On Isometry Robustness of Deep 3D Point Cloud Models under Adversarial Attacks

Yue Zhao[1], Yuwei Wu[1], Caihua Chen[2]*, Andrew Lim[1]

[1]Department of Industrial Systems Engineering and Management, National University of Singapore

[2]International Center of Management Science and Engineering,
School of Management and Engineering, Nanjing University

{yuezhao, ywwu}@u.nus.edu, chchen@nju.edu.cn, isealim@nus.edu.sg

Abstract

While deep learning in 3D domain has achieved revolutionary performance in many tasks, the robustness of these models has not been sufficiently studied or explored. Regarding the 3D adversarial samples, most existing works focus on manipulation of local points, which may fail to invoke the global geometry properties, like robustness under linear projection that preserves the Euclidean distance, i.e., isometry. In this work, we show that existing state-of-the-art deep 3D models are extremely vulnerable to isometry transformations. Armed with the Thompson Sampling, we develop a black-box attack with success rate over 95% on ModelNet40 data set. Incorporating with the Restricted Isometry Property, we propose a novel framework of white-box attack on top of spectral norm based perturbation. In contrast to previous works, our adversarial samples are experimentally shown to be strongly transferable. Evaluated on a sequence of prevailing 3D models, our white-box attack achieves success rates from 98.88% to 100%. It maintains a successful attack rate over 95% even within an imperceptible rotation range [±2.81°].

1. Introduction

Recently deep learning in 3D data sets has increasingly raised attentions due to its huge potential in real-world applications such as self-driving cars [14, 9], augmented reality [31] and medical image analysis [26]. Although techniques in 2D deep learning like multi-view approaches [39, 34] can be adopted for tackling 3D problems, it is hard to extend these approaches to some 3D tasks, for example point classification. To alleviate this difficulty, the authors of [33] proposed a novel deep neural network PointNet that directly consumes 3D point cloud data. With the proved invariance to order change, it performs well on point cloud
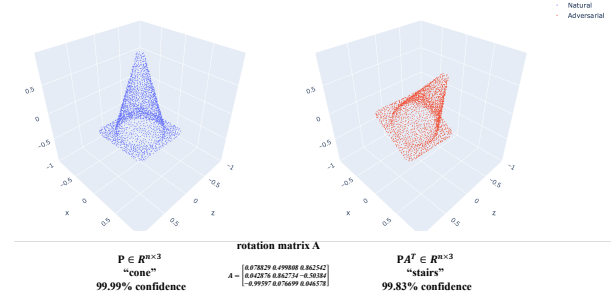
---

*Corresponding author.

Figure 1. Demonstration of Thompson Sampling Isometry attack on PointNet model [33] trained by ModelNet40 data [47]. Such classification error may cause safety hazard for autonomous vehicles, since a rotated traffic cone is commonly seen in real world scenarios.

classification and segmentation tasks. After this pioneering work, there is a surge of researches [35, 22, 23, 45, 24] on 3D neural networks that process directly on point clouds and achieve promising results.

While deep 3D models have gained remarkable success, the invariance or robustness to isometry transformation, a fundamental concern in computer vision [12], has not raised enough attention. It is also important to note that isometry represents a large class of geometry transformations including translation, rotation, and reflection as the most common examples. A misclassification case of a highway cone under our black-box attack is illustrated in Figure 1. To address this issue, in this paper we design a novel block-box attack framework based on Thompson Sampling (TS) [36] and evaluate the robustness of the existing state-of-art deep 3D models (PointNet [33], PointNet++ [35], DG-CNN [45] and RS-CNN [24]) for classification problems under linear isometry. Specifically, in the black box TS is used to generate rotation matrix parameterized by its rotation angle, and the resulted rotation transformation can be applied to nature

samples. The adoption of TS is motivated by the basic belief that the attack success rate varies in different rotation ranges. In order to generate successful attacks, a natural idea is to sample with probability to rotation angles. Compared to random sampling, TS is more efficient and explainable due to its ability to adaptively learn the prior probability of parameters. With the TS techniques, our block box attack can achieve success rates more than 95.66% on ModelNet40 data over all evaluated models by at most 10 times of iterative sampling. Heat map of attack success rates in different regions suggests certain rotation ranges are more likely to fool models, which confirms our insights. We also show links between our attacks and efficient 3D data augmentation subsequently.

Beyond isometry, we hope a deep 3D model can be robust under slightly perturbed isometry, since human beings have the capability to easily recognize a stretched 3D object from the natural one. This idea naturally leads to the Restricted Isometry Property (RIP) matrix [6], where RIP is a nice relaxation of isometry property extensively investigated in the compressed sensing field. In fact, a RIP matrix is coordinated with a small positive number called Restricted Isometric Constant (RIC), as a measurement of its distance from an isometry. In the white-box attack, we start from an initial isometry and further add perturbation on it to transform point cloud. The motivation is to preserve isometry as much as possible while misleading the model. Although finding RIC of a given matrix A is computational hard [42] in general, it is equivalent to compute the spectral norm of $A^T A - I$ in our case. Besides, spectral norm offers a fair and universal standard of perturbation level among different point cloud data sets, which is similar to the L_p norm standard in image domain. On an array of models trained by ModelNet40 data [47], we achieve attack success rates more than 98.11% while the spectral norms are less than 0.068.

The contributions of this paper are threefold:

- We utilize Thompson Sampling to design an efficient black-box attack with zero loss in the sense of isometry. It achieves success rates over 95% on ModelNet40 data set, which is the first work to fully reveal the vulnerability of deep 3D point cloud models under isometry transformations in black-box settings. To the best of our knowledge, this is the first black-box attack in 3D point cloud adversarial settings.

- Inspired by the theoretical results of Restricted Isometry Property, we introduce a novel framework of white-box attack on point cloud with success rates more than 98.11% over several state-of-the-art models. Even for imperceptible rotation angles in $[-\pi/64, \pi/64]$ ($\pm 2.81°$), it achieves a success rate over 95%.

- In contrast to previous works, 3D adversarial samples

generated by our algorithms have significantly strong transferability. This property further shows the vulnerability of current models under isometry projections. It also suggests huge potential of our approach in developing further black-box attacks.

The rest of the paper is organized as follows: We start with related works in deep 3D point cloud models, adversarial attacks in 3D settings and data augmentation. Then we offer technique details of isometry characterization, TS and RIP in preliminaries. After preliminaries, our attack methods are proposed with the pseudo-code of the algorithms. In the end, we evaluate attack performance for on state-of-the-art models trained by two bench mark data sets[1]. Furthermore, we give an analysis of the trade-off between accuracy and robustness of models from a data augmentation perspective.

2. Related Work

Deep 3D Point Cloud Models. 3D data sets are of intrinsic difference from images, thus current neural network designed for 2D domain might not be directly applied in certain 3D tasks. To address this issue, [33] proposed PointNet, a neural network which directly processes on point clouds. Due to its efficiency and invariance to order changes of points, it has been widely accepted and developed [35, 45, 24]. A concept called the critical set, is introduced in [33]. It plays an important role in enhancing model robustness, whereas [46] showed in practice it might instead be a weakness exploited by attacks. PointNet also used a mini-network called T-net and regularization term to make the feature transformation matrix near orthogonal matrix, so that it could be robust against rigid transformations (isometry). However, [24] noticed that accuracy of PointNet reduced to around 40% under rotations about Y axis by $90°$ or $180°$. We observe that the issue of isometry invariance has not been well explored in previous works. It will be our main focus in this work.

Adversarial Samples in 3D. In image domain, adversarial attacks [17, 30, 20, 7], along with the defense approaches [25, 2, 51], have been extensively studied. However, 3D point cloud data is essentially coordinates, which are completely different from pixel values, thereby it is necessary to design new approaches for adversarial attacks in 3D. By utilizing critical points, [53] assigned each point a score measuring its contribution to model training loss, then built saliency maps for attack. With an optimization approach, [48] generated adversarial point clouds by perturbation under L_p norm and adding points in local area. They also made use of critical points as the initial points for solving optimization problems in attack. [46] developed a point-wise occlusion attack by iteratively excluding points from

---

[1]https://github.com/skywalker6174/3d-isometry-robust

2

point clouds. [49] proposed a pointwise gradient-based perturbation attack and a point attachment/detachment attack, as well as statistical tools to distinguish between adversarial samples and noisy data for defense. All the above papers in 3D adversarial learning develop attacks based on local information of individual points, whereas in this work we analyse the robustness invoking the global geometry structure of point clouds.

On Data Augmentationn. For image problems, data augmentations are well-studied to improve the performance of deep learning models. Common strategies can be roughly divided into traditional methods (cropping, reflecting, rotating, scaling, and translation) and learning-based methods [11, 27, 16], see [38] for a recent survey. Data augmentation is not so well-addressed on point clouds. To our knowledge, simple strategies, such as random scaling, translation, and rotation, are most commonly used in recent studies [33, 35, 23, 54]. In this paper we show that these strategies perform far from satisfactory. Consequently, the study of data augmentation, along with the adversarial attack, may have great potential to improve both accuracy and robustness of deep 3D models.

3D Isometry Robustness. The isometry robustness considered in the past are mainly involved with 3D object retrieval [19, 41] and matching problems [4] in computer vision. Although they are different from the classification problem in this paper, we believe many strategies can be adopted and applied in deep learning settings. There are two possible ways of incorporation: 1) normalization (pose estimation), for example, the most prominent one based on PCA [44] can transform data to certain orientation, which could neutralize the effect of rotations, 2) isometry invariant descriptors based on computer vision theory, such as [19, 28], which can be embedded into the network architectures and enhance isometry robustness. Recently there are a few concurrent works [32, 52, 40] utilizing those ideas. However, in general they could suffer from more complex structures, more computational time or less accuracy. Our work fills the research gap between those attempts and prevailing deep 3D models by illustrating the vulnerability of the latter and justifying the importance of the former. Furthermore, we hope our methods can be served as an effective tool to evaluate isometry robustness and offer insights for future works.

3. Preliminaries

3.1. Linear isometry and Orthogonal matrix

Linear Isometry. An *isometry* of $\mathbb{R}^n$ is a function $f : \mathbb{R}^n \to \mathbb{R}^n$ that preserves the Euclidean distance

$$\|f(x) - f(y)\| = \|x - y\| \quad \forall x, y \in \mathbb{R}^n \quad (1)$$

Trivial examples of isometry include the identity transformation $f(x) = x$ and translation $f(x) = x + b$. An isometry that fixes the origin is indeed a linear transformation on $\mathbb{R}^n$ [43]. Because the effect of translations can always be easily eliminated by normalization, in this paper we only consider the case of linear isometry, i.e., f is a linear transformation that preserve the Euclidean distance. In the rest of this paper, we mean linear isometry when referring to isometry.

Orthogonal matrix. An $n \times n$ matrix A is said to be *orthogonal* if

$$A \in \mathrm{O}(n) := \{A \in \mathrm{GL}_n(\mathbb{R}) : A^T A = I\} \quad (2)$$

where $\mathrm{GL}_n(\mathbb{R})$ denotes the set of $n \times n$ invertible matrices of real numbers. It is easy to prove that $\det(A) = \pm 1, \forall A \in \mathrm{O}(n)$. Let $\mathrm{Iso}(n)$ denote the set of isometries of $\mathbb{R}^n$, [43] (Lemma1.7, Proposition 1.8) shows

$$\mathrm{Iso}(n) = \mathrm{O}(n) \quad (3)$$

Rotation matrices are defined as a subset of $\mathrm{O}(n)$

$$\mathrm{SO}(n) := \{A \in \mathrm{O}(n) : \det(A) = 1\} \quad (4)$$

A reflection matrix over a hyper-plane passing origin is defined by the householder matrix [18]

$$P = I - 2vv^T \quad (5)$$

where v is the unit normal vector of the hyper-plane $\{x \in \mathbb{R}^n : v^T x = 0\}$.

Parameterization of linear isometry. By CartanDieudonn theorem any matrix in $\mathrm{O}(3)$ can be described as at most three reflections, and a two reflections case in $\mathbb{R}^3$ is equal to a rotation. To further represent a rotation, the Euler's rotation theorem claims that in $\mathbb{R}^3$ we can always represent a rotation matrix by three rotation matrices about x, y, z axes

$$R = R_{\theta_x} R_{\theta_y} R_{\theta_z} \quad (6)$$

where R_{θ_i} is the rotation matrix about axis-i by angle $\theta_i \in [-\pi, \pi]$. Hence any rotation is captured by three parameters. A unit normal vector can be characterized by two angles thus by (5) any reflection over a hyper-plane passing origin can be represented by two parameters.

The parameterization of linear isometry is the foundation of our statistical-based black-box attack, indeed the parameter space constitutes the sampling space. Further details of this subsection are discussed in supplementary materials.

3.2. Thompson Sampling

Thompson sampling (TS), also known as posterior sampling and probability matching, is a powerful sampling

method that consecutively learns the prior probability of the parameters for sample distribution during the procedure. In this work, we adopt a Beta-Bernoulli finite bandit arm model [36] because of its efficiency in computation and conjugacy property.

Suppose there is an agent with K actions, each action $k \in \{1, \cdots, K\}$ on the environment results in a reward one with probability $\theta_k \in [0, 1]$ and zero with probability $1 - \theta_k$. $(\theta_1, \cdots, \theta_K)$ are unknown to the agent but fixed over time. Suppose the agent has an independent prior belief for choosing θ_k. Let the priors over θ_k to be the beta distribution with parameter α_k, β_k. Then the prior probability density function for θ_k is

$$p(\theta_k) = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha) + \Gamma(\beta)} \theta_k^{\alpha_k - 1} (1 - \theta_k)^{\beta_k - 1} \qquad (7)$$

where Γ is the gamma function. Then after each observation from the environment the distribution of θ_k is updated according to Bayes' rule. By conjugacy property of beta distribution we can derive an update rule with simple form:

$$(\alpha_k, \beta_k) \leftarrow \begin{cases} (\alpha_k, \beta_k) & , k_t \neq k \\ (\alpha_k, \beta_k) + (r_t, 1 - r_t) & , k_t = k \end{cases} \qquad (8)$$

where $t \in \{1, \cdots, T\}$ denotes the time horizon, k_t, r_t are the action and reward in time t.

Later in our black-box attack settings, the action k will be to sample a parameterized isometry in a local region corresponding to k. The reward is one if attack succeeds and zero otherwise. We assume the reward of attack follows Bernoulli distribution with success probability θ_k, which is learnt by TS during the sampling.

3.3. Restricted Isometry Property and Spectral Norm Penalty

The Restricted Isometry Property (RIP) was introduced by [6] to develop theorems in the field of compressed sensing. RIP can describe the almost orthonormal system, which makes it a perfect mathematical tool to quantify the variation of an isometry to a restricted isometry.

For all s-sparse vectors $x \in \mathbb{R}^n$, i.e., x has at most s nonzero coordinates, matrix A is said to satisfy s-restricted isometry with constant δ if

$$(1 - \delta)\|x\|^2 \leq \|Ax\|^2 \leq (1 + \delta)\|x\|^2 \qquad (9)$$

The smallest quantity of such δ is called the s-Restricted Isometry Constant (RIC) of A. We just consider the case $s = n$ since 1) there is no sparse constraint on vectors, 2) this allow us to impose orthogonality to A from an RIP perspective [3]. The RIP condition in (9) is then

$$\left|\frac{\|Ax\|^2}{\|x\|^2} - 1\right| \leq \delta, \quad \forall x \in \mathbb{R}^n \qquad (10)$$

Let $\sigma(A)$ be the spectral norm of A, which is the largest singular value of A. It can be easily shown

$$\sigma(A^T A - I) = \sup_{x \in \mathbb{R}^n, x \neq 0} \left|\frac{\|Ax\|^2}{\|x\|^2} - 1\right| \qquad (11)$$

Hence, finding the n-RIC of A is equivalent to computing $\sigma(A^T A - I)$. We call $\sigma(A^T A - I)$ the spectral norm penalty of A. It will be the core ingredient to help us develop the white-box attack in the next section.

4. Attack Methods

In this paper, we only consider adversarial attacks on classification models. Our approaches can be easily extended to other cases like segmentation models.

4.1. Point Cloud and Model Notations

Point Cloud. Point cloud is a cluster of points in $\mathbb{R}^3$ that represents an object or a scene. Each point p in a point cloud is described by its Cartesian coordinates $p = (x_p, y_p, z_p)^T \in \mathbb{R}^3$. Using a matrix notation, a point cloud P with m points is then described by $P = [p_1, p_2, \ldots, p_m]^T \in \mathbb{R}^{m \times 3}$.

Deep 3D Point Cloud Model. A deep 3D point cloud model is a function $F_\theta(\cdot) : \mathbb{R}^{m \times 3} \rightarrow \mathbb{T}^c$, where $\mathbb{T}^c = \{(a_1, \ldots, a_c) \in \mathbb{R}_+^c : \sum_{j=1}^c a_j = 1\}$ and c is the total number of classes. We will omit θ later since in our analysis the model is fixed. Suppose F is the full neural network with the softmax function and let $Z(P) = z \in \mathbb{R}^c$ be the output (called logit) except the softmax function, then $F(P) = \text{softmax}(Z(P))$.

4.2. Thompson Sampling Isometry (TSI) Attack

The intuition of our isometry attack is to find an isometry on point cloud that misleads the model. For a given point cloud P, let $C(P) = \text{argmax}_j F(P)_j$ be the classification result and $C^*(P)$ be the true label of P, then the *isometry attack* is

$$\begin{aligned} \text{find} \quad & A \\ s.t. \quad & A \in \text{Iso}(3) \\ & \text{argmax}_j F(PA^T)_j \neq C^*(P) \end{aligned} \qquad (12)$$

For better interpretation and conciseness, we will use the attack based on rotations as examples in this subsection. Following the results in subsection 3.1, we can use three rotation angles $(\theta_x, \theta_y, \theta_z) \in [a, b]^3$ to construct an arbitrary rotation matrix of $\mathbb{R}^3$. It is not efficient to choose parameters from the uniform distribution $U([a, b]^3)$, let alone capturing any useful information during the sampling procedure. To address this issue, we utilize TS to design this attack. First we divide $[a, b]^3$ in to d^3 intervals of the form

$$X_i \times Y_j \times Z_h \qquad (13)$$

for $(i, j, h) \in [d]^3 := \{1, \ldots, d\}^3$, where

$$X_i = [a + \frac{(i-1)(b-a)}{d}, a + \frac{i(b-a)}{d}]$$
$$Y_j = [a + \frac{(j-1)(b-a)}{d}, a + \frac{j(b-a)}{d}] \quad (14)$$
$$Z_h = [a + \frac{(h-1)(b-a)}{d}, a + \frac{h(b-a)}{d}]$$

Applying an action $k = (i, j, h)$ means to sample $\theta_x \sim U(X_i), \theta_y \sim U(Y_j), \theta_z \sim U(Z_h)$ and use them to create a rotation matrix by (6). Then we generate adversarial samples by this matrix and update parameters of TS. Our algorithm is as followed

---

Algorithm 1 TSI$(F, \{P_1, \ldots, P_N\}, a, b, d, S)$

Input: model F, point clouds $\{P_1, \ldots, P_N\}$, parameters range $[a, b]$, number of intervals d, maximum sampling times S.
Output: adversarial isometries $A_1^*, \ldots, A_N^*$
1: Initiate $\alpha_k \leftarrow 1, \beta_k \leftarrow 1, \forall k \in [d]^3$, where k is corresponding to partition in (14)
2: **for** $n = 1$ to N **do**
3:     **for** $s = 1$ to S **do**
4:         Sample $\hat{p}_k \sim \text{beta}(\alpha_k, \beta_k), \forall k \in [d]^3$
5:         $k_s \leftarrow \arg\max_k \hat{p}_k$
6:         Apply action k_s and get an isometry matrix $A^{(s)}$
7:         **if** $\arg\max_h F(P_n(A^{(s)})^T)_h = C^*(P_n)$ **then**
8:             $r_s \leftarrow 0$
9:         **else**
10:            $r_s \leftarrow 1$
11:         **end if**
12:         $(\alpha_{k_s}, \beta_{k_s}) \leftarrow (\alpha_{k_s} + r_s, \beta_{k_s} + 1 - r_s)$
13:     **end for**
14:     $A_n^* \leftarrow \arg\min_{A^{(s)}} F(P_n(A^{(s)})^T)_{C^*(P_n)}$
15: **end for**

We assume the prior distribution of p_s follows beta distribution $\text{Beta}(\alpha_s, \beta_s)$ with mean $\frac{\alpha_s}{\alpha_s + \beta_s}$, which represents the attack success rate in area corresponding to s. In TSI, we can adopt an early stop if the attack succeeds before reaching maximal sampling times.

To show the power of TS, in Figure 2 we present heat maps[2] of $\frac{\alpha_s}{\alpha_s + \beta_s}, s \in [d]^3$ projected to three planes. The heat map strongly testifies our insight that when the rotation angle gets closer to 0, it is less likely to achieve a successful attack.

4.3. Combined Targeted Restricted Isometry (CTRI) Attack

To further evaluate the robustness, we propose the *restricted isometry attack* and formulate it as an optimization problem

$$\min_{A \in \text{GL}_3(\mathbb{R})} \quad \sigma(A^T A - I)$$
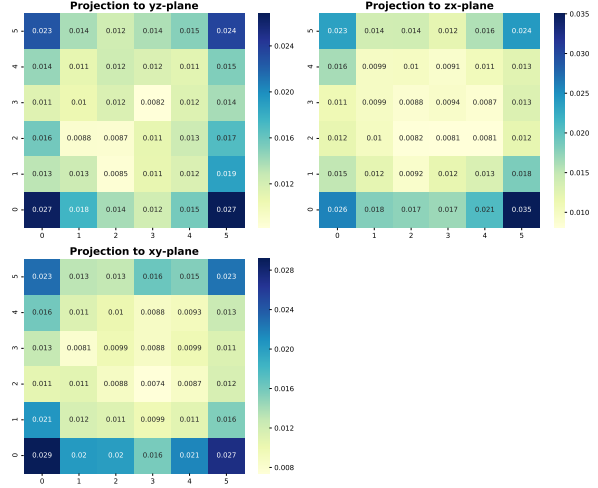$$s.t. \quad \arg\min_j F(PA^T)_j \neq C^*(P) \quad (15)$$

Figure 2. Heat map of success rate in each region learnt by TS. The number of each area represents probability of successful attack if sampling angles from this region. We project the numbers along each axis by taking means, thus they can be seen as marginal probability of a successful attack.

Based on the results from subsection 3.3, the objective function is indeed the 3-RIC of A. To impose a targeted attack, i.e., an attack that manipulates the classification result to be a targeted class t, we use the CW loss function introduced by [7]

$$g_t(P) = \max\{-\kappa, \max_{j \neq t} Z(P)_j - Z(P)_t\} \quad (16)$$

where $\kappa \geq 0$ is a chosen threshold. Then the Targeted Restricted Isometry (TRI) attack is formulated as

$$\min_{A \in \text{GL}_3(\mathbb{R})} \quad \sigma(A^T A - I) + \lambda g_t(PA^T) \quad (17)$$

where λ is a positive real number.

In practice we usually use Stochastic Gradient Descent (SGD) to iteratively solve optimization problems. To solve the problem in (17), a good initial point is of vital importance. Fortunately, since for any matrix $A \in \text{Iso}(3)$, $\sigma(A^T A - I) = 0$, TSI actually provides a fine enough starting point for TRI attack. We now propose the Combined Targeted Restricted Isometry (CTRI) attack in Algorithm 2.

CTRI attack, enjoying both the merit of TS and the theoretical interpretation of RIP matrices, is a novel framework of white-box attack in 3D settings. By isometry parameterization, TS is capable to sample any isometry and explore the weakness of 3D model under different isometry transformations. Adjusting interval division parameters of TS gives us different scopes of examination, for example we can explore to what rotation range current 3D models are robust, which may provide insights on how to efficiently conduct data augmentation. As we will show in the experiment section, relaxation of isometry to RIP matrix produce a variety of adversarial samples that preserves the global geometry shape.

Algorithm 2 CTRI Attack

Input: model F, point clouds $\{P_1, \ldots, P_N\}$, parameters range $[a, b]$, number of intervals d, maximum iteration times K, maximum sample times S, learning step η, penalty coefficient λ, target class t and κ.
Output: adversarial samples $P_1(A_1^*)^T, \ldots, P_N(A_N^*)^T$

    for $n = 1$ to N **do**
2:       Set $A_n^{(0)} \leftarrow \text{TSI}(F, \{P_n\}, a, b, d, S)$, only initiate (α, β) in TS at the first time
       $k \leftarrow 0$
4:       **while** $k < K$ and $F(P_n(A_n^{(k)})^T) = C^*(P_t)$ **do**
          $A_n^{(k+1)} \leftarrow A_n^{(k)} - \eta \nabla_{A_n^{(k)}} [\sigma((A_n^{(k)})^T A_n^{(k)} - I) + \lambda g_t(P_n(A_n^{(k)})^T)]$
6:       $k \leftarrow k + 1$
       end while
8:       $A_n^* \leftarrow A_n^{(k)}$
    end for

Moreover, the spectral norm penalty, as a unified standard directly measured on transformation, makes the perturbation levels of different adversarial samples fairly comparable regardless the difference of data sets or models.

5. Experiment Results and Analysis

In this section, we evaluate the performance of our attacks on the following models: PointNet, PointNet++, DG-CNN and RS-CNN. We conduct our experiments on two data sets: ModelNet40 [47] and part of ShapeNet [50]. For conciseness and better interpretation, we will only use rotation matrices in this section. Experiments of other isometries such as reflections will be presented in supplementary materials, indeed attacks based on reflections performs even better under certain settings.

5.1. Victim Models Setting

We retrain the above models with normalized (to a unit sphere) and augmented data as the victim models. The augmentation includes random rotation about y axis, random scale, random translation and jitter [33]. For ModelNet40, we use the official split of 9843 point clouds for training and 2468 for testing and attack, the sampling method is the same as those in [33]. For ShapeNet we use a 16 classes part with an official split to 12128 objects for training and 2874 for testing and attack. 2048 points are used for a single cloud in both data sets. The test accuracy of each retrained model is within 3% of the best reported accuracy.

In the following experiments, attacks will be exert on these retrained models. Without extra statement, in one attack evaluation, we use 1) $N = 2000$ point clouds; 2) parameter range $[a, b] = [-\pi, \pi]$ and division $d = 4$ for TS. A point cloud that victim model fails to correctly classify is not counted in statistics of attack performance.

	ModelNet40			ShapeNetPart		
	S=1	S=2	S=10	S=1	S=2	S=10
PointNet	92.69	96.92	98.28	83.10	92.61	98.11
PointNet++	83.74	91.31	95.66	79.32	89.57	92.96
DG-CNN	79.57	89.81	97.33	73.69	87.17	96.82
RS-CNN	85.68	94.19	96.81	76.46	90.01	97.41

Table 1. Attack success rates (%) of TSI with different maximum sampling times S.

5.2. TSI Attack

We first evaluate TSI attack alone as an approach of black-box attack. Table 1 shows the results of $S = 1, 2, 10$. As shown in the table, with maximum sampling times increasing to 10, TS achieves an attack success rate over 95% on ModelNet40, revealing the vulnerability of current 3D models under isometry.

Current black-box attacks in 2D are mainly through adversarial transferability [29], heuristic methods [1] and estimation of gradients [8]. These methods may not work in 3D settings due to different structure of data, for example, [48] finds that transferability is not obvious in 3D adversarial attacks. In comparison, TSI performs exceptionally well as a black-box attack.

The reason why we adopt TS is that we believe for different rotation angles the impact on victim model is different. For example, if the rotation angle is very close to 0, attack might not work, but if it is close to $\pi/4$, attack might easily succeed. This insight is strongly confirmed by heat maps in Figure 2. By understanding isometry properties of data set and model, TSI approach also shades light on how to conduct more effective and efficient data augmentation. We further guess a more delicate designed Thompson Sampling model such as an infinite-arm model might work better. Moreover, it might also perform well to generate different parameterized transformation other than isometry. We leave them for future work.

5.3. CTRI Attack

We then evaluate CTRI, where the target is chosen to be the label with second largest logit predicted by victim models, and $\kappa = 0$ is adopted. In order to find a good enough initial point, we use $S = 50$ in TSI. The attack success rate is presented in Table 2 for maximum gradient iteration times $K = 7, 50, 1000$. We adopt an early stop here, i.e., when TSI or CTRI succeeds before reaching maximum iteration times, we move to the next sample. We also show the statistics of $\sigma(A^T A - I)$ in Table 3, 4 for restricted isometry A that successfully fools the victim models.

Visualization of TSI and CTRI adversarial examples are shown in Figure 1 and 3. Figure 1 shows misclassification of a rotated highway cone, which may cause safety issues for self-driving cars. Figure 3 consists of natural and adversarial samples of a stair, where the global shape is well

	ModelNet40			ShapeNetPart		
	K=7	K=50	K=1000	K=7	K=50	K=1000
PointNet	99.44	99.66	100.00	99.38	99.69	100.00
PointNet++	97.93	98.11	98.88	96.19	96.71	99.34
DG-CNN	97.99	98.28	99.45	98.26	99.08	99.95
RS-CNN	98.21	98.38	99.33	98.23	98.84	99.95

Table 2. Attack success rates (%) of CTRI with different K , same $\lambda = 0.001$ and learning step $\eta = 0.0005$. Here we use $S = 50$ in TSI.

	Max	Mean	Var	Mean*	Var*
PointNet	0.064	8E-5	4E-6	0.033	3E-4
PointNet++	0.055	1E-4	3E-6	0.006	9E-5
DG-CNN	0.066	9E-5	4E-6	0.020	5E-4
RS-CNN	0.068	6E-5	3E-6	0.025	5E-4

Table 3. Statistics of $\sigma(A^T A - I)$ for the second column of table 2 (ModelNet40, $K = 50$). Since we adopt early stop, $\sigma(A^T A - I)$ will be zero if TSI succeeds. 'Mean*' and 'Var*' are the mean and variance over nonzero ones. They better reflect difference caused by TRI alone.

	Max	Mean	Var	Mean*	Var*
PointNet	1.11	0.002	0.001	0.331	0.107
PointNet++	1.00	0.004	0.003	0.114	0.063
DG-CNN	0.99	0.004	0.003	0.276	0.091
RS-CNN	0.84	0.003	0.002	0.318	0.073

Table 4. Statistics of $\sigma(A^T A - I)$ for the third column of table 2 (ModelNet40, $K = 1000$).
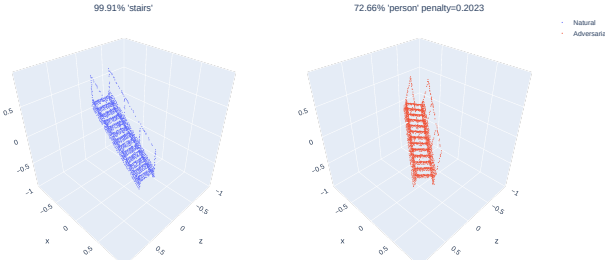

Figure 3. Illustration of CTRI on PointNet trained by Model-Net40 data. The object in blue is classified as 'stairs' with confidence 99.91%, while the grey object generated by CTRI is predicted as 'person' with confidence 72.66%. Spectral norm penalty $\sigma(A^T A - I) = 0.2033$.

preserved in the adversarial sample. More examples are demonstrated in the supplementary materials.

5.4. Transferability

Transferability is a commonly-seen property in 2D adversarial samples [29]. It refers to the case that an adversarial sample generate from one model can mislead another different model. However, in deep 3D models the adversarial samples proposed by previous works do not have strong

	PointNet	PointNet++	DG-CNN	RS-CNN
PointNet	/	86.61	80.62	85.40
PointNet++	95.76	/	85.68	92.71
DG-CNN	95.53	93.34	/	93.11
RS-CNN	95.04	92.58	84.59	/

Table 5. Transfer attack success rate (%) of CTRI with $S = 50, K = 1000$. The adversarial sample is generated from the model corresponding to the row and tested in model corresponding to the column.

transferability [48, 49].

On the contrary, we find that adversarial samples generated by CTRI attacks have strong transferability, as shown in Table 5. In fact, a transfer attack should perform at least not worse than a random generated isometry, i.e., TSI attack when $S = 1$. Such high rates indicate again that these models are vulnerable to isometry transformations. Another reason why our adversarial samples have much stronger transferability than that of other approaches is that, CTRI tries to explore the global geometry structure, while others depend much more on the property of specific model such as critical points.

5.5. Attack of Small Rotation Angles

Instead of random sampling rotation angles from $[-\pi, \pi]$, in this subsection we examine TSI and CTRI attack with a smaller range, say $[-\epsilon, \epsilon]$. The result is shown in Table 6. TSI attack maintains a success rate of 69.529% when the range shrinks to $[-\pi/16, \pi/16]$. It only fails when the range reduces to an imperceptible level ($\pi/64 = 2.81°$), indicating the model is just robust to relatively small range of rotations. Meanwhile CTRI attack succeeds easily with rates over 95% and means of penalties less than 0.3.

This experiment also provides useful information for efficient data augmentation, which we will discuss in the following subsection.

5.6. On Defense: the Price of Robustness

Data augmentation usually slightly lowers down the accuracy but helps improve the robustness [23]. However, our experiments show current data augmentation methods in 3D (random rotation about y, random scale and translation, jitter) are not enough for defense against our attacks.

To better understand the relation between robustness and accuracy from a data augmentation approach, we design an experiment: 1) First, we propose an augmentation method called *random p-rotation*. It is to randomly rotate the data by (6) with probability p. 2) Then we train PointNet model with random p-rotation ModelNet40 data 15 times for different $p \in \{0, 0.1, \ldots, 0.9, 1.0\}$. 3) Finally, we randomly choose a victim model for each p, and launch both rotation and reflection based TSI and CTRI attacks (same parameters as those in Table 2, $K = 1000$) on these models. The

	CTRI (%)	TSI (%)	Max	Mean	Var	Mean*	Var*
$[-\pi/2\ ,\pi/2\ ]$	100.000	99.335	1.012	0.001	8E-4	0.201	0.076
$[-\pi/4\ ,\pi/4\ ]$	99.944	99.386	0.443	0.001	4E-4	0.197	0.024
$[-\pi/8\ ,\pi/8\ ]$	99.944	98.153	1.460	0.003	0.003	0.214	0.101
$[-\pi/16,\pi/16]$	99.776	69.529	1.406	0.037	0.010	0.139	0.021
$[-\pi/32,\pi/32]$	98.222	18.000	1.833	0.178	0.047	0.246	0.047
$[-\pi/64,\pi/64]$	95.536	5.469	2.605	0.287	0.083	0.356	0.080

Table 6. Attack evaluation of small rotation angles on ModelNet40 data and PointNet model. The first two columns are the attack success rate of iterative CTRI and TSI with $K = 1000, S = 100$, the rest are statistics for $\sigma(A^T A - I)$. Other parameters are the same as those in Table 2.
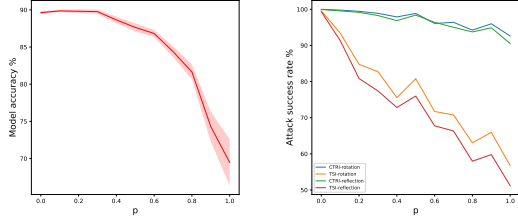

Figure 4. When p goes up, the right figure shows the attack success rates decrease, while the left figure suggests model accuracy also goes down and the variance increases. Thus, there is a trade-off between model accuracy and robustness when data augmentation approach is considered.

result is demonstrated in Figure 4.

Clearly, accuracy varies a lot with different p. The mean accuracy goes down to 69.47% when $p = 1$. TSI attack success rates decrease as p increases, but CTRI success rates still remain in a high level. Thus, the point is how to find an efficient way of data augmentation without jeopardizing the accuracy too much. As shown in Figure 4, roughly when $p \in [0.1, 0.3]$, the model accuracy remains around 90% while TSI attack success rates reduce to around 80%. Hence, random rotation with a low probability is a relatively efficient way to defense against TSI attack.

Another efficient data augmentation method could be to randomly rotate the data by several fixed angles, say $\epsilon, 2\epsilon, \ldots, k\epsilon$, where ϵ is chosen from subsection 5.5. The intuition is that since current models are robust to small rotation angle $[-\epsilon, \epsilon]$, we can take ϵ as the minimum step length to rotate the data. This approach is left for future work.

We note that 3D augmentation should draw more attention, since many data sets in 3D are often not so adequate for training large models. In particular, simple augmentations are not sufficient under adversarial attacks. On the one hand, data augmentation often brings large variance, which decreases the performance of models [21]. On the other hand, augmentations without understanding of attacks are less efficient, and can lead to computational difficulties [5]. An efficient data augmentation may be able to defense against TSI without too much cost, but against CTRI attack

a lot more computational power might be required, for example, by adversarial training [25].

To go a step further, generally it remains unknown why many deep models are vulnerable to adversarial examples [15, 37, 13]. Recently, it is shown that computation constraints might be one reason [5]. Particularly, local minima are usually believed to be satisfactory [10] so local methods such as SGD are often applied for training deep neural network. However, this might not be the case if it is facing adversarial examples. One conjecture is that different local minimum does not performs equally under adversarial examples and finding a specific local minimum is computational intractable. This explains why small perturbation leads to huge performance deterioration.

6. Conclusion

We propose a black-box attack and a white-box attack in 3D adversarial settings, which is the first work to show the extreme vulnerability of current 3D deep learning model under isometry transformations. On the one hand, there are intriguing properties of our adversarial samples, such as strong transferability. On the other hand, our approaches point out promising directions for future researches in 3D adversarial learning: 1) TS shows tremendous potential in developing black-box attack, which can also be further exploited in data augmentation. 2) CTRI achieves satisfying performance even under small rotation angles, while the geometry property of adversarial samples is well preserved. Therefore, further researches on defense against TSI and CTRI may hugely improve isometry robustness of 3D deep learning models.

7. Acknowledgement

References

[1] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B Srivastava. Genat-

tack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1111–1119. ACM, 2019. 6

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018. 2

[3] Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep cnns? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4266–4276. Curran Associates Inc., 2018. 4

[4] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. Efficient computation of isometry-invariant distances between surfaces. *SIAM Journal on Scientific Computing*, 28(5):1812–1836, 2006. 3

[5] Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 831–840, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 8

[6] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE TRANSACTIONS ON INFORMATION THEORY*, 51(12):4203, 2005. 2, 4

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 2, 5

[8] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017. 6

[9] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 1

[10] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2933–2941. Curran Associates, Inc., 2014. 8

[11] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017. 3

[12] Asi Elad and Ron Kimmel. On bending invariant signatures for surfaces. *IEEE Transactions on pattern analysis and machine intelligence*, 25(10):1285–1295, 2003. 1

[13] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *Advances in Neural Information Processing Systems*, pages 1178–1187, 2018. 8

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1

[15] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018. 8

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[18] Alston S Householder. Unitary triangularization of a nonsymmetric matrix. *Journal of the ACM (JACM)*, 5(4):339–342, 1958. 3

[19] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, volume 6, pages 156–164, 2003. 3

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 2

[21] Shiyi Lan, Ruichi Yu, Gang Yu, and Larry S Davis. Modeling local geometric structure of 3d point clouds using geo-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 998–1008, 2019. 8

[22] Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9397–9406, 2018. 1

[23] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems*, pages 820–830, 2018. 1, 3, 7

[24] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8895–8904, 2019. 1, 2

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 8

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 1

[27] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 3

[28] Marcin Novotni and Reinhard Klein. 3d zernike descriptors for content based shape retrieval. In *Proceedings of the*

eighth ACM symposium on Solid modeling and applications, pages 216–225, 2003. 3

[29] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 6, 7

[30] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram 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. 2

[31] Youngmin Park, Vincent Lepetit, and Woontack Woo. Multiple 3d object tracking for augmented reality. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 117–120. IEEE Computer Society, 2008. 1

[32] Adrien Poulenard, Marie-Julie Rakotosaona, Yann Ponty, and Maks Ovsjanikov. Effective rotation-invariant point cnn with spherical harmonics kernels. In *2019 International Conference on 3D Vision (3DV)*, pages 47–56. IEEE, 2019. 3

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 1, 2, 3, 6

[34] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 1

[35] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017. 1, 2, 3

[36] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018. 1, 4

[37] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018. 8

[38] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, Jul 2019. 3

[39] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 1

[40] Xiao Sun, Zhouhui Lian, and Jianguo Xiao. Srinet: Learning strictly rotation-invariant representations for point cloud classification and segmentation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 980–988, 2019. 3

[41] Johan WH Tangelder and Remco C Veltkamp. A survey of content based 3d shape retrieval methods. In *Proceedings*

Shape Modeling Applications, 2004., pages 145–156. IEEE, 2004. 3

[42] Andreas M Tillmann and Marc E Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2013. 2

[43] Christopher Triola. *Special orthogonal groups and rotations*. PhD thesis, Ph. D. dissertation, University of Mary Washington, 2009. 3

[44] Dejan V Vranic, Dietmar Saupe, and Jörg Richter. Tools for 3d-object retrieval: Karhunen-loeve transform and spherical harmonics. In *2001 IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564)*, pages 293–298. IEEE, 2001. 3

[45] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):146, 2019. 1, 2

[46] Matthew Wicker and Marta Kwiatkowska. Robustness of 3d deep learning in an adversarial setting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11767–11775, 2019. 2

[47] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong 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. 1, 2, 6

[48] Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9136–9144, 2019. 2, 6, 7

[49] Jiancheng Yang, Qiang Zhang, Rongyao Fang, Bingbing Ni, Jinxian Liu, and Qi Tian. Adversarial attack and defense on point sets. *arXiv preprint arXiv:1902.10899*, 2019. 3, 7

[50] Li Yi, Vladimir G Kim, Duygu Ceylan, I Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, Leonidas Guibas, et al. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (TOG)*, 35(6):210, 2016. 6

[51] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019. 2

[52] Zhiyuan Zhang, Binh-Son Hua, David W Rosen, and Sai-Kit Yeung. Rotation invariant convolutions for 3d point clouds deep learning. In *2019 International Conference on 3D Vision (3DV)*, pages 204–213. IEEE, 2019. 3

[53] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Pointcloud saliency maps. *arXiv preprint arXiv:1812.01687*, 2018. 2

[54] Y. Zhou and O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, June 2018. 3