PointCA: Evaluating the Robustness of 3D Point Cloud Completion Models against Adversarial Examples

Shengshan Hu[1,4,5,6,7], Junwei Zhang[1,4,5,6,7], Wei Liu[1,4,5,6,7], Junhui Hou[9], Minghui Li[3*], Leo Yu Zhang[10], Hai Jin[2,4,7,8], Lichao Sun[11]

[1] School of Cyber Science and Engineering, Huazhong University of Science and Technology
[2] School of Computer Science and Technology, Huazhong University of Science and Technology
[3] School of Software Engineering, Huazhong University of Science and Technology
[4] National Engineering Research Center for Big Data Technology and System
[5] Hubei Engineering Research Center on Big Data Security
[6] Hubei Key Laboratory of Distributed System Security
[7] Services Computing Technology and System Lab
[8] Cluster and Grid Computing Lab
[9] Department of Computer Science, City University of Hong Kong
[10] School of Information Technology, Deakin University
[11] Department of Computer Science and Engineering, Lehigh University
{hushengshan,jwzh,weiliu73,minghuili,hjin}@hust.edu.cn, jh.hou@cityu.edu.hk, leo.zhang@deakin.edu.au, lis221@lehigh.edu

Abstract

Point cloud completion, as the upstream procedure of 3D recognition and segmentation, has become an essential part of many tasks such as navigation and scene understanding. While various point cloud completion models have demonstrated their powerful capabilities, their robustness against adversarial attacks, which have been proven to be fatally malicious towards deep neural networks, remains unknown. In addition, existing attack approaches towards point cloud classifiers cannot be applied to the completion models due to different output forms and attack purposes. In order to evaluate the robustness of the completion models, we propose PointCA, *the first adversarial attack against 3D point cloud completion models*. PointCA can generate adversarial point clouds that maintain high similarity with the original ones, while being completed as another object with totally different semantic information. Specifically, we minimize the representation discrepancy between the adversarial example and the target point set to jointly explore the adversarial point clouds in the geometry space and the feature space. Furthermore, to launch a stealthier attack, we innovatively employ the neighbourhood density information to tailor the perturbation constraint, leading to geometry-aware and distribution-adaptive modification for each point. Extensive experiments against different premier point cloud completion networks show that PointCA can cause a performance degradation from 77.9% to 16.7%, with the structure chamfer distance kept below 0.01. We conclude that existing completion models are severely vulnerable to adversarial examples, and state-of-the-art defenses for point cloud classification will be partially invalid when applied to incomplete and uneven point cloud data.

*Corresponding Author.

Introduction

With the flourishing development of diverse 3D sensors like LiDAR, RADAR, and depth camera, point cloud data are utilized among many safety-critical fields such as autonomous driving, augmented reality, and robotics. Due to its great success in the computer vision area (He et al. 2016), deep learning techniques have been widely applied to point cloud tasks as well. Recent studies show that neural networks are vulnerable to adversarial attacks, where misclassifications can be easily triggered when facing adversarial examples (Goodfellow, Shlens, and Szegedy 2015; Hu et al. 2022a). Therefore, more and more research efforts are devoted to exploring the security and robustness of deep learning-based 3D point cloud systems.

Unfortunately, existing works only concentrate on the point cloud classification scenario, whereas point cloud completion, another important task for point cloud systems, has received no attention. The completion model is designed to restore the incomplete point cloud data induced by various real-world circumstances, and thus it has become a necessary upstream procedure in point cloud processing.

As opposed to point cloud classification, adversarial attack of point cloud completion is more challenging since it requires the manipulation of the geometric shapes rather than the object labels. Specifically, the output of the completion model is an instantial point cloud with semantic and geometric shape, instead of a hard label that directly indicates which class the object belongs to. No classification score or cross-entropy loss can be exploited to generate an adversarial point cloud. In other words, the goal of adversarial attacks on the point cloud completion models is to generate a misleading geometric shape rather than a false class label, thus a totally different loss function is needed to measure the similarity between the adversarial example and the target. Secondly, point cloud classification generally

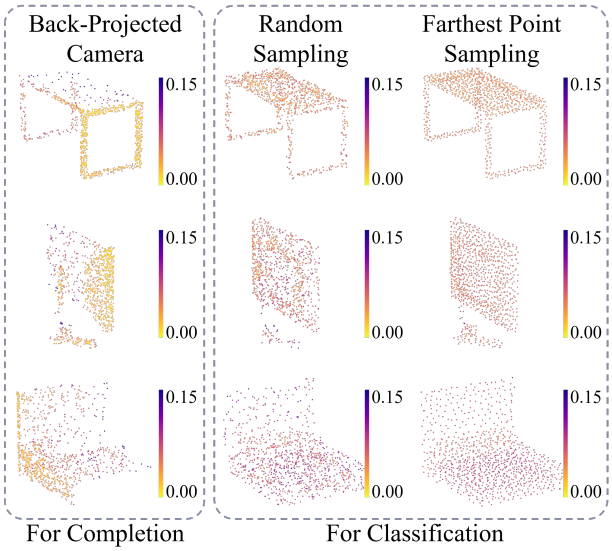


Figure 1: Geometric distribution of partial point cloud data for completion and complete point cloud data for classification. Partial point clouds have a significantly more complex and locally diverse geometric distribution.

generates adversarial examples over synthetic datasets (Wu et al. 2015), whose point clouds are complete and uniform with continuous geometric manifolds, whereas the incomplete partial point clouds in the completion tasks have a more realistic and uneven geometric distribution caused by the occlusion and limited sensor viewing angle. To verify this, Fig. 1 visualizes the difference in geometric distributions of the input point clouds from two tasks, assigning a score to each point depending on the density of its surrounding region through Eq. (10). The data distribution is more fluctuating and inconsistent in partial point clouds. How to efficiently and stealthily generate adversarial examples on more nonuniformly distributed data is a challenging problem that has been ignored by previous works in the classification, but is crucial for point cloud completion attacks.

In this paper, we propose PointCA, the first adversarial attack on 3D point cloud completion models to systematically evaluate their robustness. As shown in Fig. 2, the adversarial example generated by PointCA is visually similar to the original one, but completed as another target object with totally different semantic information. To overcome the above challenges, PointCA first explores the similarity measurement in the geometry space and the feature space respectively, formulating the adversarial example generation as an appropriate optimization problem which can be solved by the gradient-based algorithm. In addition, considering the unevenness and incompleteness of the partial point cloud data, we define and allocate local geometric neighborhood for each input point through the *k-Nearest Neighbor* (kNN) algorithm and evaluate the distribution density in the neighbor point sets, based on which the perturbation constraint will be adaptively tailored for each point to achieve a more imperceptible and efficient attack.

To summarize, the contributions of our work are as follows:

- We propose PointCA, *the first adversarial attack on 3D point cloud completion models*. By investigating the characteristics of the geometry space and the feature space, an appropriate optimization problem is formulated to measure the similarity between geometric shapes and find adversarial examples.
- We innovatively employ the neighborhood density information to tailor a perturbation constraint and design the geometry-aware and distribution-adaptive modification for each point to achieve a stealthy attack under the locally diverse geometric point cloud distribution.
- Our experiments show that PointCA can cause at least 60% performance degradation over different completion models, while keeping a low perceptibility of adversarial perturbations. We verify that the state-of-the-art defense methods based on statistic outlier removal cannot fully guard the point cloud completion model.

Related Work

Attacks on Point Cloud Classification

In the point cloud domain, Xiang et al. (2019) proposed the first point cloud attack algorithm based on the C&W framework (Carlini and Wagner 2017). Hamdi et al. (2020) then employed an autoencoder loss to strengthen the transferability of the attack across multiple classification models. To reduce the perceptibility of adversarial point clouds, Kim et al. (2021) and Shi et al. (2022) explored perturbing only a minimal number of points and preserving the original geometric shape as much as possible. LG-GAN was introduced by Zhou et al. (2020) for a more flexible point cloud attack.

Although widely studied in the literature, existing adversarial attacks focus on point cloud classification tasks, which mainly concentrate on the artificial data in virtual scenarios and ignore many realistic properties of point clouds such as nonuniform distribution and structure damage (Sun et al. 2022; Ren, Pan, and Liu 2022), making them difficult to apply to point completion tasks.

Attacks on Point Cloud Generation

Attacks on generative models have also received attention (Kos, Fischer, and Song 2018; Willetts et al. 2021). GeoAdv (Lang, Kotlicki, and Avidan 2021) first launched the attack against point cloud reconstruction model based on autoencoder structure. However, our work is significantly different from it. Compared with reconstruction, point cloud completion is a much more different and difficult task. The completion model is not explicitly enforced to retain the input in its output like an autoencoder. Instead, it needs to infer the complete structure from the partial observation with less prior knowledge. Besides, the higher data complexity in point cloud completion induces a higher standard of perturbation constraint (Fig. 1). In order to prevent excessive perturbations that do not match the local geometry manifolds and easily recognizable attacks, we create the Adaptive Geometric Constraint for each point rather than employing a globally consistent constraint threshold.

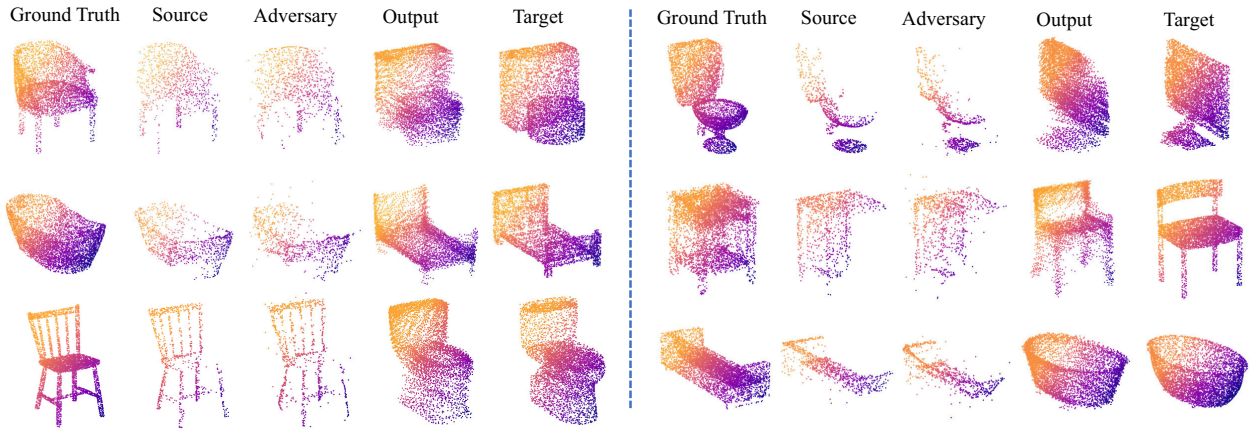


Figure 2: An illustration of targeted attack results. *Source* is the partial point cloud generated on *Ground Truth* from one view point. Tiny perturbations are added to *Source* to obtain the *Adversary*, whose completion *Output* is very close to the *target*.

Defenses on Point Clouds

Disparate methods have been developed to defend against adversarial attacks for point cloud classification. Among existing defenses including adversarial training, Gaussian noise perturbation (Yang et al. 2019), certified robustness (Liu, Jia, and Gong 2021), and point removal (Liu, Yu, and Su 2019), the *statistical outlier removal* (SOR) (Zhou et al. 2019) achieves the best performance due to its effectiveness and efficiency. Based on the fact that perturbed points in attacks are likely to become outliers off the manifold of the point cloud surface, SOR can remove adversarial points in a statistical manner. As shown in our experiments (Table 3), although SOR performs well in the classification task, it fails to fully guarantee robustness when facing incomplete partial point clouds.

Methodology

Problem Formulation

Point cloud (PC) completion models aim to predict the complete structure of a given incomplete input. Generally, a well-trained PC completion model can provide a one-way geometric mapping: $\mathbf{X}^P \in \mathbb{R}^{m\times 3} \rightarrow \mathbf{X} \in \mathbb{R}^{n\times 3}$, where $\mathbf{X}^P$ represents a partial point cloud, usually captured by the 3D sensor from a single observation angle of a real object. $\mathbf{X}$ is the ground-truth point cloud of $\mathbf{X}^P$, which can be considered as the point set uniformly scanned from the whole surface of the same original object associated with $\mathbf{X}^P$. An effective PC completion model f_θ should satisfy:

$$d\left(f_\theta\left(\mathbf{X}^P\right), \mathbf{X}\right) \leq \gamma \tag{1}$$

where $d(S_1, S_2)$ represents an appropriate metric that measures the discrepancy between two point clouds, γ is the evaluating threshold with a small value.

In this paper, we propose PointCA, the first adversarial attack towards point cloud completion models. The goal of PointCA is to delicately construct an adversarial version of $\mathbf{X}^P$, denoted as $\mathbf{X}^{P'}$, which can lead to a false completion result from the PC completion model f_θ. Formally, we have:

$$d\left(f_\theta\left(\mathbf{X}^{P'}\right), \mathbf{Y}\right) \leq \gamma \tag{2}$$

where $\mathbf{Y}$ is the target point cloud from a different category.

According to the knowledge we know about the representations of the source and target point clouds, PointCA will be investigated in the geometry space and the latent space, respectively. Here, we denote them as Geometry PointCA and Latent PointCA.

Geometry PointCA

A straightforward solution for solving Eq. (2) is the brute-force search for $\mathbf{X}^{P'}$, which is apparently infeasible in practice. We thus reformulate the problem as a rational optimization instance that can be efficiently solved by existing optimization algorithms. We denote the adversarial perturbation added to the original partial input as $\boldsymbol{\delta} = \mathbf{X}^{P'} - \mathbf{X}^P$. Therefore Eq. (2) can be formulated as:

$$\arg\min_{\boldsymbol{\delta}} \; d\left(f_\theta\left(\mathbf{X}^P + \boldsymbol{\delta}\right), \mathbf{Y}\right), \;\; \text{s.t. } \|\boldsymbol{\delta}\|_p \leq \epsilon \tag{3}$$

where $\|\cdot\|_p$ is a distance metric to measure the perceptibility of adversarial perturbation. Normally, $\|\cdot\|_p$ is instantiated as ℓ_p-norm ($p \in \{1, 2, \infty\}$).

In Geometry PointCA, we assume that the adversary knows the exact ground-truth of target point cloud. As most PC completion networks are built on the encoder-decoder structure (Yang et al. 2018), the model f_θ can be formulated as: $f_\theta(\cdot) = De(En(\cdot))$. So the optimized similarity loss based on geometry information can be defined as:

$$d_{similarity} = D_{chamfer}\left(De\left(En\left(\mathbf{X}^P + \boldsymbol{\delta}\right)\right), \mathbf{Y}\right) \tag{4}$$

where En is the encoder of the completion model, De is the decoder part, Y is the target complete point cloud. $D_{chamfer}$ represents the Chamfer distance metric.

Latent PointCA

In some strict scenarios, the point clouds of real-world objects suffer from diverse corruptions (Sun et al. 2022),

where we cannot obtain complete and detailed ground-truth samples (Yu et al. 2021). It is necessary to further explore how to generate adversarial examples that get rid of the knowledge of the target complete point cloud data. Hence we propose the Latent PointCA where the adversary only obtains the partial target point cloud.

Inspired by variational autoencoder attacks and the latent variable model (Kos, Fischer, and Song 2018), inputs whose distribution are similar to each other in the representation of the latent layer will also get similar outputs with a high probability (Hu et al. 2022b). Instead of using the target point data straightly in Eq. (4), we leverage the similarity metric in the feature space and regard the learned latent feature vector as an approximate representation of the shape manifold. To launch this latent-based attack, we combine Euclidean distance (Kos, Fischer, and Song 2018) and Distribution distance (Willetts et al. 2021) to measure the difference between features and reformulate the optimized similarity loss as:

$$\begin{aligned} d_{similarity} = &\left\| En(\mathbf{X}^P + \boldsymbol{\delta}) - En(\mathbf{Y}^P) \right\|_2 \\ &+ \lambda D_{KL}\left(En(\mathbf{Y}^P) \| En(\mathbf{X}^P + \boldsymbol{\delta}) \right) \end{aligned} \tag{5}$$

where $\mathbf{Y}^P$ is the partial point cloud in the target category different from $\mathbf{X}^P$, ℓ_2-norm $\|\cdot\|_2$ and Kullback–Leibler divergence $D_{KL}(\cdot\|\cdot)$ are jointly used as the similarity metric between the features of inputs, λ is the hyperparameter.

Adaptive Geometric Constraint

Note that the optimation formula in Eq. (3) has an important perturbation constraint: $\|\boldsymbol{\delta}\|_p$. It is obvious that a smaller perturbation metric $\|\boldsymbol{\delta}\|_p$ leads to a stealthier attack. In this paper, we design a new method to provide adaptive constraints to further improve the imperceptibility and efficiency of PointCA. Here we first describe the intuition behind the adaptive constraint and then give the detailes.

Intuition. For traditional 2D images, adversarial attacks usually use ℓ_∞-norm to limit the variation degree on each channel, which clips the redundant perturbation to limit the maximum value as: $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$. Similarly, Hamdi et al. (2020) and Liu et al. (2019) inherit this strategy to bound the perturbation on 3D coordinate of each point as:

$$-\epsilon \leq \delta_{i,j} \leq \epsilon, \quad i = 1, \cdots, m, \quad j = 1, 2, 3 \tag{6}$$

where i is the sequence number of the point x_i with the size of m, j is the coordinate parameter corresponding to one element of (x, y, z). Due to the highly structured property of point clouds, JGBA (Ma et al. 2020) considered the spatial geometry and designed a pointwise limitation to clip the perturbation amplitude in the Euclidean space: $\|\delta_i\|_2 \leq \epsilon$.

Nevertheless, these works still treat all the points equally with the same restricting threshold. According to recent studies (Sun et al. 2021) and our experiments shown in Fig. 5, the distributions of points at different portions are distinct, not all points are equally essential in the optimization for the construction of an adversarial point cloud. The same limitation could be excessively strict for certain points while being insufficient for others. Additionally, after applying too much perturbation, points in uniformly and tightly dispersed areas are more likely to produce outliers that do not match the local geometry, lowering the imperceptibility of adversarial point clouds (Zhou et al. 2019).

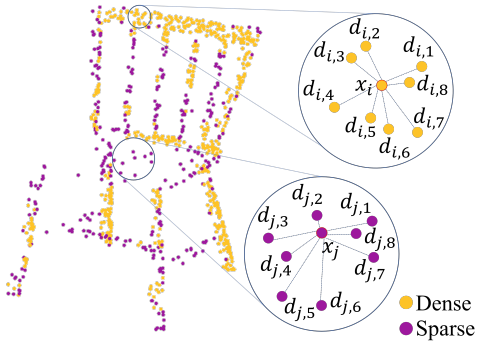


Figure 3: The illustration of local neighborhood geometry. For ease of presentation, we simply partition the points into two categories according to each point's score ρ. Points whose score ρ is below a threshold T are marked in yellow, otherwise marked in purple. The purple centroid point x_j has a sparse local neighborhood, and the yellow centroid point x_i is located in a dense local neighbor points set.

In light of this, instead of treating each independent point identically, we explore the local geometric relationship of points to tailor adversarial examples on partial point clouds.

Local geometric density. To give a more precise analysis of the geometric structure, we partition the local neighborhood set for each point through kNN algorithm. Let $\mathcal{S}(x_i,\, k)$ denote the local kNN point set of point x_i in partial point cloud $\mathbf{X}^P$. The pairwise distance $d_{i,l}$ between the set centroid point x_i and its local neighbor point x_l is defined as:

$$d_{i,l} = \|x_i - x_l\|_2, \; x_l \in \mathcal{S}(x_i,\, k), \quad l = 1, \cdots, k \tag{7}$$

The *distribution sparsity* of different local point sets $\mathcal{S}(x_i,\, k)$ is measured through the average pairwise distance:

$$d_i = \frac{1}{k} \sum_{l=1}^{k} d_{i,l}\,, \quad i = 1, \cdots, m \tag{8}$$

Meanwhile, we evaluate the *distribution uniformity* of the local point set $\mathcal{S}(x_i,\, k)$ by the standard deviation of $d_{i,l}$:

$$\sigma_i = \sqrt{\frac{1}{k-1} \sum_{l=1}^{k} (d_{i,l} - d_i)^2} \tag{9}$$

These two metrics together constitute a description score ρ about the local point cloud density for each point:

$$\rho_i = d_i + t \cdot \sigma_i \tag{10}$$

where t is an auxiliary parameter that further refines the constraint with local uniformity information. Fig. 3 gives a brief illustration for the local geometry density.

Adaptive constraint. Based on the above analysis of local geometric density, we can jointly exploit the *sparsity* and

Algorithm 1: Point Cloud Completion Attack

Input: Model f_θ; benign partial PC $\mathbf{X}^P$; target partial PC $\mathbf{Y}^P$ or target complete PC $\mathbf{Y}$.
Parameter: Iterations n; nearest neighbor range k; step size β.
Output: Completion adversarial example $\mathbf{X}^{P'}$.

Initialize $\mathbf{X}^{P'}$ with random noise $\boldsymbol{\delta}$.
Obtain the local neighbor point set $\mathcal{S}(x_i,\, k)$ that each point x_i in $\mathbf{X}^P$ has by kNN;
Compute the geometric adaptive restraint ϵ_i for each point x_i through Eq. (11);
for $q = 1$ *to* n **do**
  if *Geometry PointCA* **then**
    calculate $d_{similarity}$ through Eq. (4)
  else if *Latent PointCA* **then**
    calculate $d_{similarity}$ through Eq. (5)
  Compute the gradient $\Delta = \nabla_{\mathbf{X}^{P'}} d_{similarity}$;
  Update the point cloud as $\mathbf{X}^{P'} \leftarrow \mathbf{X}^{P'} - \beta \cdot sign(\Delta)$;
  Compute the overall perturbation $\boldsymbol{\delta} \leftarrow \mathbf{X}^{P'} - \mathbf{X}^P$;
  For each point in $\mathbf{X}^{P'}$:
  if $\|\delta_i\|_2 > \epsilon_i$ **then**
    Clip the $\delta_i \leftarrow \delta_i \cdot \frac{\epsilon_i}{\|\delta_i\|_2}$;
  Update the point cloud as $\mathbf{X}^{P'} \leftarrow \mathbf{X}^P + \boldsymbol{\delta}$;
Return Adversarial partial point cloud $\mathbf{X}^{P'}$.

uniformity to customize an adaptive geometric perturbation threshold ϵ_i independently for each point:

$$\begin{aligned}\epsilon_i &= \eta \cdot \rho_i \\ &= \frac{\eta}{k}\sum_{l=1}^{k} d_{i,l} + \eta \cdot t \cdot \sqrt{\frac{1}{k-1}\sum_{l=1}^{k}\left(d_{i,l} - d_i\right)^2} \\ & i = 1, \cdots, m\end{aligned} \tag{11}$$

where η is a scaling coefficient to set a flexible perturbation limitation. As a result, the adaptive geometric constraint will attach a larger perturbation to points which originally have a sparse distribution, and the points with a denser local neighborhood will be perturbed slightly. Thus the adversarial point cloud can maintain a similar distribution and fit the surface manifold of the clean partial point cloud better.

Finally, combining with the adaptive geometric constraint, Geometry PointCA is reformulated as:

$$\begin{aligned}\arg\min_{\boldsymbol{\delta}} \; & D_{chamfer}\left(De\left(En\left(\mathbf{X}^P + \boldsymbol{\delta}\right)\right), \mathbf{Y}\right), \\ & \text{s.t. } \|\delta_i\|_2 \le \epsilon_i, \quad i = 1, \cdots, m\end{aligned} \tag{12}$$

and Latent PointCA is reformulated as:

$$\begin{aligned}\arg\min_{\boldsymbol{\delta}} \; & \left\|En\left(\mathbf{X}^P + \boldsymbol{\delta}\right) - En\left(\mathbf{Y}^P\right)\right\|_2 \\ & + \lambda D_{KL}\left(En(\mathbf{Y}^P) \| En(\mathbf{X}^P + \boldsymbol{\delta})\right), \\ & \text{s.t. } \|\delta_i\|_2 \le \epsilon_i, \quad i = 1, \cdots, m\end{aligned} \tag{13}$$

Note that we use an iterative gradient-based strategy to minimize these two similarity losses and update the perturbation $\boldsymbol{\delta}$. A complete description of our point cloud completion attack is shown in Algorithm 1.

Experiments

Experimental Setup

Dataset and victim completion models. According to prevailing methods (Xie et al. 2021; Wang et al. 2021), we employ the back-projected depth camera (Yuan et al. 2018) to create partial point clouds on ModelNet10 (Wu et al. 2015) because of the dataset's excellent versatility in various point cloud tasks. Four well-trained point cloud completion models: PCN (Yuan et al. 2018), RFA(Zhang et al.2020), GRNet (Xie et al. 2020), and VRCNet (Pan et al. 2021) are the target models of our attacks.

Evaluation metrics. PointCA aims to alter the model's output to a target geometric shape. In order to evaluate the attack performance, we thus use *target reconstruction error* T-RE = $d(f_\theta(\mathbf{X}^{P'}), \mathbf{Y})$ to measure the similarity between outputs and targets, where $d(\cdot)$ indicates Chamfer distance (T-RE$_c$) or Earth mover's distance (T-RE$_e$) (Rubner, Tomasi, and Guibas 2000). Considering the completion models usually have an inherent error, we also investigate the relative attack effect through *target normalized reconstruction error* (Lang, Manor, and Avidan 2020) as:

$$\text{T-NRE} = \frac{d(f_\theta(\mathbf{X}^{P'}), \mathbf{Y})}{d(f_\theta(\mathbf{Y}^P), \mathbf{Y})} \tag{14}$$

If T-NRE $= 1$, the attack effect of the adversarial point cloud can be roughly equated to that of the target object's original partial point cloud.

Besides, the *perturbation budget* $d(\mathbf{X}^{P'}, \mathbf{X}^P)$ and *outliers number* under SOR algorithm (Zhou et al. 2019) are calculated to further evaluate the stealthiness of our method.

For defense countermeasure assessments on our adversarial point clouds, we exploit the *source reconstruction error* S-RE = $d(f_\theta(\mathbf{X}^{P'}), \mathbf{X})$ and *source normalized reconstruction error* (Lang, Kotlicki, and Avidan 2021):

$$\text{S-NRE} = \frac{d(f_\theta(\mathbf{X}^{P'}), \mathbf{X})}{d(f_\theta(\mathbf{X}^P), \mathbf{X})} \tag{15}$$

Attack Performance

Implementation details. Different from the object-to-label attack in the classification, the adversarial attack in point cloud completion is object-to-object. One target class may include many different objects for each source example, which results in an enormous expense. Similar to previous research (Lang, Kotlicki, and Avidan 2021), for each object class, we randomly select 20 point cloud pairs in the test set as source examples and each source pair will attack the other 9 object classes. For a given source pair and target label, we take 5 pairs from the target class whose ground truths are top-5 geometric neighbors nearest with source's in terms of Chamfer distance. To sum up, there are $20\times10\times9\times5$=9000 source-target pairs in PointCA. All the evaluation metrics are calculated on the average of these 9000 attacks.

Method	η	PCN			RFA			GRNet			VRCNet		
		T-RE$_e$	T-RE$_c$	T-NRE$_c$	T-RE$_e$	T-RE$_c$	T-NRE$_c$	T-RE$_e$	T-RE$_c$	T-NRE$_c$	T-RE$_e$	T-RE$_c$	T-NRE$_c$
Random Noise	2.5	0.184	0.056	4.391	0.210	0.054	**4.871**	0.182	0.054	4.619	0.172	0.053	4.796
	5	0.184	0.056	**4.388**	0.211	0.054	4.878	0.183	0.054	**4.617**	0.172	0.052	**4.789**
Classification Noise	2.5	0.182	0.058	4.544	0.200	0.057	**5.147**	0.182	0.058	4.893	0.168	0.056	5.069
	5	0.182	0.058	**4.543**	0.200	0.057	5.148	0.182	0.058	**4.891**	0.168	0.056	5.070
Geometry PointCA	2.5	0.119	0.020	1.517	0.154	0.021	1.916	0.128	0.020	1.725	0.108	0.022	2.009
	5	0.115	0.018	**1.355**	0.151	0.019	**1.743**	0.126	0.019	**1.638**	0.106	0.021	**1.924**
Latent PointCA	2.5	0.137	0.027	2.065	0.181	0.033	2.996	0.176	0.047	3.970	0.108	0.026	2.277
	5	0.132	0.023	**1.745**	0.175	0.030	**2.649**	0.176	0.046	**3.958**	0.099	0.022	**1.943**

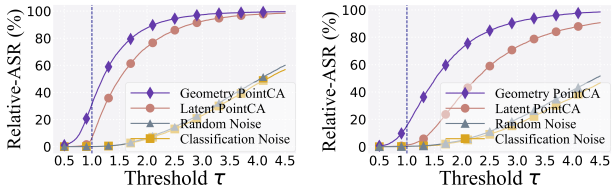
Table 1: Results of attack methods under different perturbation constraints η. A smaller value indicates a better attack effect.

Classifier		PointNet				PointNet++				DGCNN			
Model		PCN	RFA	GRNet	VRCNet	PCN	RFA	GRNet	VRCNet	PCN	RFA	GRNet	VRCNet
Benign	ACC	78.00	77.50	79.44	79.96	77.94	64.61	75.01	80.26	71.00	61.00	67.51	67.07
Geometry PointCA	ACC	18.31	18.54	**21.30**	31.80	16.72	19.93	**20.68**	31.69	16.73	17.37	**19.36**	27.29
	ASR	42.48	35.40	**33.61**	28.49	43.16	28.51	**31.01**	27.76	**35.61**	23.30	**24.72**	20.78
Latent PointCA	ACC	**16.43**	**16.51**	41.81	**12.87**	**15.14**	**17.29**	36.00	**11.71**	**15.49**	**14.42**	30.91	**13.87**
	ASR	**44.54**	**39.30**	14.56	**50.67**	**46.67**	**29.59**	12.48	**50.71**	35.14	**23.79**	12.48	**37.00**

Table 2: Semantic evaluation of the outputs reconstructed on our adversarial point clouds. The scaling coefficient η of attacks is set as 5. Lower ACC and higher ASR indicate stronger semantic information of our attack.

Baselines. *Random noise*: We add random Gaussian noise to the input partial point clouds and investigate the outputs to exclude the interference of general noise. *Classification noise*: We also verify whether the adversarial noises created on classifiers can be transferred to attack completion models.
Analysis. The detailed attack results are exhibited in Table 1. Firstly, the low results under various scaling coefficients η imply a satisfactory attack performance of our geometry attack and latent attack against all four completion models. The overall attack performance of geometry PointCA is even close to the completion effect of the original clean samples. Secondly, the high T-NRE of *random noise* shows that general noises cannot easily disturb the completion procedure towards targets. Thirdly, the adversarial perturbation created on classification model is useless when attacking completion models. Because the purpose of classification adversarial attack is merely to change the output's label, in most cases, the adversarial point clouds still resemble the original structures, making it impossible for them to successfully deceive other point cloud processing missions.
Relative attack success rate. Geometric distance metrics can be employed to measure the attack effectiveness of adversarial examples. Nevertheless, there is no precise definition of *attack success rate* (ASR) in our point cloud completion attacks. Therefore, we leverage T-NRE as a reference to dynamically assess the *relative attack success rate* (Relative-ASR). We set a threshold τ, we can roughly consider one attack is successful if its T-NRE is below this threshold.

The results of Relative-ASR on PCN and RFA are depicted in Fig. 4. We can see there is a great gap between our attacks and baselines, which obviously demonstrates the

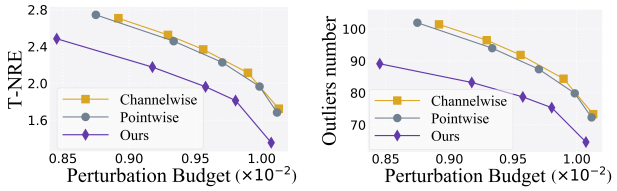


(a) Relative-ASR on PCN (b) Relative-ASR on RFA

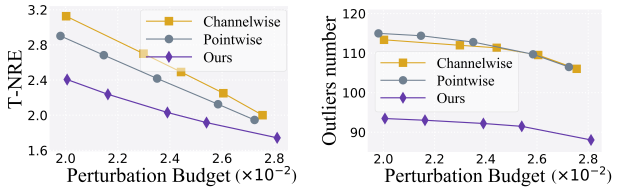
Figure 4: Relative attack success rates under different thresholds τ

strong attack ability of PointCA. Meanwhile nearly 30% of adversarial examples on PCN and 15% generated on RFA in geometry PointCA have a T-NRE below 1. The majority of examples generated by PointCA have a T-NRE below 2, further proving our attack's effectiveness.
Semantic evaluation. We train three classifiers: PointNet, PointNet++, and DGCNN, with ModelNet10 dataset. Subsequently, the complete point clouds reconstructed on our adversarial examples are fed into these models to obtain the classification results. In this way, we can comprehensively evaluate our attack methods in a point cloud processing pipeline rather than on an immutable independent model.

As shown in Table 2, the *Accuracy* (ACC) of all the three classifiers decreases drastically, e.g., PointNet++ bears a decrease from 77.9% to 16.7%. It suggests that the completion outputs of partial adversarial point clouds have changed into other shapes with semantics different from the original ones. The *Attack Success Rate* (ASR) of latent attack is



(a) T-NRE on PCN (b) Outliers on PCN



(c) T-NRE on RFA (d) Outliers on RFA

Figure 5: Comparison between different perturbation constraint strategies on PCN and RFA

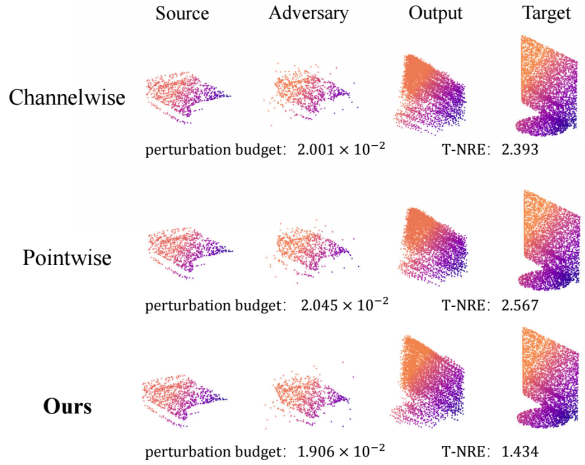


Figure 6: Visualization results of different perturbation constraint strategies. Our method can obtain a better adversarial completion performance with less perturbation budget.

higher, indicating the feature similarity might be more effectively transmitted by the decoder and has a greater influence on downstream tasks than geometry similarity. Besides, the average ASR of all reconstructed point clouds is about 32%.

Comparison Study

Implementation details. In this section, we compare our adaptive geometric constraint with two popular perturbation constraint strategies in classification attacks: channelwise l_∞-norm clip (Hamdi et al. 2020) and pointwise l_2-norm clip (Ma et al. 2020), whose basic ideas of constraining the perturbation through clipping method are similar to ours. To ensure a fair comparison, we use *perturbation budget* and *outliers number* to analyze the adaptive constraint.

Analysis. The results on PCN and RFA are shown in Fig. 5. Notably, compared with channelwise l_∞-norm and pointwise l_2-norm constraints, our method has an obvious advantage in generating a high quality of adversarial examples, i.e., a better T-NRE and fewer outliers under the same perturbation budget. It also verifies our insight that not all points contribute equally to the final attack performance. The nonuniform and locally diverse data distributions of partial point clouds lead to an anisotropic search process in generating adversarial examples, thus an adaptive geometric constraint strategy is necessary in point cloud completion attack. Although other two constraints incur more perturbation budget, they are unable to take full advantage of the budget. Fig. 6 visualizes the examples generated under three perturbation constraints for comparison.

Defense		None	SRS	OR	SOR
Model	η	S-RE	S-NRE	S-NRE	S-NRE
PCN	clean	0.010	1.004	1.021	1.170
	2.5	0.038	3.164	1.712	**1.473**
	5	0.042	**3.447**	1.675	1.470
RFA	clean	0.010	1.025	1.036	1.120
	2.5	0.029	2.910	2.448	2.324
	5	0.032	**3.199**	**2.636**	**2.508**
GRNet	clean	0.011	1.000	1.000	1.142
	2.5	0.029	2.749	2.588	2.520
	5	0.030	**2.851**	**2.852**	**2.584**
VRCNet	clean	0.010	1.009	1.001	1.083
	2.5	0.031	3.224	2.773	2.548
	5	0.032	**3.341**	**2.857**	**2.629**

Table 3: Evaluating PointCA under three defenses. Higher S-RE and higher S-NRE imply better attack strength.

Evaluation against Defenses

Implementation details. We analyze the adversarial point clouds generated by Geometry PointCA under three mainstream defenses: *Simple Random Sampling* (SRS), *Outlier Removal* (OR), and *Statistic Outlier Removal* (SOR) (Huang et al. 2022; Shi et al. 2022).

Analysis. Table 3 shows that although adversarial attacks can be somewhat mitigated, these defenses cannot fully recover the original shapes. SOR defense achieves a relatively good result, while in most cases, the reconstructed results of SOR denoised adversarial point clouds still have at least 50% errors compared with normally restored samples.

Conclusion

In this paper, we propose the first adversarial attack towards point cloud completion model, namely PointCA. The representation similarity in the geometry space and the latent space is exploited to generate adversarial point clouds. We further design an adaptive geometric constraint depending on local density information for each point to improve the imperceptibility of PointCA. The comprehensive experiments verify the effectiveness and efficiency of our attack.

Acknowledgments

Shengshan's work is supported in part by the National Natural Science Foundation of China (Grant No. U20A20177) and Hubei Province Key R&D Technology Special Innovation Project under Grant No. 2021BAA032.

References

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP'17)*, 39–57.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*.

Hamdi, A.; Rojas, S.; Thabet, A.; and Ghanem, B. 2020. Advpc: Transferable adversarial perturbations on 3D point clouds. In *Proceedings of the 16th European Conference on Computer Vision (ECCV'20)*, 241–257.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 770–778.

Hu, S.; Liu, X.; Zhang, Y.; Li, M.; Zhang, L. Y.; Jin, H.; and Wu, L. 2022a. Protecting Facial Privacy: Generating Adversarial Identity Masks via Style-robust Makeup Transfer. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, 15014–15023.

Hu, S.; Zhou, Z.; Zhang, Y.; Zhang, L. Y.; Zheng, Y.; He, Y.; and Jin, H. 2022b. BadHash: Invisible Backdoor Attacks against Deep Hashing with Clean Label. In *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*, 678–686.

Huang, Q.; Dong, X.; Chen, D.; Zhou, H.; Zhang, W.; and Yu, N. 2022. Shape-invariant 3D Adversarial Point Clouds. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, 15335–15344.

Kim, J.; Hua, B.-S.; Nguyen, T.; and Yeung, S.-K. 2021. Minimal adversarial examples for deep learning on 3D point clouds. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV'21)*, 7797–7806.

Kos, J.; Fischer, I.; and Song, D. 2018. Adversarial examples for generative models. In *Proceedings of the 2018 IEEE Symposium on Security and Privacy Workshops (SPW'18)*, 36–42.

Lang, I.; Kotlicki, U.; and Avidan, S. 2021. Geometric adversarial attacks and defenses on 3D point clouds. In *Proceedings of the 2021 International Conference on 3D Vision (3DV'21)*, 1196–1205.

Lang, I.; Manor, A.; and Avidan, S. 2020. Samplenet: Differentiable point cloud sampling. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*, 7578–7588.

Liu, D.; Yu, R.; and Su, H. 2019. Extending adversarial attacks and defenses to deep 3D point cloud classifiers. In *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP'19)*, 2279–2283.

Liu, H.; Jia, J.; and Gong, N. Z. 2021. Pointguard: Provably robust 3D point cloud classification. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 6186–6195.

Ma, C.; Meng, W.; Wu, B.; Xu, S.; and Zhang, X. 2020. Efficient joint gradient based attack against SOR defense for 3D point cloud classification. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*, 1819–1827.

Pan, L.; Chen, X.; Cai, Z.; Zhang, J.; Zhao, H.; Yi, S.; and Liu, Z. 2021. Variational relational point completion network. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 8524–8533.

Ren, J.; Pan, L.; and Liu, Z. 2022. Benchmarking and analyzing point cloud classification under corruptions. arXiv:2202.03377.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2): 99–121.

Shi, Z.; Chen, Z.; Xu, Z.; Yang, W.; Yu, Z.; and Huang, L. 2022. Shape Prior Guided Attack: Sparser Perturbations on 3D Point Clouds. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI'22)*, 8277–8285.

Sun, J.; Zhang, Q.; Kailkhura, B.; Yu, Z.; Xiao, C.; and Mao, Z. M. 2022. Benchmarking Robustness of 3D Point Cloud Recognition Against Common Corruptions. arXiv:2201.12296.

Sun, Y.; Chen, F.; Chen, Z.; and Wang, M. 2021. Local Aggressive Adversarial Attacks on 3D Point Cloud. In *Proceedings of the 2021 Asian Conference on Machine Learning (ACML'21)*, 65–80.

Wang, H.; Liu, Q.; Yue, X.; Lasenby, J.; and Kusner, M. J. 2021. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV'21)*, 9782–9792.

Willetts, M.; Camuto, A.; Rainforth, T.; Roberts, S.; and Holmes, C. 2021. Improving VAEs' Robustness to Adversarial Attack. In *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3D shapenets: A deep representation for volumetric shapes. In *Proceedings of the 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'15)*, 1912–1920.

Xiang, C.; Qi, C. R.; and Li, B. 2019. Generating 3D adversarial point clouds. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*, 9136–9144.

Xie, C.; Wang, C.; Zhang, B.; Yang, H.; Chen, D.; and Wen, F. 2021. Style-based point generator with adversarial rendering for point cloud completion. In *Proceedings of the*

2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21), 4619–4628.

Xie, H.; Yao, H.; Zhou, S.; Mao, J.; Zhang, S.; and Sun, W. 2020. Grnet: Gridding residual network for dense point cloud completion. In *Proceedings of the 16th European Conference on Computer Vision (ECCV'20)*, 365–381.

Yang, J.; Zhang, Q.; Fang, R.; Ni, B.; Liu, J.; and Tian, Q. 2019. Adversarial attack and defense on point sets. arXiv:1902.10899.

Yang, Y.; Feng, C.; Shen, Y.; and Tian, D. 2018. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*, 206–215.

Yu, X.; Rao, Y.; Wang, Z.; Liu, Z.; Lu, J.; and Zhou, J. 2021. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV'21)*, 12498–12507.

Yuan, W.; Khot, T.; Held, D.; Mertz, C.; and Hebert, M. 2018. PCN: Point Completion Network. In *Proceedings of the 2018 International Conference on 3D Vision (3DV'18)*, 728–737.

Zhou, H.; Chen, D.; Liao, J.; Chen, K.; Dong, X.; Liu, K.; Zhang, W.; Hua, G.; and Yu, N. 2020. LG-GAN: Label Guided Adversarial Network for flexible targeted attack of point cloud based deep networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*, 10356–10365.

Zhou, H.; Chen, K.; Zhang, W.; Fang, H.; Zhou, W.; and Yu, N. 2019. DUP-Net: Denoiser and Upsampler Network for 3D adversarial point clouds defense. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV'19)*, 1961–1970.