CAP: Robust Point Cloud Classification via Semantic and Structural Modeling

Daizong Ding, Erling Jiang, Yuanmin Huang, Mi Zhang, Wenxuan Li, Min Yang*
School of Computer Science, Fudan University, China
{17110240010@,eljiang21@m.,yuanminhuang21@m.,mi_zhang@,22210240091@m.,m_yang@}fudan.edu.cn

Abstract

Recently, deep neural networks have shown great success on 3D point cloud classification tasks, which simultaneously raises the concern of adversarial attacks that cause severe damage to real-world applications. Moreover, defending against adversarial examples in point cloud data is extremely difficult due to the emergence of various attack strategies. In this work, with the insight of the fact that the adversarial examples in this task still preserve the same semantic and structural information as the original input, we design a novel defense framework for improving the robustness of existing classification models, which consists of two main modules: the attention-based pooling and the dynamic contrastive learning. In addition, we also develop an algorithm to theoretically certify the robustness of the proposed framework. Extensive empirical results on two datasets and three classification models show the robustness of our approach against various attacks, e.g., the averaged attack success rate of PointNet decreases from 70.2% *to* 2.7% *on the ModelNet40 dataset under* 9 *common attacks.*

1. Introduction

With the rapid development of 3D sensors such as LiDAR used in autonomous vehicles, point cloud data, which represents real-world objects by a set of 3D coordinates of points, has been widely applied in various 3D vision applications [30]. Powered by the deep and non-linear structures, a number of deep learning models have proved to be effective in modeling the geometric pattern underlying point cloud data, such as multi-layer perceptron (MLP) [36], convolutional neural network (CNN) [24] and graph neural network (GNN) [47]. Despite the effectiveness, the extensive usage of DNN also raises the concern of *adversarial examples*, where the input point clouds are slightly manipulated by an adversary to cause the misbehavior of a model [22, 48, 52]. Considering its severe consequences and damage to real-world applications, the study of adversarial examples on point cloud data has been attracting more and more attention from both industry and academia.

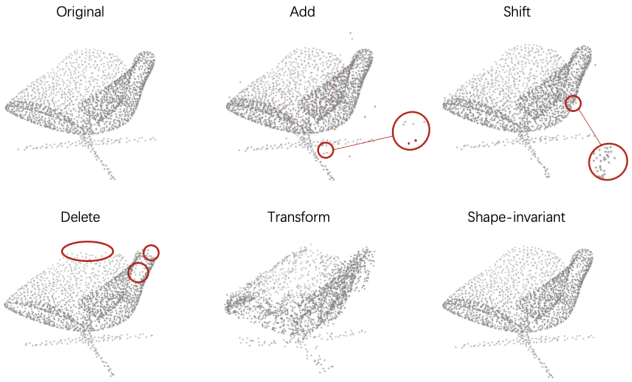

Figure 1. The demonstration of different adversarial attack strategies in point cloud classification.

Owing to the unique data format of point cloud, i.e., a set of 3D coordinates, the design of adversarial attacks varies in multiple aspects [30]. From the view of perturbations, the adversaries could shift existing points to create adversarial examples [52], which is similar to the adversarial attacks in images [13]. Besides, the adversaries could also delete [49, 63] or add points [28, 52] to conduct the attack. Recent studies show that the generative model, i.e., transforming the original point cloud into a new one [15, 64], is also effective to find adversarial examples. From the view of restrictions, the constraints of the perturbations may differ in different approaches, e.g., limiting the number of altered points [18, 22], restricting the maximal/averaged distance of shifted points [40] and constraining the shape similarity between the adversarial examples and original ones [52].

Recently, many efforts have been made to mitigate potential adversarial examples in point cloud data [26, 65], which mainly fall into two categories,

- **Adversarial Training-based (AT):** this line of research takes inspiration from the work in the image domain [32], which proposes to pair adversarial examples with correct labels and put them into the training set [26, 40]. In the context of point cloud data, the main drawback of AT-

*Corresponding authors: Mi Zhang and Min Yang

based methods is that they are only robust to certain kinds of seen adversarial examples [61]. For instance, when the adversaries leverage an attack method that differs from the methods used in AT, the attack success rate could raise from 0.6% to 100.0% [1]. Considering the diverse attack strategies in this task, it is difficult to find adversarial examples for AT-based methods that could generalize to various kinds of attacks.

- **Recovery-based:** this line of research reveals that adversarial examples in point cloud data often contain outlier points [52]. Based on this, recovery-based methods propose to restore a clean sample from an adversarial example before feeding it to the classification model. For instance, SOR utilizes a rule-based strategy to filter outlier points [65], while DUP-Net [65] leverages a deep generative network to recover the samples better [57]. Different from AT-based methods, the recovery-based methods do not focus on certain kinds of attack strategies, however, they could be evaded by *shape-invariant attacks* [42, 48] which take the geometric pattern of the perturbations into account, e.g., generating points that smoothly lie on the surface of the object [48], to make the recovery less effective.

Our Work. Despite the variety of attack strategies, the semantic and structural information of different adversarial examples can hardly change [18, 22, 48, 52, 63]. For instance, an adversarial example of a car should still look like a car no matter how adversaries choose the attack strategy, e.g., adding, deleting, shifting or transforming. However, previous works point out that existing classifiers often pay attention to limited segments or local features of the whole object to conduct the prediction, leading to the potential risk of different adversarial attacks [49, 50, 63]. This motivates us to improve the robustness of existing classifiers by enhancing the modeling of semantic and structural information. Based on this, we develop a novel defense framework called contrastive and attentional point cloud learning (CAP), which is mainly composed of two modules: (1) the attention-based feature pooling and (2) the dynamic contrastive learning paradigm. The first module aims to capture the global structural information of the object by recognizing critical points among the point cloud data. To this end, we design a multi-head attention layer to assign critical points with higher weights, which will be used for obtaining the global representation of the input. We also introduce the temperature coefficient and random sampling techniques to prevent the module from focusing on a few fixed segments. The second module aims to characterize the semantic information of different objects by disentangling the features of objects with different labels while gathering those with the same label. To this end, we design an interesting dynamic contrastive learning paradigm, which divides the learning goal into a coarse-to-fine process and helps the learning better converge.

With the aid of the proposed CAP, we can significantly improve the robustness against various adversarial examples for existing classification models such as PointNet/PointNet++ [37], DGCNN [47] and PointCNN [24]. Furthermore, we show that the robustness of CAP is theoretically certified. Specifically, given a certain constraint of the perturbations, we could evaluate whether the trained model is robust under arbitrary attack strategies, e.g., if an adversary would be able to add perturbations to a *chair* to obtain a prediction of a *car*. To this end, we first measure the changes in features after adding perturbations based on the manifold learning theory. Then we leverage the extreme value theory to estimate the upper bound of the potential changes, which indicates the optimal attack that aims to move the adversarial example across the decision boundary. After that, the robustness of the model can be measured based on the estimated upper bound. With the proposed certified defense, a user could estimate the potential risk of adversarial attacks in real-world applications before deployment. We validate the proposed CAP on two benchmark datasets and seven attack methods. In summary, the main contributions of this work are:

- We propose a novel and general solution for improving the robustness of existing point cloud classification models by modeling the semantic and structural information, which is able to train robust models against various kinds of adversarial attacks.
- We present an algorithm to theoretically certify the robustness of the proposed framework. With the aid of the manifold learning and the extreme value theory, the estimated robustness is highly consistent with the actual empirical results, i.e., attack success rate.
- Extensive experiments on two benchmark datasets show that our CAP can significantly improve the robustness of different classification models (PointNet/PointNet++, DGCNN and PointCNN), e.g., the attack success rate of PointNet decreases from 70.2% to 2.7% on average on the ModelNet40 dataset.

2. Background and Related Work

In this section, we first formulate the task of point cloud classification and common modules used in recently proposed high-performance models. Then we summarize the concept of adversarial examples in this task.

2.1. Point Cloud Classification

In this work, we focus on the 3D point cloud classification task, i.e., assigning an expected label to a 3D object

[1] For detailed experimental setting and results please refer to Sec. 5

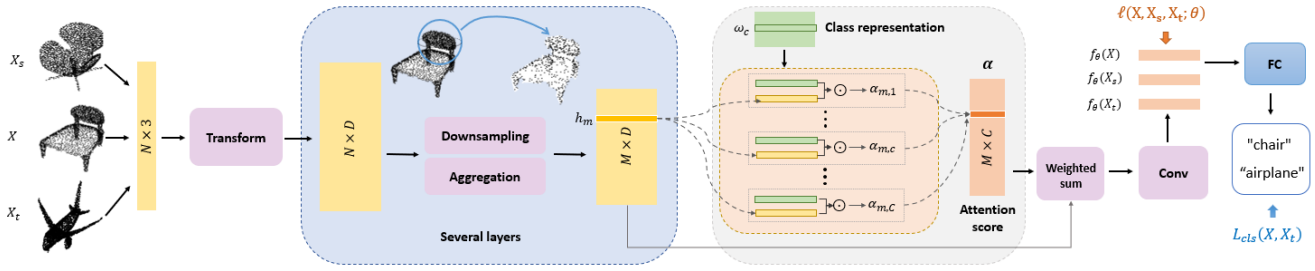

Figure 2. The demonstration of the proposed defense framework CAP.

represented by a collection of points. Formally, given an input $X \in \mathbb{R}^{N\times 3}$ consisting of N points with 3D coordinates and a label $y \in \{1, \cdots, C\}$ representing the corresponding category, the classification model can be denoted as $F : X \to \{1, \cdots, C\}$. Numerous deep-learning models have been proposed in recent years for point cloud classification [36, 38, 53]. In this work, we mainly consider the point-based models [62, 66] owing to their effectiveness in characterizing geometry patterns [8, 58], which directly receive the 3D coordinates as the model input instead of preprocessed point cloud data [36, 37, 47].

There are three key modules in point-based DNN classification models: feature transformation, neighborhood aggregation and feature pooling. Firstly, the models transform the input point cloud $X \in \mathbb{R}^{N\times 3}$ to N latent factors through a non-linear feature extraction operation, e.g., MLP used in PointNet/PointNet++ [36, 37], CNN in PointCNN [24], and the graph neural network in DGCNN [47]. Then the neighborhood aggregation module is designed to aggregate the features of nearby points given a reference point to characterize its local pattern [24, 29, 37, 47]. Before the output layer, max-pooling over the features of points is performed, whose output vector is regarded as the global representation of the point cloud, which is fed into a fully-connected layer for final classification. For a more comprehensive study please refer to [30].

2.2. Adversarial Examples

Despite the effectiveness of these approaches, the deep and non-linear structure also brings up the concern of *adversarial examples*, where the adversaries slightly perturb the input point cloud to cause the model to predict wrong labels [15, 48, 52]. The attack goal can be formed by,

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

where $\tilde{X} = X \oplus \eta$ and $\tilde{y}$ are the perturbed input and the target label, respectively, the operation $\oplus$ could be adding [28, 52], deleting [49, 63], or shifting points [22, 31, 48, 52] on the original inputs. The term $\|\eta\|_p$ denotes the L_p constraint of the perturbation, e.g., L_0 and L_∞ represent the number of points perturbed [18, 22] and the maximal value of the perturbation [40] should be smaller than δ, respectively, while L_2 represents the minimal transformation cost from one point cloud to another, e.g., the Chamfer Distance [52]. Some work also proposes to impose additional shape-invariant regularizations, e.g., the surface of the generated adversarial examples should be smooth [48]. Owing to the unique form of point cloud data, i.e., a set of coordinates, the design of adversarial examples shows strong diversity in this task.

2.3. Defending Against Adversarial Examples

Recently, several defense strategies have been proposed against adversarial attacks in this task, which roughly fall into two categories: adversarial training [27, 61] such as PAGN [26], ART-Point [46] and PointCutMix [60], and recovery-based approaches such as DUP-Net [65] and LPF-Defense [33]. Adversarial training proposes to put adversarial examples labeled correctly into the training data, which naturally lacks the ability to generalize to various attack strategies as we have discussed. On the other hand, the recovery-based approaches could not deal with shape-invariant attacks such as KNN [42] and GeoA^3 [48] effectively. In this work, we develop a novel solution to defend against various kinds of attack strategies in point cloud classification.

3. Methodology

In this section, we present the proposed defense framework for point cloud classification.

3.1. Problem Analysis

Though the design of adversarial attacks varies from different attack strategies, the perturbed point cloud retains similar structural and semantic information as the original one due to the constraints on the perturbations (Eq. 1), i.e., objects containing noise still look like the original ones in human perception. Despite the numerous efforts made in developing more effective point cloud classifications [24, 47, 58], recent studies reveal that existing classification models often pay attention to trivial parts of the objects to make predictions, e.g., limited critical sets [49],

small segments [63] and local point distributions [50]. As such, small perturbations could make the model predict wrong labels. To improve the robustness of existing classification models, we propose to enhance the modeling of structural and semantic information for different models. To this end, we develop a novel defense framework called CAP which mainly composes of two parts: attention-based pooling and dynamic contrastive learning.

3.2. Attention-based Feature Pooling

The motivation of the first module is to help the model recognize critical points that describe an object. To this end, we develop a multi-head attention layer to characterize the structural information, which learns to assign weights for points when we aggregate their features. Considering that the number of points is often large, e.g., $N = 2048$ [36], it would be difficult to conduct attention operations over all points. To address this issue, we propose to leverage the downsampling technique [24, 37, 54], which downsamples points by certain strategies, e.g., random sampling [24] and farthest point sampling [37]. Formally, after several layers of downsampling and neighborhood aggregation (Fig. 2), the features of points at the last layer could be represented as $h \in \mathbb{R}^{N' \times D}$, where the N' and D are the number of remaining points and the dimension of latent vectors respectively. Then we compute the following attention score,

$$\alpha_{i,c} = \frac{\exp(h_i^T \omega_c / \tau)}{\sum_{c'} \exp(h_i^T \omega_{c'} / \tau)}, \tag{2}$$

where $\omega_c \in \mathbb{R}^D$ is the query vector, which is a learnable hidden representation of class c, $\alpha_{i,c}$ is the attention score between the i-th key point and the label c, and τ is the temperature coefficient that prevents the model from focusing on only several points. Given the latent vector h_i of point i, we could evaluate its correlation corresponding to class c, i.e. $\alpha_{i,c}$, based on which we could obtain the global representation of the input by,

$$\hat{h}_c = \sum_{i=1}^{N} \alpha_{i,c} h_i, \quad z = \text{conv}([\hat{h}_1, \cdots, \hat{h}_C], W^h), \tag{3}$$

The feature $\hat{h}_c \in \mathbb{R}^D$ can be interpreted as the latent vector of a point cloud under the view of class c. Then we leverage a normal convolution operation on all $\hat{h}_c$ to obtain the global representation $z \in \mathbb{R}^D$, where $W^h \in \mathbb{R}^{C \times 1}$ is the parameter of the convolution layer. The global feature z will then be fed to the classification layer to predict the label of the object. After optimizing the parameters $\{w, W^h\}$, the proposed attention module could help the model to figure out which points play important roles for different kinds of objects. With the aid of the temperature coefficient, the attention module tends to focus on more points and outlines the sketch of the object, i.e., the structural information.

3.3. Dynamic Contrastive Learning

The second module aims to help the model distinguish the features from different classes of objects. To this end, we take inspiration from the *contrastive learning* [5, 16, 44] and develop the following triplet loss. Specifically, given a sample X with label s,

$$\ell(X, X_s, X_t; \theta) = \max(d(X, X_s) - d(X, X_t) + \epsilon, 0), \tag{4}$$

where X_s and X_t denote two other samples from label s and t respectively ($t \neq s$), $d(\cdot, \cdot)$ is the Euclidean distance between the global features of samples, and ϵ denotes a predefined threshold. By minimizing the triplet loss, the model will distinguish the features of samples with different labels while gathering those with the same label. This mechanism helps the model find general characteristics of similar objects, i.e., the semantic information. Based on this, we propose the following loss to better separate the features for training,

$$L_{\text{margin}}(X; \theta) = \max_{X_s, X_t \in B(X)} \ell(X, X_s, X_t; \theta), \tag{5}$$

where $B(X)$ denotes the training batch. Given X with label s, the batch is constructed by sampling several samples X_s and X_t with label s and t respectively. However, directly optimizing the maximum loss in Eq. 5 is too difficult at the early stage of the model training. As a result, the model can hardly converge in practice. To address the issue, we design the following learning paradigm,

$$\min_\theta L_{\text{all}}(\theta; X) + \exp[\epsilon - \bar{L}_{\text{all}}(\theta; X)] \cdot L_{\text{margin}}(\theta; X), \tag{6}$$

where L_{all} is the averaged triplet losses (Eq. 4 for all pairs (X_s, X_t) in the training batch), and $\bar{L}$ represents the value of the loss function with no backward gradient. At the early stage of the learning, the model minimizes the averaged triplet loss, i.e., the averaged distances of samples from two labels. When the expected triplet loss tends to converge, the model turns to focus on the borderline between features of samples from two labels and only minimizes the triplet loss for samples near the borderline. With the aid of the proposed coarse-to-fine process, the model will better characterize the semantic information of different classes of objects.

3.4. Summary

We now show how to incorporate the proposed two modules into the training of various classification models. Given a vanilla classification model which transforms all 3D coordinates to feature vectors, we first perform downsampling on all features for models that don't have this module, e.g., DGCNN, then we replace the existing max pooling operation with the proposed attention module. During the training, we combine Eq. 6 with the common classification loss

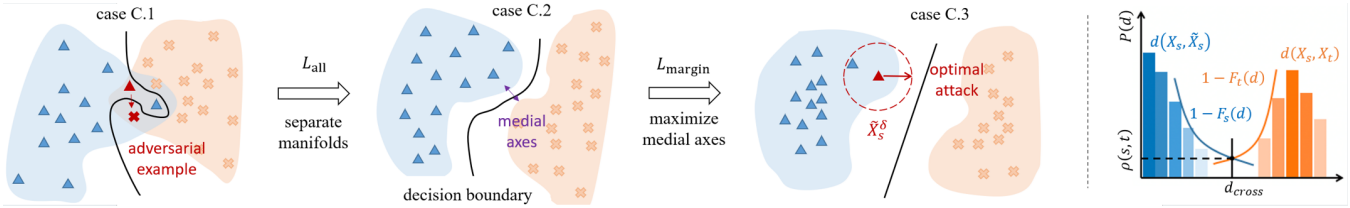

Figure 3. The demonstration of the proposed contrastive learning (left) and the estimation of $\rho(s,t)$ (right).

to jointly train the model. The overall framework is depicted in Fig. 2. For more implementation details of the design please refer to Appendix B.

4. Certified Robustness

In this section, we develop an algorithm to show that the robustness of the proposed framework can be theoretically certified. To this end, we follow the previous study on certified robustness [23] and try to answer the following research question: will the proposed model be robust against adversarial examples under certain constraints?

4.1. Manifold Learning

To answer the question above, we introduce the manifold learning, which assumes that the data samples in high-dimensional space often lie on low-dimensional manifolds. It plays an important role in explaining the behavior of deep learning models in various applications [10, 43]. Specifically, suppose the manifold $\mathcal{M}^s \subset \mathbb{R}^{N\times 3}$ consists of a set of samples with label s, and the manifold of all samples can be denoted by $\mathcal{M} = \cup_s \mathcal{M}^s$. Then the actual distance between two samples $X_i, X_j \in \mathcal{M}$ should be measured by the metric on the manifold instead of the Euclidean distance between inputs [39, 59], while deep neural networks have been found to have superior performance in modeling the metric on the manifold [6, 55]. Formally, suppose a classification model can be divided into two consecutive parts, i.e., the feature extractor f_θ and the classification layer g_ϕ, respectively. Then given two samples on the manifold $X_i, X_j \in \mathcal{M}$, the distance between them can be calculated as $d(X_i, X_j) = \|z_i - z_j\|_2$, where $z_i = f_\theta(X_i), z_j = f_\theta(X_j), z_i, z_j \in \mathbb{R}^D$ are the feature vectors learned by the feature extractor. As such, we could denote the manifold learned by a classification model as $(\mathcal{M}, f_\theta)$.

Since different models with different model structures or learning algorithms extract features from different perspectives, the learned manifold $(\mathcal{M}, f_\theta)$ can be quite different, leading to various prediction behaviors of the classification models, i.e., decision boundaries [2, 11, 56]. For instance, if two manifolds $\mathcal{M}^s$ and $\mathcal{M}^t$ are highly entangled with each other in the feature space given a feature extractor f_θ, there exist samples close to the decision boundary [17, 19–21]. Thus, an adversary can easily find small perturbations that make an input sample cross the decision boundary, i.e., generating an adversarial example [1, 14, 34, 35].

4.2. Robustness Certification

Based on the above definition, to certify the robustness, we need to study whether the optimal attack could move one sample from manifold $\mathcal{M}^s$ to $\mathcal{M}^t$ with perturbations (Fig. 5). To this end, we propose to observe the following two distances,

$$\begin{aligned} m(\mathcal{M}^s; \delta) &= \sup_{X_s \in \mathcal{M}, \|\eta\|_p \le \delta} d(X_s, \tilde{X}_s) \\ r(\mathcal{M}^s, \mathcal{M}^t) &= \inf_{X_i \in \mathcal{M}^s, X_j \in \mathcal{M}^t} d(X_i, X_j), \end{aligned} \tag{7}$$

where $\tilde{X}_s = X_s \oplus \eta$. The term $m(\mathcal{M}^s; \delta)$ measures the maximal bias of features under certain constraints on perturbations, while the medial axes $r(\mathcal{M}^s, \mathcal{M}^t)$ describes the minimal distance between two manifolds. As such, if the medial axes $r(\mathcal{M}^s, \mathcal{M}^t)$ is much larger than $m(\mathcal{M}^s; \delta)$, we could conclude that the model is robust under the constraints $\|\eta\|_p \le \delta$, i.e., the case C.3 in Fig. 5. Otherwise, there may exist case C.1 or C.2, where the optimal attack could find a sample near the decision boundary and generate the perturbations to force the model to mispredict.

Nevertheless, since the attack strategy is unseen during the certification, we should enumerate all potential η to compute the distances, which is time-consuming and impractical. To tackle the issue, we introduce the extreme value theory, which has been widely used for forecasting the potential maximal values given several observations [7, 12]. Practically, we randomly sample $X_s \in \mathcal{M}^s$ and perturbations η to compute the distances $d(X_s, \tilde{X}_s)$. To generate $\tilde{X}_s$ for point cloud data, we consider three kinds of perturbations: adding, deleting and shifting existing points. For details of generating the perturbations please refer to Appendix B. Then we could estimate the cumulative density function (CDF) of $d(X_s, \tilde{X}_s)$ by,

$$P(d_s > d) \approx \frac{v}{V} \cdot \left[1 + \gamma_s \cdot \frac{d_s^{(v)} - d}{\beta_s}\right]^{-1/\gamma_s}, \quad d > d_s^{(v)}, \tag{8}$$

where $d_s^{(v)}$ is the v^{th}-largest distance among sampled $d(X_s, \tilde{X}_s)$, V is the sampling size, and γ_s, β_s are the shape and scale parameters of the distribution. For the derivation of the CDF and the inference of the parameters, please refer to Appendix A. Intuitively, as shown in Fig. 5, the value

Table 1. Attack success rate of ModelNet40 normal, adversarial and recovered samples under various attack and defense methods on three classification models w/ and w/o our CAP.

	PointNet									DGCNN								PointCNN							
	Vanilla	SOR	DUP-Net	AT	AT-PGD	EAT	PAGN	GvG	Ours	Vanilla	SOR	DUP-Net	AT	AT-PGD	EAT	PAGN	Ours	Vanilla	SOR	DUP-Net	AT	AT-PGD	EAT	PAGN	Ours
Minimal	26.2%	7.4%	6.8%	7.6%	10.6%	10.0%	5.7%	10.6%	1.8%	7.5%	2.7%	1.9%	3.3%	1.5%	9.7%	4.6%	1.6%	1.3%	1.2%	1.5%	1.3%	0.8%	1.7%	1.2%	0.4%
Smooth	48.7%	4.7%	3.8%	29.6%	36.6%	33.3%	7.1%	36.6%	2.3%	81.7%	23.4%	2.8%	75.0%	35.7%	77.7%	23.4%	3.3%	3.3%	2.2%	2.1%	2.8%	2.5%	14.5%	1.5%	0.9%
IFGM	67.3%	3.8%	2.8%	61.1%	64.8%	42.9%	5.3%	64.8%	2.5%	97.7%	1.4%	1.1%	85.2%	95.4%	67.4%	1.3%	2.2%	6.3%	2.5%	1.5%	6.4%	5.8%	17.9%	1.5%	0.7%
PGD	73.0%	4.9%	3.8%	56.2%	53.0%	45.3%	6.5%	53.0%	1.1%	62.1%	4.0%	2.4%	74.0%	59.0%	89.0%	6.1%	0.9%	4.8%	3.4%	2.3%	2.4%	1.8%	8.9%	1.1%	0.2%
Gen3D-Add	60.4%	5.7%	4.9%	56.4%	56.1%	32.0%	5.9%	56.1%	1.6%	38.9%	3.0%	2.1%	33.3%	8.0%	30.2%	4.2%	2.8%	0.9%	0.9%	1.8%	0.8%	0.7%	1.3%	0.9%	0.4%
Gen3D-Pert	98.4%	7.1%	5.1%	0.6%	70.2%	0.5%	3.8%	70.2%	1.5%	96.8%	4.9%	1.7%	89.1%	81.6%	0.9%	5.0%	3.3%	57.4%	23.1%	4.6%	5.6%	5.1%	0.7%	3.1%	1.8%
KNN	90.6%	23.1%	15.8%	76.3%	99.3%	62.7%	31.6%	99.3%	4.0%	98.4%	12.9%	2.5%	98.7%	96.3%	98.2%	19.1%	6.4%	55.1%	27.4%	5.5%	12.1%	13.9%	36.8%	10.0%	3.3%
GeoA3	100.0%	19.1%	15.8%	100.0%	100.0%	100.0%	30.7%	100.0%	5.0%	97.1%	7.6%	2.2%	100.0%	98.9%	100.0%	14.1%	7.6%	16.8%	11.5%	4.7%	6.2%	4.4%	9.3%	11.1%	5.3%
SI	67.2%	7.3%	6.8%	64.0%	65.4%	48.7%	7.6%	65.4%	4.4%	8.2%	2.4%	2.3%	9.4%	14.2%	31.4%	2.2%	3.2%	4.1%	3.7%	2.2%	3.0%	2.6%	3.0%	3.5%	5.0%
Avg.	70.2%	9.2%	7.3%	50.2%	61.8%	41.7%	11.6%	63.1%	2.7%	65.4%	6.9%	2.1%	63.1%	54.5%	56.1%	8.9%	3.5%	16.7%	8.4%	2.9%	4.5%	4.2%	10.5%	3.8%	2.0%

Table 2. Mean test accuracy on ModelNet40 and ShapeNet.

		PointNet	DGCNN	PointCNN
ModelNet40	Vanilla	86.2%	88.9%	86.8%
	CAP	86.8%	90.1%	87.9%
ShapeNet	Vanilla	78.6%	80.5%	77.9%
	CAP	77.1%	80.5%	79.4%

$P(d_s > d)$ stands for the probability of $d(X_s, \tilde{X}_s) > d$. Similarly, we can randomly sample pairs of $X_s \in \mathcal{M}^s$, $X_t \in \mathcal{M}^t$ and compute the distances $d(X_s, X_t)$ for estimating the probability of $P(d_t < d)$. After that, we estimate the intersection of the two CDFs, i.e., d_{cross}. If $\rho(s,t) = P(d_s > d_{\text{cross}}) = P(d_t < d_{\text{cross}})$ is small enough, the condition $m(\mathcal{M}^s; \delta) < r(\tilde{M}^s, \tilde{M}^t)$ will be held for a large probability, which means that the model is robust under the constraint $\|\eta\|_p \leq \delta$ between label s and t, and vice versa. We empirically validate this claim in Sec. 5.3. In real-world applications, a user of a classification model could leverage the proposed algorithm to measure the potential risk of adversarial attacks before deployment, i.e., the robustness certification.

5. Empirical Results

Experimental Setting. We validate the effectiveness of our defense framework on three point-based classification models: the MLP-based PointNet/PointNet++ [36, 37], the convolution-based PointCNN [24] and the graph-based DGCNN [47]. The experiments are conducted on two commonly used benchmark datasets: ModelNet40 [51] and ShapeNet [4]. For both datasets, we uniformly sample 2048 points for each point cloud and normalize them into a unit sphere. The mean accuracy of the models above on both datasets is demonstrated in Table 2. For the attacks, we consider the following nine representative methods: Minimal [22], Smooth [31], IFGM [27], PGD [32], Gen3D-Add [52], Gen3D-Pert [52], KNN [42], GeoA3 [48], and ShapeInvariant (SI) [18]. These attacks cover adding and shifting points with various distance regularizations on the perturbations, including L_2 Distance [41], Chamfer Distance [52], and other shape-invariant related regularizations [18,48]. As for comparison, we employ two recovery-based defense approaches, SOR and DUP-Net [65], and four adversarial training-based approaches, Vanilla AT based on FGSM (AT) and PGD (AT-PGD) [31], Ensemble AT (EAT) and PAGN [26]. Specifically, EAT leverages adversarial examples generated by multiple attack methods for training. In addition, Gather-vector Guidance (GvG) [9] is considered for PointNet particularly. We consider both targeted attacks and untargeted attacks when generating adversarial examples. For the targeted attacks, the attack effectiveness is estimated by the attack success rate (ASR), i.e., given a sample, whether the model predicts our pre-assigned label. For the untargeted attacks, we leverage the accuracy to estimate the robustness. For more details, including the dataset statistics, the implementation of the models and baseline methods, the design of the attack and the hyperparameters, please refer to Appendix B.

5.1. The Robustness of the Proposed Framework

Table 1 and 3 illustrate the ASR of three classification models trained with and without the proposed CAP under various attack and defense methods on ModelNet40 and ShapeNet, respectively. Note that the *vanilla* column represents the results of vanilla models without any defense. The *CAP* column represents the results of models trained with our framework. The Avg. row describes the average results of all rows above. For more results on the untargeted attack please refer to Appendix C.

From the empirical results, we observe that training with CAP can greatly improve the robustness of vanilla models against all kinds of attacks, e.g., the average ASR drops from 70.2% to 2.7% on PointNet for ModelNet40 and from 54.4% to 1.3% for ShapeNet. Besides, there are several important findings. First, the adversarial training-based approaches often fail for certain kinds of attack strategies. For instance, although ASR is below 1% for PointNet under the Gen3D-Pert attack for ModelNet40, the ASR raises to 100% under the GeoA3 attack. The root reason lies in that the adversarial training could not generalize well to different kinds of attack strategies as we have discussed. As a comparison, the ASRs are all below 5% with our defense strategy for PointNet.

Second, as for the performance of existing recovery-based defense methods, we observe that SOR and DUP-Net can filter out most of the malicious perturbations generated by attack methods under the constraint of common distance regularizations, e.g. Minimal, Smooth, IFGM, Gen3D-Add and Gen3D-Pert, where the ASR can be restricted to relatively low on all models. Despite their effectiveness on

Table 3. Attack success rate of ShapeNet normal, adversarial and recovered samples under various attack and defense methods on three classification models w/ and w/o our CAP.

	PointNet							DGCNN							PointCNN						
	Vanilla	SOR	DUP-Net	AT	EAT	PAGN	CAP	Vanilla	SOR	DUP-Net	AT	EAT	PAGN	CAP	Vanilla	SOR	DUP-Net	AT	EAT	PAGN	CAP
Minimal	29.6%	6.0%	5.6%	6.7%	11.0%	2.8%	1.2%	8.6%	1.8%	1.0%	2.4%	2.7%	2.1%	1.4%	0.5%	0.6%	0.7%	0.9%	1.3%	0.9%	0.5%
Smooth	15.6%	5.0%	5.1%	30.1%	40.5%	3.8%	1.1%	6.8%	2.1%	1.7%	62.3%	71.4%	11.5%	0.7%	1.1%	0.7%	1.2%	2.6%	3.9%	1.1%	0.5%
IFGM	61.2%	4.1%	3.2%	57.6%	72.0%	2.2%	1.2%	85.4%	1.4%	0.6%	95.0%	98.9%	1.0%	3.9%	2.8%	1.4%	1.7%	3.7%	5.0%	1.5%	1.5%
Gen3D-Add	41.8%	4.5%	4.3%	34.5%	47.7%	2.8%	0.9%	13.1%	1.5%	1.1%	10.5%	18.9%	2.5%	2.9%	0.7%	0.6%	1.1%	1.3%	1.6%	0.7%	0.7%
Gen3D-Pert	97.2%	5.3%	4.0%	62.2%	67.0%	2.5%	0.8%	87.4%	3.8%	1.3%	84.0%	90.7%	2.2%	3.7%	13.9%	6.9%	3.0%	3.4%	5.1%	2.2%	1.9%
KNN	80.9%	29.2%	24.7%	92.0%	96.0%	23.0%	3.6%	94.1%	19.6%	1.3%	94.1%	98.8%	12.4%	7.5%	23.9%	12.6%	4.8%	14.9%	11.6%	9.2%	5.6%
Avg.	54.4%	9.0%	7.8%	47.2%	55.7%	6.2%	1.5%	49.2%	5.0%	1.2%	58.1%	63.6%	5.3%	3.4%	7.2%	3.8%	2.1%	4.5%	4.8%	2.6%	1.8%

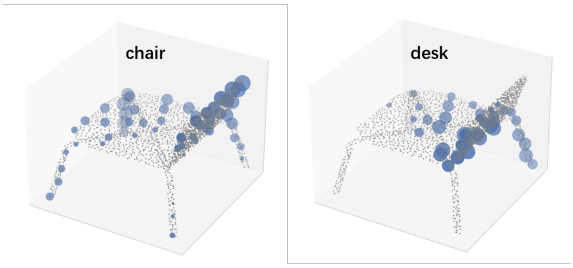

Figure 4. The visualization of attention of a chair for chair and desk as labels, a larger circle means a higher attention score.

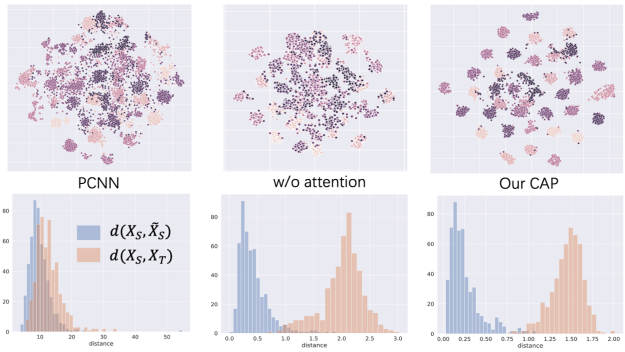

Figure 5. The visualization of features and distances $d(X_s, \tilde{X}_s), d(X_s, X_t)$ w/o and w/ our proposed CAP.

the aforementioned attack baselines, we point out that these defense methods can be evaded by shape-invariant attacks which take geometric-related regularizations into consideration, such as KNN and GeoA3. Specifically, for ModelNet40, the ASR of GeoA3 on PointNet is 19.1% and 15.8% after SOR and DUP-Net preprocessing, respectively. This can be explained by the carefully designed shape-invariant losses used in these attacks, including the k-NN distance in KNN and the curvature loss in GeoA3, which prevent the generated adversarial perturbations from being filtered out by these defenses. In comparison, our solution aims to learn a robust classification model that could defend against various attack strategies. Besides, we find that vanilla models with different structures exhibit varying degrees of robustness. For instance, the average ASR of original adversarial examples on PointCNN (16.7%) is much lower than the other two vanilla models (70.2% on PointNet, 65.4% on DGCNN) for ModelNet40. The results of ShapeNet exhibit a consistent trend as well. We believe that such robustness comes with the in-nature of the vanilla model design, where a higher aggregation degree could lead to the spread of negative impact caused by perturbations.

5.2. The Attention-based Pooling

To validate the effectiveness of our proposed attention-based pooling module, we visualize the attention score, i.e., the $\alpha_{i,c}$ in Eq. 3, of the same object regarding two labels. We visualize the attention in Fig. 4. As illustrated in the figure, when the model extracts features for label *chair*, the sketch of the chair object could be recognized. Furthermore, the chair back is assigned with higher weights. On the contrary, when the model turns to view the chair as a *desk*, the attention scores focus on small segments of the object, which indicates that the model does not recognize the full picture of the chair. Instead, most attention is laid on the key points on the chair surface. This result further explains why models trained with our framework could better characterize the structural information of objects.

5.3. Visualization of Learned Features

We also conduct two experiments to validate whether our proposed CAP can better learn semantic information. We first visualize the features extracted by PointCNN trained w/ and w/o CAP by the t-SNE [45] algorithm in Fig. 5 (first row). As we can see from the figure, the features from the vanilla model are highly clustered. In other words, the underlying manifolds of different labels overlap with each other. After applying the proposed framework, the manifolds become disentangled, which states that CAP could help models effectively distinguish different kinds of objects. To further explore this statement, we sample several pairs of $d(X, X_s)$, $d(X, X_t)$ and visualize the distribution of the feature distances in Fig. 5 (second row), where the label s and t are *chair* and *desk* respectively. The result shows that, with the aid of our framework, the distances between samples with the same label are close to 0, while those with different labels are much larger. We also validate the effectiveness of the proposed attention module, we find that the features are still entangled if we remove the attention module. For more visualization results of the other two models, please refer to Appendix C.

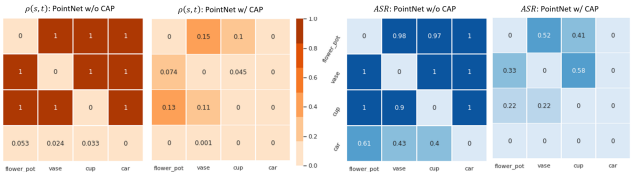

Figure 6. Left: Estimation of $\rho(s,t)$. Right: ASR of target label (row) from ground truth label (column).

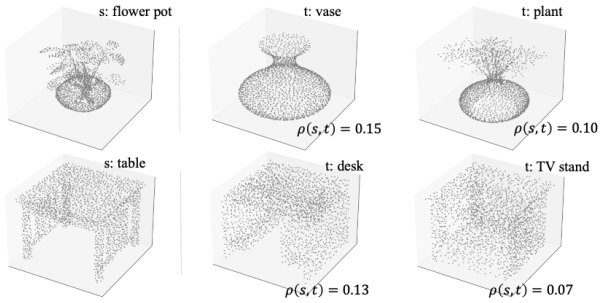

Figure 7. The visualization of sample pairs with large $\rho(s,t)$ on PointNet with our CAP.

5.4. Robustness Certification

Furthermore, we present the estimation of the robustness certification described in our theoretical analysis. Specifically, we estimate the $\rho(s,t)$ for different pairs of samples from four selected classes and perform the IFGM attack on these pairs. The averaged results on PointNet trained with our framework are reported in Fig. 6. The first finding is that the estimation is highly consistent with the actual ASR, i.e., for those pairs with extremely small $\rho(s,t)$, the actual ASR is also 0. The second finding is that our model can better learn semantic information of the point cloud data. For instance, the objects of *flower pot* are more likely to be recognized as *vase* after adding perturbation, however, the $\rho(s,t)$ and the ASR are both close to 0 when the adversary tries to force the model to predict them as *car*. A more illustrative visualization of samples of label pairs (s,t) with a large $\rho(s,t)$ without CAP, including *(vase, flower pot)* and *(table, desk)*, is presented in Fig. 7. This experiment also states the robustness of our model from another perspective, that is, the potential falsely predicted labels won't be much different with the original label visually, e.g., the damage of recognizing a flower pot as a vase will not be severe.

5.5. Go Beyond Optimal Attack

Finally, we try to investigate what kind of attacks can evade our defense. We demonstrate the generated samples in Fig. 8 by enhancing the perturbation of the Gen3D-Pert attack, i.e. lowering the weight of distance loss. As a result, a sample with small perturbations ($\|\delta\|_2 = 0.6$) can make the vanilla PointNet predict the target label *table*, while the model trained with CAP still obtains the correct prediction. To evade our defense, the perturbation needs to be

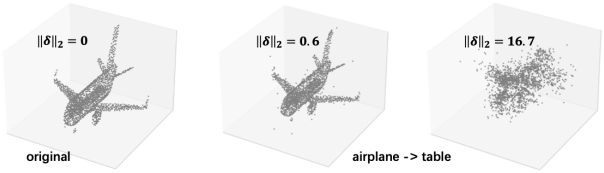

Figure 8. The results of increasing the perturbations w/ and w/o proposed CAP.

extremely large ($\|\delta\|_2 = 16.7$), as shown in the right figure, where the sample has been transformed into a totally different and unrecognizable one from the original. Considering that an adversary is only allowed to slightly manipulate the point cloud in real-world applications [3, 25], such an attack is actually unrealistic to be accomplished. For more results such as the influence of different hyper-parameters and more case studies please refer to Appendix C.

6. Conclusion

In this work, we propose to mitigate adversarial attacks in point cloud classification by enhancing the modeling of semantic and structural information, which can be applied to various classification models and generalize to different kinds of attack strategies. Furthermore, with the aid of manifold learning and extreme value theory, we could certify the robustness against potential adversarial attacks, which is empirically shown to be highly related to the actual attack success rate. As for future work, we may validate our defense framework on more kinds of classification models such as voxel-based and transformer-based approaches. In addition, we may extend the defense framework to other point cloud applications such as 3D object detection and segmentation. Lastly, it would also be interesting to apply the proposed framework to other domains.

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 (61972099, U1736208, U1836210, U1836213, 62172104, 62172105, 61902374, 62102093, 62102091), Natural Science Foundation of Shanghai (19ZR1404800). Min Yang is a faculty of Shanghai Institute of Intelligent Electronics & Systems, Shanghai Collaborative Innovation Center of Intelligent Visual Computing and Engineering Research Center of Cyber Security Auditing and Monitoring, Ministry of Education, China. Mi Zhang and Min Yang are the corresponding authors.

References

[1] Naveed Akhtar and Ajmal S. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 5

[2] Motasem Alfarra, Adel Bibi, Hasan Abed Al Kader Hammoud, Mohamed Gaafar, and Bernard Ghanem. On the decision boundaries of deep neural networks: A tropical geometry perspective. *ArXiv*, abs/2002.08838, 2020. 5

[3] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 2267–2281, 2019. 8

[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] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[6] Jen-Tzung Chien and Ching-Huai Chen. Deep discriminative manifold learning. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2672–2676. IEEE, 2016. 5

[7] Laurens de Haan and Ana Ferreira. Extreme value theory : an introduction. 2006. 5

[8] Paulo de Oliveira Rente, Catarina Brites, Joao Ascenso, and Fernando Pereira. Graph-based static 3d point clouds geometry coding. *IEEE Transactions on Multimedia*, 21(2):284–299, 2018. 3

[9] 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. 6

[10] Zhen Dong, Su Jia, Chi Zhang, Mingtao Pei, and Yuwei Wu. Deep manifold learning of symmetric positive definite matrices with application to face recognition. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 5

[11] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2018. 5

[12] Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pages 180–190. Cambridge University Press, 1928. 5

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015. 5

[15] 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

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4

[17] Warren He, Bo Li, and Dawn Xiaodong Song. Decision boundary analysis of adversarial examples. In *ICLR*, 2018. 5

[18] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, and Nenghai Yu. Shape-invariant 3d adversarial point clouds. *arXiv preprint arXiv:2203.04041*, 2022. 1, 2, 3, 6

[19] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019. 5

[20] Wen jun Zhang, Yikai Zhang, Xiaoling Hu, Mayank Goswami, Chao Chen, and Dimitris N. Metaxas. A manifold view of adversarial risk. *ArXiv*, abs/2203.13277, 2022. 5

[21] Hamid Karimi, Tyler Derr, and Jiliang Tang. Characterizing the decision boundary of deep neural networks. *ArXiv*, abs/1912.11460, 2019. 5

[22] 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, 2, 3, 6

[23] Linyi Li, Xiangyu Qi, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. *arXiv preprint arXiv:2009.04131*, 2020. 5

[24] 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. 1, 2, 3, 4, 6

[25] Yiming Li, Congcong Wen, Felix Juefei-Xu, and Chen Feng. Fooling lidar perception via adversarial trajectory perturbation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7898–7907, 2021. 8

[26] 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, 3, 6

[27] 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. 3, 6

[28] 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. 1, 3

[29] Ze Liu, Han Hu, Yue Cao, Zheng Zhang, and Xin Tong. A closer look at local aggregation operators in point cloud analysis. In *European Conference on Computer Vision*, pages 326–342. Springer, 2020. 3

[30] Haoming Lu and Humphrey Shi. Deep learning for 3d point cloud understanding: a survey. *arXiv preprint arXiv:2009.08920*, 2020. 1, 3

[31] 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. 3, 6

[32] 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. 1, 6

[33] Hanieh Naderi, Arian Etemadi, Kimia Noorbakhsh, and Shohreh Kasaei. Lpf-defense: 3d adversarial defense based on frequency analysis. *arXiv preprint arXiv:2202.11287*, 2022. 3

[34] Anh M Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2015. 5

[35] Nicolas Papernot, Patrick Mcdaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016. 5

[36] 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, 3, 4, 6

[37] 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. 2, 3, 4, 6

[38] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3577–3586, 2017. 3

[39] Lawrence K. Saul and Sam T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifold. *J. Mach. Learn. Res.*, 4:119–155, 2003. 5

[40] Jiachen Sun, Karl Koenig, Yulong Cao, Qi Alfred Chen, and Zhuoqing Mao. On the adversarial robustness of 3d point cloud classification. 2020. 1, 3

[41] Yiming Sun, Feng Chen, Zhiyu Chen, and Mingjie Wang. Local aggressive adversarial attacks on 3d point cloud. In *Asian Conference on Machine Learning*, pages 65–80. PMLR, 2021. 6

[42] 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*, volume 34, pages 954–962, 2020. 2, 3, 6

[43] Pavan Turaga, Rushil Anirudh, and Rama Chellappa. Manifold learning. *Computer Vision: A Reference Guide*, pages 1–6, 2020. 5

[44] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018. 4

[45] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[46] Robin Wang, Yibo Yang, and Dacheng Tao. Art-point: Improving rotation robustness of point cloud classifiers via adversarial rotation. *arXiv preprint arXiv:2203.03888*, 2022. 3

[47] 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, 3, 6

[48] 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. 1, 2, 3, 6

[49] 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. 1, 2, 3

[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. 2, 4

[51] 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. 6

[52] 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

[53] 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. 3

[54] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2020. 4

[55] Shijie Yang, Liang Li, Shuhui Wang, Weigang Zhang, and Qingming Huang. A graph regularized deep neural network for unsupervised image representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1203–1211, 2017. 5

[56] Roozbeh Yousefzadeh and Dianne P. O'Leary. Investigating decision boundaries of trained neural networks. *ArXiv*, abs/1908.02802, 2019. 5

[57] 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

[58] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12498–12507, 2021. 3

[59] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In *NeurIPS*, 2019. 5

[60] Jinlai Zhang, Lyujie Chen, Bo Ouyang, Binbin Liu, Jihong Zhu, Yujing Chen, Yanmei Meng, and Danfeng Wu. Pointcutmix: Regularization strategy for point cloud classification. *arXiv preprint arXiv:2101.01461*, 2021. 3

[61] Yu Zhang, Gongbo Liang, Tawfiq Salem, and Nathan Jacobs. Defense-pointnet: Protecting pointnet against adversarial attacks. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5654–5660. IEEE, 2019. 2, 3

[62] Yingxue Zhang and Michael Rabbat. A graph-cnn for 3d point cloud classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6279–6283. IEEE, 2018. 3

[63] 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, 2, 3, 4

[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

[65] 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, 2, 3, 6

[66] Haoran Zhou, Yidan Feng, Mingsheng Fang, Mingqiang Wei, Jing Qin, and Tong Lu. Adaptive graph convolution for point cloud analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4965–4974, 2021. 3