] suffers from *gradient obfuscation* in the point cloud recognition models as the unstructured point cloud data format requires unique architectural designs to digest. We also extensively evaluate IF-Defense and LPC to show that their purification strategies are actually vulnerable to stronger attacks (§ 4.3).

Furthermore, we propose *PointDP*, an adversarial purification method that leverages a diffusion model as a pre-processing module to defend against 3D adversaries. As shown in Figure 1, *PointDP* consists of two components (1) an off-the-shelf 3D point cloud diffusion model and (2) a classifier. Given an input point cloud, *PointDP* take two steps: (i) adding noise to the input data gradually via the diffusion process of the diffusion model, (ii) purifying the noised data step by step to get the reversed sample via the reverse process of a diffusion model (§ 3.1), and (iii) feeding the reversed sample to the final classifier. Since *PointDP* does not rely on any types of pre-defined adversarial examples for training, it can defend against diverse unseen threats.

We rigorously evaluate *PointDP* with six representative point cloud models and sixteen attacks, including PGD [50, 33], C&W [68, 8], and point cloud-specific attacks [88, 21] with ℓ_0, ℓ_2, and ℓ_∞ norms. *PointDP* on average achieves 75.9% robust accuracy while maintaining similar clean accuracy to the original models, outperforming existing studies by a significant margin. In a nutshell, our contributions are summarized as *two-fold*:

- We are the first to demonstrate that standard adversarial training [33, 50], the most longstanding defense in the 2D image recognition task, has a major limitation in its application in 3D point cloud models due to architecture designs. We launch black-box attacks to validate our claim that degrades adversarially trained models' robust accuracy to merely $\sim$10%, which is no longer useful for 3D point cloud recognition.
- We propose *PointDP* that leverages diffusion models to purify adversarial 3D point clouds. *PointDP* is a general framework that is independent of the diffusion model used. We also formulate rigorous adaptive attacks on *PointDP*. We conduct extensive evaluation on six representative models with numerous attacks to comprehensively understand the robustness of *PointDP*. Our evaluation shows that *PointDP* outperforms previous state-of-the-arts purification methods, IF-Defense [66] and LPC [25], by 12.6% and 40.3% on average, respectively. We also set up a rigorous protocol for 3D robustness evaluation to benefit future research.

2 Related Work

In this section, we review the current progress of deep learning, adversarial attacks, and defenses for 3D point cloud recognition tasks.

2.1 Deep Learning on 3D Point Cloud Recognition

2D computer vision has achieved stellar progress on architectural designs of convolutional neural networks [22], followed by vision transformers [15]. However, there is currently no consensus on the architecture of 3D perception models since there is no standard data format for 3D perception [53]. As raw data from both 3D scanners and triangular meshes can be efficiently transformed into point clouds, they are becoming the most often utilized data format in 3D perception. 3D networks at the early stage use dense voxel grids for perception [59, 34, 47, 54], which discretize a point cloud to voxel cells for classification, segmentation, and object detection. PointNet pioneered to leverage global pooling help achieve memory-efficient permutation invariance in an end-to-end manner. PointNet++ [42] and DGCNN [61] followed up to add sophisticated local clustering operations to advance the performance. Sparse tensors are the other direction in 3D network designs [19, 9] to use 3D convolutions to improve 3D perception performance. PointCNN and RSCNN reformed the classic pyramid CNN to improve the local feature generation [26, 29]. PointConv and KPConv designed new convolution operation for point cloud learning [65, 55]. PointTransformer and PCT advanced self-attention blocks in the 3D space and achieved good performance [87, 20]. Various novel local clustering operations [69, 32] also show enhancements on the clean performance. In this work, we focus on PointNet, PointNet++, DGCNN, PCT, CurveNet, and PointMLP as our evaluation backbones since they are representative and widely used and achieve state-of-the-art results in point cloud recognition [1].

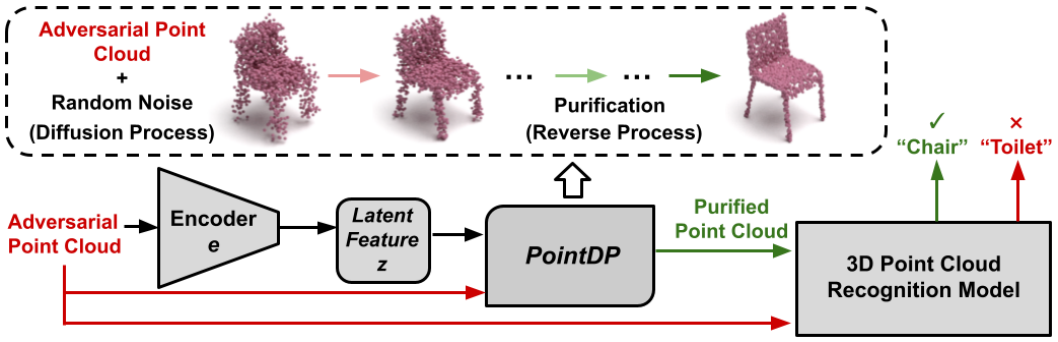

Figure 1: Illustration of *PointDP*, where *PointDP* serve as a purification module. We leverage [31] as the diffusion model in our study. The adversarial point cloud will be incorrectly classified as "toilet" by the recognition model if not purified by our *PointDP*.

2.2 Adversarial Attacks and Defenses

Adversarial attacks have become the main obstacle that hinder deep learning models from real-world deployments, especially in safety-critical applications [16, 49, 5, 86, 85]. There are a lot of adversarial attacks proposed in the 2D space to break the various vision models [8, 72, 79, 75, 23, 24, 73, 52]. To fill this gap between standard and robust accuracies, many mitigation solutions have been studied and presented to improve the robustness against adversarial attacks [80, 78, 4, 38, 35, 82, 71, 83, 70]. However, most of them including adding randomization [28, 13, 14], model distillation [38], adversarial detection [35], and input transformation [80, 78, 37, 4, 90] have been compromised by adaptive attacks [56, 2]. Adversarial training (AT) [33, 18, 64, 44], in contrast, delivered a more longstanding mitigation strategy [74, 76, 84]. However, the robust accuracy achieved by AT is still not satisfactory enough to be used in practice. Most recently, Nie *et al.* proposed DiffPure [36] that leverages diffusion models to defend against adversarial attacks, and following-up studies to extend it to certified defenses [7].

Adversarial attacks and defenses also extend to 3D point clouds. Xiang *et al.* [72] first demonstrated that point cloud recognition models are vulnerable to adversarial attacks. They also introduced different threat models like point shifting and point adding attacks. Wen *et al.* [62] enhanced the loss function in C&W attack to achieve attacks with smaller perturbations and Hamdi *et al.* [21] presented transferable black-box attacks on point cloud recognition. [63] pioneered to study the point dropping attack under both withe- and black-box settings. Zhou *et al.* [90] and Dong *et al.* [14] proposed to purify the adversarial point clouds by input transformation and adversarial detection. However, these methods have been successfully by [51] through adaptive attacks. Moreover, Liu *et al.* [28] made a preliminary investigation on extending countermeasures in the 2D space to defend against simple attacks like FGSM [18] on point cloud data. Sun *et al.* [50] conducted a more thorough study on the application of self-supervised learning in adversarial training for 3D point clodu recognition. Besides adversarial training, advanced purification methods IF-Defense [66] and LPC [25] were proposed to transform the adversarial examples to the clean manifold. Certified defenses for point clouds have been focusing on the adversarial transformations and deformations [30, 10, 40]. In this work, we present *PointDP*, that utilizes 3D diffusion models to purify adversarial point clouds that delivers both state-of-the-art (SOTA) empirical and certified robustness. We also demonstrate that standard adversarial training suffer from strong black-box attacks and SOTA purification methods (*i.e.,* IF-Defense and LPC) are vulnerable to PGD-styled adversaries (§ 4.3).

3 *PointDP*: Diffusion-driven Purification against 3D Adversaries

We first introduce the preliminaries of diffusion models and then propose *PointDP* that first introduces noise to the adversarial 3D point clouds, followed by the forward process of diffusion models to get diffused point clouds. Purified point clouds are recovered through the reverse process (§-3.2). Next, we follow [36] to apply the adjoint method to backward propagate through SDE for efficient gradient evaluation with strong adaptive attacks (§ 3.3).

3.1 Preliminaries

In this section, we briefly review the background of conditional diffusion models in 3D vision tasks. Following [31], we use the discrete-time formulation of the forward and reverse processes.

Given a clean point cloud sampled from the unknown data distribution $\boldsymbol{x}_0 \sim q(\boldsymbol{x})$, the forward process of the diffusion model leverages a fixed Markov chain to gradually add Gaussian noise to the clean point cloud $\boldsymbol{x}_0$ over a pre-defined N time steps, resulting in a number of noisy point clouds $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$. Mathematically, the forward process is defined as:

$$\begin{aligned} q(\boldsymbol{x}_{1:N}|\boldsymbol{x}_0) &:= \prod_{n=1}^{N} q(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}), \\ q(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) &:= \mathcal{N}(\boldsymbol{x}_n; \sqrt{1-\beta_n}\boldsymbol{x}_{n-1}, \beta_n \mathbf{I}) \end{aligned} \tag{1}$$

where β_n is a scheduling function of the added Gaussian noise, satisfying $0 < \beta_1, \cdots, \beta_N < 1$.

The reverse process, in contrast, is trained to recover the diffused point cloud in an iterative manner. 3D Point clouds have less semantics than 2D images due to the lack of texture information. Therefore, point cloud diffusion models leverage a separate encoder e to as a latent feature $\boldsymbol{z}_{\boldsymbol{x}} = e(\boldsymbol{x})$ as a condition to help recover the clean point cloud:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:N}|\boldsymbol{z}) &:= p(\boldsymbol{x}_N) \prod_{n=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n, \boldsymbol{z}), \\ p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n, \boldsymbol{z}) &:= \mathcal{N}(\boldsymbol{x}_{n-1}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n, \boldsymbol{z}), \beta_n \mathbf{I}) \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ denotes the approximated mean value parameterized by a neural network. The training objective is to learn the variational bound of the negative log-likelihood [31]. In practice, we jointly train the encoder e with the noise predictor $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n, \boldsymbol{z})$. Similar to the DDPM model [12], we can conduct the sampling by reparameterizing $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ as

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n, \boldsymbol{z}) = \frac{1}{\sqrt{1-\beta_n}} \left(\boldsymbol{x}_n - \frac{\beta_n}{\sqrt{1-\overline{\alpha}_n}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n, \boldsymbol{z}) \right) \tag{3}$$

where $\overline{\alpha}_n = \prod_{i=1}^{n}(1-\beta_i)$. It is worth noting that point cloud diffusion models have recently achieved SOTA performance on generating and autoencoding 3D point clouds, which provides us with opportunities for adversarial point cloud purification.

3.2 Design of *PointDP*

Overview. Figure 1 illustrates the pipeline of *PointDP*. Different from Nie *et al.* [36] use unconditional diffusion model to remove the adversarial effect for 2D images, we use the conditional diffusion models as mentioned in § 3.1. Specifically, *PointDP* first adds pre-quantified Gaussian noise to the input data and then leverage a well-trained diffusion model to purify the noisy point cloud step by step to recover the clean point cloud. The reversed point cloud will be finally fed into the recognition model for the classification task. Note that we do not aim at designing new point cloud diffusion models, but instead propose a novel purification pipeline with rigorous evaluations as our main contributions.

Following [36], in order to backpropagate through the forward and reverse processes for computing gradients, we first convert the discrete-time formulation defined in Eqs. (1) and (2) to its continuous-time counterpart, *i.e.*, the forward and reverse stochastic differential equations (SDEs) [48]. Let $\boldsymbol{x}_a$ be an adversarial example *w.r.t.* the pristine classifier f, we initialize the input of the forward diffusion process as $\boldsymbol{x}_a$, *i.e.,* $\boldsymbol{x}_0 = \boldsymbol{x}_a$. Also, let $\boldsymbol{x}(\frac{n}{N}) := \boldsymbol{x}_n$, $\beta(\frac{n}{N}) := \beta_n$, $\alpha(\frac{n}{N}) := \overline{\alpha}_n$, and $t \in \{0, 1, \cdots, \frac{N-1}{N}\}$. The forward diffusion process from $t = 0$ to $t = t^* \in (0, 1)$ can be solved by:

$$\boldsymbol{x}(t^*) = \sqrt{\alpha(t^*)}\boldsymbol{x}_a + \sqrt{1-\alpha(t^*)}\boldsymbol{\epsilon} \tag{4}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. We leverage Eq. 2 to recover the clean point clouds. Equivalently, the truncated reverse process can be also solved by the SDE solver in [36] (denoted as $\mathbf{sdeint}$):

$$\hat{\boldsymbol{x}}(0) = \mathbf{sdeint}(\boldsymbol{x}(t^*), \boldsymbol{f}_{\text{rev}}, g_{\text{rev}}, \boldsymbol{w}, t^*, 0) \tag{5}$$

where the six inputs are initial value, drift coefficient, diffusion coefficient, Wiener process, initial time, and end time [36], with the definitions:

$$\boldsymbol{f}_{\text{rev}}(\boldsymbol{x}, t, \boldsymbol{z}) = -\frac{1}{2}\beta(t)[\boldsymbol{x} + 2\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t, \boldsymbol{z})], \quad g_{\text{rev}}(t) = \sqrt{\beta(t)} \tag{6}$$

and the score function $\boldsymbol{s}_{\boldsymbol{\theta}}$ is derived from $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n, \boldsymbol{z})$ in Eq. (3) by following:

$$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t, \boldsymbol{z}) = -\frac{1}{\sqrt{1-\alpha(t)}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}(t), tN, \boldsymbol{z}) \tag{7}$$

Note that the hyper-parameter t^* and N trades off the denoising performance and efficiency. We empirically choose $t^* = 0.15$ and $N = 200$ in our study, which has shown satisfactory results in our evaluation (§ 4). We also conduct ablation studies on t in § 4.2.

3.3 Adpative Attacks on *PointDP*

PointDP is a pre-processing module that purifies the adversarial perturbations. [2] have shown that input transformation-based methods can be broken by specifically designed attacks. Therefore, it is essential to model the adaptive attacks on *PointDP* to demonstrate its lower-bound adversarial robustness. We thus formulate two types of adaptive attacks on *PointDP*.

Attack on Latent Feature. As *PointDP* utilizes conditional diffusion models for adversarial purification, the latent feature z is a good candidate for adversaries to launch attacks. Concretely, adversaries can set the goal to maximize some distance metric $\mathcal{D}$ between the latent feature of the optimized adversarial examples and the oracle latent feature of clean inputs $\boldsymbol{z}_{\text{oracle}}$. Without loss of generality, the adaptive attacks can be formulated as:

$$\boldsymbol{x}_{s+1} = \text{Proj}_{\boldsymbol{x}+\mathcal{S}}(\boldsymbol{x}_s + \alpha \cdot \text{norm}(\nabla_{\boldsymbol{x}_s}\mathcal{D}(e(\boldsymbol{x}_s), \boldsymbol{z}_{\text{oracle}}))), \tag{8}$$

where $\boldsymbol{x}_s$ denotes the adversarial examples from the s-th step, Proj is the function to project the adversarial examples to the pre-defined space $\mathcal{S}$, and α is the attack step size. We choose two distance metrics in our study, where the first one is the KL divergence [17] and the other is the the ℓ_1 norm distance. In our evaluation (§ 4), we report the lowest accuracy achieved under attacks with two distance metrics.

Adaptive Attack. We follow [36] to formulate the adaptive attack as an augmented SDE process. We re-state the attack formulation as below. For the SDE in Equation 5, the augmented SDE that computes the gradient $\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(t^*)}$ of backward propagating through it is given by:

$$\begin{pmatrix} \boldsymbol{x}(t^*) \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(t^*)} \end{pmatrix} = \mathbf{sdeint}\left(\begin{pmatrix} \hat{\boldsymbol{x}}(0) \\ \frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{x}}(0)} \end{pmatrix}, \tilde{\boldsymbol{f}}, \tilde{\boldsymbol{g}}, \tilde{\boldsymbol{w}}, 0, t^*\right) \tag{9}$$

where $\frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{x}}(0)}$ is the gradient of the objective $\mathcal{L}$ w.r.t. the output $\hat{\boldsymbol{x}}(0)$ of the SDE in Equatrion 5), and

$$\tilde{\boldsymbol{f}}([\boldsymbol{x}; \boldsymbol{z}], t) = \begin{pmatrix} \boldsymbol{f}_{\text{rev}}(\boldsymbol{x}, t) \\ \frac{\partial \boldsymbol{f}_{\text{rev}}(\boldsymbol{x}, t)}{\partial \boldsymbol{x}}\boldsymbol{z} \end{pmatrix}, \quad \tilde{\boldsymbol{g}}(t) = \begin{pmatrix} -g_{\text{rev}}(t)\mathbf{1} \\ \mathbf{0} \end{pmatrix}, \quad \tilde{\boldsymbol{w}}(t) = \begin{pmatrix} -\boldsymbol{w}(1-t) \\ -\boldsymbol{w}(1-t) \end{pmatrix}$$

where $\mathbf{1}$ and $\mathbf{0}$ denote the vectors of all ones and all zeros, respectively. Nie *et al.* [36] have demonstrated that such approximation aligns well with the true gradient value. Therefore, we leverage this adaptive attack formulation for our evaluation.

4 Experiments and Results

In this section, we first introduce our experimental setups (§ 4.1). We then present the standard robustness evaluation of *PointDP*(§ 4.2). We next show that how the SOTA adversarial training and adversarial purification methods fail under various strong attacks (§ 4.3). We finally conduct stress test on *PointDP* to show its actual robustness under various stronger adaptive attacks (§ 4.4).

4.1 Experimental Setups

Datasets and Network Architectures. We conduct all the experiments on the widely used ModelNet40 point cloud classification benchmark [67], consisting of 12,311 CAD models from 40 artificial object categories. We adopt the official split with 9,843 samples for training and 2,468 for testing. We also uniformly sample 1024 points from the surface of each object and normalize them into an edge-length-2 cube, following most of the prior arts [41]. As mentioned before, there are various backbones for 3D point cloud recognition in the literature. To demonstrate the universality of *PointDP*, we select six representative model architectures including PointNet [41], PointNet++ [42], DGCNN [61], PCT [20], CurveNet [69], and PointMLP [32]. These backbones either have representative designs

Table 1: Evaluation Results of Plain Model on PA and PD (Accuracy %). Models under other attacks mostly have **0%** accuracy, and we put the detailed results in Appendix A.

	PointNet	PointNet++	DGCNN	PCT	CurveNet	PointMLP
None	90.1	92.8	92.5	92.8	93.2	93.5
PA	44.1	19.9	35.1	20.8	48.9	7.2
PD	33.3	69.8	64.5	53.0	72.6	71.1

(*e.g.*, Transformer and MLP) or achieve SOTA performance on the ModelNet40 benchmark (*e.g.*, CurveNet and PointMLP).

Adversarial Attacks. As briefly described in § 2.2, adversarial attacks could be roughly categorized into C&W- and PGD-styled attacks. C&W attacks involves the perturbation magnitude into the *objective* term of the optimization procedure by Lagrange multiplier, while PGD attacks set the perturbation magnitude as a firm *constraint* in the optimization procedure. Moreover, adversarial attacks by ℓ_p norm as the distance metric for the perturbation. Although a number of attacks measure Chamfer and Handoff "distances" in 3D point cloud [68], they are not formal distance metrics as they do not satisfy the triangular inequality. Therefore, we still leverage ℓ_2 and ℓ_∞ norm, following most defense studies in both 2D and 3D vision tasks [8, 50]. We also have designed adaptive attacks on our proposed method § 3.3. Besides naive C&W and PGD attacks, we leverage specific attacks designed to break the robustness of point cloud recognition such as kNN [57] and AdvPC [21]. We also apply strong adaptive AutoAttack [11] (*i.e.*, APGD) in our evaluation. Moreover, we use SPSA [58] and Nattack [27] as black-box adversaries, followed by the suggestion of Carlini *et al.* [6]. We also leverage EOT-AutoAttack. Point adding (PA) and dropping/dropping (PD) attacks are also evaluated in our study, followed by the setups in [50]. We set the attack steps to 200 to maximize the adversarial capability and follow the settings in [50] for other attack parameters by default.

Evaluation Metrics. We leverage two main metrics to evaluate the performance of our defense proposal, which are *standard* and *robust* accuracy. The standard accuracy measures the performance of the defense method on clean data, which is evaluated on the whole test set from ModelNet40. The robust accuracy measures the performance on adversarial examples generated by different attacks. Because of the high computational cost of applying *adaptive* and *black-box* attacks to our method, we evaluate robust accuracy for our defense on a fixed subset of 128 point clouds randomly sampled from the test set. Notably, robust accuracies of most baselines do not change much on the sampled subset, compared to the whole test set. We evaluate the robust accuracy on the whole test set for other adversarial attacks with acceptable overhead (*e.g.*, C&W and PGD attacks).

Table 2: Evaluation Results (Accuracy) of Adversarial Attacks on *PointDP* (%). Colored rows are corresponded to rows in Table 5 for clear comparisons with IF-Defense results.

		PointNet	PointNet++	DGCNN	PCT	CurveNet	PointMLP
	None	86.8	87.9	86.9	87.0	88.0	88.2
ℓ_∞ $\epsilon = 0.05$	C&W	77.9	78.6	78.9	76.8	73.1	76.2
	PGD	78.1	80.6	80.3	77.2	74.8	79.8
	AdvPC	69.7	76.6	79.1	79.4	72.6	75.2
	PA	82.1	85.1	84.8	85.5	86.3	85.8
ℓ_2 $\epsilon = 1.25$	C&W	82.4	82.9	81.9	80.9	81.5	82.6
	PGD	80.1	75.0	74.6	72.0	71.7	76.4
	AdvPC	69.1	76.3	79.0	74.2	74.1	75.6
	kNN	83.5	82.9	83.3	82.3	81.5	83.1
ℓ_0 $\epsilon = 200$	PD	68.9	74.1	77.3	76.3	76.8	77.4

Baseline. Without any defense applied to the original recognition models, the robust accuracy is mostly **0%** for all models under ℓ_2 and ℓ_∞ based attacks (see Appendix A). DGCNN exceptionally achieves 64% on ℓ_2-based PGD, AutoAttack, respectively, due to its dynamic clustering design, which adaptively discards outlier points. PA and PD are two weaker attacks and Table 1 presents the robust accuracy against these two attacks.

4.2 Experiment Results of *PointDP*

In this section, we first present the evaluation results of *PointDP* under attacks on the plain models. We train the diffusion and 3D point cloud recognition models in a sequential order. Table 2 presents the detailed results of *PointDP* against attacks on six models. We find that *PointDP* overall achieves satisfactory results across all models and attacks. The average robust accuracy against adversarial

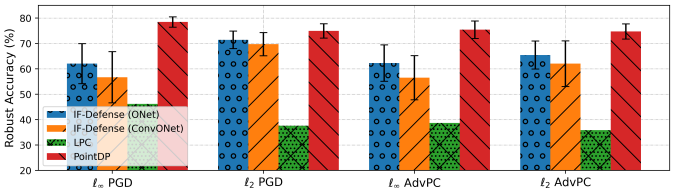

Figure 2: Compare among SOTA Adversarial Purification Strategies (*i.e.,* IF-Defense [66], LPC [25], and *PointDP*). The results of IF-Defense and *PointDP* are averaged from six models.

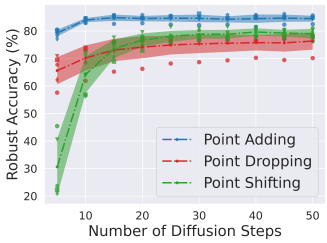

Figure 3: Ablation on Discrete Diffusion Steps in *PointDP*.

attacks is above 75%. We observe a drop on the clean accuracy for the chosen models due to the imperfect reconstruction of diffusion models. As mentioned before, diffusion models for 3D point cloud is a more difficult task than 2D image diffusion, which may lead to partial semantic loss. The average drop of standard accuracy is 4.9%. We find that DGCNN still achieves the best robustness combined with *PointDP*, which has a 79.9% of robust accuracy. We further compare the performance of *PointDP* with adversarial training, IF-Defense, and LPC in the next section.

We also ablate the effect of diffusion steps in *PointDP*. Figure 3 shows the averaged evaluation results of point shifting, adding, and dropping attacks with PGD adversary over the selected models. Point shifting attack is much stronger than point adding and dropping attacks. It is, thus, more sensitive to the diffusion steps in *PointDP*. We find that the robust accuracy converges after the number of diffusion steps $n \geq 30$ (or equivalently $t \geq 0.15$). Therefore, we choose to use $t^* = 0.15$ in the main evaluation of our study. Adversarial purification inevitably introduces overhead during model inference, we benchmark the computation of *PointDP* and other baselines using an RTX3080 GPU and a batch size of 32. Table 3 presents the results, where *PointDP* achieves the most negligible cost than existing SOTA methods, which is a 27× speed-up than IF-Defense.

Table 3: Ablation Study on Overhead Introduced by Adversarial Purification Methods.

	DUP-Net	IF-Defense	***PointDP***
Time (s)	1.33	2.60	**0.097**

4.3 Failure of State-of-the-Art Defenses

In this section, we demonstrate how lately proposed defense solutions fail when encountered with stronger (adaptive) adversarial attacks on 3D point cloud recognition models.

```
1  def knn(x, k):
2      inner = -2*torch.matmul(x.transpose(2, 1), x)
3      xx = torch.sum(x**2, dim=1, keepdim=True)
4      pairwise_distance = -xx - inner - xx.transpose
        (2, 1)
5      idx = pairwise_distance.topk(k=k, dim=-1)[1]
6      # (batch_size, num_points, k)
7      return idx
8
9  def get_graph_feature(x, k):
10     #x's shape is (batch_size,num_dims,num_points)
11     idx = knn(x, k=k) # (batch_size,num_points,k)
12     ...... # shape transformation here
13     feature = x.view(batch_size*num_points, -1)[
        idx, :]
14     # idx is used as index to select features
15     ......
16     return feature
17
18 # forward function for EdgeConv
19 def forward(self, x):
20     ......
21     x = get_graph_feature(x, k=self.k)
22     x = self.conv1(x) # convolution
23     ......
```

Figure 4: PyTorch [39]-Style Code Snippet of EdgeConv [61] in Point Cloud Recognition Models. Adversarial training fails since the kNN layers leverage the top-k function where the gradient propagate to the index, resulting in gradient obfuscation.

Adversarial training (AT) has been applied to PointNet, DGCNN, and PCT with the help of self-supervised learning [50] that achieves satisfactory robustness. Such observations are consistent with the performance of AT for 2D perception models. However, we find that AT is, in fact, a weak defense solution in 3D perception models. First, as acknowledged by [50], point cloud models (*e.g.,* PointNet++ and CurveNet) often leverage different sampling strategies to select anchor points, like furthest point sampling (FPS). Such sampling involves high randomness. AT either cannot converge with different random seeds in each iteration or overfits to a single random seed. Therefore, AT cannot fit these models. Moreover, we discover that the kNN layers will cause severe *gradient obfuscation* in point cloud models as well. Different from standard training process that only needs the gradient of model parameters *w.r.t.* the loss function $\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}}$, AT additionally requires the gradient flow to the input $\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}}$. As shown in Line 5 from Figure 4, kNN essentially applies top-k for point selection. Top-k is a general case for $\max$ pooling that does not have trainable model parameters. Therefore, it will not affect the standard training. However, top-k is not differentiable *w.r.t.* the input

Table 5: Evaluation Results (Accuracy) of Adversarial Attacks on IF-Defense (%). Colored rows are corresponded to rows in Table 2 for clear comparisons with *PointDP* results.

		PointNet	PointNet++	DGCNN	PCT	CurveNet	PointMLP
ONet	None	90.0	92.8	92.4	92.8	93.1	93.5
ℓ_∞ $\epsilon = 0.05$	PGD	69.9	74.0	61.0	54.1	51.9	61.6
	AdvPC	69.4	72.8	61.6	53.9	53.6	62.5
ℓ_2 $\epsilon = 1.25$	PGD	74.2	77.5	70.5	67.2	68.7	70.5
	AdvPC	69.0	72.9	63.0	64.5	55.4	67.9
ConvONet	None	90.1	92.8	92.5	92.8	93.2	93.5
ℓ_∞ $\epsilon = 0.05$	PGD	66.4	73.2	52.9	46.8	45.3	55.7
	AdvPC	63.7	71.2	55.5	47.2	46.7	55.0
ℓ_2 $\epsilon = 1.25$	PGD	72.2	76.7	69.8	65.6	62.7	71.4
	AdvPC	63.4	74.3	56.6	59.8	47.2	71.0

$\boldsymbol{x}$. Therefore, the implementation simplifies the gradient backward propagation through the top-k function as an indexing function to make the chain propagation smooth:

$$\{\boldsymbol{y}\}_1^k = \text{top-}k(\{\boldsymbol{x}\}_1^n) \quad \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}_i} = \begin{cases} 1 & \text{if } i \in \arg\text{top-}k(\{\boldsymbol{x}\}_1^n) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

However, such simplification still cannot resolve the differentibility issue of the top-k function [77].

Different from 2D model usually at most use one layer of $\max$ pooling, the heavy usage of kNN layers in DGCNN and PCT will drastically hinder the actual gradient flow. As mentioned in § 4.1, we exploit black-box SPSA and Nattack to validate our findings. Table 4 presents the results of AT. SPSA and Nattack can greatly lower the average robust accuracy (7.8%) than white-box attacks (55.6%) on DGCNN and PCT. This phenomenon exactly reveals *gradient obfuscation* as white-box attacks rely on the backward propagated gradient to succeed. The results demonstrated that the approximated gradients from black-box attacks are more accurate than the propagated ones. PointNet, however, achieves better robustness under black-box attacks because it only has one max pooling layer and does not employ kNN layers.

Table 4: Evaluation Results (Accuracy) of Standard Adversarial Training (%) with ℓ_∞ norm $\epsilon = 0.05$.

	PointNet	DGCNN	PCT
None	87.8	90.6	89.7
PGD	52.1	67.4	51.3
AutoAttack	40.5	56.4	47.2
SPSA	56.7	7.8	11.4
Nattack	55.1	5.4	6.5

Existing purification-based defenses against 3D adversarial point clouds mainly leverage C&W-styled attacks in their evaluation. C&W attacks utilize the method of Lagrange multipliers to find tractable adversarial examples while minimizing the magnitudes of the perturbation. From the perspective of adversary, such attacks are desirable due to their stealthiness, while this does not hold from a defensive view. Defense methods should be evaluated against strong adaptive attacks [6]. DUP-Net [90] is a pioneer study that uses statistical outlier removal and a upsampler network for purification, but it was adaptively attacked by [51]. We thus present the evaluation results of DUP-Net in Appendix A. IF-Defense and LPC are the SOTA adversarial purification methods for 3D point cloud models. We leverage PGD and AdvPC attacks, which assign constant adversarial budget in the adversarial optimization stage. We follow the original setups of IF-Defense and LPC in our study. Such evaluation is stronger than C&W attacks, while we note that they are not strict adaptive attacks since the adversarial target is still the classifier itself. Similar to *PointDP* , IF-Defense can be pre-pended to any point cloud classifier, but LPC uses a specific backbone. Table 5 presents the detailed evaluation results of IF-Defense under various settings and attacks. We find that *PointDP* achieves much better robustness than IF-Defense, which is on average an 12.6% improvements. However, IF-Defense achieves slightly higher clean accuracy (4.9%). This is because IF-Defense leverages SOR to smooth the point cloud [90]. However, such an operation has been demonstrated to be vulnerable [51]. With specific adaptive attacks, there will be a even larger drop of robust accuracy for IF-Defense.

Figure 2 shows the comparison among *PointDP* and existing methods. *PointDP* overall achieves the best performance than prior arts, which are 12.6% and 40.3% improvements than IF-Defense and LPC, respectively. We find that even without adaptive attacks, adversaries with constant budgets can already hurt the robust accuracy by a significant gap. This suggests that IF-Defense and LPC fail to deliver strong robustness to 3D point cloud recognition models. Especially, LPC appears in the proceedings of CVPR 2022, but actually achieves trivial robustness, indicating that a rigorous evaluation protocol is highly required in this community.

4.4 DEFENSE AGAINST ADAPTIVE THREATS

We have so far illustrated that state-of-the-art defenses can be easily broken by (adaptive) adversarial attacks and *PointDP* consistently achieves the best robustness. In this section, we further extensively evaluate the robustness of *PointDP* on even stronger adaptive attacks to demonstrate the actual robustness realized by *PointDP*. As mentions in § 4.1, we leverage two types of adaptive attacks in our study, and Table 6 presents their results. We also leverage black-box SPSA and Nattack to validate our results. We find that BPDA-PGD the strongest adaptive attacks, which align well with previous study on 2D diffusion-driven purification [36]. Even though with strong adaptive attacks, *PointDP* still achieves much better robustness. Besides, black-box attacks are much less effective. Although we admit that *PointDP* still relies on gradient obfuscation, the extremely high randomness will hinder the black-box adversaries finding correct gradients. We also ablate the effectiveness of *PointDP* with larger attack budgets in Appendix A.

Table 6: Evaluation Results (Accuracy) of **Strong Adaptive** Attacks on *PointDP* (%).

		PointNet	PointNet++	DGCNN	PCT	CurveNet	PointMLP
	None	86.8	87.9	86.9	87.0	88.0	88.2
ℓ_∞ $\epsilon = 0.05$	BPDA-PGD	77.1	78.6	79.2	76.1	73.9	77.7
	EOT-AutoAttack	78.0	79.9	79.1	76.5	75.9	78.9
	PGD	80.8	80.7	82.9	82.5	80.8	79.9
	AdvPC	69.9	76.8	79.4	79.8	72.9	75.4
	SPSA	76.6	78.9	74.9	78.5	76.4	80.9
	Nattack	75.2	77.9	74.4	78.0	76.1	78.9
	PA	81.7	84.7	84.1	84.5	84.8	85.2
ℓ_2 $\epsilon = 1.25$	BPDA-PGD	78.9	73.3	73.3	71.2	70.7	75.1
	EOT-AutoAttack	79.6	74.4	74.2	71.3	71.3	75.9
	PGD	86.1	87.5	82.5	86.3	87.7	87.8
	AdvPC	69.1	76.9	79.2	74.5	74.3	76.1
	SPSA	76.1	77.0	74.4	74.5	77.0	78.9
	Nattack	74.9	76.5	73.9	74.0	76.3	77.2
ℓ_0 $\epsilon = 200$	PD	61.3	72.1	73.5	75.9	74.1	74.4

5 DISCUSSION AND CONCLUSION

Adversarial robustness has been well-established in 2D vision tasks, where Carlini *et al.* [6] and many other researchers have devoted significant efforts to setting up a rigorous evaluation protocol. In this study, we emphasize that this evaluation protocol should be strictly followed in the 3D point cloud robustness study as well. Counter-intuitively, we have demonstrated that standard adversarial training (AT) is not a good candidate to deliver robustness against strong black-box adversaries because *gradient obfuscation* in 3D point cloud architectures will hinder the inner maximization stage from making real progress in AT. We propose *PointDP* as an adversarial purification strategy to mitigate the robustness loss in the 3D space. We want to clarify that almost all purification methods (including *PointDP*) still depend on *gradient obfuscation* to mislead adaptive attackers. However, we argue that proper usage of *gradient obfuscation* could still serve as a good defense, as long as the obfuscation is sophisticated enough. The multi-step purification in diffusion models adds extremely high-level randomness that EOT [3] and BPDA [2] attacks are hard to model. Therefore, we believe our extensive evaluation reveals the actual robustness of *PointDP*. Our evaluation also unveils a concerning fact that existing defenses in the 3D domain could be easily broken by strong attacks. Therefore, we hope our evaluation protocol sets a standard for robustness assessment in this community, *i.e.,* a defense study should strictly follow a formal distance metric and leverage strong attacks including PGD, black-box, and adaptive attacks to evaluate its actual robustness.

Limitation. Mitigation solutions to adversarial attacks are critical and essential for modern machine learning systems. Given that 3D point cloud is heavily adopted in safety-critical applications, we believe our study is valuable in demonstrating the vulnerabilities of existing SOTA defenses. *PointDP* also. On the other hand, diffusion models needs multiple steps in the reverse process to recover the point cloud and hinder adaptive attacks, which will incur additional computational overhead. *PointDP* also limits itself to empirical robustness without theoretical guarantees.

In this paper, we propose *PointDP*, an adversarial purification method against attacks on 3D point cloud recognition. We showed that adversarial training and prior purification methods are vulnerable to strong attacks. We then performed extensive rigorous evaluations to validate that *PointDP* outperforms existing SOTA methods by a significant margin (12.6%-40.3%) in robust accuracy.

ETHICS STATEMENT

As we continuously mentioned in our main paper and acknowledged in other studies [8, 6], adversarial robustness is critical to the real-world deployment of machine learning models, especially for safety-related applications. Point cloud data is heavily used in many such applications like autonomous driving, robotics, and medical imaging. Although various defenses were proposed in the literature, even in top-tier conferences like ICCV [90], CVPR [25], and NeurIPS [50], we find that they actually can be broken by carefully-designed strong attacks. Therefore, the first contribution of study is very beneficial for the 3D point cloud community to illustrate how existing state-of-the-art fail to deliver real robustness, as we have raised the attention for the 3D point cloud community to focus on actual robustness under strongest adaptive attacks. *PointDP* is also beneficial since we have leveraged the most rigorous evaluation protocol to test its robustness. We follow the licenses of usage for all the public models and datasets in our study.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we have provided our codebase in the supplementary materials and all of our results are based on consistent random seed in our implementation.

REFERENCES

[1] 3D Point Cloud Classification Benchmark on ModelNet40. `https://paperswithcode.com/sota/3d-point-cloud-classification-on-modelnet40`, 2021.

[2] A. Athalye, N. Carlini, and D. 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] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 284–293. PMLR, 10–15 Jul 2018.

[4] M. Bafna, J. Murtagh, and N. Vyas. Thwarting adversarial examples: An l_0-robustsparse fourier transform. *arXiv preprint arXiv:1812.05013*, 2018.

[5] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 2267–2281, 2019.

[6] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[7] N. Carlini, F. Tramer, J. Z. Kolter, et al. (certified!!) adversarial robustness for free! *arXiv preprint arXiv:2206.10550*, 2022.

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

[9] C. Choy, J. Gwak, and S. Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[10] W. Chu, L. Li, and B. Li. Tpc: Transformation-specific smoothing for point cloud models. *arXiv preprint arXiv:2201.12733*, 2022.

[11] F. Croce and M. 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.

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

[13] G. S. Dhillon, K. Azizzadenesheli, Z. C. Lipton, J. Bernstein, J. Kossaifi, A. Khanna, and A. Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

[14] X. Dong, D. Chen, H. Zhou, G. Hua, W. Zhang, and N. Yu. Self-robust 3d point recognition via gather-vector guidance. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11513–11521. IEEE, 2020.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[17] J. Goldberger, S. Gordon, H. Greenspan, et al. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *ICCV*, volume 3, pages 487–493, 2003.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[19] B. Graham and L. van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017.

[20] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu. Pct: Point cloud transformer. *arXiv preprint arXiv:2012.09688*, 2020.

[21] A. Hamdi, S. Rojas, A. Thabet, and B. Ghanem. Advpc: Transferable adversarial perturbations on 3d point clouds. In *European Conference on Computer Vision*, pages 241–257. Springer, 2020.

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

[23] L. Huang, C. Gao, Y. Zhou, C. Xie, A. Yuille, C. Zou, and N. Liu. Universal physical camouflage attacks on object detectors, 2019.

[24] L. Huang, C. Gao, Y. Zhou, C. Xie, A. L. Yuille, C. Zou, and N. Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 720–729, 2020.

[25] K. Li, Z. Zhang, C. Zhong, and G. Wang. Robust structured declarative classifiers for 3d point clouds: Defending adversarial attacks with implicit gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15294–15304, 2022.

[26] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31:820–830, 2018.

[27] Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In *International Conference on Machine Learning*, pages 3866–3876. PMLR, 2019.

[28] D. Liu, R. Yu, and H. Su. Extending adversarial attacks and defenses to deep 3d point cloud classifiers. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2279–2283. IEEE, 2019.

[29] Y. Liu, B. Fan, S. Xiang, and C. Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8895–8904, 2019.

[30] T. Lorenz, A. Ruoss, M. Balunović, G. Singh, and M. Vechev. Robustness certification for point cloud models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7608–7618, 2021.

[31] S. Luo and W. 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.

[32] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022.

[33] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[34] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928. IEEE, 2015.

[35] D. Meng and H. Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 135–147, 2017.

[36] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

[37] N. Papernot and P. McDaniel. Extending defensive distillation. *arXiv preprint arXiv:1705.05264*, 2017.

[38] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

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

[40] J. C. Pérez, M. Alfarra, S. Giancola, B. Ghanem, et al. 3deformrs: Certifying spatial deformations on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15169–15179, 2022.

[41] C. R. Qi, H. Su, K. Mo, and L. 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.

[42] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.

[43] G. Riegler, A. O. Ulusoy, and A. Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[44] A. Shafahi, M. Najibi, A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019.

[45] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020.

[46] S. Shi, X. Wang, and H. Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019.

[47] S. Song and J. Xiao. Deep Sliding Shapes for amodal 3D object detection in RGB-D images. In *CVPR*, 2016.

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

[49] J. Sun, Y. Cao, Q. A. Chen, and Z. M. Mao. Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 877–894. USENIX Association, Aug. 2020.

[50] J. Sun, Y. Cao, C. B. Choy, Z. Yu, A. Anandkumar, Z. M. Mao, and C. Xiao. Adversarially robust 3d point cloud recognition using self-supervisions. *Advances in Neural Information Processing Systems*, 34:15498–15512, 2021.

[51] J. Sun, K. Koenig, Y. Cao, Q. A. Chen, and Z. M. Mao. On the adversarial robustness of 3d point cloud classification, 2020.

[52] J. Sun, A. Mehra, B. Kailkhura, P.-Y. Chen, D. Hendrycks, J. Hamm, and Z. M. Mao. Certified adversarial defenses meet out-of-distribution corruptions: Benchmarking robustness and simple baselines. *arXiv preprint arXiv:2112.00659*, 2021.

[53] J. Sun, Q. Zhang, B. Kailkhura, Z. Yu, C. Xiao, and Z. M. Mao. Benchmarking robustness of 3d point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296*, 2022.

[54] L. P. Tchapmi, C. B. Choy, I. Armeni, J. Gwak, and S. Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *International Conference on 3D Vision (3DV)*, 2017.

[55] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6411–6420, 2019.

[56] F. Tramer, N. Carlini, W. Brendel, and A. Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

[57] T. Tsai, K. Yang, T.-Y. Ho, and Y. Jin. Robust adversarial objects against deep learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 954–962, 2020.

[58] J. Uesato, B. O'donoghue, P. Kohli, and A. Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pages 5025–5034. PMLR, 2018.

[59] D. Z. Wang and I. Posner. Voting for voting in online point cloud object detection. In *Robotics: Science and Systems*, volume 1, pages 10–15607. Rome, Italy, 2015.

[60] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017.

[61] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.

[62] Y. Wen, J. Lin, K. Chen, and K. Jia. Geometry-aware generation of adversarial and cooperative point clouds. 2019.

[63] M. Wicker and M. Kwiatkowska. Robustness of 3d deep learning in an adversarial setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11767–11775, 2019.

[64] E. Wong, L. Rice, and J. Z. Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

[65] W. Wu, Z. Qi, and L. 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.

[66] Z. Wu, Y. Duan, H. Wang, Q. Fan, and L. J. Guibas. If-defense: 3d adversarial point cloud defense via implicit function based restoration. *arXiv preprint arXiv:2010.05272*, 2020.

[67] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

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

[69] T. Xiang, C. Zhang, Y. Song, J. Yu, and W. Cai. Walk in the cloud: Learning curves for point clouds shape analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 915–924, 2021.

[70] C. Xiao, R. Deng, B. Li, T. Lee, B. Edwards, J. Yi, D. Song, M. Liu, and I. Molloy. Advit: Adversarial frames identifier based on temporal consistency in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3968–3977, 2019.

[71] C. Xiao, R. Deng, B. Li, F. Yu, M. Liu, and D. Song. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 217–234, 2018.

[72] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

[73] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.

[74] C. Xie, M. Tan, B. Gong, A. Yuille, and Q. V. Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

[75] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. In *International Conference on Computer Vision*. IEEE, 2017.

[76] C. Xie and A. Yuille. Intriguing properties of adversarial training at scale. In *International Conference on Learning Representations*, 2020.

[77] Y. Xie, H. Dai, M. Chen, B. Dai, T. Zhao, H. Zha, W. Wei, and T. Pfister. Differentiable top-k with optimal transport. *Advances in Neural Information Processing Systems*, 33:20520–20531, 2020.

[78] W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[79] C. Yang, A. Kortylewski, C. Xie, Y. Cao, and A. Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. In *European Conference on Computer Vision*, pages 681–698. Springer, 2020.

[80] Y. Yang, G. Zhang, D. Katabi, and Z. Xu. Me-net: Towards effective adversarial robustness with matrix estimation. *arXiv preprint arXiv:1905.11971*, 2019.

[81] T. Yin, X. Zhou, and P. Krähenbühl. Center-based 3d object detection and tracking. *CVPR*, 2021.

[82] H. Zhang, H. Chen, C. Xiao, S. Gowal, R. Stanforth, B. Li, D. Boning, and C.-J. Hsieh. Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316*, 2019.

[83] H. Zhang, H. Chen, C. Xiao, B. Li, D. S. Boning, and C.-J. Hsieh. Robust deep reinforcement learning against adversarial perturbations on observations. 2020.

[84] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019.

[85] Q. Zhang, S. Hu, J. Sun, Q. A. Chen, and Z. M. Mao. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15159–15168, June 2022.

[86] X. Zhang, A. Zhang, J. Sun, X. Zhu, Y. E. Guo, F. Qian, and Z. M. Mao. Emp: Edge-assisted multi-vehicle perception. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 545–558, 2021.

[87] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021.

[88] T. Zheng, C. Chen, J. Yuan, B. Li, and K. Ren. Pointcloud saliency maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1598–1606, 2019.

[89] H. Zhou, D. Chen, J. Liao, K. Chen, X. Dong, K. Liu, W. Zhang, G. Hua, and N. 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.

[90] H. Zhou, K. Chen, W. Zhang, H. Fang, W. Zhou, and N. 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.

A EVALUATION DETAILS

As mentioned in § 4.1, the robust accuracies of the unprotected base models are mostly 0%. Table 7 presents the detailed results.

Table 7: Evaluation Results (Accuracy) of Adversarial Attacks on Base Models(%).

		PointNet	PointNet++	DGCNN	PCT	CurveNet	PointMLP
	None	90.1	92.8	92.5	92.8	93.2	93.5
ℓ_∞ $\epsilon = 0.05$	C&W	0.0	0.0	0.0	0.0	0.0	0.0
	PGD	0.4	0.5	0.2	0.4	0.8	0.3
	AdvPC	0.4	0.3	0.0	0.2	0.6	0.3
	PA	44.1	19.9	35.1	20.8	48.9	7.2
ℓ_2 $\epsilon = 1.25$	C&W	0.0	0.0	0.0	0.0	0.0	0.0
	PGD	0.1	0.3	64.5	0.5	0.5	0.5
	AdvPC	0.0	0.5	62.7	0.4	0.3	0.5
ℓ_0 $\epsilon = 200$	PD	33.3	69.8	64.5	53.0	72.6	71.1

We also include [63] in our evaluation. [63] proposed ISO attack that iteratively drops the most salient points. This setting is very similar to our point-dropping (PD) adversary evaluated in § 4.2. The difference is that [63] leverages a heuristic way to determine critical points, but PD uses the gradient that backward propagates to each point to select the critical points. [63] only works for PointNet because i) both [63] and PointNet are very first explorations in the area of 3D point cloud recognition and ii) PointNet utilizes global max pooling so that only the critical points will affect the prediction results. We evaluate ISO under PointNet with an attack budget of 200 points; the results are shown in the Table 8.

Table 8: Robust Accuracy (%) of Different Purification Methods under ISO Attack.

	IF-Defense	***PointDP***
ISO [63]	67.3	**70.1**
PD	66.1	**68.9**

We find that ISO is a weaker attack than PD as it by design restricts its attack capability, which is good for an attack paper. However, it cannot show the worst-case robustness of a defense proposal.

We also evaluate DUP-Net with IF-Defense and *PointDP* under ℓ_∞ norm PGD attacks using different attack budgets. As Table 9 presents, DUP-Net is vulnerable to such attacks due to sensitivity of the upsampler network to ℓ_∞ norm noises [51]. The robust accuracy for LPC is 27.8% and 19.1% for $\epsilon = 0.075$ and $\epsilon = 0.1$, respectively. Even with these extremely large distortions, *PointDP* achieves the strongest robustness, outperforming existing SOTA by a very large margin.

Table 9: Evaluation Results (Accuracy) of Adversarial Attacks on Base Models(%).

		PointNet	PointNet++	DGCNN	PCT	CurveNet	PointMLP
ℓ_∞ $\epsilon = 0.05$	DUP-Net	0.0	1.3	0.9	0.9	0.6	1.0
	IF-Defense	66.4	73.2	52.9	46.8	45.3	55.7
	PointDP	**80.8**	**80.7**	**82.9**	**82.5**	**80.8**	**79.9**
ℓ_∞ $\epsilon = 0.075$	DUP-Net	0.5	0.3	0.0	0.2	0.2	0.6
	IF-Defense	60.7	67.3	47.2	40.9	39.8	50.9
	PointDP	**73.9**	**73.6**	**74.2**	**70.2**	**67.9**	**72.5**
ℓ_∞ $\epsilon = 0.1$	DUP-Net	0.0	0.0	0.0	0.2	0.1	0.3
	IF-Defense	53.9	57.1	42.0	35.1	33.3	44.7
	PointDP	**67.3**	**62.4**	**64.2**	**59.2**	**58.3**	**63.1**