A Critical Revisit of Adversarial Robustness in 3D Point Cloud Recognition with Diffusion-Driven Purification

Jiachen Sun [1] **Jiongxiao Wang** [2] **Weili Nie** [3] **Zhiding Yu** [3] **Z. Morley Mao** [1] **Chaowei Xiao** [2 3 4]

Abstract

3D point clouds serve as a crucial data representation in numerous real-world applications such as autonomous driving, robotics, and medical imaging. While the advancements in deep learning have spurred the utilization of 3D point clouds, deep models are notoriously vulnerable to adversarial attacks. Various defense solutions have been proposed to build robust models against adversarial attacks. In this work, we pinpoint a major limitation of the leading empirical defense, adversarial training, when applied to 3D point cloud models: gradient obfuscation, which significantly hampers robustness against potent 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 a variety of threats. We conduct a comprehensive evaluation of *PointDP* across six representative 3D point cloud architectures, employing sixteen strong and adaptive attacks to manifest its foundational 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 (Wang & Posner, 2015; Maturana & Scherer, 2015; Riegler et al., 2017; Wang et al., 2017; Qi et al., 2017a; Choy et al., 2019) and thus realized several safety-critical applications (*e.g.,* autonomous driving) (Yin et al., 2021; Shi et al., 2019; 2020). Although deep learning models (Qi et al., 2017a;b) have exhibited performance boosts on many challenging tasks, extensive studies show that they are notoriously vulnerable to adversarial attacks (Cao et al., 2019; Sun et al., 2020a; Xiang et al., 2019), where attackers manipulate the input in an imperceptible manner, leading to incorrect predictions of the target model. Because of the broad applications of 3D point clouds in safety-critical fields (Hu et al., 2021), it is imperative to study the adversarial robustness of point cloud recognition models.

The manipulation space for 2D adversarial attacks primarily involves altering pixel-level numeric values in input images. However, the flexible representation of 3D point clouds arguably expands the attack surface. For example, adversaries could shift or detach existing points (Xiang et al., 2019; Zheng et al., 2019), add new points into the pristine point cloud (Sun et al., 2021b), or generate totally new point clouds (Zhou et al., 2020) to launch attacks. Different strategies, including limits of the number of altered points and constraints of the maximal magnitude of shifted points (Sun et al., 2021b) were proposed to make attacks less perceptible. The inherent flexibility of 3D point cloud data formats enables a variety of attacks, complicating the development of a practical and universal defense mechanism.

Given the safety-critical nature of 3D point cloud applications, numerous studies have aimed to enhance the robustness of 3D point cloud recognition models. Pioneering efforts such as DUP-Net (Zhou et al., 2019) and GvG-PointNet++ (Dong et al., 2020) incorporated statistical outlier removal (SOR) modules as pre-processing and in-network blocks, respectively, as mitigation strategies. More lately, Sun et al. (2020b) 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 for PointNet, DGCNN, and PCT (Sun et al., 2021b). Meanwhile, advanced purification strategies like IF-Defense (Wu et al.,

[1]University of Michigan, Ann Arbor, MI, USA [2]Arizona State University, AZ, USA [3]NVIDIA, USA [4]University of Wisconsin, Madison, WI, USA. Correspondence to: Jiachen Sun <jiachens@umich.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2020) and LPC (Li et al., 2022) leverage more complex modules to cleanse the adversarial point clouds. However, given that the point cloud is a sparse and unstructured data format and there are significant differences between 2D and 3D perception models, it motivates us to re-think that *whether the current adversarial training and purification-based methods are robust enough against stronger adversarial attacks.*

Our journey in this work starts with revisiting the prior studies and exploring their actual adversarial robustness. By devising various types of strong adaptive attacks, we, for the first time, demonstrate that standard adversarial training (Madry et al., 2017) suffers from *gradient obfuscation* in the deep point cloud recognition models as the unstructured data format requires unique architectural designs to digest (§ 3). We also extensively evaluate IF-Defense and LPC to show that their purification strategies are actually vulnerable to stronger attacks and limits to perfectly-structured point clouds (§ 5.3).

Furthermore, we propose *PointDP*, an adversarial purification method that leverages a diffusion model as a preprocessing module to defend against 3D adversaries. As shown in Figure 2, *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 three 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 (§ 4.1), and (iii) feeding the reversed sample to the final classifier. Applying a diffusion denoiser for point clouds is non-trivial as they have fewer semantics than 2D images. Different from DiffPure (Nie et al., 2022) that relies on unconditional diffusion models, we leverage a conditional diffusion model to improve the quality of the purified input. We, therefore, use another supervised contrastive loss term to further improve the end-to-end robustness in the latent feature space during training our *PointDP*. Since *PointDP* does not rely on any types of predefined adversarial examples for training, it can defend against diverse unseen threats.

We rigorously evaluate *PointDP* with six representative point cloud models and sixteen adversarial attacks, including PGD (Sun et al., 2021b; Madry et al., 2017), C&W (Xiang et al., 2019; Carlini & Wagner, 2017), and point cloud-specific attacks (Zheng et al., 2019; Hamdi et al., 2020) 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 *three-fold*:

- We are the first to demonstrate that standard adversarial training (Madry et al., 2017; Sun et al., 2021b), 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$10%, which is no longer useful for 3D point cloud recognition.
- We propose *PointDP* that leverages diffusion models to purify adversarial 3D point clouds with a supervised contrastive loss term. *PointDP* is a general framework that is independent of the diffusion model used. We also formulate rigorous adaptive attacks on *PointDP*. We conduct an extensive evaluation on six representative models with numerous attacks to comprehensively understand the robustness of *PointDP*. Our evaluation shows that *PointDP* outperforms previous state-of-the-art (SOTA) purification methods, IF-Defense (Wu et al., 2020) and LPC (Li et al., 2022), by 12.6% and 40.3% on average, respectively. *PointDP* also achieves 14-27$\times$ speed up than SOTA purification methods.
- Based on our extensive exploration and experimentation, we set up a rigorous protocol with diverse attacks for robustness evaluation on 3D point cloud models to benefit future research in assessing the true robustness.

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 (He et al., 2016), followed by vision transformers (Dosovitskiy et al., 2020). However, there is currently no consensus on the architecture of 3D perception models since there is no standard data format for 3D perception (Sun et al., 2022). 3D networks at the early stage use dense voxel grids for perception (Maturana & Scherer, 2015; Song & Xiao, 2016; Tchapmi et al., 2017), which discretize a point cloud to voxel cells. PointNet pioneered leveraging global pooling to help achieve memory-efficient permutation invariance in an end-to-end manner. PointNet++ (Qi et al., 2017b) and DGCNN (Wang et al., 2019) followed up to add sophisticated local clustering operations to advance the performance. Sparse tensors are the other direction in 3D network designs (Graham & van der Maaten, 2017; Choy et al., 2019) to use 3D convolutions to improve 3D perception performance. PointCNN and RSCNN reformed the classic pyramid CNN to improve the local feature generation (Li et al., 2018; Liu et al., 2019b). PointConv and KPConv designed new convolution operations for point cloud learning (Wu et al., 2019; Thomas et al., 2019). PointTransformer and PCT advanced self-attention blocks in the 3D space and achieved good performance (Zhao et al., 2021; Guo et al.,

2020). Various novel local clustering operations (Xiang et al., 2021; Ma et al., 2022) also show enhancements in 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 achieve state-of-the-art results in point cloud recognition.

2.2. Adversarial Attacks and Defenses

Adversarial attacks have become the main obstacle that hinders deep learning models from real-world deployments, especially in safety-critical applications (Eykholt et al., 2018; Sun et al., 2020a; Zhang et al., 2021; 2022c). There are a lot of adversarial attacks proposed in the 2D space to break the various vision models (Carlini & Wagner, 2017; Xiao et al., 2018b; Yang et al., 2020; Xie et al., 2017; Huang et al., 2019; 2020; Sun et al., 2021c). To fill this gap between standard and robust accuracies, many mitigation solutions have been studied and presented to improve the robustness against adversarial attacks (Yang et al., 2019; Xu et al., 2017; Bafna et al., 2018; Papernot et al., 2016; Meng & Chen, 2017; Zhang et al., 2019a; Xiao et al., 2018a; Zhang et al., 2020; Xiao et al., 2019). However, most of them including adding randomization (Liu et al., 2019a; Dhillon et al., 2018; Dong et al., 2020), model distillation (Papernot et al., 2016), adversarial detection (Meng & Chen, 2017), and input transformation (Yang et al., 2019; Xu et al., 2017; Papernot & McDaniel, 2017; Bafna et al., 2018; Zhou et al., 2019) have been compromised by adaptive attacks (Tramer et al., 2020; Athalye et al., 2018a). Adversarial training (AT) (Madry et al., 2017; Goodfellow et al., 2014; Wong et al., 2020; Shafahi et al., 2019), in contrast, delivered a more longstanding mitigation strategy (Xie et al., 2020a; Xie & Yuille, 2020; Zhang et al., 2019b). Most recently, Nie et al. (2022) proposed DiffPure that leverages diffusion models to defend against adversarial attacks, and following-up studies to extend it to certified defenses (Carlini et al., 2022).

Adversarial attacks and defenses also extend to 3D point clouds. Xiang et al. (2019) 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. (2019) enhanced the loss function in C&W attack to achieve attacks with smaller perturbations and Hamdi et al. (2020) presented transferable black-box attacks on point cloud recognition. Wicker & Kwiatkowska (2019) pioneered to study the point dropping attack under both white- and black-box settings. Zhou et al. (2019) and Dong et al. (2020) proposed to purify the adversarial point clouds by input transformation and adversarial detection. However, these methods have been successfully by Sun et al. (2020b) through adaptive attacks. Moreover, Liu et al. (2019a) made a preliminary investigation on extending countermeasures in the 2D space to defend against simple attacks like FGSM (Goodfellow et al., 2014) on point cloud data. Sun et al. (2021b;a) conducted a more thorough study on the application of self-supervised learning in adversarial training for 3D point cloud recognition. Besides adversarial training, advanced purification methods IF-Defense (Wu et al., 2020) and LPC (Li et al., 2022) were proposed to transform the adversarial examples into a clean manifold. In this work, we present *PointDP*, which utilizes 3D diffusion models to purify adversarial point clouds that deliver SOTA empirical robustness. We also demonstrate that standard adversarial training suffers from strong black-box attacks and SOTA purification methods (*i.e.,* IF-Defense and LPC) are vulnerable to PGD-styled adversaries.

```python
def knn(x, k):
    inner = -2*torch.matmul(x.transpose(2, 1), x)
    xx = torch.sum(x**2, dim=1, keepdim=True)
    pairwise_distance = -xx - inner - xx.transpose(2,
     1)
    idx = pairwise_distance.topk(k=k, dim=-1)[1]
    # (batch_size, num_points, k)
    return idx

def get_graph_feature(x, k):
    #x's shape is (batch_size,num_dims,num_points)
    idx = knn(x, k=k) # (batch_size,num_points,k)
    ...... # shape transformation here
    feature = x.view(batch_size*num_points, -1)[idx, :]
    # idx is used as index to select features
    ......
    return feature

# forward function for EdgeConv
def forward(self, x):
    ......
    x = get_graph_feature(x, k=self.k)
    x = self.conv1(x) # convolution
    ......
```

Figure 1: PyTorch-Style Code Snippet of EdgeConv (Wang et al., 2019) in Point Cloud Recognition Models. Adversarial training fails since the kNN layers leverage the top-k function where the gradient propagates to the index, resulting in gradient obfuscation.

3. Catastrophe of Adversarial Training!

Adversarial training (AT) is well-known as the most longstanding empirical defense for the 2D classification task, which has been applied to PointNet, DGCNN, and PCT with the help of self-supervised learning (Sun et al., 2021b). However, we find that AT is, in fact, a weak defense solution in 3D perception models. First, 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 does not suit these models. Moreover, we discover that the kNN layers will cause severe *gradient obfuscation* in point cloud models as well. Different from the 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 addition-

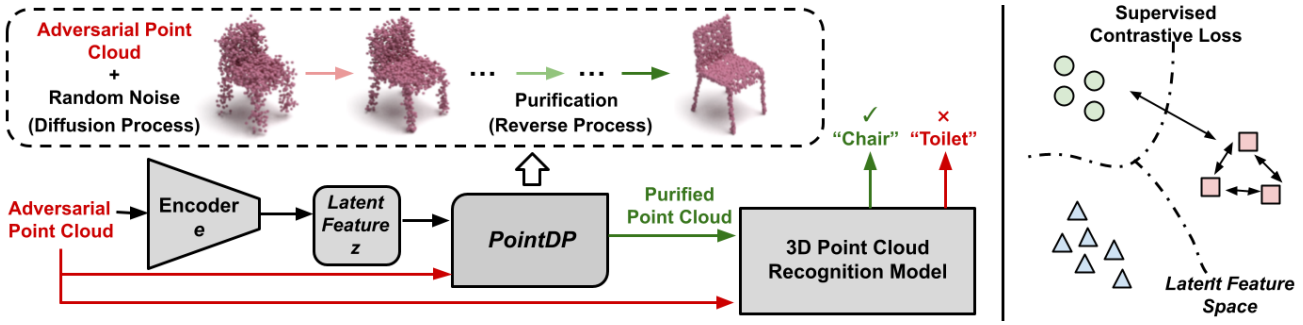

Figure 2: Illustration of *PointDP*, where *PointDP* serve as a purification module. We leverage a supervised contrastive loss term during the training of the diffusion model. The adversarial point cloud will be incorrectly classified as "toilet" by the recognition model if not purified by our *PointDP*.

Table 1: Robust Accuracy of 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

ally requires the gradient flow to the input (*i.e.,* point cloud) $\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}}$ in the inner maximization stage. As shown in Line 5 from Figure 1, kNN essentially applies top-k operation for point selection. Top-k is a general case for max pooling that does not have trainable model parameters, so it does not affect standard training. *However, top-k is not differentiable w.r.t. the input $\boldsymbol{x}$.* Therefore, the implementation simplifies the gradient flow through the top-k function as an indexing function to make the chain propagation smooth:

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

However, such simplification still cannot resolve the differentiability issue of the top-k function *w.r.t.* the input (Xie et al., 2020b).

Different from 2D models usually at most use one layer of max pooling, the heavy usage of kNN layers in 3D point cloud models like DGCNN and PCT will drastically hinder the actual gradient flow *w.r.t.* the input. As mentioned in § 5.1, we exploit black-box SPSA and Nattack to validate our findings. Table 1 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, in contrast, achieves stronger robustness under black-box attacks because it only has one max pooling layer and does not employ kNN layers. The failure of AT demonstrates that adversarial analysis of 3D point cloud models requires extra care. Otherwise, the claimed robustness may fail with adaptive attacks. We further show the failure of other purification methods in § 5.3. All these results highlight the desire for a rigorous robustness evaluation protocol for 3D point cloud models.

4. *PointDP*: Diffusion-Driven Purification

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 (§ 4.2). Next, we follow Nie et al. (2022) to apply the adjoint method to backward propagate through SDE for efficient gradient evaluation with strong adaptive attacks (§ 4.3).

4.1. Preliminaries

In this section, we briefly review the background of conditional diffusion models in 3D vision tasks. Following Luo & Hu (2021), 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 predefined 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:

$$\begin{aligned} q(\boldsymbol{x}_{1:N}|\boldsymbol{x}_0) &:= \prod_{n=1}^{N} q(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}), \\ q(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) &:= \mathcal{N}(\boldsymbol{x}_n; \sqrt{1-\beta_n}\boldsymbol{x}_{n-1}, \beta_n \mathbf{I}) \end{aligned} \tag{2}$$

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:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:N}|\boldsymbol{z}) &:= p(\boldsymbol{x}_N) \prod_{n=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n, \boldsymbol{z}), \\ p_{\boldsymbol{\theta}}(\boldsymbol{x}_{n-1}|\boldsymbol{x}_n, \boldsymbol{z}) &:= \mathcal{N}(\boldsymbol{x}_{n-1}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n, \boldsymbol{z}), \beta_n \mathbf{I}) \end{aligned} \tag{3}$$

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 (Luo & Hu, 2021). 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 (Dhariwal & Nichol, 2021), we can conduct the sampling by reparameterizing $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ as

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n, \boldsymbol{z}) = \frac{1}{\sqrt{1-\beta_n}} \left(\boldsymbol{x}_n - \frac{\beta_n}{\sqrt{1-\overline{\alpha}_n}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n, \boldsymbol{z}) \right) \tag{4}$$

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.

4.2. Design of *PointDP*

Overview. Figure 2 illustrates the pipeline of *PointDP*. Different from Nie et al. (2022) using unconditional diffusion model to remove the adversarial effect for 2D images, we use conditional diffusion models as mentioned in § 4.1. Specifically, *PointDP* first adds pre-quantified Gaussian noise to the input data and then leverages 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 Nie et al. (2022), in order to backward propagate through the forward and reverse processes for computing gradients, we first convert the discrete-time formulation defined in Eqs. (2) and (3) to its continuous-time counterpart, *i.e.*, the forward and reverse stochastic differential equations (SDEs) (Song et al., 2021). 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:

$$\boldsymbol{x}(t^*) = \sqrt{\alpha(t^*)}\boldsymbol{x}_a + \sqrt{1-\alpha(t^*)}\boldsymbol{\epsilon} \tag{5}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. We leverage Eq. 3 to recover the clean point clouds. Equivalently, the truncated reverse process can be also solved by the SDE solver in (Nie et al., 2022) (denoted as $\mathbf{sdeint}$):

$$\hat{\boldsymbol{x}}(0) = \mathbf{sdeint}(\boldsymbol{x}(t^*), \boldsymbol{f}_{\text{rev}}, g_{\text{rev}}, \boldsymbol{w}, t^*, 0) \tag{6}$$

where the six inputs are initial value, drift coefficient, diffusion coefficient, Wiener process, initial time, and end time (Nie et al., 2022), with the definitions:

$$\boldsymbol{f}_{\text{rev}}(\boldsymbol{x}, t, \boldsymbol{z}) = -\frac{1}{2}\beta(t)[\boldsymbol{x} + 2\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t, \boldsymbol{z})], \; g_{\text{rev}}(t) = \sqrt{\beta(t)} \tag{7}$$

and the score function $\boldsymbol{s}_{\boldsymbol{\theta}}$ is derived from $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_n, n, \boldsymbol{z})$ in Eq. (4) by following:

$$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t, \boldsymbol{z}) = -\frac{1}{\sqrt{1-\alpha(t)}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}(t), tN, \boldsymbol{z}) \tag{8}$$

Note that the hyper-parameter t^* and N trade 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 (§ 5). We also conduct ablation studies on t in § 5.2.

Since we leverage conditional diffusion models, we add a contrastive loss term (Khosla et al., 2020) to further improve the robustness of the latent feature during training:

$$\mathcal{L}_{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p)}{\sum_{a \in A(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a)} \tag{9}$$

where $A(i)$ denotes the mini-batch I except for i itself and $P(i) := \{p \in A(i) : y_p = y_i\}$. The intuition is that $\mathcal{L}_{\text{SupCon}}$ could further enforce the latent feature to be clustered distantly for different classes.

4.3. Adpative Attacks on *PointDP*

PointDP is a pre-processing module that purifies the adversarial perturbations. (Athalye et al., 2018a) 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 the 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:

$$\boldsymbol{x}_{s+1} = \text{Proj}_{\boldsymbol{x}+\mathcal{S}}(\boldsymbol{x}_s + \alpha \cdot \text{norm}(\nabla_{\boldsymbol{x}_s} \mathcal{D}(e(\boldsymbol{x}_s), \boldsymbol{z}_{\text{oracle}}))), \tag{10}$$

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 (Goldberger et al., 2003) and the other is the ℓ_1 norm distance. In our evaluation (§ 5), we report the lowest accuracy achieved under attacks with two distance metrics.

Attack on the Reverse Diffusion Process. We follow Nie et al. (2022) to formulate the adaptive attack as an augmented SDE process. We re-state the attack formulation as below. For the SDE in Equation 6, the augmented SDE that computes the gradient $\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(t^*)}$ of backward propagating through it is given by:

$$\begin{pmatrix} \boldsymbol{x}(t^*) \\ \frac{\partial \mathcal{L}}{\partial \boldsymbol{x}(t^*)} \end{pmatrix} = \textbf{sdeint}\left(\begin{pmatrix} \hat{\boldsymbol{x}}(0) \\ \frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{x}}(0)} \end{pmatrix}, \tilde{\boldsymbol{f}}, \tilde{\boldsymbol{g}}, \tilde{\boldsymbol{w}}, 0, t^* \right) \tag{11}$$

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 Equation 6, and

$$\tilde{\boldsymbol{f}}([\boldsymbol{x}; \boldsymbol{z}], t) = \begin{pmatrix} \boldsymbol{f}_{\text{rev}}(\boldsymbol{x}, t) \\ \frac{\partial \boldsymbol{f}_{\text{rev}}(\boldsymbol{x}, t)}{\partial \boldsymbol{x}} \boldsymbol{z} \end{pmatrix},$$

$$\tilde{\boldsymbol{g}}(t) = \begin{pmatrix} -g_{\text{rev}}(t)\mathbf{1} \\ \mathbf{0} \end{pmatrix}, \quad \tilde{\boldsymbol{w}}(t) = \begin{pmatrix} -\boldsymbol{w}(1-t) \\ -\boldsymbol{w}(1-t) \end{pmatrix}$$

where $\mathbf{1}$ and $\mathbf{0}$ denote the vectors of all ones and all zeros, respectively. Nie *et al.* (Nie et al., 2022) have demonstrated that such approximation aligns well with the actual gradient value. Therefore, we leverage this adaptive attack formulation for our evaluation.

5. Experiments and Results

In this section, we first introduce our experimental setups (§ 5.1). We then present the standard robustness evaluation of *PointDP*(§ 5.2). We next show that how the SOTA adversarial training and adversarial purification methods fail under various strong attacks (§ 5.3). We finally conduct a stress test on *PointDP* to show its actual robustness under various stronger adaptive attacks (§ 5.4).

5.1. Experimental Setups

Datasets and Network Architectures. We conduct all the main experiments on the widely used ModelNet40 point cloud classification benchmark (Wu et al., 2015), consisting of 12,311 CAD models from 40 artificial object categories.

Table 2: Robust Accuracy (%) of Plain Model on PA and PD on ModelNet40. Models under other attacks mostly have **0.0%** accuracy, which are detailed 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

In addition, we leverage ScanObjectNN (Uy et al., 2019) to further demonstrate the superiority of *PointDP*. ScanObjectNN is a real-world dataset consisting of 2,902 point clouds within 15 classes. 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 (Qi et al., 2017a). For the ScanObjectNN dataset, we adhere to the original configuration of 2048 points and maintain experimental setups consistent with those employed for ModelNet40. 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 (Qi et al., 2017a), PointNet++ (Qi et al., 2017b), DGCNN (Wang et al., 2019), PCT (Guo et al., 2020), CurveNet (Xiang et al., 2021), and PointMLP (Ma et al., 2022). These backbones either have representative designs (*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 involve 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 (Xiang et al., 2019), 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 (Carlini & Wagner, 2017; Sun et al., 2021b). We also have designed adaptive attacks on our proposed method § 4.3. Besides naive C&W and PGD attacks, we leverage specific attacks designed to break the robustness of point cloud recognition such as kNN (Tsai et al., 2020) and AdvPC (Hamdi et al., 2020). We also apply strong adaptive AutoAttack (Croce & Hein, 2020) (*i.e.*, APGD) in our evaluation. Moreover, we use SPSA (Uesato et al., 2018) and Nattack (Li et al., 2019) as black-box adversaries, followed by the suggestion of Carlini *et al.* (Carlini et al., 2019). We also leverage EOT-AutoAttack. Point adding (PA) and dropping/detaching (PD) attacks are also evaluated in our study, followed

Table 3: Robust Accuracy (%) of Adversarial Attacks on *PointDP* on ModelNet40. Colored rows correspond 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

Table 5: Robust Accuracy (%) of Adversarial Attacks on IF-Defense on ModelNet40. Colored rows correspond to rows in Table 3 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

by the setups in (Sun et al., 2021b). We set the attack steps to 200 to maximize the adversarial capability and follow the settings in (Sun et al., 2021b) for other attack parameters.

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).

Baseline. Without any defense applied to the original recognition models, the robust accuracy is mostly **0.0%** for all models under ℓ_2 and ℓ_∞ based attacks (see Appendix A). DGCNN exceptionally achieves 64% on ℓ_2-based PGD and AutoAttack, respectively, due to its dynamic clustering design, which adaptively discards outlier points. PA and PD are two weaker attacks and Table 2 presents robust accuracy against these two attacks.

5.2. 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 sequential order. Table 3 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 attacks is above 75%. We observe a drop in the clean accuracy for the chosen models due to the imperfect reconstruction of diffusion models. As mentioned before, designing diffusion models for 3D point clouds is a more difficult task than 2D image diffusion, which may lead to partial semantic loss. The average drop in 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*.

Table 4: Ablation Study on Overhead Introduced by Adversarial Purification Methods.

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

Figure 4 shows the averaged evaluation results of point shifting, adding, and dropping attacks with PGD-styled adversaries 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 cost of *PointDP* and other baselines using an RTX3080 GPU and a batch size of 32 over 100 runs. Table 4 presents the results, where *PointDP* achieves the lowest cost than existing SOTA methods, which is a **27**$\times$ speed-up than IF-Defense.

5.3. Comparison with State-of-the-Art Defenses

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 an 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 (Carlini et al., 2019). DUP-Net (Zhou et al., 2019) is a pioneer study that uses statistical outlier removal

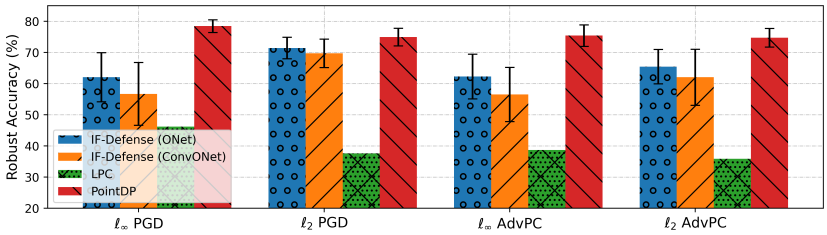

Figure 3: Compare among SOTA Adversarial Purification Strategies (*i.e.,* IF-Defense (Wu et al., 2020), LPC (Li et al., 2022), and *PointDP*). The results of IF-Defense and *PointDP* are averaged from six models on ModelNet40.

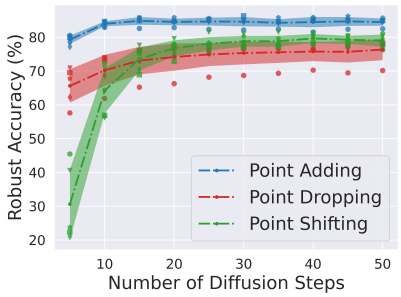

Figure 4: Ablation on Discrete Diffusion Steps in *PointDP* on ModelNet40.

and an upsampler network for purification, but it was adaptively attacked by (Sun et al., 2020b). 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 budgets 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 a 12.6% improvement. However, IF-Defense achieves slightly higher clean accuracy (4.9%). This is because IF-Defense leverages SOR to smooth the point cloud (Zhou et al., 2019). However, such an operation has been demonstrated to be vulnerable (Sun et al., 2020b). With specific adaptive attacks, there will be an even larger drop in robust accuracy for IF-Defense.

In addition, we have performed supplementary experiments on ScanObjectNN. The results, outlined in Table 6, underscore the effectiveness of *PointDP*. IF-Defense necessitates a pristine dataset for estimating the 3D occupancy field, which becomes infeasible with ScanObjectNN due to its real-world origin and partial visibility caused by occlusion. We use the pretrained ConvONet for these experiments. On the other hand, LPC transposes the point cloud into 2D space, thereby disrupting the native point cloud structure, rendering it inadequate for ScanObjectNN by default. *PointDP* manifests reasonable robust accuracies of 65.7% and 66.7% under ℓ_∞ and ℓ_2 norm PGD attacks, respectively. In contrast, other methods fail to maintain any level of robustness when applied to real-world ScanObjectNN.

Figure 3 shows the comparison among *PointDP* and existing methods. *PointDP* overall achieves the best performance than prior arts, demonstrating 12.6% and 40.3% improvements over IF-Defense and LPC, respectively. We find that

Table 6: Robust Accuracy (%) of PGD-styled Attacks on *PointDP* with Baselines on ScanObjectNN.

		PCT	CurveNet	PointMLP
ℓ_∞ $\epsilon = 0.05$	None	0.0	0.0	0.0
	DUP-Net	0.0	0.0	0.0
	IF-Defense	3.5	4.1	3.1
	LPC	-	-	-
	PointDP	**63.7**	**64.3**	**69.2**
ℓ_2 $\epsilon = 1.25$	None	0.0	0.0	0.0
	DUP-Net	8.2	7.9	10.1
	IF-Defense	5.1	4.5	4.9
	LPC	-	-	-
	PointDP	**64.0**	**65.9**	**70.1**

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 achieves trivial robustness, emphasizing that a rigorous evaluation protocol is highly required in this community. Evaluation results of ScanObjectNN further highlight the limitations of existing methods and substantiate the superior effectiveness of *PointDP*.

5.4. 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 mentioned in § 5.1, we leverage two types of adaptive attacks in our study, and Table 7 presents their results. We also leverage black-box SPSA and Nattack to validate our results. We find that BPDA-PGD has the strongest adaptive attacks, which aligns well with the previous study on 2D diffusion-driven purification (Nie et al., 2022). Even though with strong adaptive attacks, *PointDP* still achieves much stronger 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

Table 7: Robust Accuracy (%) of **Strong Adaptive** Attacks on Our Plain *PointDP*on ModelNet40.

		PointNet	PointNet++	DGCNN	PCT	CurveNet	PointMLP
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-Latent	80.8	80.7	82.9	82.5	80.8	79.9
	AdvPC-Latent	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-Latent	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-Latent	85.1	86.6	82.0	85.3	86.7	86.8
	AdvPC-Latent	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-Latent	61.3	72.1	73.5	75.9	74.1	74.4

black-box adversaries from finding correct gradients. We further ablate the usage of $\mathcal{L}_{\text{SupCon}}$ and Table 8 shows that it will bring additional $\sim 1.2\%$ robustness under attacks on the latent feature. We also ablate the effectiveness of *PointDP* with larger attack budgets in Appendix A, where *PointDP* consistently achieves the strongest robustness. In addition, we employ attacks with greater ℓ_∞ norm distance to dissect the extra robustness provided by $\mathcal{L}_{\text{SupCon}}$. The evaluation results, as presented in Table 9, indicate that $\mathcal{L}_{\text{SupCon}}$ is even more helpful in enhancing robustness with the increase in the attack budget.

Table 8: Robust Accuracy (%) of **Strong Adaptive** Attacks on *PointDP* with $\mathcal{L}_{\text{SupCon}}$ on ModelNet40.

		PCT	CurveNet	PointMLP
ℓ_∞ $\epsilon = 0.05$	PGD-Latent	83.9	81.8	81.0
	AdvPC-Latent	80.5	74.1	76.7
	PA-Latent	84.9	85.0	85.7
ℓ_2 $\epsilon = 1.25$	PGD-Latent	86.3	87.6	87.7
	AdvPC-Latent	76.3	76.0	77.7
ℓ_0 $\epsilon = 200$	PD-Latent	76.8	75.5	76.4

Further experiments, including larger attack budgets and additional adversaries, are conducted and detailed in Appendix A. Across different setups, *PointDP* consistently achieves the highest robust accuracy.

6. A Rigorous Robustness Evaluation Protocol

Our evaluation unveils a concerning fact that existing defenses in the 3D domain could be easily broken by strong attacks. Therefore, we follow Carlini et al. (2019) to set up a rigorous evaluation protocol to help future robustness assessment in the 3D point cloud community:

• A defense study should strictly follow formal distance metrics with quantified budgets. FGSM and C&W attacks are not designed for robustness evaluation. As mentioned in § 5.3, those attacks were proposed to minimize perturbations, which are not suitable for the defense evaluation. We strongly suggest future research use PGD-styled adversaries to test the real robustness with the claimed budget.

Table 9: Robust Accuracy (%) of PGD-styled Attacks on *PointDP* with $\mathcal{L}_{\text{SupCon}}$ on ModelNet40.

		PCT	CurveNet	PointMLP
ℓ_∞ $\epsilon = 0.075$	*PointDP*	65.1	65.2	67.8
	+$\mathcal{L}_{\text{SupCon}}$	**67.8**	**67.2**	**70.9**
ℓ_∞ $\epsilon = 0.1$	*PointDP*	53.2	53.2	57.4
	+$\mathcal{L}_{\text{SupCon}}$	**57.5**	**56.5**	**60.3**
ℓ_∞ $\epsilon = 0.125$	*PointDP*	40.5	40.0	43.7
	+$\mathcal{L}_{\text{SupCon}}$	**45.0**	**44.4**	**48.3**

• It is crucial to perform adaptive attacks on the proposed defense and verify that the adaptive attacks are effective. BPDA and EOT techniques are good methods to formulate adaptive attacks. Adaptive attacks usually should reflect the lower-bound robustness of a defense.

• Evaluation of black-box attacks is indeed necessary. As shown in § 3, the results of black-box attacks are a good indicator of severe gradient obfuscation. Should black-box attacks yield lower robust accuracy compared to white-box attacks, there's a significant likelihood that the defense is considerably less potent than its claimed efficacy.

• It is also suggested to perform point-cloud-specific attacks like point adding and dropping to demonstrate the generalization of the proposed defense. Other attacks include GeoA3 (Wen et al., 2020), AOF (Liu et al., 2022), and SS (Zhang et al., 2022a). It's worth acknowledging that our *PointDP* may be susceptible to transformation-based attacks like SS, given that SS lies outside our assumed ℓ_p norm threat model, as discussed in Appendix B.

7. Conclusion

In this paper, we propose *PointDP*, an adversarial purification method against attacks on 3D point cloud recognition. Our study exposes the vulnerability of adversarial training and current purification techniques under 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, while achieving 14-27$\times$ speed-up in purification.

Acknowledgment

We thank our area chairs and anonymous reviewers for their insightful comments and feedback. This work was partially supported by NSF under CNS-1930041, the National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks, Grant # 2112562, DHS No. 17STQAC00001-06-00, and a grant from Mcity.

References

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

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 284–293. PMLR, 10–15 Jul 2018b. URL https://proceedings.mlr.press/v80/athalye18b.html.

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

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

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

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

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

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

Croce, F. and 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.

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

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

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

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

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

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

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

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

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

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

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

Hu, S., Chen, Q. A., Sun, J., Feng, Y., Mao, Z. M., and Liu, H. X. Automated Discovery of Denial-of-Service Vulnerabilities in Connected Vehicle Protocols. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security '21)*, 2021.

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

Huang, L., Gao, C., Zhou, Y., Xie, C., Yuille, A. L., Zou, C., and Liu, N. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pp. 720–729, 2020.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Li, K., Zhang, Z., Zhong, C., and Wang, G. 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*, pp. 15294–15304, 2022.

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

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

Liu, B., Zhang, J., and Zhu, J. Boosting 3d adversarial attacks with attacking on frequency. *IEEE Access*, 10: 50974–50984, 2022.

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

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

Luo, S. and 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.

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

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

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

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

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

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

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

Qi, C. R., Su, H., Mo, K., and 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, 2017a.

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

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

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

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

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

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

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

Sun, J., Cao, Y., Chen, Q. A., and Mao, Z. M. Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and

countermeasures. In *29th USENIX Security Symposium (USENIX Security 20)*, pp. 877–894. USENIX Association, August 2020a. ISBN 978-1-939133-17-5. URL https://www.usenix.org/conference/usenixsecurity20/presentation/sun.

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

Sun, J., Cao, Y., Choy, C., Yu, Z., Xiao, C., Anandkumar, A., and Mao, Z. M. Improving adversarial robustness in 3d point cloud classification via self-supervisions. In *International Conference on Machine Learning Workshop (ICMLW)*, volume 1, 2021a.

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

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

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

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

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

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

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

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

Uy, M. A., Pham, Q.-H., Hua, B.-S., Nguyen, T., and Yeung, S.-K. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1588–1597, 2019.

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

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

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

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

Wen, Y., Lin, J., Chen, K., Chen, C. P., and Jia, K. Geometry-aware generation of adversarial point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (6):2984–2999, 2020.

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

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

Wu, W., Qi, Z., and 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.

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

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

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

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

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

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

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

Xie, C. and Yuille, A. Intriguing properties of adversarial training at scale. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HyxJhCEFDS.

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

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

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

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

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

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

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

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

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

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

Zhang, J., Dong, Y., Zhu, J., Zhu, J., Kuang, M., and Yuan, X. Improving transferability of 3d adversarial attacks with scale and shear transformations. *arXiv preprint arXiv:2211.01093*, 2022a.

Zhang, K., Zhou, H., Zhang, J., Huang, Q., Zhang, W., and Yu, N. Ada3diff: Defending against 3d adversarial point clouds via adaptive diffusion. *arXiv preprint arXiv:2211.16247*, 2022b.

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

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

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

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

Zhou, H., Chen, K., Zhang, W., Fang, H., Zhou, W., and 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.

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

A. Evaluation Details

As mentioned in § 5.1, the robust accuracies of the unprotected base models are mostly 0%. Table 10 presents the detailed results.

Table 10: Robust 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 (Wicker & Kwiatkowska, 2019) in our evaluation. (Wicker & Kwiatkowska, 2019) proposed ISO attack that iterative drops the most salient points. This setting is very similar to our point-dropping (PD) adversary evaluated in § 5.2. The difference is that (Wicker & Kwiatkowska, 2019) leverages a heuristic way to determine critical points, but PD uses the gradient that backward propagates to each point to select the critical points. (Wicker & Kwiatkowska, 2019) only works for PointNet because i) both (Wicker & Kwiatkowska, 2019) 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 Table 11.

Table 11: Robust Accuracy (%) of Different Purification Methods under the ISO Attack.

	IF-Defense	***PointDP***
ISO	67.3	**70.1**
PD	66.1	**68.9**

We observe that ISO is a less potent attack compared to PD as it inherently restricts its attack capability. While this may be advantageous for an attack paper, it fails to showcase 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 12 presents, DUP-Net is vulnerable to such attacks due to the sensitivity of the upsampler network to ℓ_∞ norm noises (Sun et al., 2020b). 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 an extremely large margin. Comparable improvements are observed under PGD attacks with larger ℓ_2 norms. We limit our selection to three models due to time constraints.

B. Discussion

Adversarial robustness has been well-established in 2D vision tasks, where Carlini *et al.* (Carlini et al., 2019) 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 (Athalye et al., 2018b) and BPDA (Athalye et al., 2018a) 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

Table 12: Robust Accuracy (%) of Adversarial Attacks on Different Purification Methods.

		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**
ℓ_2 $\epsilon = 2.0$	DUP-Net	-	-	-	40.1	39.8	44.7
	IF-Defense	-	-	-	50.9	51.4	56.3
	PointDP	-	-	-	**61.5**	**61.1**	**65.2**
ℓ_2 $\epsilon = 2.5$	DUP-Net	-	-	-	24.6	24.3	29.5
	IF-Defense	-	-	-	39.2	38.9	47.0
	PointDP	-	-	-	**46.9**	**44.8**	**53.1**

robustness. We notice a concurrent work (Zhang et al., 2022b) that primarily focuses on defending against 3D adversarial attacks using diffusion models under Chamfer distance. In contrast, our study proposes to address the formal ℓ_p norm-based adversarial robustness. We believe these two studies are complementary to each other.

Limitation. Mitigation solutions to adversarial attacks are critical and essential for modern machine learning systems. Given that the 3D point cloud is heavily adopted in safety-critical applications, we believe our study is valuable in demonstrating the vulnerabilities of existing SOTA defenses. On the other hand, diffusion models need multiple steps in the reverse process to recover the point cloud and hinder adaptive attacks, which will incur additional computational overhead, although *PointDP* has demonstrated to achieve the lowest cost. *PointDP* also limits itself to empirical robustness without theoretical guarantees. As previously stated, *PointDP* is currently designed to offer robustness against adversaries based on ℓ_p norms. Developing a more general defense mechanism is left for challenging future research.