Invisible Backdoor Attack against 3D Point Cloud Classifier in Graph Spectral Domain

Linkun Fan[1], **Fazhi He**[*1], **Tongzhen Si**[2], **Wei Tang**[1], **Bing Li**[13]

[1]School of Computer Science, Wuhan University, Wuhan, Hubei, China
[2] School of Information Science and Engineering, University of Jinan, Jinan, China
[3] Hubei Luojia Laboratory
flinkun@whu.edu.cn, fzhe@whu.edu.cn, tjsitongzhen@gmail.com, weitang2021@whu.edu.cn, bingli@whu.edu.cn

Abstract

3D point cloud has been wildly used in security crucial domains, such as self-driving and 3D face recognition. Backdoor attack is a serious threat that usually destroy Deep Neural Networks (DNN) in the training stage. Though a few 3D backdoor attacks are designed to achieve guaranteed attack efficiency, their deformation will alarm human inspection. To obtain invisible backdoored point cloud, this paper proposes a novel 3D backdoor attack, named IBAPC, which generates backdoor trigger in the graph spectral domain. The effectiveness is grounded by the advantage of graph spectral signal that it can induce both global structure and local points to be responsible for the caused deformation in spatial domain. In detail, a new backdoor implanting function is proposed whose aim is to transform point cloud to graph spectral signal for conducting backdoor trigger. Then, we design a backdoor training procedure which updates the parameter of backdoor implanting function and victim 3D DNN alternately. Finally, the backdoored 3D DNN and its associated backdoor implanting function is obtained by finishing the backdoor training procedure. Experiment results suggest that IBAPC achieves SOTA attack stealthiness from three aspects including objective distance measurement, subjective human evaluation, graph spectral signal residual. At the same time, it obtains competitive attack efficiency. The code is available at https://github.com/f-lk/IBAPC.

Introduction

3D point cloud plays an important role in representing 3D shape. It has been deeply applied in many security-crucial domains, including self-driving, 3D face recognition, and service robot. Take self-driving as an example, a vehicle usually perceives the environment and recognizes objects by point clouds. Thus, the security concern about point cloud should be solved.

For efficiency, many 3D Deep Neural Networks (3D DNN) are developed to understand point cloud (Qi et al. 2017a; Liu et al. 2019). However, these popular 3D DNNs have been proved vulnerable to carefully designed attacks. Specifically, adversarial attack and backdoor attack are two main threats. Recently, many 3D adversarial attacks are designed for 3D DNN and achieve guaranteed performance

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

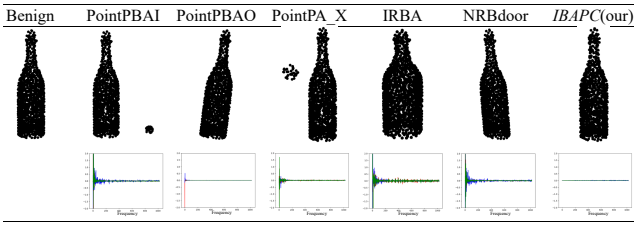

Figure 1: Backdoored 3D point clouds from different attacks. The second line represent residual of graph spectral signal, lower is better. IBAPC is invisible to human vision, and achieves SOTA stealthiness from three perspectives including objective distance measurement, subjective human evaluation, graph spectral signal residual.

(Liu and Hu 2022; Xiang, Qi, and Li 2019). They aim to deceive 3D DNN in the inference stage.

Backdoor attack is another serious security threat that destroys DNN in training stage. It aims to implant hidden malicious behaviors into victim DNN. One typical scenario is that users upload their 3D DNN and dataset to a third-party platform for sufficient computing resources. Hence, it enables the attacker (third-party platform) to implant backdoor into the victim DNN. The backdoored DNN works normally on benign samples, but causes mis-classification on samples containing a specific "trigger". Such attack has the characteristics of high time efficiency in the inference stage, strong stealthiness, and universality. Therefore, lots of efforts are motivated to investigate backdoor attacks and their defenses in 2D image domain (Gu, Dolan-Gavitt, and Garg 2017; Wu et al. 2022; Cheng et al. 2023).

However, backdoor attacks designed for 2D image will disable for 3D point cloud due to data format restriction. As a result, a few 3D backdoor attacks are designed recently. Though they obtain high Attack Success Rate (*ASR*), the deformation is so serious that they are unavailing, since they will be detected even by human observing, shown as Fig 1. Therefore, *this paper aims to design an Invisible Backdoor Attack against 3D Point Cloud (IBAPC).*

We firstly analyze why the stealthiness of existing attacks are poor. In detail, the first reason is that the conducted deformation is concentrated on just one type of deformation, such as added points or shape transformation. In detail,

PointPBA-I (Li et al. 2021) and PointBA_X (Xiang et al. 2021) implant backdoor trigger by a ball with several points. It is quite obtrusive for human vision. Though the triggers designed by IRBA (Gao et al. 2022), PointPBA-O (Li et al. 2021) and NRBdoor (Fan et al. 2022) are distributed on all points, the shape variation caused by overly transformation is visible. Besides, the existing backdoor triggers are manually designed in spatial domain no matter by adding points, rotation, or transformation. It leaves traces in graph spectral domain that facilitates defenders.

To address the above issues, we consider implanting backdoor trigger in graph spectral domain. In particular, we find that low-frequency represents the global structure of point cloud, and high-frequency represents the local structure. Therefore, (**motivation:**) *perturbing point cloud graph frequency signal can induce both global shape variation and local points perturbation to be responsible for the deformation caused by backdoor trigger.* The dispersed deformation in spatial domain is difficult noticed by human vision. Furthermore, to conduct backdoor trigger on graph frequency signal, the backdoor implanting function $T_\xi(\cdot)$ with parameter ξ is designed. It firstly transforms point cloud from spatial domain to graph spectral domain. Then, the backdoor trigger ξ is added to the benign graph spectral signal. Finally, $T_\xi(\cdot)$ transforms the backdoored graph spectral signal to spatial domain to obtain backdoored point cloud.

Moreover, it is a tricky problem about how to decide the specific perturbation on benign graph spectral signal, since only a numerical vector can be added to it. For efficiency, we argue to learn its parameter ξ instead of manually designing. Specifically, a backdoor training procedure is designed following Doan's work (Doan et al. 2021), which updates the parameter ξ of backdoor implanting function and θ of the victim 3D DNN alternately. By finishing the learning process, a backdoored 3D DNN $f(\cdot, \theta^*)$ and a associated backdoor implanting function $T_{\xi^*}(\cdot)$ are obtained.

Extensive experiments suggest that IBAPC achieves SOTA attack stealthiness from three perspectives including objective distance between benign point cloud and backdoored point cloud, human subjective evaluation, and residual in frequency domain. Meanwhile, IBAPC achieves the competitive attack effectiveness. Furthermore, defense experiments exhibit that IBAPC is able to evade several defense methods. Our main contributions can be described as follows:

- An invisible backdoor attack against 3D point cloud classifier named IBAPC is proposed.
- A new backdoor implanting function in graph spectral domain is proposed, which induces both global shape variation and local points perturbation to share the caused deformation.
- A training procedure is designed. The parameters of the backdoor implanting function and the victim 3D DNN are trained alternately.
- Experiment results suggest that IBAPC achieves SOTA attack stealthiness and competitive attack efficiency. Meanwhile, it can evade several backdoor defenses.

Related Works

Backdoor attack is firstly proposed in 2D image domain (Gu, Dolan-Gavitt, and Garg 2017). Then, a lot of attacks was developed (Zhao et al. 2022; Wu et al. 2023; Feng et al. 2022; Yuan et al. 2023; Doan et al. 2023). However, due to data format restriction, attacks for 2D image cannot directly apply to 3D point cloud. Therefore, some spatial 3D backdoor attacks have been proposed recently which can be classified as points addition attacks and shape transformation attacks. Point addition attacks conduct attack by building mapping from a certain geometrical pattern to the target class. Specifically, PointPBA-I (Li et al. 2021) regards a points set with a certain shape and location as the backdoor trigger. PointBA_X (Xiang et al. 2021) optimizes the shape and location of a points set as the trigger for higher efficiency. Besides, shape transformation attacks implant backdoor into point cloud by 3D transformation such as rotation, scaling, and affine. In particular, IRBA (Gao et al. 2022) processes point cloud by the designed weighted local transformation. PointPBA-O (Li et al. 2021) and NRBdoor (Fan et al. 2022) propose to implant backdoor trigger by conducting clean rotation and noisy rotation, respectively. Though existing 3D backdoor attacks achieve high *ASR*, the caused deformation are too serious to evade human inspection. Therefore, this paper aims to pursue invisible 3D backdoor attack.

The Proposed Invisible Backdoor Attack

Threat Model

We focus on the typical threat model that wildly studied by many works (Doan et al. 2021; Zhao et al. 2022). In detail, users design a 3D DNN and collect a training dataset. However, there is no sufficient computing resource. Therefore, they upload their model and dataset to a third-part platform for training. Then, the attacker (third-part platform) returns a well trained but infected 3D DNN to the users. Under such scenario, attacker can access both 3D DNN structure and training dataset.

Problem Definition

A 3D point cloud $P \in R^{n*3}$ is constituted by a set of n unordered 3D points, whose point $x_i \in R^3$ is represented by a 3D coordinates (x_i, y_i, z_i). 3D DNN described as $f(P, \theta) \rightarrow y$ aims to classify P to its label y. θ represents parameters of the classifier, which is learned by minimizing the following loss:

$$\mathcal{L}_{clean}(\theta) = \sum_{P \in D_{clean}} L\{f(P, \theta), y\} \tag{1}$$

Where, D_{clean} means the clean training dataset without poisoned data. L represents the loss function. Different from the benign training process, backdoor attack aims to induce the victim 3D DNN to learn not only the θ of $f(\cdot)$ but also the mapping from backdoor trigger to the target classes y^t by minimizing:

$$\min_{\theta} \sum_{P \in D_{clean}} L\{f(P, \theta), y\} + \sum_{P \in D_{poison}} L\{f(T(M), \theta), y^t\} \tag{2}$$

Where D_{poison} means the poisoned training dataset. $T(\cdot)$ represents the trigger implanting function which is the critical process of a successful backdoor. It decides the attack efficiency and stealthiness. Therefore, this paper aims to develop an efficient $T(\cdot)$ in the graph spectral domain.

Analysis of Graph Spectral Signal

The main aim of this paper is to design a backdoor attack with high stealthiness. To achieve this, it is crucial to control the way of perturbing point cloud during implanting backdoor trigger. Considering the spectral signal in image domain, it provides a new perspective for image processing (Yue et al. 2022; Wang et al. 2022; Zeng et al. 2021). Such transformation facilitates the control of deformation conducted by backdoor attack. As a result, this paper aims to design invisible backdoor attack for 3D point cloud in graph spectral domain.

Graph Fourier Transform. Firstly, we illustrate the process of Graph Fourier Transform (GFT). Given a signal x over a graph $G(V, E, W, D)$, where V is the vertices set, E is the edges set that connect each vertex, W is the adjacency matrix whose element represents the weight of each edge, and D is the degree matrix whose element denotes the degree of each vertex. The procedure for transforming $G(V, E, W, D)$ to graph spectral signal can be described as follows. 1) Obtain the Laplacian matrix L by $L = D - W$. 2) Conduct eigen-decomposition to L, where $L = U\Lambda U^T$. In detail, $U = [u_1, ..., u_n]$ consists of eigenvectors u_i, and $\Lambda = diag(\lambda_1, ..., \lambda_n)$ consists of eigenvalues λ_i. 3) The coefficient vector $\hat{x}$ is acquired by:

$$\hat{x} = GFT(x) = U^{-1}x \tag{3}$$

Moreover, given a coefficient vector $\hat{x}$, the corresponding signal in spatial domain is obtained by inverse GFT:

$$x = IGFT(x) = U\hat{x} \tag{4}$$

GFT for 3D Point Cloud. Representing 3D point cloud by a graph $G(P, E, W, D)$ is an efficient way for shape analyzing. To achieve this, we firstly connect each point with its k-nearest neighbors. The weight of each vertex is set to 1. Furthermore, the vertices position $P = (p_i)_{i=1}^{n} \in R^{n*3}$ of 3D point cloud is regarded as the graph signal. Therefore, its coefficient vector is obtained by $\hat{P} = GFT(P) = U^{-1}P$. Besides, the inverse GFT from coefficient vector to spatial position is $P = IGFT(P) = U\hat{P}$.

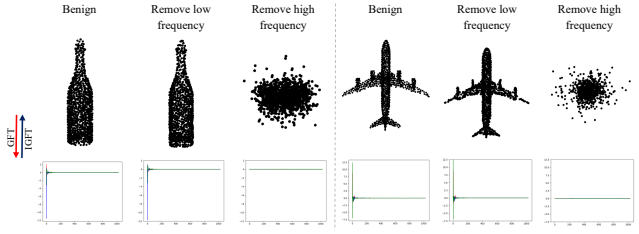

Figure 2: Point clouds generated by removing low frequency or high frequency. It shows that low frequency represents global structure, and high frequency represents local details.

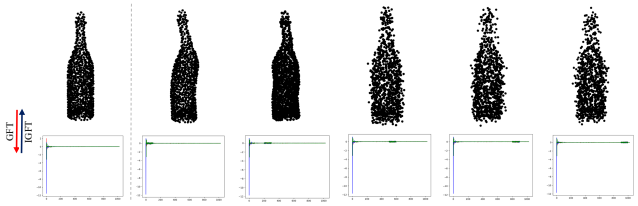

Figure 3: Point cloud deformation caused by perturbing graph spectral signal from different frequency. Perturbing low frequency leads to global shape variation. Besides, perturbing high frequency leads to local points perturbation. To guide both global structure and local points to share the caused deformation, we argue to implant backdoor by perturbing over the entire frequency domain.

Influence of Perturbing Graph Spectral Signal. To decide the way of perturbing graph spectral signal, the spatial shape deformation caused by different graph frequency is analyzed. We empirically define the range from 0 to 200 as low frequency, and range from 200 to 1024 as high frequency. Fig 2 illustrates that removing low frequency erases the global structure of point cloud. Besides, removing high frequency disable point cloud from expressing local details. Furthermore, we take the point cloud *bottle* as an example. Firstly, a random perturbation with a fixed size is conducted to its graph spectral signal from low to high. Then, we transform the perturbed graph spectral signal back to spatial domain to observe its deformation. Results shown in Fig 3 suggest that disturbing low frequency leads to global structure variation, and disturbing high frequency leads to local points perturbation. To pursue the invisibility of backdoored point cloud, we argue that the conducted deformation should be shared by global structure and local points. Therefore, the proposed IBAPC implants backdoor by perturbing over the entire frequency domain.

Learning the Graph Spectral Trigger

To generate stealthy backdoor trigger in graph spectral domain, we newly formulate the trigger implanting function $T_\xi(\cdot)$ with parameter ξ as follows:

$$T_\xi(P) = IGFT(GFT(P) + \xi) \tag{5}$$

$T_\xi(P)$ takes the benign point cloud as the input and returns a backdoored point cloud. Specifically, the benign point cloud is firstly transformed to graph spectral signal by GFT. Then, the backdoor trigger ξ is added to the benign graph spectral signal. Finally, the backdoored point cloud is obtained by IGFT.

One tricky problem is how to decide the added trigger ξ to achieve high stealthiness as well as high attack efficiency. Unlike point cloud in spatial domain where transformation and perturbation can be applied, only numerical vector can be conducted in graph spectral domain. Besides, if we regard a uniform vector as the trigger as does for 2D images, the corresponding information in spatial is unmeaning and will not be learned by the victim 3D DNN. As a result, we generate the backdoor trigger on graph spectral signal by

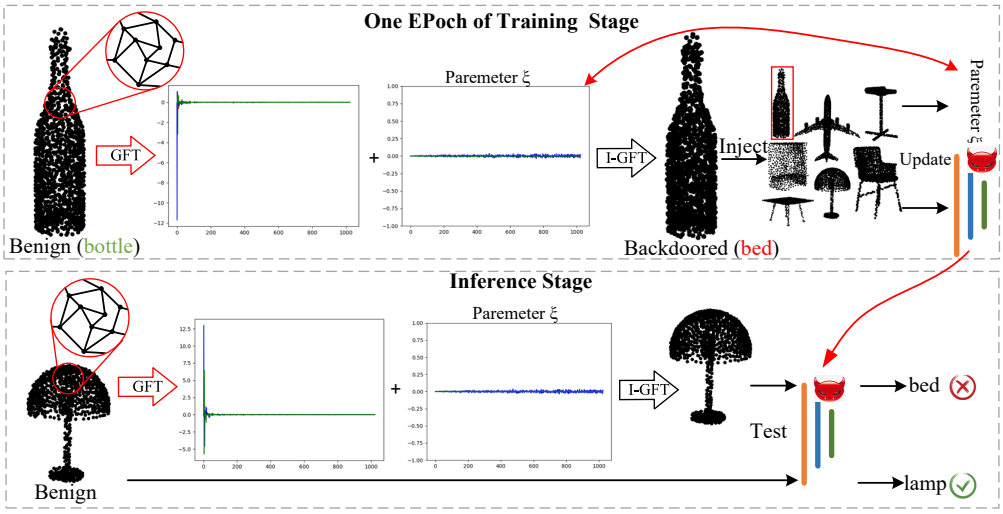

Figure 4: Framework of IBAPC. The parameters ξ and θ for backdoor implanting function $T_\xi(\cdot)$ and victim 3D DNN $f_\theta(\cdot)$ are updated alternately during the training stage. Finally, a backdoored victim 3D DNN $f_{\theta^*}(\cdot)$ with well trained θ^*, and a associated backdoor implanting function $T_\xi(x_i)$ with well trained ξ^* are obtained. In the inference stage, victim 3D DNN will classify the benign point cloud as its corresponding label. However, the backdoored point cloud transformed by $T_{\xi^*}(cdot)$ will be classified as the target label y^t.

learning rather than manually designing. The framework is shown as Fig 4.

Recalling the target of backdoor learning shown as equation 2 which requires the backdoored 3D DNN performs normally on benign point cloud, and achieves high ASR. We further consider the requirement of trigger implanting function $T_\xi(\cdot)$, and redefine the backdoor learning process shown in equation 2 as:

$$\min_{\theta} \sum_{(P_i,y)\in D_{clean}} L(f_\theta(P_i), y_i) + \alpha \sum_{(P_i,y)\in D_{poison}} L(f_\theta(T_{\xi^*}(P_i)), y^t) \tag{6}$$

$$s.t. \quad \xi^* = argmin \sum_{(P_i,y)\in D_{clean}} (L(f_\theta(T_\xi(P_i)), y^t) + \beta D(P_i, T_\xi(P_i)))$$

Where, α and β are weights to control the mixing strengths. $D(P_i, T_\xi(P_i))$ measures the euclidean distance between backdoored point cloud and benign point cloud. By finishing this optimization, three main targets are achieved. In detail, 1) Victim 3D DNN obtains high accuracy on benign point cloud by minimizing $\sum_{(P_i,y)\in D_{clean}} L(f_\theta(P_i), y_i)$; 2) Victim 3D DNN obtains high *ASR* by minimizing $\sum_{(P_i,y)\in D_{poison}} L(f_\theta(T_{\xi^*}(P_i)), y^t)$ and $\sum_{(P_i,y)\in D_{clean}} L(f_\theta(T_\xi(P_i)), y^t)$; 3) Trigger implanting function $T_{\xi^*}(P_i)$ obtains high stealthiness by minimizing $\sum_{(P_i,y)\in D_{clean}} D(P_i, T_\xi(P_i))$. Finally, a well trained 3D DNN with parameter θ^* and its corresponding optimal trigger implanting function $T_{\xi^*}(\cdot)$ are obtained.

Model Training and Inference

The optimization problem shown in equation 6 are non-convex and constrained. The parameters θ and ξ can not be updated at the same time. Therefore, we alternately update θ and ξ while fixing the other one. The backdoor will be implanted into victim 3D DNN by finishing the backdoor training process.

In the inference stage, given a benign point cloud P, attackers simply obtain a backdoored point cloud by the obtained $T_{\xi^*}(P)$. In detail, the victim 3D DNN will misclassify the backdoored point cloud to the target class y^t.

Experiments

Experiments Setting

Datasets and victim 3D DNNs. We select ModelNet40 (MN40), ModelNet10 (MN10) (Wu et al. 2015) and ShapeNetPart (SNP) (Yi et al. 2016) as the evaluating datasets. In particular, there are 12311 models for 40 categories in MN40, where 9843 models are utilized for training and 2468 models for testing. MN10 is down-sampled from MN40, which contains 10 categories. There are 3911 models for training, and 908 models for testing. The SNP with 16 categories is a part of ShapeNet, which contains 12128 and 2874 objects for training and testing, respectively. For fairness, we uniformly sample 1,024 points from the surface of each object and re-scale them into a unit cube. The selected victim 3D DNNs are PointNet (Qi et al. 2017a), PointNet++ (Qi et al. 2017b), DGCNN (Wang et al. 2019), and RSCNN (Liu et al. 2019). They are typical 3D DNNs designed for 3D point cloud classification.

Baseline 3D Backdoor Attacks. This paper regards most existing SOTA 3D backdoor attacks as benchmarks, including PointPBA (Li et al. 2021), PointBA_X (Xiang et al. 2021), IRBA (Gao et al. 2022), NRBdoor (Fan et al. 2022). In detail, PointPBA implants backdoor by conducting rotation (PointPBA-O) and adding a ball (PointPBA-I), respec-

tively. PointBA_X regards a ball as the backdoor trigger, whose location and shape are obtained by optimization. Furthermore, NRBdoor and IRBA both conduct attack by linear transformation.

Evaluating Metrics. Three standard metrics are used for evaluating IBAPC, including Attack Success Rate(*ASR*), Benign Accuracy(*BAc*), and stealthiness. In detail, *ASR* is the main indicator that measures the ability of backdoor attack to mislead victim 3D DNN to output the target label y^t, once backdoor trigger are attached on the inferences point cloud. *BAc* means the inference accuracy of backdoored 3D DNN on the clean samples. For ease of reading, the *BAc* of original 3D DNN is referred as *oBAc* (original Benign Accuracy). Besides, stealthiness measures the ability of backdoor trigger to evade defenders. It is evaluated by quantitative measurement (L_2 distance) and human observation. During the comparison, higher *ASR*, lower descent of *BAc*, and better stealthiness means better backdoor attack.

Attack Experiments

Attack Effectiveness: Results shown in Table 1 illustrate that IBAPC achieves competitive *ASR*. In detail, *ASR* of IBAPC is about 2% less than that of the best attack in most case. Though it can't achieve the best attack efficiency all the time, the loses on *ASR* is significantly less than the benefit on stealthiness.

Attack Stealthiness. We evaluate the stealthiness of IBAPC from the objective measurement and subjective observation respectively. Shown as Table 1, IBAPC achieves the minimum deformation in most cases (11/12). Moreover, the over-performance of IBAPC is great. In particular, the L_2 distance of IBAPC is only 14.29% of PointPBA-O (the second best algorithm), with PointNet++ on MN40. Moreover, Fig 5 shows the backdoored point clouds and the corresponding residual in graph spectral domain. The perturbation conducted on IBAPC is the minimal no matter in spatial domain or graph spectral domain. Besides, the performance of PointPBA-O is close to IBAPC since it does not introduces disturbance on point position, and only a rotation is caused. However, the uniform rotation is usually inspected by human vision, and defended by data augmentation.

Human Inspection. We conduct subjective evaluation on several backdoored point cloud. In detail, for each object instance, we upload five point cloud snapshots including the benign one, the backdoored point clouds generated by PointPBA-I, IRBA, NRBdoor, and IBAPC. The victim 3D DNN is DGCNN trained on MN40. All participants are asked to choose one point cloud that most similar to the benign one, shown as Fig 6. To make the objective evaluation fair, we shuffle the order of the backdoored point clouds, and show each participant with 8 trials. In total, we collect 816 trials from 102 participants. Results suggest that participants argue that IBAPC is the most similar point cloud in 66.79% cases. It is much higher than 14.33% of PoinPBA-I, 3.92% of IRBA, and 14.95% of NRBdoor. Therefore, IBAPC is the most stealthy backdoor attack for human inspection.

All-to-all Attack. Different from all-to-one attack, all-to-all attack assigns different label to different sample. We evaluate the performance of IBAPC under all-to-all attack scenario following (Doan et al. 2021). The victim 3D DNNs are DGCNN and PointNet++ trained on MN40. Results in Fig 7 suggest that the performance of IBAPC is guaranteed.

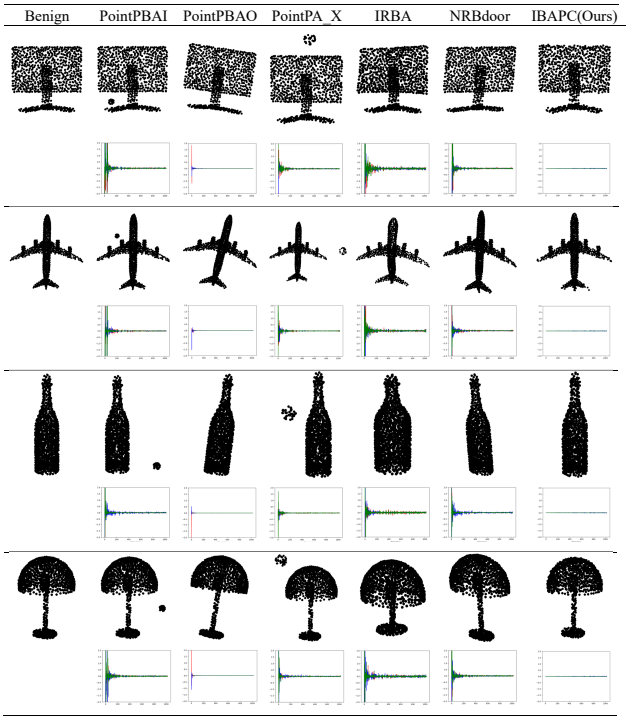

Figure 5: Disturbance conducted by different backdoor attacks in spatial domain and graph spectral domain. The second line below point clouds represents the residual of graph spectral signal, lower is better. IBAPC achieves the best stealthiness.

Defense Experiments

Resistance to STRIP. STRIP (Gao et al. 2019) fuses the detected sample with multiple benign samples to obtain a prediction distribution. Observing the overlap of two distributions from backdoored samples and benign samples, the larger the overlap, the more difficult it is for the backdoored sample to be detected. We compare IBAPC with PointPBA-I. The victim 3D DNNs are DGCNN trained on MN40 and ShapeNet, and PointNet trained on MN10. Results in Fig 8 illustrate that the distribution overlap with benign distribution of IBAPC is larger than that of PointPBA-I. Therefore, IBAPC obtains better resistance performance.

Resistance to Saliency-based Defense. Such defense evades backdoor attack by locating and removing attached trigger utilizing saliency score (Huang, Alzantot, and Srivastava 2019). Fan *et al.* (Fan et al. 2022) extend them to 3D point cloud domain. The evaluation with DGCNN on MN40 in Fig 9 illustrates that IBAPC can resist to saliency-based defense. In contrast, the attached trigger by PoinPBA-I and PoinBA_X are removed. Furthermore, *ASR* of IBAPC slightly decreases from 95.63% to 87.45%. However, that of PointPBA-I and PointBA_X decrease from 100% and

Dataset ↓	Victim → / Attack ↓	PointNet			PointNet++			DGCNN			RSCNN		
		BAc	ASR	L_2	BAc	ASR	L_2	BAc	ASR	L_2	BAc	ASR	L_2
ModelNet40	Clean	89.13	-	-	91.46	-	-	92.01	-	-	93.28	-	-
	PointPBA-I	88.18	90.48	5.558	91.46	**100**	5.558	**92.42**	**100**	5.558	89.50	88.33	5.558
	PointPBA-O	88.53	96.54	**2.533**	**91.93**	97.55	2.533	92.01	89.83	2.533	88.93	88.16	2.533
	NRBdoor	84.10	**97.40**	2.346	89.85	99.27	3.735	92.00	97.10	2.152	89.14	88.16	2.152
	PointBA_X	88.56	81.75	6.301	90.77	78.68	15.16	92.03	93.55	2.825	89.46	83.16	8.811
	IRBA	87.70	87.62	35.17	90.71	89.82	35.17	91.73	99.61	35.17	89.95	97.16	35.17
	IBAPC(our)	**88.94**	93.16	2.754	91.31	92.08	**0.362**	91.69	94.50	**0.372**	**90.07**	**99.83**	**1.290**
ModelNet10	Clean	93.82	-	-	94.75	-	-	94.53	-	-	94.67	-	-
	PointPBA-I	**93.64**	90.57	5.610	**94.97**	99.17	5.610	**98.95**	**100**	5.610	91.74	86.00	5.610
	PointPBA-O	93.57	96.21	2.203	95.07	99.17	2.203	94.49	97.33	2.203	**95.66**	92.29	2.203
	NRBdoor	93.31	**99.67**	2.386	**94.97**	99.67	2.356	94.38	100	2.218	92.95	97.66	2.442
	PointBA_X	92.07	94.46	6.944	93.50	86.00	8.753	93.30	52.30	4.490	93.10	87.95	9.486
	IRBA	91.37	95.31	35.95	93.09	99.38	35.95	94.27	100	35.95	93.50	**100**	35.95
	IBAPC(our)	93.15	93.16	**1.618**	94.38	**99.83**	**0.321**	92.84	96.16	**0.311**	91.40	90.50	**0.985**
ShapeNetPart	Clean	98.37	-	-	99.22	-	-	99.08	-	-	99.03	-	-
	PointPBA-I	98.64	93.36	5.432	99.05	99.01	5.432	98.95	**100**	5.432	98.64	96.33	5.432
	PointPBA-O	**98.75**	95.55	2.414	**99.08**	97.53	2.414	98.60	91.50	2.414	98.71	93.16	2.414
	NRBdoor	98.57	98.68	2.129	99.05	99.34	1.936	98.08	95.16	2.221	98.50	92.50	2.092
	PointBA_X	92.03	93.55	2.825	93.30	52.30	4.490	**98.91**	95.55	3.137	97.44	90.73	9.761
	IRBA	98.05	91.84	33.81	98.70	93.41	33.81	99.23	99.86	33.81	**99.19**	97.96	33.81
	IBAPC(our)	98.71	**99.06**	**1.301**	98.46	**99.76**	**0.530**	98.81	98.16	**0.768**	98.46	**98.00**	**1.002**

Table 1: Performance of comparison backdoor attacks. The best value is shown in bold. IBAPC achieves competitive *ASR* and the minimal deformation measured by L_2 distance. *ASR* of IBAPC is about 2% less than that of the best attack in most case. However, the loses on *ASR* is significantly less than the benefit on stealthiness.

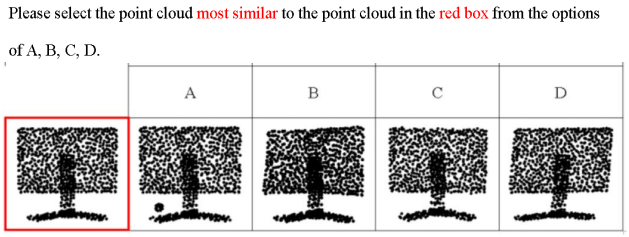

Figure 6: One trail of the designed human inspection evaluation. 816 trials from 102 participants are collected. 66.79% cases support that IBAPC is the most similar point cloud with the benign one, which is higher than the comparison attacks.

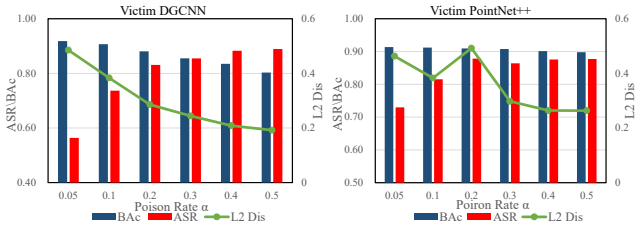

Figure 7: Performance of IBAPC under all-to-all attack scenario. Attack efficiency and stealthiness enhanced according to the increase of poison rate α.

93.55% to 4.90% and 10.36%, respectively.

Resistance to Data Augmentation. Data augmentation is a popular way to increase the robustness of DNN. We augment the backdoored data by random rotation, jitter, random scaling, and shift. The victim 3D DNN is DGCNN trained on MN40. Results in Table 2 illustrate that IBAPC can evade data augmentation. Its *ASR* descent is 7.14% when conducting all augmentation at once. In contrast, *ASR* descent of PointPBA-O is 87.66%.

Resistance to SOR. Statistical Outlier Removal (SOR) (Zhou et al. 2019) removes the noisy points by obtaining the distribution of point distance. Our evaluation varies hyper-parameter k of k-nearest neighbors from 2 to 65, and sets standard deviation σ to 1.0. The victim 3D DNNs are DGCNN and PointNet++ trained on MN40 and MN10, respectively. Results comparing with PoinPBA-I shown in Fig 10 suggest that IBAPC can resist SOR. Specifically, its *ASR* changes slightly according to the variation of k. In contrast, that of PointPBA-I decreases to 50.51% with DGCNN on MN40 when k = 65.

Resistance to SSD. Spectral Signature Defense (SSD) (Tran, Li, and Madry 2018) obtains a cleaned training dataset by detecting and removing the backdoored samples

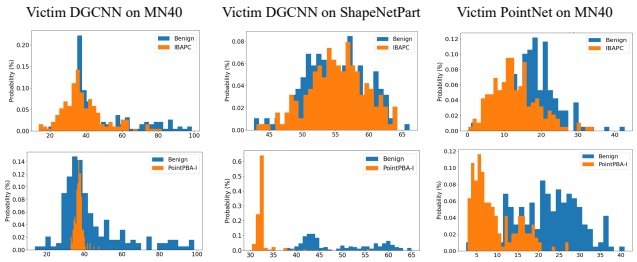

Figure 8: Defense results against STRIP. IBAPC achieves larger overlap than PoinPBA-I. Hence, it obtains the better resistance performance.

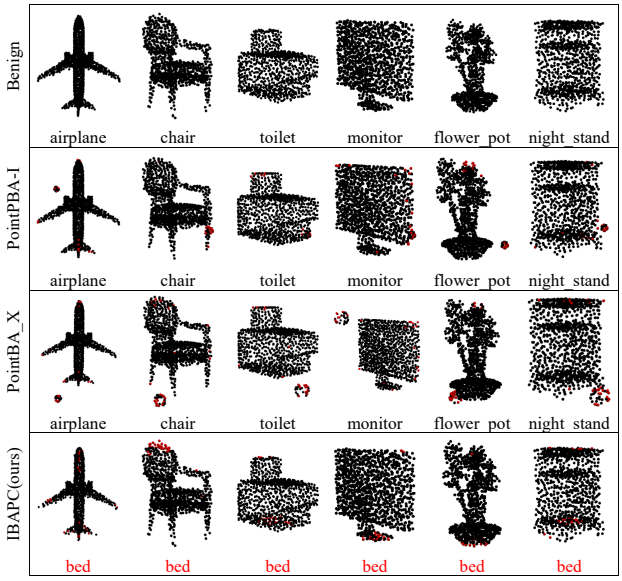

Figure 9: Resistance results against saliency-based defense. The trigger attached by IBAPC cannot be detected by saliency-based defense. In contrast, that of PointPBA-I and PointBA_X are defended.

from each class. Then a detoxified DNN is acquired by retraining on the cleaned training dataset. Our evaluation regards DGCNN trained on MN40 as the victim 3D DNN. Poison rate is set to 0.05. Results suggest that 50.21% backdoored samples are detected by SSD. However, the retrained 3D DNN still acquires 88.81% *ASR*.

Ablation Experiments

Poison Rate α and Target Class y^t. In practical scenario, poison rate α and target class y^t both vary largely. Fig 11 shows the attack efficiency of IBAPC with DGCNN according to the variation of poison rate and target class. In particular, target class has slightly influence on the attack performance. Such characteristic enables IBAPC to meet different requirement. Besides, *ASR* is enhanced with the increase of poison rate. It is reasonable since higher poison rate brings more backdoored samples for learning backdoor trigger.

Necessity for Implanting Backdoor Trigger in Graph

Table 2: *ASR* of IBAPC and PointPBA-O against data augmentation. IBAPC is able to resist data augmentation.

	None	Rotation	Jitter	Scaling	Shifting	All
IBAPC	94.50	92.54	90.27	96.91	96.91	87.36
PointPBA-O	89.83	2.66	93.33	93.00	92.67	2.17

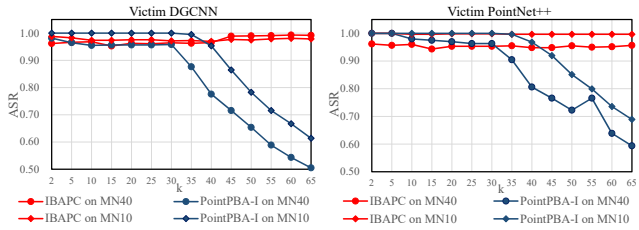

Figure 10: Resistance results against SOR. ASR of PointPBA-I decreases sharply when $k = 30$. By contrast, that of IBAPC changes slightly according to the variation of k.

Spectral Domain. To illustrate the necessity for implanting backdoor trigger through graph spectral domain rather than spatial domain, we directly conduct trigger on point position by replacing $T_\xi(P)$ in equation 6 with P. Results with DGCNN trained on MN40 suggest that the modified attack achieves 49.62% *ASR*, which is much lower than 94.50% of virgin IBAPC.

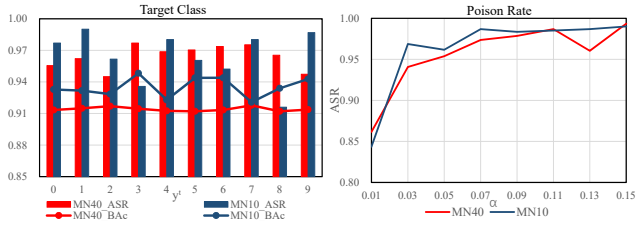

Figure 11: Influence of target class y^t and poison rate α. Attack efficiency changes slightly with the variation of y^t. Besides, increasing α can enhance *ASR* since more backdoored samples are generated for learning backdoor trigger.

Conclusions and Future Works

Aiming to generate invisible backdoored 3D point cloud, this paper proposes IBAPC which conducts backdoor trigger in graph spectral domain. Specifically, a new backdoor implanting function is designed. Moreover, the parameter of backdoor implanting function and the victim 3D DNN are alternately updated. Experiment results suggest that IBAPC achieves SOTA attack stealthiness and competitive attack efficiency. Though its efficiency, the altered label will alarm defender. Hence, the invisible 3D backdoor attack under clean label should be studied in the future works.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62072348 and China Yunnan province major science and technology special plan project No. 202202AF080004. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Cheng, Z.; Wu, B.; Zhang, Z.; and Zhao, J. 2023. TAT: Targeted backdoor attacks against visual object tracking. *Pattern Recognition*, 142: 109629.

Doan, K.; Lao, Y.; Zhao, W.; and Li, P. 2021. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11966–11976.

Doan, K. D.; Lao, Y.; Yang, P.; and Li, P. 2023. Defending backdoor attacks on vision transformer via patch processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 506–515.

Fan, L.; He, F.; Guo, Q.; Tang, W.; Hong, X.; and Li, B. 2022. Be Careful with Rotation: A Uniform Backdoor Pattern for 3D Shape. *arXiv preprint arXiv:2211.16192*.

Feng, Y.; Ma, B.; Zhang, J.; Zhao, S.; Xia, Y.; and Tao, D. 2022. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20876–20885.

Gao, K.; Bai, J.; Wu, B.; Ya, M.; and Xia, S.-T. 2022. Imperceptible and Robust Backdoor Attack in 3D Point Cloud. *arXiv preprint arXiv:2208.08052*.

Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D. C.; and Nepal, S. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, 113–125.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Huang, X.; Alzantot, M.; and Srivastava, M. 2019. Neuroninspect: Detecting backdoors in neural networks via output explanations. *arXiv preprint arXiv:1911.07399*.

Li, X.; Chen, Z.; Zhao, Y.; Tong, Z.; Zhao, Y.; Lim, A.; and Zhou, J. T. 2021. Pointba: Towards backdoor attacks in 3d point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16492–16501.

Liu, D.; and Hu, W. 2022. Imperceptible transfer attack and defense on 3d point cloud classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4727–4746.

Liu, Y.; Fan, B.; Xiang, S.; and Pan, C. 2019. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8895–8904.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Tran, B.; Li, J.; and Madry, A. 2018. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31.

Wang, T.; Yao, Y.; Xu, F.; An, S.; Tong, H.; and Wang, T. 2022. An invisible black-box backdoor attack through frequency domain. In *European Conference on Computer Vision*, 396–413. Springer.

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

Wu, B.; Chen, H.; Zhang, M.; Zhu, Z.; Wei, S.; Yuan, D.; and Shen, C. 2022. Backdoorbench: A comprehensive benchmark of backdoor learning. *Advances in Neural Information Processing Systems*, 35: 10546–10559.

Wu, B.; Liu, L.; Zhu, Z.; Liu, Q.; He, Z.; and Lyu, S. 2023. Adversarial machine learning: A systematic survey of backdoor attack, weight attack and adversarial example. *arXiv preprint arXiv:2302.09457*.

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 IEEE conference on computer vision and pattern recognition*, 1912–1920.

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

Xiang, Z.; Miller, D. J.; Chen, S.; Li, X.; and Kesidis, G. 2021. A backdoor attack against 3d point cloud classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7597–7607.

Yi, L.; Kim, V. G.; Ceylan, D.; Shen, I.-C.; Yan, M.; Su, H.; Lu, C.; Huang, Q.; Sheffer, A.; and Guibas, L. 2016. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6): 1–12.

Yuan, Z.; Zhou, P.; Zou, K.; and Cheng, Y. 2023. You Are Catching My Attention: Are Vision Transformers Bad Learners Under Backdoor Attacks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24605–24615.

Yue, C.; Lv, P.; Liang, R.; and Chen, K. 2022. Invisible backdoor attacks using data poisoning in the frequency domain. *arXiv preprint arXiv:2207.04209*.

Zeng, Y.; Park, W.; Mao, Z. M.; and Jia, R. 2021. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16473–16481.

Zhao, Z.; Chen, X.; Xuan, Y.; Dong, Y.; Wang, D.; and Liang, K. 2022. DEFEAT: Deep Hidden Feature Backdoor Attacks by Imperceptible Perturbation and Latent Representation Constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15213–15222.

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 IEEE/CVF International Conference on Computer Vision*, 1961–1970.