

RDPA: Real-Time Distributed-Concentrated Penetration Attack for Point Cloud Learning

Youtong Shi¹, Lixin Chen^{1,2}, Yu Zang^{1,2*}, Chenhui Yang¹, Cheng Wang^{1,2}

¹Fujian Key Lab of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University (XMU), China

²Key Laboratory of Multimedia Trusted Perception and Efficient Computing, XMU, China
{shiyoutong, chenlixin}@stu.xmu.edu.cn, zangyu7@126.com, {chyang, cwang}@xmu.edu.cn

Abstract

Partial point attack approaches focus on leveraging the fewest points to achieve the best attack efficiency for easy implementation in the physical domain. For the first time, this paper proposes that the partial point attack strategy should pay attention to not only the selection and disturbance of points, but also the penetration of current defense methods. By re-examining characteristics of previous partial point attack approaches leading to performance improvement, we discover two fundamental principles: first, the selection of attacked points should consider not only the favourable visual salience but also the proper position concentration, thus to acquire effective structural destruction on the basis of remaining imperceptible; second, the perturbation of target points should form meaningful structures rather than outliers. To achieve this, we first propose a novel distributed-concentrated point selection (DPS) strategy, which is easier to concentrate salient points containing rich local information in a few tiny regions. Additionally, to enhance the penetration efficacy and real-time performance of attack point clouds against defenses, we further design a perturbation network based on the multi-scale penetration loss (L_{msp}), which can generate adversarial samples with as few outliers as possible only through a single forward propagation. Experimental results demonstrate that the real-time distributed-concentrated penetration attack (RDPA) framework can achieve state-of-the-art (SOTA) success rates by perturbing only 3.5% of points, and have the best penetration for mainstream defense methods such as SRS and SOR.

1 Introduction

Numerous studies have demonstrated that 3D point cloud deep neural networks are extremely vulnerable to deception by adversarial examples [Xiang *et al.*, 2019; Lee *et al.*, 2020; Cao *et al.*, 2021]. Most existing attack methods primarily perturb the entire point cloud, offering the advantage of

not necessitating complex region selection or local processing. However, the added noise may propagate throughout the whole point cloud, thus making it challenging to implement in the physical domain.

Partial point attack approaches [Kim *et al.*, 2021; Zheng *et al.*, 2019; Wicker and Kwiatkowska, 2019; Zheng *et al.*, 2023] focus on leveraging the fewest points to achieve the best attack efficiency. A common attack framework usually consists of two parts: point selection and point perturbation. Specific to these two aspects, by re-examining the characteristics of previous approaches, we discover two fundamental principles:

- The selection of attacked points should consider not only the favourable visual salience but also the proper position concentration, thus to acquire effective structural destruction on the basis of remaining imperceptible.
- The perturbation of target points should form meaningful structures rather than outliers.

An example of the first criterion is shown in the top row of Figure 1. Assuming a set of point clouds distributed in squares, point selection strategies that solely focus on visual salience may result in attacked points being evenly distributed along contours [Qi *et al.*, 2017a; Zheng *et al.*, 2019]. After removing the contour points with salient features, the point cloud still remains square in shape, but making it difficult to fool the robust neural network. On the other hand, when choosing only one salient region to attack, it indeed leads the neural network to make erroneous decisions. However, this approach causes particularly obvious damage to the structure because the selected points are too concentrated. For the second criterion, as shown in the bottom row of Figure 1, if we do not constrain the system to execute perturbation, the adversarial sample will have a large number of individual outliers whose attack success rate can be rapidly decreased by only imposing a simple Gaussian filter. Specifically, when applying two widely used defense methods, simple random sampling (SRS) and statistical outlier removal (SOR) [Zhou *et al.*, 2019], the attack success rates against PointNet [Qi *et al.*, 2017a] on ModelNet40 [Wu *et al.*, 2015] using current SOTA attack approaches [Kim *et al.*, 2021; Zheng *et al.*, 2023] significantly declines from 89.41% and 92.35% to 17.60%, 3.77% and 50.32%, 16.73% respectively. Thus, the attack penetration must be considered in the frame-

*Corresponding author.

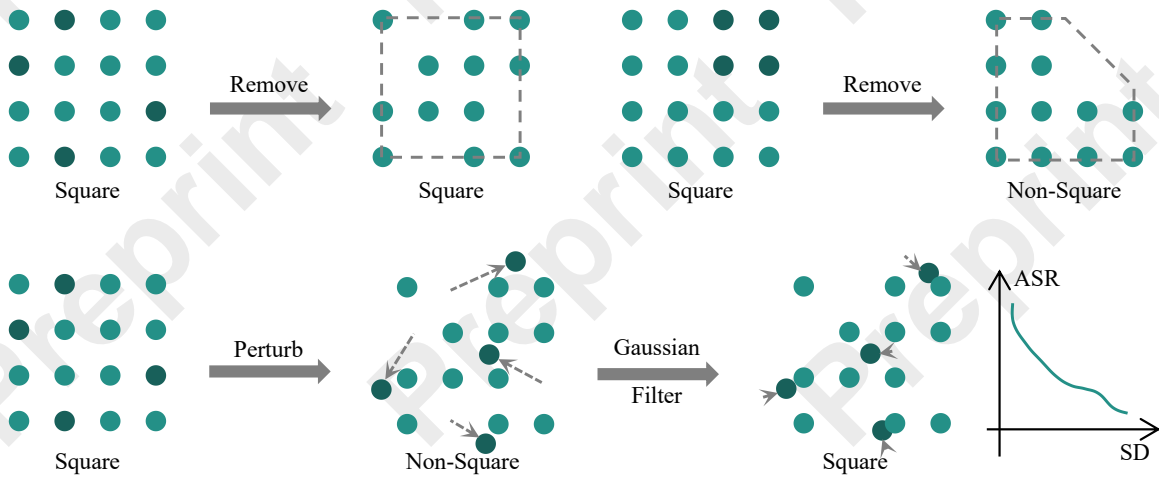


Figure 1: Illustrating the importance of the two fundamental principles which are critical for the partial point attack strategy.

work.

Based on the analysis of above two principles, we propose a novel real-time distributed-concentrated penetration attack (RDPA) framework for point cloud learning, which not only can quickly generate high-quality adversarial point cloud but also show penetration ability against mainstream defenses. Specifically, in the point selection phase, different from previous works mainly focusing on the salience of the individual point or region, we propose a distributed-concentrated point selection (DPS) strategy which is easier to concentrate salient points containing rich local information in a few tiny regions. In the point disturbance phase, we design a perturbation network based on the multi-scale penetration loss (L_{msp}), which can generate adversarial samples with as few outliers as possible only through a single forward propagation.

In general, the contributions of this paper are as follows:

- By re-examining the characteristics of previous partial point attack approaches leading to the performance improvement, we propose two fundamental principles for an effective attack scheme, in which the penetration ability is taken into account for the first time.
- A novel distributed-concentrated point selection (DPS) strategy and multi-scale penetration loss (L_{msp}) are proposed to improve the attack efficiency and penetration ability against mainstream defenses.
- Comprehensive experiments show that the proposed attack framework is capable of providing SOTA results on various benchmarks. We also provide extensive ablation studies to examine the effect and robustness of different parts.

2 Related Works

2.1 3D Point Cloud Attacks

Adversarial attacks on 3D point cloud models can be broadly classified into optimization-based, gradient-based, and drop-based methods. Optimization-based attacks, such as those of [Xiang *et al.*, 2019] and [Carlini and Wagner, 2017], focus

on perturbing point clouds by optimizing objective functions. Techniques such as adding custom loss terms [Tsai *et al.*, 2020], using binary random vectors [Kim *et al.*, 2021], and encoding high-dimensional data into a compact latent space [Lee *et al.*, 2020] have also been explored. [Zhou *et al.*, 2020] applied generative networks for the generation of adversarial examples.

Gradient-based methods modify the input by using model gradients. Early works introduced curvature losses [Wen *et al.*, 2020], while [Goodfellow *et al.*, 2014; Madry *et al.*, 2017] refined gradient computation. Autoencoders were used by [Hamdi *et al.*, 2020] to generate perturbations, and [Moosavi-Dezfooli *et al.*, 2017; Naseer *et al.*, 2019; Poursaeed *et al.*, 2018] extended this approach. [Ma *et al.*, 2020] proposed the Joint Gradient Based Attack (JGBA) and [Yang *et al.*, 2019] introduced the Pointwise Gradient Enhanced Momentum (MPG) to improve the transferability of the attack. [Huang *et al.*, 2022] adapted I-FGSM [Zhao *et al.*, 2020] and [Tao *et al.*, 2023] developed a black-box attack based on decision boundaries. [Zheng *et al.*, 2023] targeted salient regions in point clouds for more effective attacks.

In dropping-based attacks, methods like Drop200 [Zheng *et al.*, 2019] discard points based on importance scores. [Yang *et al.*, 2019] further refined this by calculating gradients before maximum pooling in PointNet, while [Wicker and Kwiatkowska, 2019] classified models and used latent translations for keypoint definition to improve attack precision.

2.2 3D Point Cloud Defenses

Defensive strategies aim to purify adversarial samples or enhance model robustness. Purification methods include SOR, which uses upsampling networks [Zhou *et al.*, 2019], IF-Defense with implicit function networks [Wu *et al.*, 2020], Adv3Diff combining diffusion and denoising [Zhang *et al.*, 2023b], and IT-Defense, which employs invariant transformations [Zhang *et al.*, 2023a].

To strengthen the robustness of the classifier, various data augmentation techniques have been proposed. PointMixup [Chen *et al.*, 2020] and PointCutMix [Zhang *et al.*, 2022]

blend point clouds, while RSMix [Lee *et al.*, 2021] combines rigid subsets. Other methods include experiments on data augmentation [Sun *et al.*, 2022], self-attention for feature extraction [Dong *et al.*, 2020], and lattice classifiers [Li *et al.*, 2022]. [Ding *et al.*, 2023] introduced CAP, a framework that enhances both semantic and structural information in point clouds.

3 Method

As illustrated in Figure 3, the novel real-time distributed-concentrated penetration attack (RDPA) framework begins with employing the distributed-concentrated point selection (DPS) strategy as Separator to separate the attacked target points P_h and the remains P_l . Subsequently, both the original point cloud P and P_h are fed into the perturbation network named Disruptor, which can generate the adversarial subset P_{h-adv} only through a single forward propagation. Then, we merge P_{h-adv} and P_l to get the final adversarial point cloud P_{adv} . By inputting P_{adv} into the attacked classifier, we can calculate the misclassification loss L_{mis} , while comparing P and P_{adv} determines the disturbance distance loss L_{dis} . Additionally, we design a novel multi-scale penetration loss L_{msp} according to P_{adv} .

3.1 Distributed-Concentrated Point Selection Strategy

Existing point selection strategies only focus on a single salient point or a single salient region. As shown in Figure 2, if only consider the salience when selecting points to attack, they are basically scattered in point cloud surface, such as critical subset strategy [Qi *et al.*, 2017a] and saliency map algorithm [Zheng *et al.*, 2019]. The disadvantage of these methods are that once disturbed points are defended, original points located on the subsurface will form a new profile with their neighboring points, which is why the attack success rate of method [Kim *et al.*, 2021] will rapidly decline in the face of defenses. Another drawback is that the points to be attacked are too dispersed, resulting in minimal changes to the point cloud structure. In addition, as shown in the visualization result of local adversarial method [Zheng *et al.*, 2023], choosing to attack a single salient region will cause serious deformation of the point cloud structure, which violates the principle of imperceptibility because the salient points are too concentrated.

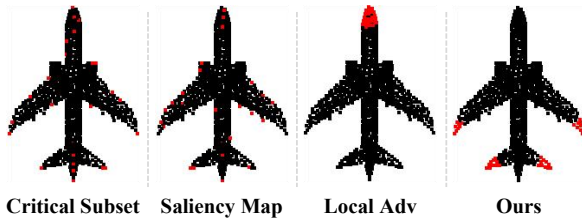


Figure 2: Comparison of point selection strategies highlighting the balance of DPS between salience and concentration for effective and imperceptible attacks.

Based on the analysis of above deficiencies, we consider both salience and concentration when selecting the attacked

target points. Meanwhile, in order to prevent salient points from being concentrated in a certain region, we propose a distributed-centralized point selection (DPS) strategy which fully follows the first principle of the partial point attack. As shown in the final part of Figure 2, salient points have both the distributed characteristic and the advantage of centralized attacks. The unique advantage of DPS is that it can be effectively against minor defenses without causing excessive damage to the point cloud structure.

Specifically, we employ the covariance matrix to measure information in the nearest region of a point. By choosing appropriate neighbor numbers, we are able to concentrate information-rich salient points within a few tiny regions, which is a simple but effective method. Formally, a point cloud P is defined as a collection of N unordered points, denoted as $P = \{p_i | i = 1, 2, \dots, N\}$, $P \in R^{N \times 3}$, where each point p_i is represented by a vector containing its (x, y, z) coordinates. To identify the n nearest points to a given point p_i , we employ the k -nearest neighbor (knn) algorithm, resulting in a set denoted as $knn(p_i, n) = \{p_{ij} | j = 1, 2, \dots, n\}$. The mean vector of point p_i is computed as

$$\bar{p}_i = \left(\frac{1}{n} \sum_{j=1}^n x_{ij}, \frac{1}{n} \sum_{j=1}^n y_{ij}, \frac{1}{n} \sum_{j=1}^n z_{ij} \right)^T, j = 1, 2, \dots, n, \quad (1)$$

while the covariance matrix of the local region is defined as

$$C_i = \frac{1}{n-1} \sum_{j=1}^n (p_{ij} - \bar{p}_i)(p_{ij} - \bar{p}_i)^T, \quad (2)$$

which is a 3 by 3 real symmetric matrix.

Analyzing the eigenvalues and eigenvectors of the covariance matrix allows us to discern the primary magnitude and direction of change within the local structure of the point cloud. Let $\lambda_{i1}, \lambda_{i2}$, and λ_{i3} be the descending eigenvalues of the matrix respectively, where λ_{i1} provides the richest local information because it indicates the magnitude of change along the major direction. To offer more nuanced insights into structural changes, we consider both the ratio of three eigenvalues and the max eigenvalue. Therefore, we define the salient score of point p_i as

$$s_i = \frac{\sum_{k=1}^3 \sum_{t=k+1}^3 (\lambda_{ik} - \lambda_{it})^2}{\sum_{k=1}^3 \lambda_{ik}^2 + \epsilon} \cdot \max(\lambda_{ik}^2), \quad (3)$$

$k < t, k, t = 1, 2, 3,$

where ϵ is a constant set to prevent the denominator from being 0, and we set $\epsilon = 1$ for consistency. When the differences between the three eigenvalues are significant, the left-hand side of the dot product ensures that the overall expression increases as these differences widen. Simultaneously, the right-hand side ensures that s_i increases with the growth of the max eigenvalue.

After calculating the salient score s_i for each point p_i , we first arrange them in descending order. Then, we define the top M points as the attacked target points P_h and the remaining $N - M$ low-score points as P_l , where $P_h \in R^{M \times 3}$, $P_l \in$

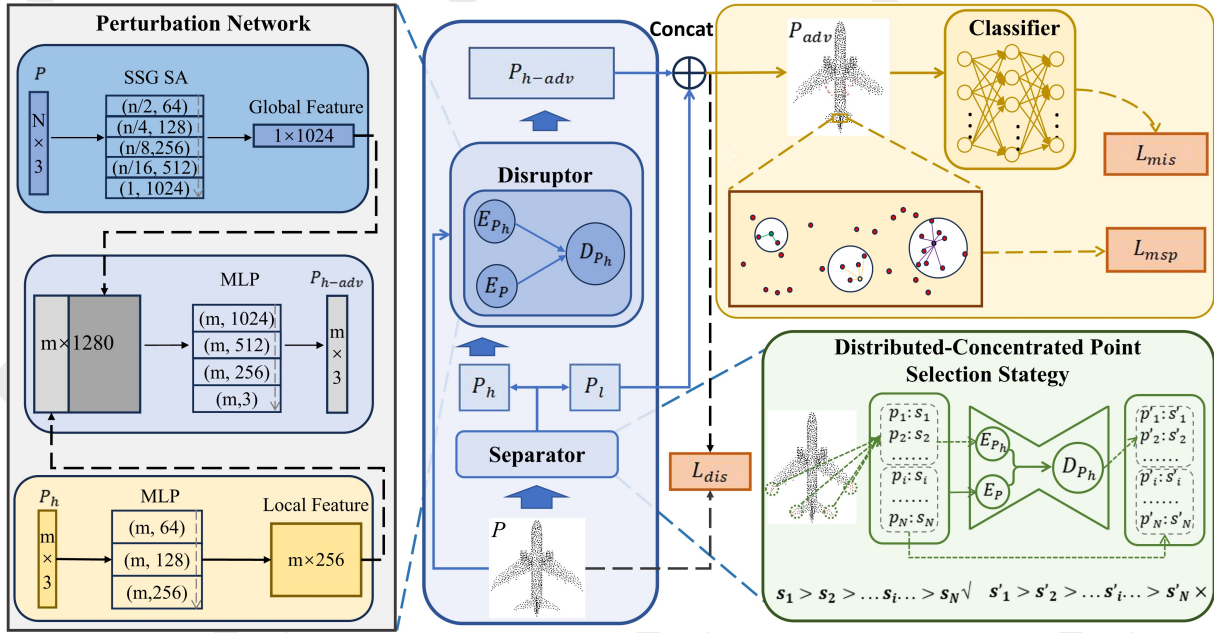


Figure 3: The framework of RDPA which consists of a separator for attacked points selection and a disruptor for target points disturbance.

$R^{(N-M) \times 3}$. They have the following relationship:

$$\begin{aligned} P_h \cup P_l &= P, \\ P_h \cap P_l &= \emptyset \end{aligned} \quad (4)$$

As shown in the lower right part of Figure 3, for $\forall p_h \in P_h, p_l \in P_l$, we have $s_h \geq s_l$.

3.2 Perturbation Network Based on the Multi-Scale Penetration Loss

After the separator extracts P_h and P_l from the original point cloud, we aim to design a network to generate P_{h-adv} , which can effectively meet the flexibility requirements for real-time attacks because it only requires a single forward pass. Meanwhile, to meet the second principle of our proposed partial point attacks, we design a novel multi-scale penetration loss to train the perturbation network. The specific structure of the perturbation network and its loss function is as follows:

Design of Disruptor

Regarding the design of the disruptor, we address the learning of both global and local features. This process involves three components: the original point cloud encoder E_p , the attacked target points encoder E_{P_h} , and the decoder D_{P_h} . Specifically, we employ the single-scale grouping set abstraction (SSGSA) algorithm [Qi et al., 2017b] for learning global features which involved a carefully designed coarse-to-fine five-layer farthest point sampling processes. In the i -th layer, the number of sampling points is $\frac{N}{2^i}$, and the number of aggregated feature channels is $64 \cdot 2^{i-1}$. For learning local features, we employ a simple three-layer perceptron exclusively, with each layer outputting a shape of $M \times (64 \cdot 2^{i-1})$. For the decoding process, the final output is the P_{h-adv} after perturbation. The entire encoder-decoder process is illustrated in

Figure 3, its mathematical expression as follows:

$$P_{h-adv} = D_{P_h}(E_p(P), E_{P_h}(P_h)). \quad (5)$$

Multi-Scale Penetration Loss

In order to make our attack framework more penetrating in the face of mainstream defense strategies, we propose L_{msp} to constrain adversarial examples with as few outliers as possible. To identify outliers more accurately in local ranges of different densities, we begin with computing the multi-scale average k -nearest neighbor distance D_i for each point p_i . In the following equations 6 to 9, we uniformly set $n = 2^t$, $t = 1, 2, 3$:

$$D_i = \left\{ d_{in} \mid d_{in} = \frac{1}{n} \sum_{p_{ij} \in k\text{nn}(p_i, n)} \|p_{ij} - p_i\|_2 \right\}, \quad (6)$$

followed by the calculation of the mean \bar{D} and standard deviation Σ of all distances on every scale within the point cloud:

$$\bar{D} = \left\{ \bar{d}_n \mid \bar{d}_n = \frac{1}{N} \sum_{i=1}^N d_{in} \right\}, \quad (7)$$

$$\Sigma = \left\{ \sigma_n \mid \sigma_n = \sqrt{\frac{1}{N} \sum_{i=1}^N (d_{in} - \bar{d}_n)^2} \right\}. \quad (8)$$

Based on the multi-scale means and standard deviations, we define the multi-scale penetration loss L_{msp} as follows:

$$L_{msp} = \frac{1}{N} \sum_{\{d \mid d > \bar{d}_n + \gamma \cdot \sigma_n\}} d \quad (9)$$

Joint Loss Function

On the premise of ensuring the penetration of adversarial samples, we must balance the high attack success rate and good visual quality. Therefore, we propose a novel joint loss function, which consists of L_{mis} , L_{dis} , and L_{msp} three terms:

$$L = L_{mis} + \alpha \cdot L_{dis} + \beta \cdot L_{msp}, \quad (10)$$

where α, β are weight parameters.

Our work primarily focus on untargeted attacks rather than targeted because these are two different patterns. Untargeted attacks are mainly to deceive the network, while target attacks are to induce. Specifically, let F represents the classifier network for the point cloud, and $F_i(P)$ denotes the likelihood of the point cloud being classified into the i -th class. Our attack can increase the likelihood of misclassification by the classifier, that is $\arg\max_i F_i(P) \neq \arg\max_{i'} F_{i'}(P_{adv})$. The definition of L_{mis} is shown as follows:

$$L_{mis} = \max \left\{ \theta, F_{i^t}(P_{adv}) - \max_{i' \neq i^t} F_{i'}(P_{adv}) \right\}, \quad (11)$$

where θ represents the optimization factor, usually with $\theta = 0$; i^t denotes the correct label of the original point cloud P . Given the limited number of points affected by perturbation, we employ the Hausdorff distance to determine L_{dis} , defined as follows:

$$L_{dis} = \max \left\{ \max_{p_i \in P} \left\{ \min_{p_j \in P_{adv}} \|p_i - p_j\|_2 \right\}, \max_{p_j \in P_{adv}} \left\{ \min_{p_i \in P} \|p_j - p_i\|_2 \right\} \right\}. \quad (12)$$

4 Experiments

4.1 Experiment Setup

Dataset and Attacked Classifier

We take ModelNet40 [Wu *et al.*, 2015] and PointNet [Qi *et al.*, 2017a] as the basic experimental dataset and classification network. Specifically, ModelNet40 comprises 12,311 CAD models across 40 common object categories, where 9,843 objects were allocated for training and 2,468 served as the testing set. Similar to [Qi *et al.*, 2017a], we initially sample 1,024 points by the farthest point sampling (FPS) for each point cloud and then normalize them to fit within a sphere with a radius of 1. Due to the unified processing of data input, the disruptor can conduct additional training on multiple point cloud datasets to satisfy the needs of specific adversarial samples.

Implementation Details

When calculating salient scores, we utilize *knn* to select a local point count of 30 and define the attacked target subset comprising 36 points. For Disruptor, we employ Adam optimizer and CosineAnnealingLR scheduler, setting the learning rate to 0.001 and the cosine annealing cycle to 20. We train Disruptor on PyTorch using an NVIDIA RTX 3090 Ti GPU, over 400 epochs with a batch size of 64. Within the joint loss terms, we conducted an exhaustive search [Kim *et al.*, 2021] for the parameters in Eq. 10 and empirically set them as $\alpha = 5$ and $\beta = 50$. For the definition of L_{msp} , we take $\gamma=1$.

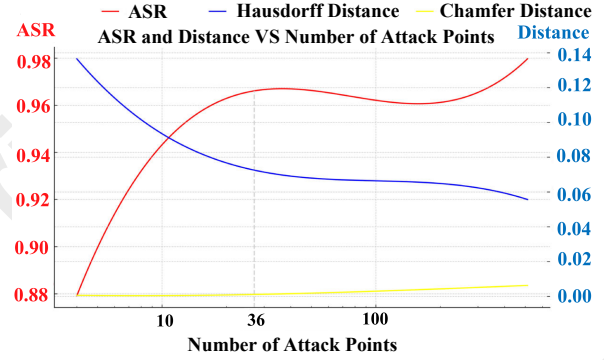


Figure 4: The performance of RDPA method with different attacked point numbers on ModelNet40 when the victim is PointNet.

4.2 Comparison and Evaluation

We evaluate RDPA from four primary dimensions: attack success rate (ASR), distance, generation time and disturbance point number. When evaluating ASR, we consider not only the performance of the classification network without any defense, but also the performance after incorporating mainstream defenses: SRS and SOR [Zhou *et al.*, 2019], where SRS randomly samples and discards 500 points and SOR follows the same setup as [Zhou *et al.*, 2019]. We measure distance employing Hausdorff and Chamfer distances, with the generation time excluding perturbation network training but including saliency score computation. In addition, we compare RDPA with six SOTA partial point attack methods: random occlusion (RO) [Wicker and Kwiatkowska, 2019], iterative significant occlusion (ISO) [Wicker and Kwiatkowska, 2019], critical subset based point dropping (CS-D) [Zheng *et al.*, 2019], saliency map based point dropping (SM-D) [Zheng *et al.*, 2019], minimal attack examples (MAE) [Kim *et al.*, 2021] and adaptive local adversarial (Al-Adv) [Zheng *et al.*, 2023].

As depicted in Table 1, in the absence of any defense, our RDPA exhibits the highest attack success rate at 93.52%, which is 1.17% higher than the second place method Al-Adv and 4.11% higher than the third place method MAE. Even when facing the defenses SRS and SOR, the attack success rates of RDPA still remain high at 89.58% and 81.39%. However, in such a condition, the attack success rates of Al-Adv and MAE drop significantly from 92.35% and 89.41% to 50.32%, 16.73% and 17.60%, 13.77%. Furthermore, RDPA takes an average of only 0.03 seconds to generate a single adversarial sample, allowing it to attack 2000 point cloud frames per second, which fully meets the flexibility requirements of real-time dynamic scenarios. At the same time, the minimal cost of 36 perturbation points provides significant guidance for practical attacks in the physical domain. While RDPA slightly lags behind ISO and MAE in Hausdorff distance and Chamfer distance, this discrepancy arises from their iterative querying of the classification network for enhanced results, which is a challenging process to the real-time attack because it is very time consuming.

Additionally, we validate RDPA on ScanObjectNN [Uy *et*

Methods	ASR(%) \uparrow			Distance(meter) \downarrow		Time(second) \downarrow	Points \downarrow
	No Defense	SRS	SOR	Chamfer	Hausdorff		
RO	25.79/29.54	18.09/19.95	16.68/19.26	0.0019/0.0020	0.0158/0.0182	3.62/3.64	488/482
ISO	38.35/41.33	24.57/29.61	11.06/14.04	0.0006/0.0007	0.0122/0.0145	10.79/15.49	201/242
CS-D	38.50/40.43	50.90/53.18	39.23/41.27	0.0063/0.0055	0.1575/0.1766	0.34/0.29	200/200
SM-D	62.87/68.56	65.72/69.39	60.39/67.94	0.0043/0.0066	0.1196/0.1700	0.09/0.08	200/200
MAE	89.41/91.83	17.60/19.62	13.77/16.33	0.0002/0.0002	0.0201/0.0198	15.21/18.56	38/36
AI-Adv	92.35/94.83	50.32/55.48	16.73/20.22	0.0003/ 0.0002	0.0482/0.0211	33.39/32.75	41/37
RDPA	93.52/95.15	89.58/90.80	81.39/85.41	0.0019/0.0018	0.0953/0.0724	0.03/0.03	36/36

Table 1: Attack performance to PointNet on two datasets (ModelNet40/ScanObjectNN).

al., 2019] setting as in [Kim *et al.*, 2021], which is a high quality real-world dataset and include 15000 objects. As illustrated in Table 1, the comparison results on ScanObjectNN are similar to those on ModelNet40, which demonstrates the robust scalability of RDPA across diverse datasets. Compared with the hand-crafted dataset, the scanned dataset in the real world is more likely to achieve a high attack success rate, because the data itself has certain irregularity.

DPS	L_{msp}	E_P	No Defense	SRS	SOR
✓	✓		79.92	76.89	63.88
✓		✓	95.41	87.45	57.44
	✓	✓	90.32	83.33	76.93
✓	✓	✓	93.52	89.58	81.39

Table 2: The impact of three parts: check marks indicate presence; blanks indicate absence. Right columns of the vertical line show attack success rates (%) under three defense states.

4.3 Ablation Study

The Effect of Each Part

In order to evaluate the proposed DPS, L_{msp} and E_P , we carry out a study on ModelNet40 against PointNet. As can be seen from Table 2, when the encoder E_P is omitted, the attack success rate decreases drastically by more than 10%. Removing the loss term L_{msp} slightly improves the attack success rate without defense, but reduces it by 23.95% against SOR defense. To validate DPS, we replace it with the point selection strategy based on saliency map [Zheng *et al.*, 2019], resulting in decreases of 3.20%, 5.25%, and 4.46% under no defense, SRS defense, and SOR defense. This is because DPS considers both point salience and appropriate concentration, while L_{msp} avoids the occurrence of individual outliers, and E_P captures the global feature of the original point cloud.

The Size of KNN for DPS

The k value of knn employed in DPS plays a crucial role in the distribution of salient points. As shown in table 3, when k is set to a small value, such as 5, salient points tend to disperse along the contour surface, whose attack success rate is similar to that of the point selection strategy based on saliency map [Zheng *et al.*, 2019]. However, as k increases, salient points begin to cluster. Imagine that when k becomes

infinitely large, i.e., $k = N$ (where N is the total point number of a point cloud), salient points will cluster within a local region because each point uses the entire point cloud in calculating, where RDPA had been degraded to AI-Adv method. Therefore, to achieve both distributed characteristics and concentrated attack advantages for salient points, clustering them into several tiny regions is most appropriate. From the experimental data in table 3, setting k to 30 results the best distribution and attack success rate.

The Analysis of Attacked Point Number

As shown in Figure 4, the experiment reveals that as the number of attack points increases, the attack success rate (ASR) rises, the Hausdorff distance decreases, and the Chamfer distance slightly increases. However, when the number of attack points exceeds 36, the success rate begins to decline. Beyond 128 points, although the success rate improves slightly, the excessive number of attack points becomes impractical and unnecessary. Consequently, selecting 36 points as the optimal attack number ensures a high success rate while maintaining low interference costs and imperceptibility.

4.4 Robustness Analysis

Transferability

To validate the transferability of adversarial samples, we perform black-box attacks on four other classification networks, namely PointNet++ [Qi *et al.*, 2017b], DGCNN [Wang *et al.*, 2019], CurveNet [Xiang *et al.*, 2021], and PCT [Guo *et al.*, 2021]. At the same time, we compare RDPA with AI-Adv. PointNet is the pioneer designed to address the disorder nature and rigid transformation invariance of point cloud, while PointNet++ builds on PointNet as a foundational layer and incorporates sampling-grouping layers. To further strengthen the geometric relationship, DGCNN performs edge convolution on the neighborhood graph and CurveNet groups the point cloud into multiple curves. PCT applies attention mechanisms to point cloud based on Transformer.

As demonstrated in table 4, when no defense, the black-box attack success rates of RDPA consistently exceed 15%, while that of AI-Adv is below 3%. This is because DPS is based on analyzing the structure of the point cloud, rather than relying on the specific classification network being attacked. And, during the point perturbation phase, AI-Adv employs an iterative query method, which means the generated adversarial samples lack generality. In the face of SRS,









K	5	10	15	20	25	30	35	...	1024
Visualization								...	
No Defense	90.73	92.27	92.31	93.10	93.43	93.52	92.18	...	92.16
SRS	85.05	86.71	87.27	87.41	88.94	89.58	86.90	...	86.37
SOR	77.58	78.29	79.98	80.60	80.87	81.39	76.81	...	78.34

Table 3: Effect of k in DPS. The top row indicates the value of k , the second row shows the visual depiction of salient points (red), and the subsequent rows display the attack success rates (%) under no defense, SRS defense, and SOR defense.

the black-box attack success rates surpasses that of no defense because the removal of points disrupts the structure of the point cloud, which can be exploited by adversarial samples. Therefore, SRS is not suitable as the black-box defense mechanism. In addition, despite lower black-box attack success rates with SOR defense compared to no defense, RDPA still outperforme AI-Adv.

ISO tend to drop points from the entire point cloud, whereas methods CS-D and SM-D are more inclined to drop points from specific local regions, such as the head of an airplane or the legs of a table. MAE is inclined to attack contour points and generates a large number of individual outliers. AI-Adv targets individual salient region, causing severe disruption to the point cloud structure and similarly generating outliers. In contrast, RDPA chooses to attack multiple tiny regions, enhancing the imperceptibility of adversarial samples. Furthermore, the perturbed points form meaningful structures rather than outliers, which makes RDPA equally penetrating against mainstream defenses. These analyses highlight the impacts of diverse methods, demonstrating the trade-off of RDPA between structural disruption and imperceptibility.

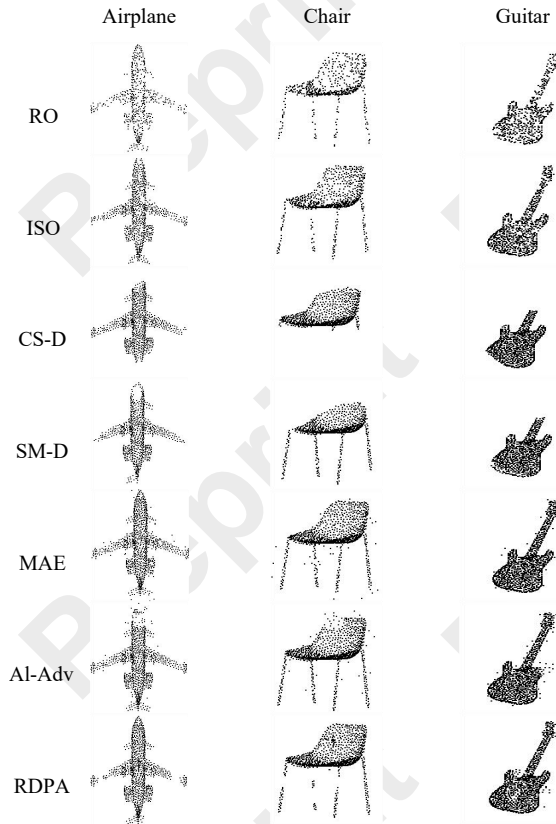


Figure 5: Adversarial samples generated by various attack methods against PointNet on ModelNet40.

Visual Quality

Figure 5 illustrates typical instances of adversarial samples for three classes of point clouds: airplane, chair, and guitar. Since the attack methods RO, ISO, CS-D, and SM-D are based on dropping mechanisms, their adversarial samples appear relatively sparse. Specifically, methods RO and

Classifier	No Defense	SRS	SOR
PointNet	93.52/92.35	89.58/50.32	81.39/16.73
PointNet++	23.68/2.53	47.71/21.06	21.06/2.58
DGCNN	15.58/1.26	54.94/28.77	12.20/2.77
CurveNet	19.65/1.55	46.07/21.23	20.79/1.65
PCT	27.05/1.79	52.04/12.2	16.54/2.37

Table 4: Comparison of black-box attacks on different networks between RDPA and AI-Adv (RDPA/AI-Adv) using adversarial samples generated on PointNet with a white-box attack on ModelNet40.

5 Conclusion

In conclusion, this paper first proposes the partial point attack strategy should pay attention to not only the selection and disturbance of points but also the penetration of defenses, which identifies two fundamental principles: first, the selection of attacked points should consider both the visual salience and the proper position concentration; second, the perturbation of target points should form meaningful structures rather than outliers. Further, we introduce the novel RDPA framework leveraging the DPS strategy and the Disruptor design based on L_{msp} , which achieves a 93.52% attack success rate by perturbing only 3.5% of points. Even when faced with the mainstream defenses SRS and SOR, the attack success rates still remain at 89.58% and 81.39%. In addition, RDPA requires an average of only 0.03 seconds to generate a single adversarial sample, fully meeting the requirements for real-time attacks. In the future, we plan to extend RDPA to real-world applications, particularly focusing on enhancing the robustness of large-scale models.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (No. 62471415); the Natural Science Foundation of Fujian Province of China (No. 2023J01004).

References

- [Cao *et al.*, 2021] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *2021 IEEE symposium on security and privacy (SP)*, pages 176–194. IEEE, 2021.
- [Carlini and Wagner, 2017] 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.
- [Chen *et al.*, 2020] Yunlu Chen, Vincent Tao Hu, Efstratios Gavves, Thomas Mensink, Pascal Mettes, Pengwan Yang, and Cees GM Snoek. Pointmixup: Augmentation for point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 330–345. Springer, 2020.
- [Ding *et al.*, 2023] Daizong Ding, Erling Jiang, Yuanmin Huang, Mi Zhang, Wenxuan Li, and Min Yang. Cap: Robust point cloud classification via semantic and structural modeling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12260–12270, 2023.
- [Dong *et al.*, 2020] Xiaoyi Dong, Dongdong Chen, Hang Zhou, Gang Hua, Weiming Zhang, and Nenghai Yu. Self-robust 3d point recognition via gather-vector guidance. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11513–11521, 2020.
- [Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [Guo *et al.*, 2021] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021.
- [Hamdi *et al.*, 2020] Abdullah Hamdi, Sara Rojas, Ali Thabet, and Bernard Ghanem. Advpc: Transferable adversarial perturbations on 3d point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 241–257. Springer, 2020.
- [Huang *et al.*, 2022] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, and Nenghai Yu. Shape-invariant 3d adversarial point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15335–15344, 2022.
- [Kim *et al.*, 2021] Jaeyeon Kim, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Minimal adversarial examples for deep learning on 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7797–7806, 2021.
- [Lee *et al.*, 2020] Kibok Lee, Zhuoyuan Chen, Xinchun Yan, Raquel Urtasun, and Ersin Yumer. Shapeadv: Generating shape-aware adversarial 3d point clouds. *arXiv preprint arXiv:2005.11626*, 2020.
- [Lee *et al.*, 2021] Dogyoon Lee, Jaeha Lee, Junhyeop Lee, Hyeonmin Lee, Minhyeok Lee, Sungmin Woo, and Sangyoun Lee. Regularization strategy for point cloud via rigidly mixed sample. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15900–15909, 2021.
- [Li *et al.*, 2022] Kaidong Li, Ziming Zhang, Cuncong Zhong, and Guanghui Wang. Robust structured declarative classifiers for 3d point clouds: Defending adversarial attacks with implicit gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15294–15304, 2022.
- [Ma *et al.*, 2020] Chengcheng Ma, Weiliang Meng, Baoyuan Wu, Shibiao Xu, and Xiaopeng Zhang. Efficient joint gradient based attack against sor defense for 3d point cloud classification. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1819–1827, 2020.
- [Madry *et al.*, 2017] 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.
- [Moosavi-Dezfooli *et al.*, 2017] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.
- [Naseer *et al.*, 2019] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 32, 2019.
- [Poursaeed *et al.*, 2018] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4422–4431, 2018.
- [Qi *et al.*, 2017a] 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.
- [Qi *et al.*, 2017b] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

- [Sun *et al.*, 2022] Jiachen Sun, Qingzhao Zhang, Bhavya Kailkhura, Zhiding Yu, Chaowei Xiao, and Z Morley Mao. Benchmarking robustness of 3d point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296*, 2022.
- [Tao *et al.*, 2023] Yunbo Tao, Daizong Liu, Pan Zhou, Yulai Xie, Wei Du, and Wei Hu. 3dhacker: Spectrum-based decision boundary generation for hard-label 3d point cloud attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14340–14350, 2023.
- [Tsai *et al.*, 2020] Tzungyu Tsai, Kaichen Yang, Tsung-Yi Ho, and Yier Jin. Robust adversarial objects against deep learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 01, pages 954–962, 2020.
- [Uy *et al.*, 2019] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597, 2019.
- [Wang *et al.*, 2019] 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):1–12, 2019.
- [Wen *et al.*, 2020] Yuxin Wen, Jiehong Lin, Ke Chen, CL Philip Chen, and Kui Jia. Geometry-aware generation of adversarial point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2984–2999, 2020.
- [Wicker and Kwiatkowska, 2019] Matthew Wicker and Marta Kwiatkowska. Robustness of 3d deep learning in an adversarial setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11767–11775, 2019.
- [Wu *et al.*, 2015] 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.
- [Wu *et al.*, 2020] Ziyi Wu, Yueqi Duan, He Wang, Qingnan Fan, and Leonidas J Guibas. If-defense: 3d adversarial point cloud defense via implicit function based restoration. *arXiv preprint arXiv:2010.05272*, 2020.
- [Xiang *et al.*, 2019] Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9136–9144, 2019.
- [Xiang *et al.*, 2021] Tiange Xiang, Chaoyi Zhang, Yang Song, Jianhui Yu, and Weidong Cai. Walk in the cloud: Learning curves for point clouds shape analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 915–924, 2021.
- [Yang *et al.*, 2019] 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.
- [Zhang *et al.*, 2022] Jinlai Zhang, Lyujie Chen, Bo Ouyang, Binbin Liu, Jihong Zhu, Yujin Chen, Yanmei Meng, and Danfeng Wu. Pointcutmix: Regularization strategy for point cloud classification. *Neurocomputing*, 505:58–67, 2022.
- [Zhang *et al.*, 2023a] Jinlai Zhang, Yinpeng Dong, Minchi Kuang, Binbin Liu, Bo Ouyang, Jihong Zhu, Houqing Wang, and Yanmei Meng. The art of defense: Letting networks fool the attacker. *IEEE Transactions on Information Forensics and Security*, 18:3267–3276, 2023.
- [Zhang *et al.*, 2023b] Kui Zhang, Hang Zhou, Jie Zhang, Qidong Huang, Weiming Zhang, and Nenghai Yu. Ada3diff: Defending against 3d adversarial point clouds via adaptive diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8849–8859, 2023.
- [Zhao *et al.*, 2020] Yue Zhao, Yuwei Wu, Caihua Chen, and Andrew Lim. On isometry robustness of deep 3d point cloud models under adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1201–1210, 2020.
- [Zheng *et al.*, 2019] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Pointcloud saliency maps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1598–1606, 2019.
- [Zheng *et al.*, 2023] Shijun Zheng, Weiquan Liu, Siqi Shen, Yu Zang, Chenglu Wen, Ming Cheng, and Cheng Wang. Adaptive local adversarial attacks on 3d point clouds. *Pattern Recognition*, 144:109825, 2023.
- [Zhou *et al.*, 2019] Hang Zhou, Kejiang Chen, Weiming Zhang, Han Fang, Wenbo Zhou, and Nenghai Yu. Dupnet: Denoiser and upsampler network for 3d adversarial point clouds defense. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1961–1970, 2019.
- [Zhou *et al.*, 2020] Hang Zhou, Dongdong Chen, Jing Liao, Kejiang Chen, Xiaoyi Dong, Kunlin Liu, Weiming Zhang, Gang Hua, and Nenghai Yu. Lg-gan: Label guided adversarial network for flexible targeted attack of point cloud based deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10356–10365, 2020.