CausalPC: Improving the Robustness of Point Cloud Classification by Causal Effect Identification

Yuanmin Huang, Mi Zhang*, Daizong Ding, Erling Jiang, Zhaoxiang Wang, Min Yang
School of Computer Science, Fudan University, China
{yuanminhuang23@m.,mi_zhang@,17110240010@,eljiang21@m.,wangzx23@m.,m_yang@}fudan.edu.cn

Abstract

Deep neural networks have demonstrated remarkable performance in point cloud classification. However, previous works show they are vulnerable to adversarial perturbations that can manipulate their predictions. Given the distinctive modality of point clouds, various attack strategies have emerged, posing challenges for existing defenses to achieve effective generalization. In this study, we for the first time introduce causal modeling to enhance the robustness of point cloud classification models. Our insight is from the observation that adversarial examples closely resemble benign point clouds from the human perspective. In our causal modeling, we incorporate two critical variables, the structural information, (standing for the key feature leading to the classification) and the hidden confounders, (standing for the noise interfering with the classification). The resulting overall framework CausalPC *consists of three sub-modules to identify the causal effect for robust classification. The framework is model-agnostic and adaptable for integration with various point cloud classifiers. Our approach significantly improves the adversarial robustness of three mainstream point cloud classification models on two benchmark datasets. For instance, the classification accuracy for DGCNN on ModelNet40 increases from* 29.2% *to* 72.0% *with* CausalPC*, whereas the best-performing baseline achieves only* 42.4%*.*

1. Introduction

Recent years have witnessed tremendous development in autonomous driving, where understanding 3D point clouds plays an indispensable role. Although deep neural networks (DNN) have achieved extraordinary performance in point cloud classification [32, 33, 45], the deep and non-linear structure also raises the concern of *adversarial attacks*. For instance, an attacker can slightly modify a point cloud by human-unnoticeable perturbations to mislead the prediction of a DNN. Since such misprediction can lead to severe consequences in real-world scenarios, the study of adversarial examples in point cloud classification has become an intense topic in 3D vision.

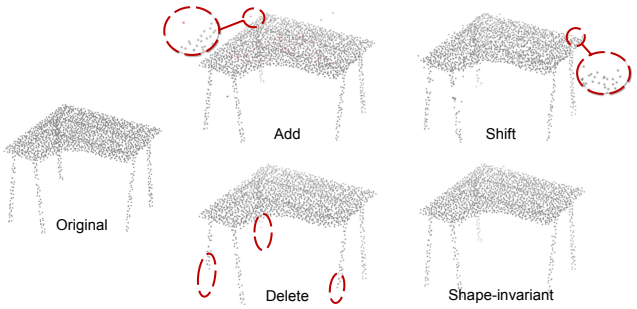

Figure 1. The demonstration of adversarial point clouds generated by various adversarial attacks.

Existing adversarial attacks for point cloud data are summarized in Fig.1. Each attack holds a distinct geometric pattern. For instance, beyond the conventional shifting points attacks [15, 51] (i.e., slight perturbation on the position of existing points), previous works also propose adding [51] or deleting [62] points as their attack strategies. Recent variants of shifting points attacks use generative models like GANs and AutoEncoders to yield perturbed point clouds [11, 64]. Following the definition of adversarial examples, these attacks produce such slight perturbations to a point cloud that are imperceptible to human eyes, which however become the key for us to design an effective defense.

Existing defenses against adversarial examples in point cloud data are roughly categorized into *adversarial training* and *input-oriented defense*. The former incorporates the generated adversarial examples into the training set, aiding classifiers in correctly identifying perturbed point clouds [20, 21]. In contrast, the latter explicitly recognize specific patterns, such as outliers within adversarial point clouds, and mitigate these identified patterns [7, 63]. However, both the above approaches share the common limitation: they fail to generalize in defending against more sophisti-

*Corresponding author: Mi Zhang

cated attacks. For instance, recent findings [12] reveal that a model trained with PGD adversarial training can be successfully attacked by GeoA3 [46], an intricately designed attack. Moreover, newly proposed shape-invariant attacks [12, 46] can evade input-oriented defenses as they introduce significantly fewer outlier biases during the attack.

When faced with emerging attack strategies, developing a defense framework that can effectively mitigate diverse adversarial examples becomes a primary challenge. To tackle this issue, we direct our attention to the following observation: *Regardless of the attack strategy to generate perturbations, the resulting adversarial examples still bear a strong resemblance to benign ones for human eyes.* This can be observed in Fig.1, where the perturbed point cloud remains identifiable as a table. Motivated by this finding, we propose to incorporate the causal language [8, 29] into the formulation of point cloud classification. This helps investigate the causal relation between variables which aligns with human reasoning about the natural world.

Technically, incorporating the causal language in point cloud applications is challenging: the unique modality of point cloud data makes it impractical to apply existing causal modeling and effect identification techniques. In Section 3, we first summarize two causal variables that influence point cloud classification, i.e., the *structural information* that humans usually extract to recognize a point cloud and the *hidden confounders* that may influence both the point cloud and prediction, e.g., the potential adversarial perturbations and the noises caused by LiDAR sensors. Based on the above variables, we build a causal graph to identify the causal effect from the point cloud data to the label. We note that humans could correctly identify the causal effect naturally, i.e., neglecting the influence of the hidden confounders and paying attention to the structural information given the point cloud data. Therefore, humans are insensitive to noises or adversarial perturbations in most cases. In contrast, existing classifiers are designed to characterize the statistical correlations instead of the causal effect between the input and the label. As a result, existing classifiers are sensitive to noises and adversarial perturbations, though they seem to be effective in modeling the structural information in some cases [22, 51].

Following the above analysis, our defense neglects the influences caused by hidden confounders and identifies the correct causal effect from the point cloud data to the label. In Section 4, we propose a novel defense framework called *CausalPC* to achieve the goal. Faced with the challenge that the hidden confounders are difficult to measure or even observe in this problem, we develop a causal inference algorithm based on *front-door adjustment*, which could effectively approximate the expectation of predictions over all potential hidden confounders. Furthermore, we take inspiration from the mesh representation of real-world objects and propose an effective module to characterize the structural information. Finally, we develop a novel attention module to incorporate the structural information into the prediction.

In summary, we mainly make the following contributions:

- We propose *CausalPC*, the first solution to mitigate adversarial examples in point cloud classification through a causal lens. Instead of focusing on certain kinds of attack techniques or input patterns, we take inspiration from human recognition systems and propose to identify the causal effect to make correct classification under various kinds of adversarial perturbations.

- *CausalPC* is a model-agnostic defense framework, which could be incorporated with different point cloud classifiers. With the aid of the proposed causal inference algorithm, we could correctly identify the causal effect with only slight structure modifications and partial finetuning on existing classification models.

- *CausalPC* substantially improves the adversarial robustness of three mainstream PC classification models on two benchmark datasets (Section 5). For example, the classification accuracy is increased from 29.2% to 72.0% for DGCNN on ModelNet40, while the best baseline only achieves 42.4%.

2. Preliminary and Related Work

2.1. Point Cloud Classification

Point cloud classification aims to categorize objects consisting of a set of points with 3D coordinates into different classes. Formally, the task can be viewed as learning a mapping function $F : X \mapsto y$, where $X \in \mathbb{R}^{N\times 3}$ denotes a point cloud with N points with 3D coordinates and $y \in \{1, 2, \ldots, C\}$ denotes the ground truth label among C categories that X belongs to. Recently, deep neural networks have achieved great success in point cloud classification with their deep non-linear feature extraction capability [24]. In this paper, we mainly consider point-based models [32, 33] due to the superior effectiveness of characterizing geometry patterns. These models directly take 3D coordinates of point clouds as input without further pre-processing procedures such as voxelization [52] or pillarization [17]. Generally, a point-based model first extracts latent features for each of the N points in an input point cloud $X \in \mathbb{R}^{N\times 3}$ with its non-linear feature extractor f, e.g., MLP in PointNet/PointNet++ [32, 33], GNN in DGCNN [45], and CNN in PointCNN [19]. Then, a fully-connected layer is used as the classification head given the extracted global feature h of the point cloud. For a more comprehensive study, please refer to the work [24].

2.2. Adversarial Examples

Adversarial attacks aim to mislead the output of DNN models by introducing minor perturbations to the input samples. Following attacks targeting image classification models [9, 39], adversarial attacks have begun to emerge in the context of point cloud classification as well [46, 51]. The objective of adversarial attacks can be formulated as,

$$\min_{\eta} \mathcal{L}(F(\tilde{X}, \tilde{y})), \quad \text{s.t.}, ||\eta||_p \leq \delta, \tag{1}$$

where F and $\mathcal{L}$ denote the model and classification loss function, η, $\tilde{X} = X \oplus \eta$ and $\tilde{y}$ are the perturbation, adversarial example and target label, respectively. Adversarial attacks for point clouds are diverse in methodology. In terms of the form of perturbations, i.e., $\oplus$ in Eq.1, shifting all the points by a certain distance [51], adding [22, 51] or deleting [47, 62] points or transforming the entire point cloud [11, 64] are all alternatives. As for the perturbation budget $||\eta||_p$, multiple distance metrics including L_0, L_2, L_∞ have been utilized, e.g., Chamfer Distance as a representative for L_2 distance [51].

Adversarial defenses for point cloud classifiers develop as well. Recent works can be roughly categorized into the adversarial training (AT)-based ones [21, 59] and the input-oriented-based ones [63]. The former line takes inspiration from works in image classification [26], where the defender augments the training data of a classifier with adversarial examples. On the other hand, input-oriented-based defenses pay attention to the particular characteristics introduced by specific attack strategies. To name a few, SOR [63] filters out outlier points introduced by adversarial perturbations, while DUP-Net [63] further copes with the inconsistent point density within a point cloud caused by attacks. GvG [7] tends to correct the relative position bias of local parts of a perturbed point cloud.

Despite the effectiveness in defending against certain kinds of attacks, both lines of defense share a similar drawback. They generalize poorly to unseen attacks with characteristics other than attacks that have been considered. In comparison, our work focuses on the inherent characteristic of point cloud adversarial examples that they preserve the similarity to normal samples from the perspective of human perception. Subsequently, we propose a defense method with the capability to counter various types of attacks.

2.3. Causal Inference

Causality refers to the modeling of relationships between factors in a task from a human perspective [8, 29]. Although deep neural networks (DNNs) have shown superior performance over humans in certain tasks, recent studies have revealed that they may rely on spurious correlations, as their optimization objective is to learn statistical correlations only [14]. Since DNNs lack the ability to perform causal reasoning, they are more susceptible to paying attention to irrelevant features, such as a model relying on the wrong parts of input to make classification decisions [34, 44]. This vulnerability to external noise decreases model performance, reinforcing the importance of incorporating causal reasoning into machine learning models.

As a result, several researches have been proposed to study the causal effect between variables through causal inference [1, 29, 31] to improve the performance of DNNs from various aspects, such as fairness [10, 16], generalization [57], and adversarial robustness [35, 61]. Causal inference helps models learn the correct causal relations between variables and focus on those causal factors only.

Although several works have focused on enhancing the robustness of image classification models against adversarial [35, 61] or natural noises [42] with causal inference, it is nontrivial to adapt these methods to point cloud classification due to the inherent differences in human modeling between the two tasks, which leads to completely different solutions. For instance, previous works characterize image style information [61] and background patterns [35] to investigate the causal relations between images and classification decisions, which are not applicable to point cloud data. In this work, we propose a specific causal modeling approach based on human observations for *point cloud classification* for the first time.

3. The Causal Modeling

3.1. Causal Modeling of Point Cloud Classification

To quantify the underlying logic behind human decision-making, we introduce the use of causal graph [8, 29], which employs structured paths (i.e., *causal reasoning paths*) between nodes (i.e., *causal variables*). Specifically, we construct a causal graph to formalize the process of point cloud classification, as illustrated in Fig.2 (a), consisting of three causal variables during mapping point cloud data X to classification results Y: the physical object O corresponding to X, the structural information Z involved in X, and external noises U.[1] We now show the four causal reasoning paths regarding the decision process for this task:

- *Point cloud generation*: We first consider the generation process from a real-world object to point cloud data. From the physical world perspective, given an object O, we can obtain point cloud data X through multiple LiDAR scans [48] or modeling mesh data and sampling [49]. This process can be described as $O \rightarrow X$, where the arrow represents a causal reasoning path.

[1] Throughout the paper in the following sections, we use capital letters like X to denote random variables, while lowercase letters like x to denote a specific value taken for the variable.

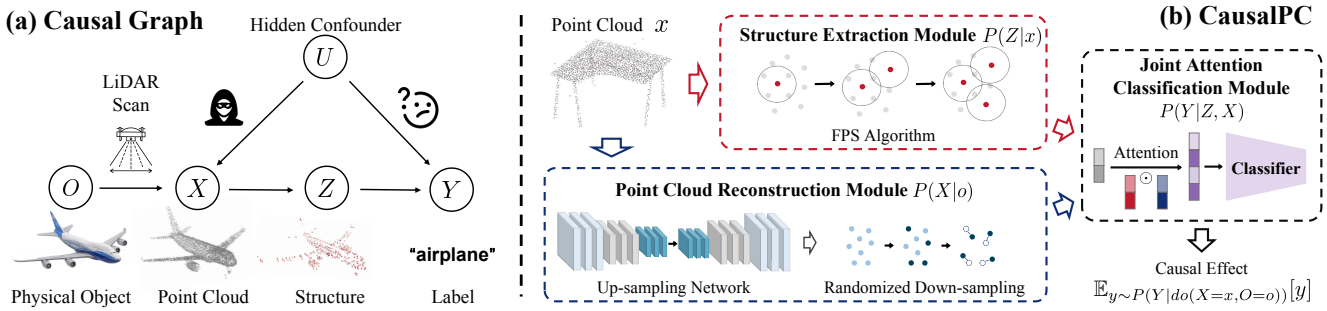

Figure 2. The proposed point cloud classification through a causal lens: (a) the causal graph; (b) the proposed framework.

- *Structural information extraction*: At first glance, humans typically recognize the structural information of an object, i.e., the framework [13]. This process involves identifying the main parts of the object and their relative positions, as well as the symmetry and other features that determine the overall structure. We use Z to represent such structural information and path $X \rightarrow Z$ to describe such a process.

- *Classification*: With the accurate structural information Z, the classification of a point cloud can be correctly performed. Specifically, humans recall what kinds of objects share a similar structure, i.e., the path $Z \rightarrow Y$.

- *Hidden confounders*: Various hidden variables influence the generation of point cloud X, including sampling noises from different LiDAR sensors and adversarial noise from potential attackers with specific strategies. We represent these potential noises collectively with variable U and refine the generation process as $(O, U) \rightarrow X$ in the causal graph. In real-world scenarios, where Y may itself be subject to spurious noises like dataset annotation errors [43] and recognition bias introduced by humans [27], we also introduce the variable U to denote such noises for convenience, resulting in the causal path $(Z, U) \rightarrow Y$. In this context, variable U serves as *hidden confounders* in causal modeling, influencing both the input X and the output Y through the causal reasoning path $X \leftarrow U \rightarrow Y$.

3.2. Flaws of Existing Models and Defenses

Based on the built causal graph, we now analyze why DNN-based point classification models are vulnerable to adversarial perturbations. In related literature, DNNs tend to model the statistical correlations between point cloud X and label Y, given the optimization goal of maximizing the classification accuracy [14]. With the absence of the modeling of hidden confounders, such a naive optimization goal would lead DNNs to overfit the potential spurious features, e.g., the local pattern of normal point clouds, instead of structural information [47, 50]. Therefore, DNNs may rely on a shortcut solution where the negative impact of U spreads to the classification through the indirect causal reasoning path $X \leftarrow U \rightarrow Y$. As a result, when facing adversarial point clouds, DNNs would be confused by a previously unseen statistical pattern, which leads to a wrong classification. In comparison, recent research has also pointed out that humans are able to model the causal effect between data instead of statistical ones, which therefore helps avoid the influence of the confounding variables within the data [37, 58]. For instance, in point cloud classification, we can notice abnormal points in Fig.1 and still recognize the structure of the table, which helps us make the correct classification decision.

Through the lens of causal reasoning, existing defenses [7, 21, 63] in point cloud classification attempt to optimize the modeling process of the classification task with only partial observation of U. Identifying the perturbation patterns of points introduced by particular types of attacks, these methods improve the modeling between X and Y by explicitly considering certain kinds of U, e.g., specific input patterns or attack techniques. However, due to the incomplete modeling of U, these defenses can not generalize well to new attacks with other patterns, leading to poor robustness. For instance, our empirical results (in Section 5) show that the adversarial training baseline trained with IFGM [21] adversarial examples still reaches an attack success rate of 100% when attacked by GeoA3 [46], an attack with distinct geometry patterns from IFGM, on PointNet and DGCNN.

4. The Proposed CausalPC

4.1. Causal Effect Identification

To avoid making the classification process overfit to specific hidden confounders, we propose to perform robust modeling according to the causal graph in Fig.2 (a). To this end, we leverage the techniques of the graphical causal model (GCM) framework [29] to formulate the causal graph and identify the causal effect, where conditional probabilities are used to represent the causal effects between variables.

For instance, the distribution $P(Z|X)$ stands for the causal effect $X \rightarrow Z$ in the graph. Based on the definition, we formulate the classification as follows,

$$P(Y|do(X=x)) = \int P(Y|Z,U)P(Z|x)P(U)dUdZ. \tag{2}$$

where the operation do-calculus is the *causal intervention* in related literature [8]. In our problem, the causal intervention stands for performing classification given a certain point cloud x. As such, O and U-related terms on x are omitted because the observed x has deterministic O and U.

Eq.2 stands with our motivation that the causal modeling of point cloud classification should take all possible values of confounding noise U into account. In addition, the proper structural information Z should also be extracted to perform the classification indicated by the term $P(Z|x)$, where x may contain the information of hidden confounders. However, an obvious challenge present is that such a requirement for integration on U is unrealistic during the implementation due to the unobservable U that cannot be exhaustively enumerated. To tackle this challenge, we propose the following theorem based on the front-door adjustment [8].

Theorem 1. *Under the causal graph in Fig.2 (a), suppose the real-world object of a point cloud data x is o. The causal effect is transformed into the following equation:*

$$\begin{aligned} P(Y|do(X=x)) &= P(Y|do(X=x,O=o)) \\ &= \int P(Z|x)\left[\int P(Y|Z,X)P(X|o)dX\right]dZ. \end{aligned} \tag{3}$$

Theorem 1 states that we derive Eq.3 from Eq.2, which excludes the explicit modeling for the unobservable U. Intuitively, the conditional probability $P(X|o)$ describes the distribution of all sampled point cloud data X given physical object o, which implicitly enumerates the hidden confounders U. The complete proof for the theorem can be found in the supplementary material.

4.2. Causal Inference

Based on Eq.3, we now show how we build up a practical causal inference framework to identify the causal effect. First of all, we propose to leverage the Monte Carlo sampling to approximate the integral term,

$$\begin{aligned} P(Y|do(X=x,O=o)) &= \mathbb{E}_{z\sim P(Z|x),x'\sim P(X|o)}\left[P(Y|z,x')\right] \\ &\approx \frac{1}{M_z \cdot M_x}\sum_{i=1}^{M_z}\sum_{j=1}^{M_x} P(Y|z_i,x'_j), \end{aligned} \tag{4}$$

where x'_j and z_i are sampled from $P(X|o)$ and $P(Z|x)$, M_x and M_z are the sampling size of the two variables, respectively. To implement the sampling process and the causal inference, the practical modeling of the three distributions $P(Z|x)$, $P(X|o)$, and $P(Y|Z,X)$ becomes a technical challenge. To fill the gap, we develop three submodules, i.e., a structure extraction module, a point cloud reconstruction module, and a joint attention classification module. They work together as illustrated in Fig.2 (b). We detail their design as follows.

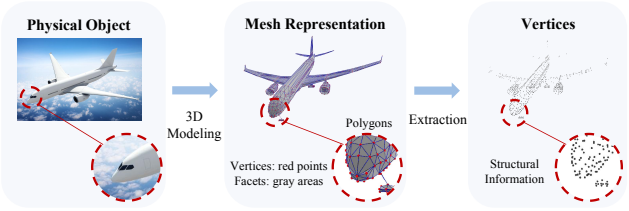

Figure 3. The mesh representation of physical objects and the demonstration of framework vertices.

Structure Extraction Module. For $z \sim P(Z|x)$, i.e., extracting the structural information of an observed point cloud x, we take inspiration from the mesh representation of real-world 3D objects. Specifically, a mesh is a set of small polygons as an approximation of 3D objects, where each one of the polygons is typically built by several vertices and a facet, as shown in Fig.3. In related literature, the vertices and facets describe the structural and textural information of a 3D object [49]. Motivated by this, we propose to use the farthest point sampling (FPS) strategy [33] to sample a subset of points from the observed point cloud x as an approximation of vertices in mesh representation. FPS starts from a randomly selected point and iteratively samples new points far from the set of selected points, which shares a similar idea to the mesh construction for an object where the vertices are expected to be scattered along the object surface [2].

Point Cloud Reconstruction Module. For $x \sim P(X|o)$, the observed point clouds given a physical object o, we need to simulate the sampling process of modern point cloud scanners like a LiDAR sensor ideally. However, such a simulation is impractical due to the unaccessible original object o. Therefore, we first seek for a solution to recover o from x. As our causal graph shows, the point cloud generation process $(O,U) \rightarrow X$ inevitably leaves out some of the detailed features of the physical object to represent a surface with limited points. This motivates us to develop an up-sampling network to recover the lost details. To this end, we train an up-sampling model based on PU-Net [56], which effectively transforms a point cloud into a denser one. With the reconstructed point cloud data, we utilize random subsampling to approximate the point cloud sampling process,

i.e., $x \sim P(X|o)$. Moreover, to simulate the random noises introduced during point cloud sampling, we add additional noises of uniform distribution to the sampled point clouds.

Joint Attention Classification Module. The estimation of $P(Y|Z,X)$ denotes the overall classification distribution given a resampled point cloud x and the structural information z. Different from existing works that leverage a similar front-door adjustment technique in other tasks [38, 54], where typically an extra network is dedicated for modeling $P(Y|Z,X)$, we propose to directly adopt existing point cloud classifiers, e.g., PointNet [32], DGCNN [45] and PointCNN [19]. We show that one advantage of the mainstream models is that they can take an arbitrary number of points for a point cloud as input, making inference valid with the union of x and z. Particularly, we leverage the feature extractor f of existing models to extract feature h_{zx} for the concatenated z and x. In addition to h_{zx}, we also introduce $h_z = f(z)$, $h_x = f(x)$ for joint classification. Now that h_{zx} encodes the global feature of both structural and reconstructed object information, we design extra cross-attentions to better extract the structural feature in z and object feature in x, where h_{zx} acts as the query for point-wise features in h_z and h_x. The attention features h_z^{att} and h_x^{att} are then fused with h_{zx} as the overall point cloud feature h, which is later classified by the classification heads of existing models. In practice, we reuse pre-trained models and fine-tune the cross-attention and classification heads for only a few epochs. The design of the joint module is described in Fig.4. More design details can be found in the supplementary material.

4.3. Summary

As a recap, to develop a defense framework for various adversarial attacks, our work first models point cloud classification from humans' perspective and constructs the causal graph accordingly. To estimate the unobservable confounding variable U during causal effect identification, we propose Theorem 1 to find a solution. For the implementation of Eq.4, we propose three sub-modules to realize the causal inference with only slight modifications to existing point cloud classifiers.

5. Experiments

5.1. Experimental Settings

Models and datasets. To validate the effectiveness of our proposed method, we conduct experiments on three representative point cloud classification models: PointNet [32], DGCNN [45] and PointCNN [19] with two commonly used benchmark datasets: ShapeNet [4] and ModelNet40 [49], each consisting of CAD models belonging to 55 and 40 human-labeled object classes, respectively. The ShapeNet dataset has $35,708$ training samples and $15,429$ test samples, while the ModelNet40 dataset official split has $9,843$ training samples and $2,468$ test samples initially. For both datasets, we uniformly sample $2,048$ points from each object and normalize them into a unit sphere. All baselines are evaluated on the test set of each dataset. To improve efficiency, we follow the approach of previous work [46] and perform sub-sampling on the dataset. Specifically, we randomly select a subset of $2,732$ samples from the test set of ShapeNet, with at most 60 samples from each class.

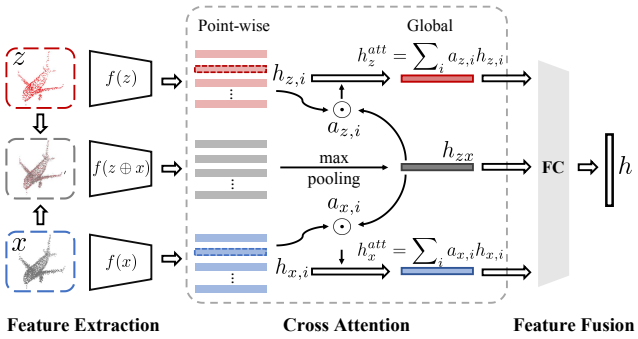

Figure 4. The attention module for modeling $P(Y|Z,X)$.

Baseline attacks. We consider ten representative adversarial attack methods in this paper, ranging from adding, shifting, deleting, and shape-invariant attacks, including Minimal [15], Smooth [25], IFGM [21], Gen3D-Add [51], Gen3D-Pert [51], AdvPC [11], Drop [62], KNN [40], GeoA3 [46], and ShapeInvariant [12]. Among these attacks, Gen3D-Add adds additional adversarial points to perform attacks, Drop removes existing points, AdvPC is based on autoencoders, and the others shift existing points in a benign point cloud. Various regulations are leveraged by these methods to perform stealthy attacks, e.g., KNN and GeoA3 use k-NN distance loss and local curvature loss respectively to perform shape-invariant attacks. Specifically, as GeoA3 requires normal vectors of a point cloud to perform, it is only evaluated on the ModelNet40 dataset. For a complete comparison, we consider both targeted and untargeted attack settings, i.e., an attacker either promotes the prediction of a specifically designated label or suppresses the prediction of the ground truth label. Note that most of the existing adversarial attacks are targeted ones. Therefore, we take IFGM, Minimal, Gen3D-Pert, and Drop as the untargeted attack baselines, where we modify the objective function accordingly to perform the specific type of attack.

Baseline defenses. As for defense baselines, we choose three effective input-oriented defenses: SOR [63], DUP-Net [63] and GvG [7], and two adversarial training (AT)-based defenses: AT [21] and PAGN [20]. SOR and DUP-Net preprocess the adversarial point cloud by filtering outliers to recover the benign one. GvG leverages the predicted gather

Table 1. Classification accuracy (%) of untargeted attack strategies against different defense methods on ModelNet40 dataset. The best among all defenses is in bold.

	PointNet							DGCNN						PointCNN					
	Vanilla	SOR	DUP-Net	GvG	AT	PAGN	Ours	Vanilla	SOR	DUP-Net	AT	PAGN	Ours	Vanilla	SOR	DUP-Net	AT	PAGN	Ours
IFGM	1.6	23.6	24.8	1.6	15.1	24.1	**49.0**	0.0	11.9	32.1	22.7	10.5	**75.9**	40.8	59.9	54.2	47.2	63.2	**68.9**
Minimal	32.5	63.0	61.4	32.1	59.6	61.8	**77.5**	12.6	37.9	35.9	59.0	41.0	**80.7**	66.6	67.8	40.3	57.7	68.8	**70.0**
Gen3D-Pert	**64.4**	63.4	63.0	**64.4**	48.1	63.5	**64.4**	34.6	35.2	37.1	28.6	34.9	**48.4**	48.9	51.2	37.1	42.3	45.4	**56.0**
Drop	65.0	69.0	67.8	60.5	36.3	69.5	**78.6**	69.6	71.6	49.2	59.1	71.2	**83.1**	**82.1**	77.6	42.7	72.2	73.7	71.7
Avg.	40.9	54.8	54.3	39.7	39.8	54.7	**67.4**	29.2	39.2	38.6	42.4	39.4	**72.0**	59.6	64.1	43.6	54.9	62.8	**66.6**

Table 2. Attack success rates (%) of targeted attack strategies against different defense methods on ModelNet40 dataset. The best among all defenses is in bold.

	PointNet							DGCNN						PointCNN					
	Vanilla	SOR	DUP-Net	GvG	AT	PAGN	Ours	Vanilla	SOR	DUP-Net	AT	PAGN	Ours	Vanilla	SOR	DUP-Net	AT	PAGN	Ours
Minimal	26.2	7.4	6.8	12.8	7.6	5.7	**1.5**	7.5	2.7	1.9	3.3	4.6	**0.6**	1.3	1.2	1.5	1.3	1.2	**0.8**
Smooth	48.7	4.7	3.8	42.5	29.6	7.1	**1.2**	81.7	23.4	2.8	75.0	23.4	**1.0**	3.3	2.2	2.1	2.8	1.5	**1.1**
IFGM	67.3	3.8	2.8	61.1	61.1	5.3	**0.8**	97.7	1.4	1.1	85.2	1.3	**0.5**	6.3	2.5	1.5	6.4	1.5	**1.1**
Gen3D-Add	60.4	5.7	4.9	49.1	56.4	5.9	**2.2**	38.9	3.0	2.1	33.3	4.2	**0.5**	0.9	0.9	1.8	0.8	0.9	**0.6**
Gen3D-Pert	98.4	7.1	5.1	77.8	**0.6**	3.8	1.1	96.8	4.9	1.7	89.1	5.0	**0.4**	57.4	23.1	4.6	5.6	3.1	**1.3**
AdvPC	99.0	5.2	4.3	99.7	100.0	6.4	**1.9**	84.2	3.3	1.7	82.5	4.9	**0.7**	5.2	2.6	2.4	**2.2**	2.8	**2.2**
KNN	90.6	23.1	15.8	97.1	76.3	31.6	**4.7**	98.4	12.9	2.5	98.7	19.1	**0.4**	55.1	27.4	5.5	12.1	10.0	**4.1**
GeoA3	100.0	19.1	15.8	100.0	100.0	30.7	**3.2**	97.1	7.6	2.2	100.0	14.1	**1.1**	16.8	11.5	4.7	6.2	11.1	**3.9**
ShapeInvariant	67.2	7.3	6.8	60.0	64.0	7.6	**3.6**	8.2	2.4	2.3	9.4	2.2	**1.5**	4.1	3.7	**2.2**	3.0	3.5	2.7
Avg.	73.1	9.3	7.3	66.7	55.1	11.6	**2.2**	67.8	6.8	2.0	64.1	8.8	**0.7**	16.7	8.3	2.9	4.5	4.0	**2.0**

vector for each point to recover the perturbed prediction. We apply GvG to PointNet only due to its model-specific design. The AT-based methods generate adversarial point clouds for model training. Specifically, we take IFGM adversarial examples for AT, and PAGN generates perturbations in the feature space.

Evaluation metrics. To evaluate the performance, we take classification accuracy (ACC) and attack success rate (ASR) as metrics for untargeted and targeted attacks. The ACC is the rate of examples that remain correctly classified, while the ASR stands for the rate of adversarial examples that are classified as the designated target labels. Therefore, a higher ACC or a lower ASR indicates a more robust classifier under specific attacks. For more details including the model training settings, hyper-parameter settings of baseline methods, the design of the attacks and more results, please refer to the supplementary material.

5.2. The Robustness of the Proposed Framework

We first show the effectiveness of CausalPC under both untargeted and targeted attack settings. The performance of our method and all baselines on ModelNet40 is shown in Table 1 and 2. The results of ShapeNet are presented in the supplementary material. We report the performance of the classifiers with no defense as the *Vanilla* columns. *Ours* columns stand for the classifiers with CausalPC. The *Avg.* row describes the averaged results of each column.

From the results, we observe that our CausalPC can substantially improve the adversarial robustness of current point cloud classifiers against various attacks. For example, we observe a remarkable increase in the average ACC of DGCNN from 29.2% to 72.0% for untargeted attacks, which greatly outperforms the current best baseline of 42.4%. Similarly, for targeted attacks, the average ASR of PointNet reduces drastically from 73.1% to 2.2%.

Current defense mechanisms often overfit to specific types of attacks. In the case of *AT-based defenses*, as reported in Table 2, it is observed that while AT achieves an ASR of 3.3% against the Minimal attack for DGCNN, due to the training regimen involving similar IFGM adversarial examples, it proves ineffective against novel unseen attacks such as KNN and GeoA3, exhibiting an ASR approaching 100%. Although PAGN demonstrates slightly improved performance owing to its adversarial training in the feature space, its robustness against shape-invariant attacks is still unsatisfying. *Input-oriented defenses* such as SOR and DUP-Net exhibit comparatively superior performance compared to AT-based ones. This is attributed to their consideration of a broader range of attack patterns, typically involving the removal of outlier points and the restoration of objects. The average ASR of both methods can be reduced to below 10%. However, DUP-Net performs poorly when facing the Drop attack which doesn't produce outliers, where the ACC is even lower than Vanilla for PointCNN against untargeted attacks. For shape-invariant attacks, whose attack goals are carefully designed to generate imperceptible adversarial examples, all input-oriented defenses tend to be breached. The ASR against GvG even reaches near 100% for KNN and GeoA3 attacks.

CausalPC outperforms these baselines by employing causal modeling in the context of point cloud classification tasks. When utilized for inference, CausalPC enables existing classifiers to accurately identify causal effects in classification, leading to enhanced adversarial robustness.

Table 3. Classification accuracy (%) of untargeted attacks against CausalPC with different modules on ModelNet40 dataset, DGCNN. The best among all baselines is in bold.

Method	$P(Z\|x)$	$P(X\|o)$	Attention	IFGM	Minimal	Gen3D-Pert	Drop	Avg.
Vanilla				0.0	12.6	34.6	69.6	29.2
no Att, x	✓			30.9	33.1	27.8	33.8	31.4
no Att, z		✓		37.9	45.8	38.1	51.1	43.2
no Att	✓	✓		72.9	78.6	43.0	**85.7**	70.1
Ours	✓	✓	✓	**75.9**	**80.7**	**48.4**	83.1	**72.0**

For instance, for Drop attack, the ACC exceeds 70%, and for shape-invariant attacks, the ASR consistently remains below 5%. In terms of overall performance, the average ASR for PointNet on targeted attacks is only 2.2%, compared to 7.3% for the best-performing baseline, DUP-Net.

5.3. Ablation Study

Subsequently, we investigate the relative impacts of the three modules within CausalPC. An ablation study is conducted on ModelNet40 using DGCNN, as presented in Table 3. This study assesses the robustness of CausalPC with the sub-modules analyzed individually.

The results indicate that different modules contribute diversely to robustness. For instance, when utilizing only structural information z (the *no Att, x* row), CausalPC exhibits only a marginal improvement over Vanilla. This is because z alone lacks detailed information essential for effective classification, albeit reducing susceptibility to adversarial perturbations. Similarly, employing only the reconstructed point cloud x (the *no Att, z* row) results in a modest enhancement of robustness. In the *no Att* row, it is demonstrated that through the causal modeling of both z and x, CausalPC achieves exceptional robustness, reflected in an ACC of 70.1%. Finally, the incorporation of the joint attention module in modeling $P(Y|Z, X)$ further elevates the robustness of our framework to an ACC of 72.0%.

5.4. Visualization

To further illustrate the difference between the extracted structure information z and the reconstructed point cloud x, we randomly pick an adversarial example generated by GeoA3 on PointNet in ModelNet40, whose ground truth label is *airplane*. We visualize an extracted z and a reconstructed x from the sample in Fig.5.

As shown in the figure, the contour of an airplane can be recognized from z at first glance, which demonstrates that the FPS algorithm is effective in extracting the structural information of a point cloud. However, upon closer inspection of details such as the zoomed-in wing part, we observe that z preserves only the framework of the original object and has lost the details. In contrast, the reconstructed x provides a denser point cloud with recovered details. By combining z and x as input, the joint classification module is able to leverage both the structural information and surface details

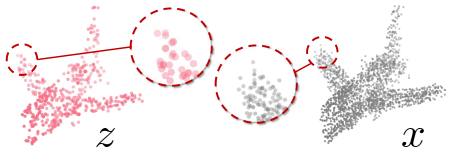

Figure 5. Visualization of z and x.

to achieve a more robust classification. For a more intuitive visualization of the adversarial robustness gained from joint classification, please refer to the supplementary material.

6. Conclusion & Limitations

In this study, driven by the pivotal observation that adversarial point clouds closely resemble benign ones to the human eye, we introduce a causal framework to enhance the robustness of point cloud classification. Specifically, we formulate the modeling of two crucial causal variables, the structural information Z and the hidden confounders U. This analysis reveals that the incompletely modeled U constitutes a limitation in existing defense strategies. Guided by the proposed causal modeling, we devise three effective submodules to identify the causal effect for robust classification, yielding significant performance improvements over existing baselines.

However, despite the guidance of the causal analysis, the sub-modules in actual implementation may induce inaccuracy inevitably. Such a flaw introduces a gap between the causal framework and the algorithm design. In our work, we have tried our best to utilize suitable modules for point cloud data and achieved a balance between effectiveness and efficiency. Future works should consider more advanced implementations and provide an analytical lower bound for the robustness.

For other future investigations, we aim to: (1) deepen our insights into causal relations by focusing on real-world objects; (2) extend the application of causal inference to other point cloud-related tasks, such as 3D object detection.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments that helped improve the quality of the paper. This work was supported in part by the National Key Research and Development Program (2021YFB3101200), National Natural Science Foundation of China (U1736208, U1836210, U1836213, 62172104,62172105, 61902374, 62102093, 62102091). Min Yang is a faculty of Shanghai Institute of Intelligent Electronics & Systems, Shanghai Insitute for Advanced Communication and Data Science, and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China. Mi Zhang is the corresponding author.

References

[1] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91 (434):444–455, 1996. 3

[2] Marshall W Bern and Paul E Plassmann. Mesh generation. *Handbook of computational geometry*, 38, 2000. 5

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 2, 3

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6

[5] Wenda Chu, Linyi Li, and Bo Li. Tpc: Transformation-specific smoothing for point cloud models. In *International Conference on Machine Learning*, pages 4035–4056. PMLR, 2022. 4

[6] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019. 4

[7] 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. IEEE, 2020. 1, 3, 4, 6, 2

[8] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 3, 5

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3, 2

[10] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness without the sensitive attribute via causal variational autoencoder. *arXiv preprint arXiv:2109.04999*, 2021. 3

[11] Abdullah Hamdi, Sara Rojas, Ali Thabet, and Bernard Ghanem. Advpc: Transferable adversarial perturbations on 3d point clouds. In *European Conference on Computer Vision*, pages 241–257. Springer, 2020. 1, 3, 6, 2

[12] 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. 2, 6

[13] John E Hummel. Object recognition. *Oxford handbook of cognitive psychology*, 810:32–46, 2013. 4

[14] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 3, 4

[15] 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. 1, 6, 2

[16] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017. 3

[17] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 2

[18] Linyi Li, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. *arXiv preprint arXiv:2009.04131*, 2020. 4

[19] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018. 2, 6

[20] Qi Liang, Qiang Li, Weizhi Nie, and An-An Liu. Pagn: perturbation adaption generation network for point cloud adversarial defense. *Multimedia Systems*, pages 1–9, 2022. 1, 6, 2

[21] Daniel Liu, Ronald Yu, and Hao Su. Extending adversarial attacks and defenses to deep 3d point cloud classifiers. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2279–2283. IEEE, 2019. 1, 3, 4, 6, 2

[22] Daniel Liu, Ronald Yu, and Hao Su. Adversarial shape perturbations on 3d point clouds. In *European Conference on Computer Vision*, pages 88–104. Springer, 2020. 2, 3

[23] Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Pointguard: Provably robust 3d point cloud classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6186–6195, 2021. 3, 4, 5

[24] Haoming Lu and Humphrey Shi. Deep learning for 3d point cloud understanding: a survey. *arXiv preprint arXiv:2009.08920*, 2020. 2

[25] Chengcheng Ma, Weiliang Meng, Baoyuan Wu, Shibiao Xu, and Xiaopeng Zhang. Towards effective adversarial attack against 3d point cloud classification. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 6, 2

[26] 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. 3

[27] Slava Mikhaylov, Michael Laver, and Kenneth R Benoit. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, 20(1):78–91, 2012. 4

[28] Ronghui Mu, Wenjie Ruan, Leandro S. Marcolino, and Qiang Ni. 3DVerifier: Efficient robustness verification for 3D point cloud models. *Machine Learning*, pages 1–28, 2022. 4

[29] Judea Pearl. *Causality*. Cambridge university press, 2009. 2, 3, 4, 1

[30] Juan C. Pérez, Motasem Alfarra, Silvio Giancola, and Bernard Ghanem. 3deformrs: Certifying spatial deformations on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15169–15179, 2022. 4

[31] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. 3

[32] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1, 2, 6

[33] 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. 1, 2, 5

[34] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1025–1034, 2021. 3

[35] Qibing Ren, Yiting Chen, Yichuan Mo, Qitian Wu, and Junchi Yan. DICE: Domain-attack Invariant Causal Learning for Improved Data Privacy Protection and Adversarial Robustness. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1483–1492, 2022. 3

[36] Radu Bogdan Rusu, Zoltan Csaba Marton, Nico Blodow, Mihai Dolha, and Michael Beetz. Towards 3d point cloud based object maps for household environments. *Robotics and Autonomous Systems*, 56(11):927–941, 2008. 2

[37] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 4

[38] Hongda Sun, Shufang Xie, Shuqi Li, Yuhan Chen, Ji-Rong Wen, and Rui Yan. Debiased, Longitudinal and Coordinated Drug Recommendation through Multi-Visit Clinic Records. *Advances in Neural Information Processing Systems*, 35: 27837–27849, 2022. 6, 1

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. 3

[40] 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*, pages 954–962, 2020. 6, 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[42] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021. 3

[43] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018. 4

[44] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021. 3

[45] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 1, 2, 6

[46] 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*, 2020. 2, 3, 4, 6

[47] 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. 3, 4

[48] Yutian Wu, Yueyu Wang, Shuwei Zhang, and Harutoshi Ogai. Deep 3d object detection networks using lidar data: A review. *IEEE Sensors Journal*, 21(2):1152–1171, 2020. 3

[49] 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. 3, 5, 6

[50] 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. 4

[51] 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. 1, 2, 3, 6

[52] Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Generative voxelnet: learning energy-based models for 3d shape synthesis and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[53] Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Kui Yu. Causal Effect Estimation with Variational AutoEncoder and the Front Door Criterion. *arXiv preprint arXiv:2304.11969*, 2023. 1

[54] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857, 2021. 6, 1

[55] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2790–2799, 2018. 2

[56] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2790–2799, 2018. 5, 1

[57] Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Transporting causal mechanisms for unsupervised

domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8599–8608, 2021. 3

[58] Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. *Advances in Neural Information Processing Systems*, 33:289–301, 2020. 4

[59] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In *NeurIPS*, 2019. 3

[60] Jinghuai Zhang, Jinyuan Jia, Hongbin Liu, and Neil Zhenqiang Gong. PointCert: Point Cloud Classification with Deterministic Certified Robustness Guarantees. *arXiv preprint arXiv:2303.01959*, 2023. 4

[61] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Adversarial robustness through the lens of causality. *arXiv preprint arXiv:2106.06196*, 2021. 3

[62] 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. 1, 3, 6, 2

[63] Hang Zhou, Kejiang Chen, Weiming Zhang, Han Fang, Wenbo Zhou, and Nenghai Yu. Dup-net: 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. 1, 3, 4, 6, 2

[64] 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. 1, 3

[65] Xinyuan Zhu, Yang Zhang, Fuli Feng, Xun Yang, Dingxian Wang, and Xiangnan He. Mitigating hidden confounding effects for causal recommendation. *arXiv preprint arXiv:2205.07499*, 2022. 1