The Art of Defense: Letting Networks Fool the Attacker

Jinlai Zhang, Graduate Student Member, IEEE, Yinpeng Dong, Minchi Kuang, Binbin Liu, Bo Ouyang, Jihong Zhu, Houqing Wang, and Yanmei Meng

Abstract—3D perception of objects is critical for many real-world applications, such as autonomous cars and robots. Among them, most state-of-the-art (SOTA) 3D perception systems are based on deep learning models. Recently, the research community found that 3D object classifiers on point cloud based on deep learning are easily fooled by adversarial point cloud crafted by attackers. To overcome this, adversarial defenses are considered the most effective ways to improve the robustness of deep learning models, and most adversarial defenses on point cloud are focused on input transformation. However, all previous defense methods decrease the natural accuracy, and the nature of the point cloud classifiers itself has been overlooked. To this end, in this paper, we propose a novel adversarial defense for 3D point cloud classifiers that makes full use of the nature of the point cloud classifiers. Due to the disorder of point cloud, all point cloud classifiers have the property of permutation invariant to the input point cloud. Based on this nature, we design invariant transformations defense (IT-Defense). We show that, even after accounting for obfuscated gradients, our IT-Defense is a resilient defense against SOTA 3D attacks. Moreover, IT-Defense does not hurt clean accuracy compared to previous SOTA 3D defenses. Our code will be available at: https://github.com/cuge1995/IT-Defense.

Index Terms—Adversarial Attack, Point Cloud Classification, Adversarial Defenses.

I. INTRODUCTION

3D OBJECT perception is one of the fundamental problems in 3D vision. Recently, driven by applications such as autonomous driving [1], [2], robotics, digital cities, and virtual/mixed reality, 3D vision has received more and more attention and has become a hot topic in the research community. With the huge breakthrough made by Deep neural networks (DNN) in many fields [3], [4], [5], [6], [7], [8], the state-of-the-art 3D vision algorithms are also gradually occupied by DNNs. However, the DNNs are vulnerable to maliciously generated adversarial examples [2]. As DNN models have been implemented into various real-world applications that were build upon 3D computer vision, the DNNs are pieced by DNNs. However, the DNNs are vulnerable to maliciously generated adversarial examples [2]. Adversarial training is considered to be the most effective defense and it can generalize across different threat models [10]. However, adversarial training faces many problems. Firstly, the high cost of standard adversarial training makes it impractical. In order to reduce the cost of standard adversarial training, Shafahi et al. [11] recycled the gradient information that was computed when updating model parameters, it finally speed up adversarial training 7 to 30 times compared to the standard adversarial training. Maksym et al. [12] proposed the GradAlign that prevents catastrophic overfitting in fast gradient sign method (FGSM) training. Secondly, all adversarial training methods cannot overcome the problem that adversarial trained model decrease the recognition accuracy in clean samples. Thirdly, adversarial training is not effective as input transformation-based defense [13] for point cloud. Another promising line of defense against adversarial examples is to randomize the inputs or the model parameters [14]. The randomize of the model parameters is to sample weights of networks from some distribution [15], [16], [17], however, it needs to adapt with the target networks. The random transform to the input in 2D image as defense methods [18], [19], [20], [21] has been studied extensively, and have shown excellent robustness, but it rarely been explored in 3D point cloud. In this paper, we focus on random input transforms in 3D point cloud.

As pointed out in [14], any deterministic classifier could be outperformed by a randomized one in terms of robustness. We therefore ask, can we build a randomized classifier that with consistent clean accuracy? Motivated by this question, we observed that the 3D point cloud classifier’s property can be used to build the randomized classifier. As shown in Figure 1, due to the unordered nature of 3D point clouds, most point cloud analysis DNNs are invariant to the order...
permutation of the input point cloud. In this paper, we utilize this property to transform the input point cloud, and finally build the randomized point cloud classifier.

The main contributions of this paper are summarized as follows:

- To the best of our knowledge, invariant transformations defense (IT-Defense) is the first work that uses the networks’ property to break the strongest gradient-based attacks. It breaks the attack success rate from 100% to almost 0%.
- Our IT-Defense has no impact on clean accuracy, which is significantly better than previous defense methods.
- IT-Defense is compatible with different DNNs and adversarial defense methods, which can serve as a basic network module for 3D point cloud adversarial defense.

II. RELATED WORK

A. Adversarial Attacks on Images

The deep neural networks are vulnerable to adversarial examples, it was first found in the image domain [22]. Then a bunch of algorithms to generate adversarial examples or to attack the deep neural networks are proposed [23], [24], [25], [26]. Which can be summarized as white-box attacks and black-box attacks [10]. Most attacks in white-box setting are based on the input gradient. The fast gradient method (FGM) [27] generates adversarial examples by one-step update towards the input gradient. The iterative version of FGM (IFGM) [28] generate adversarial examples by small steps towards the gradient direction. A momentum term was introduced by [29] to stabilize the update direction during the iterations, which was known as MIFGM. The projected gradient descent method (PGD) adopts random starts during the iterations, which was served as a baseline in first-order adversary. The C&W attack [30] turn the process of obtaining adversarial examples into an optimization problem, and used the Adam [31] for optimization. This algorithm was heavily used in recent point cloud attack research.

B. Adversarial Attacks on Point Clouds

Due to the safety-critical applications of the point cloud in robotics and safe-driving cars, 3D point clouds have attracted many researchers in the computer vision community. However, its robustness is relatively under-explored compared to the image domain. Reference [32] first proposed to generate adversarial point cloud examples via C&W attack, which introduced two point cloud attacks, point perturbation and adding. The point adding attack can further divide by adding independent points, adding clusters, and adding objects. However, the generate adversarial point clouds examples are very messy, which can be easily perceivable by humans. The knn attack [33] adopted a kNN distance constraint, a clipping and a projection operation to generate more smooth and imperceptible adversarial point clouds examples. The Geometry-Aware Adversarial Attack (GeoA³) [34] further improved the imperceptible to humans. The perturbation based attacks can be removed by statistical outlier removal (SOR) [35] method that removes points with a large kNN distance if the perturbation is too large. To solve this problem, [36] developed the JGBA attack, which is an efficient attack to SOR defense. Besides, the point drop attack [37] was developed by a gradient based saliency map, which iteratively removes the most important points. Moreover, the AdvPC [38] improved the transferability of adversarial point cloud examples by utilizing a point cloud auto-encoder, and the LG-GAN [39] utilized the powerful GANs [40] to generate adversarial examples guided by the input target labels. However, most of those attacks integrate the gradient information from the input, which can be a weakness.

C. Adversarial Defenses

To overcome the threat of adversarial examples to DNNs, extensive research has been conducted on defending against adversarial attacks. Existing adversarial defense methods can be roughly divided into two classes: attacking stage defense and testing stage defense. The adversarial training [41], [42], [43], [44] is an effective way to improve the model’s robustness, which has two stages defense effect. There are other defense methods that have two stages defense effect. For example, Pang et al. [45] proposed a novel loss, the model’s adversarial robustness is increased if the model was trained with this loss. Thermometer encoding [46] encodes values in a discrete way. EMPIR [47] constructed ensemble models with mixed precision of weights and activations. Ensemble diversity [48] improved robustness with a regularization term. For the attacking stage defense, k-Winners Take All [49] developed a novel activation function that masks the backpropagated gradient, input transformations [20] and input randomization [50] utilized the backpropagated transformed gradient to fool the attacker, stochastic activation [51] replaces the dropout layer with a non-differentiable function. Those defend methods can prevent the attacker from generating adversarial examples effectively. For the testing stage defense, they are a lot of adversarial examples detection methods [52], [53], [54], [55], [56] are developed, the transformation method of Pixeldefend [57], defense-GAN [58], and sparse fourier transform [59] transforms the adversarial examples to a normal sample, the Mixup Inference [60], ME-Net [61], and Error Correcting Codes [62] mitigate the adversarial perturbations by inference the adversarial examples directly.

We note that Guo et al. [20] and Xie et al. [50] also utilized input transformation before feeding into the DNN to defend adversarial attacks, which caused the obfuscated gradients effect and can be defeated by the expectation over transformation (EOT) attacks proposed by [63]. However, our work has several key differences with Guo et al. and Xie et al. Firstly, we used the invariant transformations of DNN, which do no
harm to the DNN, but Guo et al. and Xie et al. have some kind of accuracy drop of the standard model. Secondly, we do not cause information loss between the original sample and the transformed sample. Thirdly, we cannot be defeated by the EOT attack.

### III. METHODOLOGY

#### A. An Overview of Point Cloud Attack

Let $x \in \mathbb{R}^{N \times 3}$ represents a set of clean 3D points $\{P_i| i = 1, \ldots, N\}$, and $y$ denote the corresponding true label. For a classifier $F(x) : x \rightarrow y$ that outputs the prediction for an input, the attacker wants to generate an adversarial example $x^{\text{adv}}$ which is imperceptible by humans from $x$ but fools the classifier. The adversarial attack on point cloud can be roughly divided into gradient-based attack and optimization-based attack. The FGM [27], I-FGM [28] and MIFGM [29] are gradient-based attacks, and KNN attack [33] and GeoA$^3$ attack [34] are optimization-based attack. We give a brief introduction of some famous attack algorithms on 3D point cloud classifier in this section.

**FGM [27]** generates adversarial example by one-step update.

$$
x^{\text{adv}} = x - \epsilon \cdot \text{sign} (\nabla_x J(x, y))
$$
where $\nabla_x J$ is the gradient of the loss function with respect to the input $x$, $\text{sign}(\cdot)$ is the sign function that turns the gradient value into the direction value.

**I-FGM [28]** generate adversarial example by small steps in a iterative fashion.

$$
x_t^{\text{adv}} = x_t^{\text{adv}} - \alpha \cdot \text{sign} (\nabla_x J(x_t^{\text{adv}}, y))
$$
where $\alpha = \epsilon / T$ with $T$ steps iteration, and $x_0^{\text{adv}} = x$.

**MIFGM [29]** introduced a momentum term to stabilize the update direction during the iteration.

$$
g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t^{\text{adv}}, y)}{\|\nabla_x J(x_t^{\text{adv}}, y)\|_1}
$$

$$
x_{t+1}^{\text{adv}} = x_t^{\text{adv}} - \alpha \cdot \text{sign} (g_{t+1})
$$
where $g_t$ gathers the gradient information up to the $t$-th iteration with a decay factor $\mu$. We mainly compare FGM, I-FGM and MIFGM with GvF-P [64], the first self-robust point cloud defense method.

**C&W [30]** turn the process of obtaining adversarial examples into an optimization problem.

$$
\arg \min_{x^{\text{adv}}} \|x^{\text{adv}} - x\|_p + c \cdot J(x^{\text{adv}}, y)
$$

where the loss $J$ can be different from the cross-entropy loss and many variants are proposed for point cloud attack [32], [33], [34], such as the KNN attack [33] and GeoA$^3$ attack [34].

#### B. Invariant Transformation Defense (IT-Defense)

In this paper, we make full use of the nature of point cloud classifiers, i.e., the point cloud classifiers are permutation invariant to the input point cloud’s index, and propose the invariant transformation defense (IT-Defense). The pipeline of IT-Defense is shown in Figure 2. The IT-Defense can be described as following:

$$
g(x) = \nabla_{t(x)} J(t(x), y)
$$

where $\nabla_{t(x)} J$ is the gradient of the loss function with respect to the random input transformation $t$ of $x$. Note that unlike previous random input transformation defense [19], [20], [50], our transformation is invariant to the classifier $F()$. The intuition is that by randomly perturb the index of the input point cloud, the transformed point cloud will not affect the performance of point cloud classifier due to the orderless nature of point cloud, but the real gradient have perturbed significantly, thus fool the attacker.

Following, we perform theoretical analysis to investigate why IT-Defense works.

#### C. Theoretical Analysis of IT-Defense Against Gradient-Based Attacker

In the following section, we use the theory of [65] to analyze IT-Defense against various attacks. Under the system of game theory, the relationship between the adversarial robustness of neural networks and the complexity of game interaction is modeled uniformly by [65]. They proved that the adversarial perturbations mainly affect high-order interactions. They
define the game-theoretic interactions as follows:

\[ I(i, j) = \hat{\phi}(i | N) | \text{always present} - \hat{\phi}(i | N) | \text{always absent} \tag{7} \]

where \( \hat{\phi}(i | N) | \text{always present} \) denote the importance of input variable \( i \) when \( j \) is always present and \( \hat{\phi}(i | N) | \text{always absent} \) denote the importance of input variable \( i \) when \( j \) is always absent. Then, the interactions can be decomposed into multiple orders [66] as following:

\[ I(i, j) = \frac{1}{n - 1} \sum_{m=0}^{n-2} I_{ij}^{(m)}, \quad I_{ij}^{(m)} = \mathbb{E}_{S \subseteq N \setminus \{i,j\}}[\Delta v(i, j, S)] \tag{8} \]

where \( I_{ij}^{(m)} \) represents the \( m \)th-order interaction. \( m \) represents the number of units in the background \( S \) other than input units \( i \) and \( j \), reflecting the complexity expressed by the interaction. When the background contains more input units, \( m \) is larger, and the game-theoretic interaction between \( i \) and \( j \) can be regarded as a high-order interaction. When the background contains a small number of input units, \( m \) is small, the game-theoretic interaction between \( i \) and \( j \) is a low-order interaction.

As shown in Figure 3, we perform I-FGM and MIFGM and IT-Defense against their in Pointnet [67], where the lower ratio of points represents the low-order interactions, and higher ratio of points represents the high-order interactions, the results show that the adversarial example without defense mainly affected high-order interactions, the IT-Defense mainly change the low-order interactions, thus mitigate adversarial effects. In addition, for MIFGM, IT-Defense has a larger effect on low-order interactions, significantly reducing the magnitude of low-order interactions in clean samples. IT-Defense has a significant impact on low-order interactions with point ratios of 0.1, 0.2, 0.3, and 0.4.

**D. Theoretical Analysis of IT-Defense Against Optimization-Based Attacker**

Following Sec. III-C, two optimization-based attacker are selected, the 3D-Adv [32] and kNN attack [33]. From Figure 4, we can observe similar results as Sec. III-C, which means the IT-Defense mainly affects the low-order interactions to overcome adversarial effects. Specifically, it can be observed from figure 4 that IT-Defense can reduce the low-order interaction amplitude of clean samples, mainly by reducing the low-order interaction amplitude with point ratios of 0.1 and 0.2 to reduce the effect of kNN attack. And it can be observed from figure 4 that IT-Defense also reduces the effect of kNN attack by reducing the low-order interaction amplitude with a point ratio of 0.1 and 0.2. It is worth noting that optimization-based attacks mainly affect low-order interactions with point ratios of 0.1 and 0.2, while gradient-based attacks mainly affect low-order interactions with point ratios of 0.1, 0.2, 0.3, and 0.4, but IT-Defense can effectively defend against two types of attacks.

**E. Theoretical Analysis of IT-Defense Against Adaptive Attacks**

Recent works on robust defense [63], [68], [69] suggest that the proposed new defense algorithms should take a further evaluation against the corresponding adaptive attack. Since our method transforms the index of the points, the expectation over transformation (EOT) attacks proposed by [63] is expected as the adaptive attacker, due to its excellent performance in adaptive attack to the input transform-based methods [19]. Following [19], we use the EOT to build a strong attacker. The EOT [70] is defined as follows:

\[ \nabla_x \mathbb{E}_r \left[ f_r(x) \right] = \mathbb{E}_r \left[ \nabla_x f_r(x) \right] \approx \frac{1}{n} \sum_{i=1}^{n} \nabla_x f_r(x) \tag{9} \]
where \( f_r(x) \) is the randomized classifier, and the \( r_i \) are independent draws of the randomness transformation. But our transformation is big enough (up to \( 1024^i \) or \( 2048^i \) depends on \( n \)), given one transformation \( t(.) \) of our defender, the probability of correctly sampling this \( t(.) \) by an attacker is \( 1/n! \), which is sufficiently small, simply repeat the transformation multiple times (usually 10 to 30 times in most EOT attacks literatures [68]) cannot recover the true gradient. As shown in Figure 5, the EOT attack further increased the gap in low-order interactions compare to without EOT attack, thus it is not effective to attack the IT-Defense. We further verified this by experiment.

IV. EXPERIMENTS

In this section, we present the experimental results to demonstrate the effectiveness of the proposed method. We first specify the experimental settings. Then we conduct extensive experiments to study the defense effects of IT-Defense, and perform comparative experiments to analyze IT-Defense in detail.

A. Experimental Settings

We implement all experiments on a Linux server with 8 Nvidia RTX 3090 GPUs. For point cloud attacks, we use the ModelNet40 [71] test set kindly provided by IF-Defense [72], which also contained the targeted label for targeted attacks. We also select three commonly used networks in 3D computer vision area [73], [74] for evaluation, \( \text{i.e.} \text{let@tokenonestdot} \), PointNet [67], PointNet++ [5], and DGCNN [75], the SOTA networks such as PointConv [76], PACNN [77], Point Cloud Transformer (PCT) [78] and CurveNet [79] are also selected for KNN attack evaluation. The FGM [27], I-FGM [28], MIFGM [29] and PGD [80] attacks were selected as the gray-box attacker. We follow the same settings as [64] for point cloud attack. Moreover, the untargeted point dropping attack [37], and the C&W attack’s [30] variant kNN attack [33] and 3D-Adv attack [32] were further verified our effective defense ability. The JGBA attack and GeoA³ attack were implemented by their open-sourced code and dataset. Experiments are repeated 5 times. The estimated gradient of EOT is averaged by 10 random transformations.

B. Results

1) IT-Defense Does Not Reduce Clean Accuracy: IT-Defense only change the order of points, it does not reduce clean accuracy. In this section, we validate this via experiments. As shown in Table I, unlike previous SOTA defenses that reduce the clean accuracy, such as the Simple Random Sampling [81] (SRS), Statistical Outlier Removal [82] (SOR), DUP-Net [35] and the IF-Defense [72] are reduced the clean accuracy up to 4% for PointNet. IT-Defense also performs best for the PointNet++ model, with an accuracy of 88.94%. Other defense methods show lower performance, with accuracies between 81.32% (IF-Defense) and 87.40% (SOR). For DGCNN, IT-Defense outperforms the other methods as well, with an accuracy of 89.26%. The other methods have lower accuracies, ranging from 78.40% (SRS) to 87.93% (SOR). For the PointConv model, the proposed defense method (Ours) achieves the highest accuracy of 88.49%, equal to the clean model’s accuracy. The other defense methods have lower accuracies, varying between 81.12% (IF-Defense) and 87.36% (SOR).

In summary, the IT-Defense builds upon the property of the point cloud recognition model, resulting in matching the accuracy of the clean models without any defense. The other defense methods generally exhibit lower performance. The fact that IT-Defense does not reduce clean accuracy is an important property, which indicates that IT-Defense can be seamlessly integrated into any point cloud recognition model.

2) IT-Defense Against Various Attackers: In this section, we show the experimental results of the proposed invariant transformation defense (IT-Defense) method with different attackers. We firstly perform the classical adversarial attacks to our IT-Defense, the results are shown in Table II. We report the success rates of FGM, I-FGM, I-FGM+EOT, PGD, and MIFGM attacks against our defense method and in no defense setting. The I-FGM+EOT attack cannot break our defense and getting worse results than I-FGM attack. The results shown that our defense improve the robustness of the model against various attacks significantly, our method break the success rate from 98.87 to 0.45 in some case.

The point dropping attack [37] is based on the saliency maps of the input point cloud, where also used the gradient in some degree, we therefore perform this attack to our defense. For fairness, we used the same settings with [73] and [72] and used classification accuracy to compare with state of the art defense algorithms, the results are shown in Table III. We can see that for Drop 200 and Drop 100 attacks, the IT-defense leads to better results than the IF-Defense. The results verified
that our defense method can be applied to any kind of attack based on gradient information.

We further verified our defense on kNN attack [33], 3D-Adv attack [32], JGBA attack [36] and GeoA3 attack [34], three of them are the C&W attack’s [30] variant in point cloud. The C&W attack turns the process of obtaining adversarial examples into an optimization problem, but the gradient information is needed in every iteration step within the optimization process. We report the success rates of the kNN attack [33], the 3D-Adv attack [32], the JGBA attack [36] and the GeoA3 attack [34] against IT-Defense in Table IV.

In general, the IT-Defense consistently breaks the success rate from a high level (near 100%) to near 0%, which means that we can totally break the attack effect caused by the attacker.

### C. Comparative Experiments

In this section, we explore different attack settings on IT-Defense.

1) Influence of Perturbation Budget: As suggested in [10], the perturbation budget has a significant impact on the attack performance. We therefore perform experiments with different perturbation budget in the range of [0.05, 0.40], with point cloud in [0, 1]. As shown in Figure 6, our IT-Defense is robust within the 0.25 perturbation budget, it’s big enough to be easily detected by humans.

2) Influence of Attack Steps: The number of attack steps of gradient-based attack is another important factor that affects the attack performance. The results are shown in Figure 6, for Pointnet and DGCNN, the Attack Success Rate(%) increased with the number of attack steps for IFGM without our IT-Defense, but maintain almost constant zero Attack Success Rate(%) with IT-Defense.

3) Influence of Attack Iterations: The number of iterations of optimization-based attacks is a vital variable during attack. We perform KNN attack [33] on Pointnet and DGCNN with the iterations within [500, 2500]. As shown in Figure 6, for Pointnet and DGCNN, similar to the number of steps for IFGM attack, the Attack Success Rate(%) increased slightly with the number of attack iterations for KNN attack without our IT-Defense, but maintain almost constant zero Attack Success Rate(%) with IT-Defense.

4) Influence of EOTs: In order to validate our IT-Defense resistance to EOTs, we scale the number of EOTs up to 100. To the best of our knowledge, this is the largest EOT that has been performed. The results are shown in Table V, for permutation transformation, the Attack Success Rate(%)
increased slightly with the number of EOTs initially, but decreased as the EOTs bigger than 50. The success rate for scale-based EOT attacks fluctuates across different numbers of EOTs. The highest success rate is observed at 0.98% for 30 EOTs, followed by 0.89% for 50 EOTs. The success rate declines to 0.40% for 100 EOTs. The success rates for rotation-based EOT attacks also show variations across different numbers of EOTs. The highest success rate of 1.36% is achieved for 20 EOTs. The success rate generally fluctuates between 0.69% and 1.30% for other values of EOTs, with the exception of 1.22% for 100 EOTs. Overall, the success rate is generally low, with the highest value observed at 1.36% for rotation-based EOT attacks with 20 EOTs, indicating that EOT attack is not an effective attack on our IT-Defense.

5) Understanding the IT-Defense: To figure out what IT-Defense brings to the adversarial point cloud examples, we visualised the airplane and chair of clean sample and adversarial samples without and with IT-Defense under different perturbation budget. From Figure 7, we can conclude that IT-Defense can help models escape the ‘adversarial region’ when the perturbation budget is small, and making adversarial point cloud examples human recognizable when the perturbation budget is large.

V. DISCUSSION

In addition to the points mentioned above, IT-Defense also has some flaws. For example, the random permutation may not provide substantial robustness improvement against strong adversaries. However, the use of multiple defense mechanisms can further improve robustness by making it harder for the...
attacker to devise effective attacks, such as adversarial training [83], implicit gradients [84] and diffusion models [85]. Moreover, IT-Defense requires the deep neural networks to have some invariant transformations of input.

VI. CONCLUSION

In this paper, we propose a defense strategy that uses the networks’ properties to break adversarial attacks. Our findings are insightful, the network’s property is utilized to defend against attacks, and the results show that our defense can break most of the existing point cloud attacks. Note that our method can only resist attacks for finding adversarial samples, but the adversarial samples generated from the original model can be transferred well to IT-Defense. However, our method can easily incorporate with the model with higher robustness, such as models trained with ensemble adversarial training [86] and PointCutMix [73], and combined with other defense methods, thus not only making the attacker hard to generate effective adversarial examples but also letting our AI system more robust resist to adversarial samples generated by other models.

REFERENCES

[1] S. Shi et al., “PV-RCNN: Point-voxel feature set abstraction for 3D object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10529–10538.

[2] X. Xu, J. Zhang, Y. Li, Y. Wang, Y. Yang, and H. T. Shen, “Adversarial attack against urban scene segmentation for autonomous vehicles,” IEEE Trans. Ind. Informat., vol. 17, no. 6, pp. 4117–4126, Jun. 2021.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.

[4] T. B. Brown et al., “Language models are few-shot learners,” 2020, arXiv:2005.14165.

[5] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep hierarchical feature learning on point sets in a metric space,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 5099–5108.

[6] Z. Wu, C. Shen, and A. van den Hengel, “Wider or deeper: Revisiting the ResNet model for visual recognition,” Pattern Recognit., vol. 90, pp. 119–133, Jun. 2019.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[8] J. Zhang et al., “Monitoring sugar crystallization with deep neural networks,” J. Food Eng., vol. 280, Sep. 2020, Art. no. 109965.

[9] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, “Learning meta face recognition in unseen domains,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6163–6172.

[10] Y. Dong et al., “Benchmarking adversarial robustness on image classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 321–331.

[11] A. Shafahi et al., “Adversarial training for free!” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019.

[12] M. Andriushchenko and N. Flammarion, “Understanding and improving fast adversarial training,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 16048–16059.

[13] Q. Huang, X. Dong, D. Chen, H. Zhou, W. Zhang, and N. Yu, “Shape-invariant 3D adversarial point clouds,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 15335–15344.

[14] R. Pinot, R. Ettedgui, G. Rizk, Y. Chevaleyre, and J. Atif, “Randomization matters how to defend against strong adversarial attacks,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 7717–7727.

[15] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh, “Towards robust neural networks via random self-ensemble,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 369–385.

[16] X. Liu, Y. Li, C. Wu, and C.-J. Hsieh, “Adv-BNN: Improved adversarial defense through robust Bayesian neural network,” 2018, arXiv:1810.01279.

[17] S. Lee, H. Kim, and J. Lee, “GradDiv: Adversarial robustness of randomized neural networks via gradient diversity regularization,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 2, pp. 2645–2651, Feb. 2023.

[18] O. Taran, S. Rezaeifar, T. Holotyak, and S. Voloshynovskiy, “Defending against adversarial attacks by randomized diversification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 684–693.

[19] E. Raff, J. Sylvester, S. Forsyth, and M. McLean, “Barrage of random transforms for adversarially robust defense,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6528–6537.

[20] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, “Countering adversarial images using input transformations,” 2017, arXiv:1711.00117.

[21] J. Cohen, E. Rosenfeld, and Z. Kohler, “Certified adversarial robustness via randomized smoothing,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 1310–1320.

[22] C. Szegedy et al., “Intriguing properties of neural networks,” 2013, arXiv:1312.6199.

[23] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “DeepFool: A simple and accurate method to fool deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2574–2582.

[24] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1765–1773.

[25] J. Su, D. V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” IEEE Trans. Evol. Comput., vol. 23, no. 5, pp. 828–841, Oct. 2019.

[26] K. Eykholt et al., “Robust physical-world attacks on deep learning visual classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1625–1634.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014, arXiv:1412.6572.

[28] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” 2016, arXiv:1611.01236.

[29] Y. Dong et al., “Boosting adversarial attacks with momentum,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 9185–9193.

[30] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in Proc. IEEE Symp. Secur. Privacy (SP), May 2017, pp. 39–57.

[31] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[32] C. Xiang, C. R. Qi, and B. Li, “Generating 3D adversarial point clouds,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 9136–9144.
[33] T. Tsai, K. Yang, Y.-T. Ho, and Y. Jin, “Robust adversarial objects against deep learning models,” in Proc. AAAI Conf. Artif. Intell., vol. 34, 2020, pp. 954–962.

[34] Y. Wen, J. Lin, K. Chen, C. L. P. Chen, and K. Jia, “Geometry-aware generation of adversarial point clouds,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 2984–2999, Jun. 2022.

[35] H. Zhou, K. Chen, W. Zhang, H. Fang, W. Zhou, and N. Yu, “DUP-Net: Denoiser and upsampler network for 3D adversarial point clouds defense,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1961–1970.

[36] C. Ma, W. Meng, B. Wu, S. Xu, and X. Zhang, “Efficient joint gradient based attack against SOR defense for 3D point cloud classification,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 1819–1827.

[37] T. Zheng, C. Chen, J. Yuan, B. Li, and K. Ren, “PointCloud saliency maps,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1598–1606.

[38] A. Hanidi, S. Rojas, A. Thabet, and B. Ghanem, “AdvPC: Transferable adversarial perturbations on 3D point clouds,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 241–257.

[39] H. Zhou et al., “LG-GAN: Label guided adversarial network for flexible targeted attack of point cloud based deep networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10556–10565.

[40] I. J. Goodfellow et al., “Generative adversarial networks,” 2014, arXiv:1406.2661.

[41] A. Shafahi et al., “Adversarial training for free!” 2019, arXiv:1904.12843.

[42] D. Zhang, T. Zhang, Y. Lu, Z. Zhu, and B. Dong, “You only propagate once: Accelerating adversarial training via maximal principle,” 2019, arXiv:1906.08077.

[43] E. Wong, L. Rice, and J. Z. Kolter, “Fast is better than free: Revisiting adversarial training,” 2020, arXiv:2001.03994.

[44] T. Na, J. Hwan Ko, and S. Mukhopadhyay, “Cascade adversarial machine learning regularized with a unified embedding,” 2017, arXiv:1708.02582.

[45] T. Pang, K. Xu, Y. Dong, C. Du, N. Chen, and J. Zhu, “Rethinking softmax cross-entropy loss for adversarial robustness,” 2019, arXiv:1905.10626.

[46] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, “Thermometer encoding: One hot way to resist adversarial examples,” in Proc. Int. Conf. Learn. Represent., 2018.

[47] S. Sen, B. Ravindran, and A. Raghunathan, “EMPIR: Ensembles of mixed precision deep networks for increased robustness against adversarial attacks,” 2020, arXiv:2004.10162.

[48] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu, “Improving adversarial robustness via promoting ensemble diversity,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 4970–4979.

[49] C. Xiao, P. Zhong, and C. Zheng, “Enhancing adversarial defense by k-winners-take-all,” 2019, arXiv:1905.10510.

[50] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Vuille, “Mitigating adversarial effects through randomization,” 2017, arXiv:1711.01991.

[51] G. S. Dhillon et al., “Stochastic activation pruning for robust adversarial defense,” 2018, arXiv:1803.01442.

[52] K. Roth, Y. Kilcher, and T. Hofmann, “The odds are odd: A statistical test for detecting adversarial examples,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 5498–5507.

[53] X. Yin, S. Kolouri, and G. K. Rohde, “GAT: Generative adversarial training for adversarial example detection and robust classification,” 2019, arXiv:1905.11475.

[54] T. Yu, S. Hu, C. Guo, W.-L. Chao, and K. Q. Weinberger, “A new defense against adversarial images: Turning a weakness into a strength,” 2019, arXiv:1910.07629.

[55] Y. Li, J. Bradshaw, and Y. Sharma, “Are generative classifiers more robust to adversarial attacks?” in Proc. Int. Conf. Mach. Learn., 2019, pp. 3804–3814.

[56] X. Mu et al., “Characterizing adversarial subspace using local intrinsic dimensionality,” 2018, arXiv:1801.02613.

[57] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, “PixelDefend: Leveraging generative models to understand and defend against adversarial examples,” 2017, arXiv:1710.10766.

[58] P. Samangouei, M. Kabbab, and R. Chellappa, “Defense-GAN: Protecting classifiers against adversarial attacks using generative models,” 2018, arXiv:1805.06677.

[59] M. Bafna, J. Murtagh, and N. Vyas, “Thwarting adversarial examples: An L_0-robust sparse Fourier transform,” 2018, arXiv:1812.05013.
[84] K. Li, Z. Zhang, C. Zhong, and G. Wang, “Robust structured declarative classifiers for 3D point clouds: Defending adversarial attacks with implicit gradients,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 15294–15304.
[85] J. Sun, W. Nie, Z. Yu, Z. M. Mao, and C. Xiao, “PointDP: Diffusion-driven purification against adversarial attacks on 3D point cloud recognition,” in Proc. 40th Int. Conf. Mach. Learn., 2023.
[86] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” 2017, arXiv:1705.07204.

Jinlai Zhang (Graduate Student Member, IEEE) received the B.S. degree from the Changsha University of Science and Technology, Changsha, China, in 2017. He is currently pursuing the Ph.D. degree with the College of Mechanical Engineering, Guangxi University. His research interests include 3D deep learning and time series forecasting.

Yinpeng Dong received the B.S. and Ph.D. degrees from the Department of Computer Science and Technology, Tsinghua University. He is currently a Post-Doctoral Researcher with the Department of Computer Science and Technology, Tsinghua University. His research interests include the adversarial robustness of machine learning and deep learning. He received the Microsoft Research Asia Fellowship and the Baidu Fellowship.

Minchi Kuang received the Ph.D. degree from Tsinghua University in 2017. He is currently an Assistant Researcher with the Department of Precision Instruments, Tsinghua University. His main research interests include artificial intelligence and intelligent control.

Binbin Liu received the B.S. degree from the University of Science and Technology of China in 2019. He is currently pursuing the master’s degree with the Department of Computer Science and Technology, Tsinghua University. His research interests include 3D deep learning and robustness of object detection.

Bo Ouyang received the B.E. degree from the National University of Defense Technology in 2011. He is currently pursuing the M.A. degree with the Department of Computer Science and Technology, Tsinghua University. His research interests include computer vision, deep learning, and natural language processing.

Jinhong Zhu received the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, in 1995. He is currently a Professor with the Department of Precision Instrument, Tsinghua University. His research interests include flight control and robotics.

Houqing Wang received the B.S. degree in ship electrical and electronic engineering and the M.S. and Ph.D. degrees in power electronics and power drives from Shanghai Maritime University, Shanghai, China, in 2015, 2017, and 2020, respectively. From April 2019 to September 2019, he was a Research Assistant with the Centre for Smart Energy Conversion and Utilization Research, City University of Hong Kong. Since January 2021, he has been with Tsinghua University, Beijing, China, where he is currently a Post-Doctoral Researcher with the Department of Precision Instrument. His research interests include digital control techniques of power converters, renewable energy generation systems, and battery energy storage systems.

Yanmei Meng received the Ph.D. degree from the Hefei University of Technology, Hefei, China, in 2003. She is currently a Professor with the College of Mechanical Engineering, Guangxi University. Her research interests include intelligent detection and control technology, and computer vision.