The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

# Multimodal Robust Prompt Distillation for 3D Point Cloud Models

**Xiang Gu**[1]*, **Liming Lu**[1]*, **Xu Zheng**[2,3]†, **Anan Du**[4], **Yongbin Zhou**[1], **Shuchao Pang**[1]†
[1]Nanjing University of Science and Technology
[2]The Hong Kong University of Science and Technology (Guangzhou)
[3]INSAIT, Sofia University, St. Kliment Ohridski
[4]Nanjing University of Industry Technology
{eminentguxiang, luliming, zhouyongbin, pangshuchao}@njust.edu.cn, zhengxu128@gmail.com, anan.du@niit.edu.cn

## Abstract

Adversarial attacks pose a significant threat to learning-based 3D point cloud models, critically undermining their reliability in security-sensitive applications. Existing defense methods often suffer from (1) high computational overhead and (2) poor generalization ability across diverse attack types. To bridge these gaps, we propose a novel yet efficient teacher-student framework, namely **M**ultimodal **R**obust **P**rompt **D**istillation (MRPD) for distilling robust 3D point cloud model. It learns lightweight prompts by aligning student point cloud model's features with robust embeddings from three distinct teachers: a vision model processing depth projections, a high-performance 3D model, and a text encoder. To ensure a reliable knowledge transfer, this distillation is guided by a confidence-gated mechanism which dynamically balances the contribution of all input modalities. Notably, since the distillation is all during the training stage, there is no additional computational cost at inference. Extensive experiments demonstrate that MRPD substantially outperforms state-of-the-art defense methods against a wide range of white-box and black-box attacks, while even achieving better performance on clean data. Our work presents a new, practical paradigm for building robust 3D vision systems by efficiently harnessing multimodal knowledge.

**Code** — https://github.com/eminentgu/MRPD

## Introduction

Deep learning(Lu et al. 2024) has significantly advanced 3D perception, enabling autonomous systems, such as self-driving cars, to understand their environments with impressive accuracy (Sohail et al. 2024). However, these models remain highly vulnerable to adversarial attacks (Zhang et al. 2023a; Pang et al. 2025a), where small, often imperceptible perturbations cause severe prediction failures. This fragility undermines their reliability in open-world settings and raises serious safety concerns in high-stakes applications. For example, in autonomous driving, a maliciously altered point cloud could lead to misidentifying a pedestrian, with potentially fatal consequences (Wang et al. 2021).

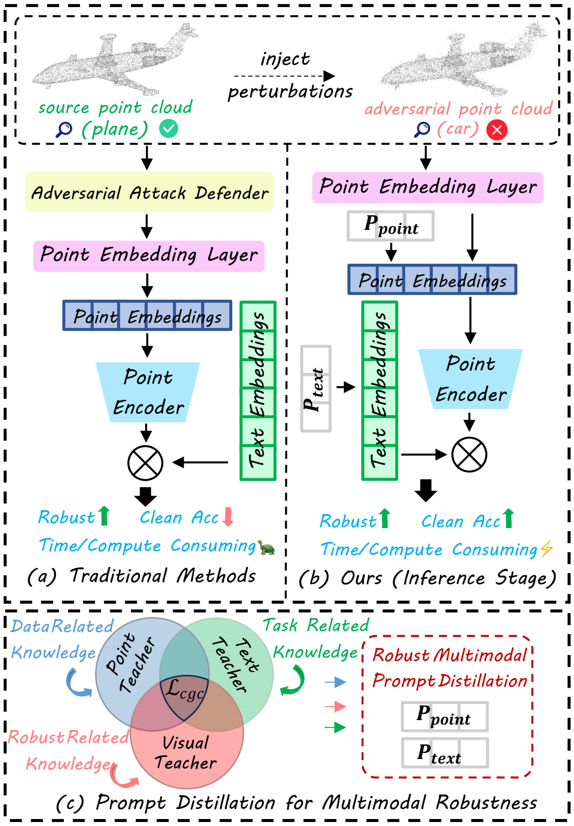

Figure 1: (a) Traditional Defenses: Heavy modules with increasing inference costs. (b) Our Inference: The optimized prompts provide robustness with **zero** computational overhead. (c) Our Training: We distill robust, multimodal knowledge into lightweight prompts.

This vulnerability is especially pronounced in 3D point clouds. Although they provide rich geometric cues essential for robotics and autonomous vehicles, their sparse and unstructured nature creates a broad attack surface, making them particularly susceptible to adversarial manipulation (Xiao et al. 2018; Liu, Yu, and Su 2019). In response, various defense methods have been proposed (Wicker and Kwiatkowska 2019; Wu et al. 2020; Zhang et al. 2023b, 2024), but they still face a fundamental trade-off. Many

---

[*] These authors contributed equally.

[†] Corresponding author.

are highly specialized and fail to generalize across attacks, while more robust methods, such as adversarial training or architectural modifications, sincur significant computational costs. This makes them unsuitable for real-time, latency-sensitive scenarios(Wang et al. 2024; Pang et al. 2025b; Lu et al. 2025). As illustrated in Figure 1(a), these conventional defenses often rely on cumbersome modules that increase inference latency, limiting their practical deployment.

To break this paradigm of relying on computationally expensive, single-modality defenses, recent advances in large-scale Vision Language Models (VLMs)(Xia et al. 2025) offer a compelling alternative. These models have demonstrated remarkable robustness and knowledge transfer capabilities, inspiring efforts to adapt their pretrained knowledge to the 3D domain, with methods like PointCLIP(Zhang et al. 2022b; Zhu et al. 2023) and ULIP(Xue et al. 2023) showing great promise. However, these pioneering efforts have overwhelmingly focused on boosting standard task accuracy, largely overlooking the crucial opportunity to transfer the robustness of these powerful models. Consequently, a systematic mechanism to endow 3D perception with the robustness advantages of 2D vision, without incurring the heavy costs of traditional defenders, remains largely absent.

To address these challenges, we propose a novel yet efficient teacher-student framework: Multimodal Robust Prompt Distillation (MRPD). Instead of relying on costly adversarial examples or altering the model's architecture, our MRPD distills robustness into lightweight, learnable prompts and further improve robustness of 3D point cloud models. As shown in Figure 1(c), during the training phase, we leverage a combination of three robust "teachers": (1) a frozen image encoder that provides stable supervision, as 3D adversarial perturbations often lose efficacy when projected to 2D; (2) a text encoder guided by learnable prompts to discover a more robust semantic space; and (3) a powerful point cloud teacher model that offers high-quality geometric guidance on clean data. This multimodal knowledge from these distinct teachers is transferred to the student model's learnable prompts through a carefully designed distillation process, featuring a confidence-gated mechanism to filter unreliable teacher signals and a dynamic weighting strategy to balance the different knowledge sources.

Crucially, the complex distillation process is discarded after training. As shown in Figure 1(b), only the lightweight, optimized prompts are retained, providing adversarial resilience with **zero** additional computational overhead. Our main contributions are summarized as follows: (I) We propose Multimodal Robust Prompt Distillation (MRPD), a novel framework that efficiently transfers multimodal robustness from image, text, and teacher models into lightweight prompts, thereby leaving the student model architecture untouched. (II) We propose a highly efficient paradigm that achieves robust defense with zero inference overhead, effectively breaking the longstanding trade-off between adversarial robustness and clean data accuracy. (III) Through extensive experiments, we demonstrate that MRPD establishes a new state-of-the-art in 3D adversarial defense, consistently outperforming computationally expensive methods against a diverse suite of attacks.

## Related Work

**Adversarial Defense for Point Clouds.** In response to this diverse threat landscape, a similarly varied array of defense mechanisms has been proposed for 3D point cloud models. A common strategy involves pre-processing the input point cloud to remove or repair adversarial perturbations. This includes early methods like Statistical Outlier Removal (SOR(Rusu et al. 2008)) and their more advanced successors, such as DUP-Net(Zhou et al. 2019) and IF-Defense(Wu et al. 2020), which focus on input reconstruction and purification. Rather than pre-processing inputs at inference time, another line of thought focuses on augmenting the training data to build more intrinsically robust models. For instance, PointGuard(Liu, Jia, and Gong 2021) utilizes random subsampling with a majority voting scheme, while Point-CutMix(Zhang et al. 2022a) employs a mixup-based strategy for regularization. More recently, diffusion models like Ada3Diff(Zhang et al. 2023c) have emerged as powerful tools for purifying adversarially perturbed inputs. Despite their effectiveness against certain threats, these defense paradigms often suffer from two major drawbacks: limited generalizability across attack types and significant computational overhead. By contrast, our MRPD charts a different course by aiming for a general-purpose, efficient defense that enhances the model's intrinsic robustness without costly pre-processing or architectural changes at inference time.

**Vision-Language Guided Point Cloud Models.** Beyond these traditional defense paradigms, a promising new direction emerges from leveraging the rich knowledge within large-scale VLMs like CLIP(Radford et al. 2021). Initial approaches, such as PointCLIP(Zhang et al. 2022b) and PartSLIP(Liu et al. 2023), sought to bridge the 2D and 3D domains by projecting point clouds into multi-view depth maps, thereby capitalizing on powerful pre-trained 2D encoders. To avoid the geometric information loss inherent in 2D projection, a second type of models emerged. These models, including CLIP2Point(Huang et al. 2023) and ULIP(Xue et al. 2023), focus on directly aligning 3D encoders with the VLM feature space using large-scale text-image-point cloud triplet datasets, a paradigm that recent works like UNI3D(Zhou et al. 2023) have pushed to billion-parameter scales. However, a critical blind spot in this line of research has been its overwhelming focus on improving performance on standard downstream tasks (*e.g.*, classification), leaving the crucial opportunity to transfer robustness from the 2D domain largely unexplored. Our work directly addresses this gap, proposing a systematic teacher-student framework to distill this untapped multimodal knowledge into the 3D models.

## Methodology

As shown in Figure 2, we propose a framework to enhance the adversarial robustness of VLM-guided point cloud models via parameter-efficient multimodal prompt tuning. The key idea is to distill multimodal knowledge from image, text, and point cloud modalities into lightweight, learnable prompts within the point cloud student and its text encoder. We first introduce our **MultiModal Robust Knowl-**

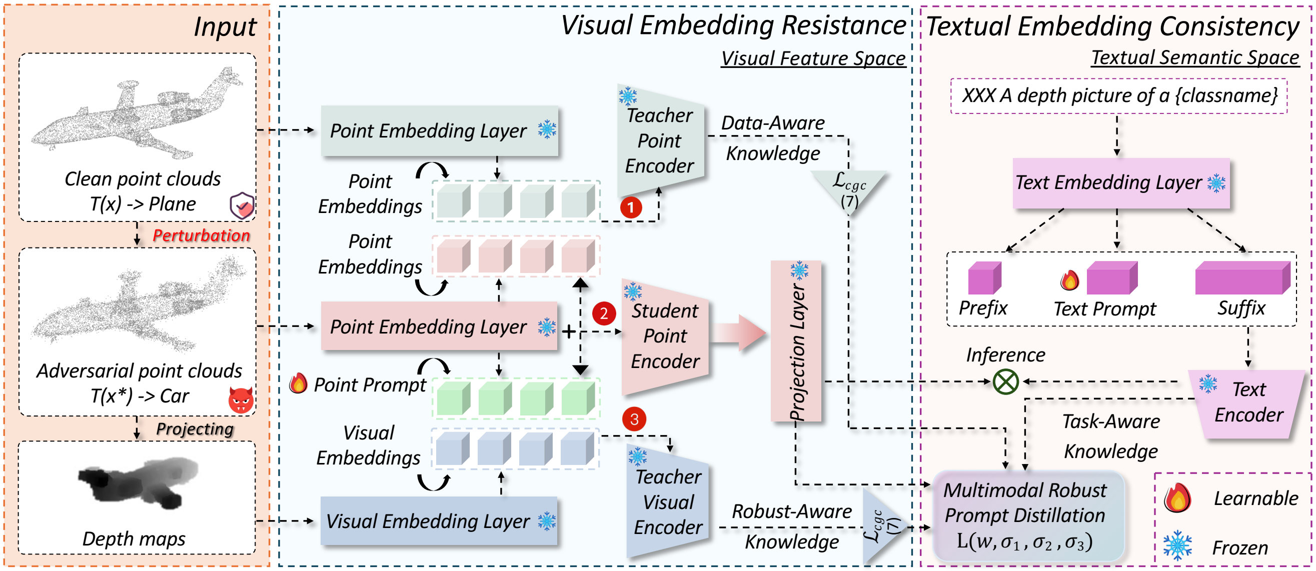

Figure 2: The proposed Multimodal Robust Prompt Distillation (MRPD) framework. During training, robust knowledge from three teachers (image, text, and a 3D model) is distilled into lightweight prompts. At inference, these prompts enhance the student model's robustness with zero additional computational cost.

**edge Collector**, which sources stable supervision signals from each modality. Subsequently, we describe the **Multi-Prompt Distillation** architecture, where these signals guide the optimization of the student model's prompts.

## MultiModal Knowledge Collector

Enhancing the robustness of VLM-guided 3D models is a formidable challenge. Unlike in the 2D domain, the field of 3D deep learning lacks a wealth of powerful, pre-trained robust models that can serve as a foundation. Meanwhile, point cloud data is inherently vulnerable due to its sparsity, unstructured nature, and sensitivity to geometric perturbations. These characteristics make standard robust training techniques, such as adversarial training, not only computationally expensive but also prone to failure, as they rely on large-scale, high-quality data which is scarce for 3D shapes. Given these domain-specific limitations, sourcing robust knowledge solely from the 3D modality is insufficient. We therefore propose to collect and aggregate multimodal supervision signals from three complementary modalities: the 2D projection domain, the semantic text domain, and 3D point cloud domain.

**Image Modality: A Source of Cross-Domain Stability** We observe a helpful domain gap between 3D point clouds and their 2D image-space representations. *Adversarial attacks meticulously crafted to perturb a 3D model's prediction often fail to transfer effectively when the point cloud is projected into a 2D image.* The geometric perturbations become subtle textural changes or are lost entirely during rendering. We leverage this phenomenon by using the frozen image encoder of the VLM, $f_I$, as a stable source of cross-modal robust knowledge.

Specifically, for a given 3D point cloud $P$, we first render it into a multi-view depth image set via a projection operator $\Pi(\cdot)$. This process generates a view-invariant representation. The resulting image is then fed into the image encoder to extract a robust feature vector, $\mathbf{z}_I$:

$$\mathbf{z}_I = \text{normalize}(f_I(\Pi(P))) \tag{1}$$

This feature $\mathbf{z}_I$ remains remarkably stable even when $P$ is adversarially attacked, providing a consistent supervision signal for the student model.

**Text Modality: Learnable Prompts for Semantic Priors** We introduce a learnable textual prompt to capture robust, task-specific semantic priors. Instead of relying on static, hand-crafted templates (*e.g.*, "a photo of a {class}"), we design a prompt that can be optimized to absorb semantic cues for our multimodal distillation process.

Let $T = \{t_1, t_2, ..., t_N\}$ be the names of the downstream classes. We define a set of learnable prompt vectors $V_{ctx} = \{\mathbf{v}_1, ..., \mathbf{v}_M\}$. These vectors are concatenated with each class name to form a dynamic prompt:

$$T'_{cls} = \text{concat}(V_{ctx}, t_{cls}) \tag{2}$$

The complete set of dynamic prompts is then passed through the text encoder $f_T$ to produce robust class-level semantic embeddings $\mathbf{z}_T$:

$$\mathbf{z}_T = \text{normalize}(f_T(\{T'_1, T'_2, ..., T'_N\})) \tag{3}$$

By making the prompt context $V_{ctx}$ learnable, our model can adaptively find a semantic space that is more resilient to adversarial shifts.

**Point Cloud Modality: Structural Guidance from a Teacher** While the 3D modality is vulnerable, a stronger, more capable point cloud model can still provide valuable geometric and structural knowledge. We employ a pre-trained point cloud model, denoted as the "teacher" encoder $f_P^{\text{teacher}}$, which remains frozen throughout training.

This teacher model's role is to provide a consistent and high-quality representation of the original, *clean* 3D shape.

For a clean point cloud $P$, the teacher produces a stable reference feature $\mathbf{z}_P^{\text{teacher}}$:

$$\mathbf{z}_P^{\text{teacher}} = \text{normalize}(f_P^{\text{teacher}}(P)) \tag{4}$$

This feature serves as a geometric anchor, guiding the student model to maintain structural integrity even when its input is perturbed. Collectively, these three components provide a rich and diverse set of robust features, $\mathbf{z}_I$, $\mathbf{z}_T$, and $\mathbf{z}_P^{\text{teacher}}$, that form the foundation of our MRPD framework. These supervision signals are then utilized to optimize the student model's learnable prompt.

## Multimodal Robust Prompt Distillation

Having established the collection of multimodal knowledge from image, text, and point cloud modalities, we now detail the core of our framework: the distillation architecture. Our goal is to transfer this multi-source knowledge into a lightweight "student" point cloud model, $f_P^{\text{student}}$, by optimizing a set of learnable prompts integrated within it. This process is designed to harden the student model against adversarial attacks without altering its backbone, ensuring zero overhead at inference. The distillation is governed by two key principles. First, to mitigate potential semantic conflicts and prevent the student from learning wrong teacher predictions, we introduce a **Confidence-Gated Distillation** loss. Second, to optimally balance the influence of the three diverse modalities, we employ a **Dynamic Weighting** mechanism that learns the relative importance of each teacher signal during training.

**Confidence-Gated Distillation for Reliable Knowledge Transfer** A key challenge in multi-teacher distillation is handling cases where teachers disagree or are incorrect, especially on ambiguous or out-of-distribution samples. Blindly forcing the student to mimic all teachers can propagate errors and degrade performance. To address this, we propose a **Confidence-Gated Contrastive Loss**. This loss function acts as a quality filter, ensuring that knowledge is only distilled from a teacher when it is both correct and confident in its prediction.

Let $\mathbf{z}_{\text{stu}} \in \mathbb{R}^{B\times D}$ be the feature outputs from the student model for a batch of $B$ inputs, and $\mathbf{z}_{\text{ref}} \in \mathbb{R}^{B\times D}$ be the corresponding features from one of the teacher modalities (image, text, or point cloud). Let $\mathbf{z}_T \in \mathbb{R}^{C\times D}$ be the set of $C$ class-level text embeddings from (3), which serve as the semantic ground truth for classification. First, we assess the teacher's prediction confidence by computing the logits between its features and the class embeddings:

$$\mathcal{L}_{\text{ref}} = \frac{\mathbf{z}_{\text{ref}} \cdot \mathbf{z}_T^\top}{\tau} \tag{5}$$

where $\tau$ is a temperature hyperparameter. Next, for each sample $i$ in the batch with ground-truth label $y_i$, we create a binary mask $\mathcal{M}$ that validates the teacher's prediction. A sample is considered valid for distillation only if its true label is within the top-$k$ predictions made by the teacher:

$$\mathcal{M}_i = \begin{cases} 1 & \text{if } y_i \in \text{top-k}(\mathcal{L}_{\text{ref},i}) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

This mask $\mathcal{M}$ effectively filters out instances where the teacher is likely incorrect, preventing negative knowledge transfer. Finally, we compute a symmetric contrastive loss only on the subset of samples where $\mathcal{M}_i = 1$. Let $\mathbf{z}_{\text{stu}}^{\text{sel}}$ and $\mathbf{z}_{\text{ref}}^{\text{sel}}$ be the student and reference features selected by the mask. The Confidence-Gated Contrastive Loss, $\mathcal{L}_{\text{CGC}}$ is:

$$\mathcal{L}_{\text{CGC}}(\mathbf{z}_{\text{stu}}, \mathbf{z}_{\text{ref}}) = \frac{1}{2}\left(\text{CE}(\mathbf{S}, \mathbf{y}) + \text{CE}(\mathbf{S}^\top, \mathbf{y})\right) \tag{7}$$

where $\mathbf{S} = (\mathbf{z}_{\text{stu}}^{\text{sel}} \cdot (\mathbf{z}_{\text{ref}}^{\text{sel}})^\top)/\tau$ is the similarity matrix between the selected student and reference features, $\mathbf{y}$ is a vector of identity labels $[0, 1, \ldots, B_{sel} - 1]$, and CE denotes the Cross-Entropy loss. This loss pulls the student's representation of a valid sample towards its teacher's representation, while pushing it away from other samples in the batch.

**Dynamic Loss Weighting for Multimodal Balancing** The three teacher modalities provide complementary but distinct types of knowledge. The image teacher offers holistic visual cues, the text teacher provides semantic priors, and the point cloud teacher gives fine-grained geometric guidance. Statically assigning weights to their respective distillation losses is suboptimal, as their relative importance may vary across different data samples and training stages. To this end, we employ a dynamic weighting strategy inspired by multi-task learning (Kendall, Gal, and Cipolla 2018) to automatically balance their contributions. This method frames the total loss as a multi-task objective where each modality's loss has an associated learnable uncertainty parameter. The model learns to down-weigh modalities with higher uncertainty (*i.e.*, less reliable signals).

Let $\mathcal{L}_I$, $\mathcal{L}_P$, and $\mathcal{L}_T$ be the Confidence-Gated Contrastive Losses calculated using the image, point cloud, and text teacher features, respectively, as defined in Eq. 7. We introduce three learnable log-variance parameters, $\lambda_I, \lambda_P, \lambda_T$. The final, combined distillation loss $\mathcal{L}_{\text{total}}$ is formulated as:

$$\mathcal{L}_{\text{total}} = \sum_{k\in\{I,P,T\}} \left(e^{-\lambda_k}\mathcal{L}_k + \lambda_k\right) \tag{8}$$

In this formulation, each $\exp(-\lambda_k)\mathcal{L}_k$ term is an uncertainty-weighted loss, and $\lambda_k$ acts as a regularizer to prevent the weights from growing infinitely. This allows the model to learn an optimal, data-driven balance between the three robust supervision signals, leading to more stable and effective distillation.

## Training and Evaluation

Our framework consists of two phases. **During training**, we perform Multimodal Robust Prompt Distillation. Specifically, we optimize a set of lightweight prompts for the student's point cloud encoder and the text encoder. This process is guided by a dynamic, confidence-gated distillation loss that aggregates supervision signals from image, text, and 3D teacher models. **At inference**, the entire distillation apparatus (teachers and loss mechanism) is discarded. The student model operates solely with its optimized prompts, achieving enhanced robustness with **no additional computational overhead or architectural changes**. This efficiency makes our method highly practical for deployment. The full procedure is outlined in *Algorithm 1 in supplementary materials*.

| **Method** | **Clean** | **PGD** | **Perturb** | **KNN** | **ADD-CD** | **ADD-HD** | **AOF** | **Drop-200** | **AdvPC** | **Avg. R** | **+Params** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Dataset: ModelNet40*** | | | | | | | | | | | |
| Clean Model | 70.99 | 63.82 | 0.00 | 51.86 | 0.04 | 0.00 | 0.00 | 53.04 | 0.00 | 21.10 | 0 |
| SRS | 55.75 | 54.90 | 44.12 | 53.00 | 61.83 | 46.11 | 13.25 | 31.36 | 16.57 | 40.14 | 0 |
| SOR | 65.56 | 61.95 | 50.12 | 59.20 | 60.86 | 55.06 | 17.26 | 41.82 | 19.81 | 45.76 | 0 |
| DUP | 61.67 | 61.06 | 52.35 | 62.44 | 54.54 | 43.96 | 25.61 | 39.30 | 20.54 | 44.98 | 814 307 |
| IF-Defense | 62.93 | 62.40 | 63.37 | 62.80 | 63.09 | 62.40 | 50.36 | 63.57 | **47.65** | 59.46 | 1 978 209 |
| Adv Training | 89.95 | 88.01 | 68.15 | 85.70 | 68.60 | 56.81 | 54.74 | 72.45 | 32.62 | 65.89 | 40 960 |
| **MRPD (Ours)** | **90.52** | **89.14** | **80.79** | **87.88** | **81.36** | **68.31** | **54.86** | **78.32** | 39.99 | **72.58** | 0 |
| ***Dataset: ScanObjectNN*** | | | | | | | | | | | |
| Clean Model | 52.74 | 25.99 | 0.00 | 4.86 | 0.00 | 0.00 | 0.00 | 49.03 | 0.00 | 9.99 | 0 |
| SRS | 51.73 | 41.81 | 39.14 | 35.88 | 43.30 | 20.12 | 7.43 | 46.63 | 14.54 | 31.11 | 0 |
| SOR | 45.98 | 38.83 | 32.58 | 33.17 | 37.30 | 31.71 | 20.16 | 41.50 | 21.17 | 32.05 | 0 |
| DUP | 32.69 | 31.99 | 29.49 | 31.71 | 31.68 | 28.42 | 17.14 | 27.55 | 17.35 | 26.92 | 814 307 |
| IF-Defense | 39.87 | 39.73 | 39.83 | 40.49 | 39.73 | 40.60 | 34.80 | 36.71 | 33.62 | 38.19 | 1 978 209 |
| Adv Training | **83.48** | **79.15** | 69.81 | 63.43 | 74.15 | 48.82 | 57.29 | **80.40** | 40.39 | 64.18 | 15 360 |
| **MRPD (Ours)** | 78.80 | 76.13 | **72.24** | **72.52** | **74.74** | **57.08** | **60.06** | 77.72 | **48.65** | **67.39** | 0 |

Table 1: Classification accuracy (%) of different defense strategies under various **white-box attacks** on ModelNet40 and ScanObjectNN. Best results in each column are in **bold**. **Avg. R** means average robust accuracy of all attacks listed.

# Experiments

## Experimental Setup

**Datasets and Metrics.** We evaluate our method on two standard benchmarks: the synthetic **ModelNet40** (Wu et al. 2015) (40 classes, 1024 points/object) and the real-world **ScanObjectNN** (Uy et al. 2019) (15 classes, with background noise and occlusions). We report classification accuracy (%) on both clean data (**Clean Accuracy**) and adversarially perturbed data (**Robust Accuracy**) to measure performance and resilience. **Attack Scenarios.** We test our defense against a comprehensive suite of attacks under both **white-box** (full model access) and **black-box** (transfer-based) settings. The attacks include point-wise perturbations (**PGD** (Liu, Yu, and Su 2019), **Perturb**, **KNN** (Xiao et al. 2018)), point additions (**ADD-CD/HD** (Wen et al. 2020), **AOF** (Zhang et al. 2023b)), **AdvPC** (Hamdi et al. 2020), and point removal (**Drop-200**). This diverse set of threats allows for a rigorous evaluation of our model's robustness.

## Main Experimental Results

**White-Box Attack Analysis.** Table 1 showcases the white-box robustness of our MRPD framework. The results confirm its ability to achieve superior defense across a broad spectrum of attacks while introducing zero inference overhead. On ModelNet40, MRPD establishes a new state-of-the-art, achieving the highest average robustness (72.58%) and even surpassing the strong adversarial training baseline on clean data (90.52% vs. 89.95%). Its significant gains against diverse attacks like Perturb (+12.64%) and ADD-HD (+11.50%) highlight that our multimodal distillation learns a more generalizable feature space, avoiding the overfitting common to adversarial training. This advantage extends to the challenging real-world ScanObjectNN dataset, where MRPD again delivers a higher average robust accuracy (67.39% vs. 64.18%). Its strong performance on noisy and occluded data validates our core premise: distilling knowledge from stable 2D and text teachers enables the model to preserve essential features against both adversarial attacks and real-world imperfections. Ultimately, by embedding multimodal robustness into lightweight prompts, MRPD offers a powerful and practical defense that breaks the trade-off between security and efficiency.

**Generalization Against Black-Box Attacks.** We further assess MRPD's generalizability in a challenging black-box setting, where attacks are transferred from a known, third-party model. As shown in Table 2, our method demonstrates remarkable resilience, confirming its robustness extends beyond specific threat models. On ModelNet40, MRPD again achieves superior performance with the highest average robust accuracy (65.72%). It significantly outperforms the adversarial training baseline on a majority of transferable attacks, including ADD-CD (+9.16%) and ADD-HD (+9.93%). This strong generalization suggests that by learning from diverse 2D, 3D, and text teachers, MRPD develops a decision boundary less correlated with standard 3D architectures, making it inherently more resistant to transferred attacks. This pattern is amplified on the real-world ScanObjectNN dataset, where MRPD once again secures the highest average robustness (67.47%). Its notable performance across a wide spectrum of threats, especially against challenging attacks like AOF and AdvPC, shows that the distilled multimodal knowledge provides resilience against both transferred perturbations and the inherent noise of real-world data. This black-box evaluation confirms that MRPD's multimodal distillation paradigm yields a fundamentally more generalizable defense.

**Qualitative Analysis.** To provide a qualitative understanding of our method's effectiveness, we visualize the feature embeddings of adversarial examples from ModelNet40 using t-SNE (Van der Maaten and Hinton 2008), as shown in Figure 3. The visualization reveals a stark contrast. Without defense (Figure 3(a)), adversarial attacks cause the feature space to collapse, resulting in chaotic and overlap-

| Method | Clean | PGD | Perturb | KNN | ADD-CD | ADD-HD | AOF | Drop-200 | AdvPC | Avg. R | +Params |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Dataset: ModelNet40*** | | | | | | | | | | | |
| Clean Model | 70.99 | 69.37 | 52.63 | 58.39 | 57.37 | 41.49 | 20.62 | 56.65 | 18.64 | 46.90 | 0 |
| SRS | 55.75 | 54.66 | 42.54 | 47.93 | 51.78 | 37.52 | 17.38 | 33.55 | 16.98 | 37.79 | 0 |
| SOR | 65.56 | 62.76 | 48.99 | 54.01 | 59.93 | 51.99 | 20.18 | 43.07 | 18.40 | 44.92 | 814 307 |
| DUP | 61.67 | 61.06 | 45.66 | 53.61 | 46.43 | 39.91 | 16.73 | 41.73 | 15.76 | 40.11 | 0 |
| IF-Defense | 62.93 | 62.76 | 61.59 | 61.55 | 58.43 | 50.28 | **40.80** | 52.96 | **38.09** | 53.31 | 1 978 209 |
| Adv Training | 89.95 | 89.42 | 74.03 | 82.13 | 63.94 | 44.69 | 38.09 | 77.67 | 29.34 | 62.41 | 40 960 |
| **MRPD (Ours)** | **90.52** | **89.55** | **77.84** | **84.52** | **73.10** | **54.62** | 36.22 | **83.51** | 26.42 | **65.72** | 0 |
| ***Dataset: ScanObjectNN*** | | | | | | | | | | | |
| Clean Model | 52.74 | 50.90 | 37.82 | 40.42 | 39.63 | 30.85 | 15.13 | 47.88 | 14.26 | 34.61 | 0 |
| SRS | 51.73 | 49.62 | 37.44 | 41.43 | 39.52 | 30.50 | 15.27 | 46.15 | 14.23 | 34.27 | 0 |
| SOR | 45.98 | 44.62 | 34.00 | 37.23 | 39.49 | 37.99 | 20.89 | 41.57 | 20.51 | 34.54 | 814 307 |
| DUP | 32.69 | 32.96 | 29.63 | 31.37 | 30.36 | 29.01 | 16.48 | 28.14 | 16.83 | 26.85 | 0 |
| IF-Defense | 39.87 | 40.28 | 40.56 | 40.46 | 39.63 | 38.65 | 32.34 | 37.13 | 33.52 | 37.82 | 1 978 209 |
| Adv Training | **83.48** | **82.55** | 74.05 | 73.46 | 71.24 | 61.55 | 40.18 | **81.44** | 41.57 | 65.76 | 15 360 |
| **MRPD (Ours)** | 78.80 | 78.87 | **75.16** | **75.09** | **74.81** | **67.28** | **44.07** | 77.93 | **46.56** | **67.47** | 0 |

Table 2: Classification accuracy (%) of different defense strategies under various **blackbox attacks** on ModelNet40 and ScanObjectNN. Best results in each column are in **bold**. **Avg. R** means average robust accuracy of all attacks listed.

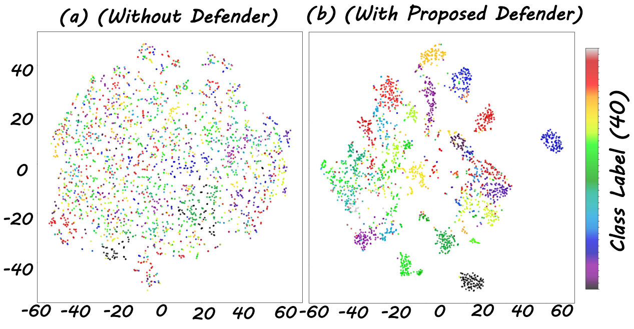

Figure 3: MRPD preserves feature space integrity under adversarial attack on ModelNet40. (a) Without defense, features from different classes become indistinguishable. (b) With MRPD, features remain well-separated, ensuring robust classification.

ping clusters where class distinctions are lost. This leads to widespread misclassification. In sharp contrast, the feature space of our MRPD-protected model (Figure 3(b)) remains highly structured even under the same attacks. The features form compact, well-separated clusters corresponding to their true classes. This demonstrates that our method successfully preserves the semantic structure of the feature space, ensuring robust class separability and correct classification against adversarial manipulations. Similar improvements are observed on the more challenging real-world ScanObjectNN dataset, as detailed in *Figure 1 in the supplementary materials*.

## Ablation Studies

To validate the contributions of each component in our MRPD framework, we conduct a series of ablation studies. We analyze the effectiveness of the multimodal prompts, the dynamic loss weighting, the confidence-gated distillation loss, and the prompt parameterization. For brevity, we present the key results here, while a more comprehensive breakdown of results across all attack types and additional analyses are provided in the *supplementary materials*.

| Method | Clean | PGD | ADD-CD | Drop | Avg. R |
|---|---|---|---|---|---|
| ***Dataset: ModelNet40*** | | | | | |
| Baseline | 70.99 | 63.82 | 0.04 | 53.04 | 21.10 |
| + Text Prompt | 86.18 | 80.83 | 52.67 | 70.54 | 52.37 |
| + Point Prompt | 75.89 | 73.82 | 67.95 | 61.59 | 55.16 |
| **MRPD (Full)** | **90.52** | **89.14** | **81.36** | **78.32** | **72.58** |
| ***Dataset: ScanObjectNN*** | | | | | |
| Baseline | 52.74 | 25.99 | 0.00 | 49.03 | 9.99 |
| + Text Prompt | 69.26 | 55.24 | 49.41 | 64.26 | 44.56 |
| + Point Prompt | 53.37 | 46.01 | 45.80 | 49.48 | 35.67 |
| **MRPD (Full)** | **78.80** | **76.13** | **74.74** | **77.72** | **67.39** |

Table 3: Ablation study of MRPD components on ModelNet40 and ScanObjectNN. We show accuracy (%) on clean data, three representative attacks, and the recalculated average robustness. **Avg. R** denotes the average accuracy over all white-box attacks.

## Effectiveness of MultiModal Prompts

To isolate the contributions of our multimodal prompts, we compare our full MRPD model against a Baseline (no prompts) and single-prompt variants in Table 3. The Baseline is extremely vulnerable, with accuracy collapsing on attacks like 'ADD-CD', highlighting the need for defense. Introducing either the text or point prompt individually yields significant gains in both clean and robust accuracy, confirming their value in providing semantic or geometric resilience. Crucially, our full MRPD model, integrating both prompt types, substantially outperforms all other configurations. Its average robust accuracy on ModelNet40 (72.58%) and

ScanObjectNN (67.39%) demonstrates a massive improvement over single-prompt variants. This performance leap, far exceeding the sum of individual contributions, confirms a powerful synergistic effect. Fusing multimodal knowledge into both point and text encoders via our distillation is therefore essential for achieving a comprehensive, sota defense.

## Analysis of Dynamic Loss Weighting

To validate our dynamic loss weighting, we analyzed the evolution of the learned weights ($w_k = \exp(-\lambda_k)$) during training on ModelNet40. As shown in Figure 4, the model automatically learns a sophisticated balancing strategy. The weights for the point and image teachers increase dramatically in early epochs, stabilizing at high values ($w_P \approx 55, w_I \approx 47$), while the text teacher's weight ($w_T$) remains consistently low ($\approx$ 1.2). This learned hierarchy reveals that the model prioritizes rich geometric guidance from the point and image teachers to build a robust feature foundation, while treating the text teacher as a high-level semantic regularizer. This emergent behavior confirms the efficacy of our dynamic scheme in achieving a more effective and stable balance than manually fixed weights could provide.

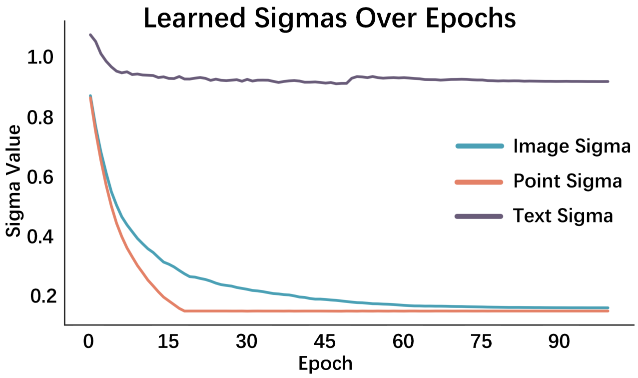

Figure 4: **Evolution of learned loss weights** ($1/\sigma^2$) **for each distillation task.** The model learns to heavily prioritize the point and image teachers while using the text teacher as a low-weight semantic regularizer.

## Ablation on Confidence-Gated Distillation Loss

| Method | Clean | PGD | Perturb | AOF | Avg. R |
|---|---|---|---|---|---|
| w/o CGC | 90.40 | **89.47** | 79.21 | 52.88 | 70.99 |
| **MRPD (Ours)** | **90.52** | 89.14 | **80.79** | **54.86** | **72.58** |

Table 4: Ablation of the Confidence-Gated Contrastive (CGC) loss on ModelNet40. **Avg. R** denotes the average accuracy over all white-box attacks. Best results are in **bold**.

We perform an ablation study on the Confidence-Gated Contrastive (CGC) loss to highlight its importance. We compare our full MRPD with a variant (w/o CGC) that uses a standard, ungated multi-teacher distillation. As shown in Table 4, the w/o CGC variant performs marginally better on the standard PGD attack, likely due to its simple averaging of teacher knowledge. However, this naive approach falters on more complex attacks. Our full model with CGC shows superior performance on Perturb (+1.58%) and AOF (+1.98%), where CGC's ability to filter conflicting teacher signals in ambiguous cases is critical. This strategic gating is key to achieving higher overall robustness (Avg. R), confirming CGC as a vital component of our framework.

## Ablation on Prompt Parameters

| Prompt Size | | Accuracy (%) | | | | |
|---|---|---|---|---|---|---|
| Point | Text | Clean | PGD | ADD-CD | Drop | Avg. R |
| 5 | 3 | 89.71 | 87.93 | 78.97 | 78.97 | 69.91 |
| **10** | **3** | **90.52** | **89.14** | **81.36** | **78.32** | **72.58** |
| 10 | 5 | 89.71 | 88.70 | 79.38 | 77.55 | 70.78 |
| 15 | 3 | 89.38 | 88.41 | 79.70 | 76.74 | 69.94 |

Table 5: Ablation on prompt size using ModelNet40. We report accuracy (%) on clean data and representative white-box attacks. **Avg. R** denotes the average accuracy over all white-box attacks. The optimal configuration is highlighted.

To determine the optimal prompt configuration, we ablate the number of point and text tokens. As shown in Table 5, a configuration of 10 point tokens and 3 text tokens strikes the best balance, achieving the highest average robustness (72.58%) without compromising clean accuracy. Increasing the point prompt size from 5 to 10 significantly boosts performance, but further increasing it to 15 leads to a decline, suggesting potential overfitting. Similarly, a concise 3-token text prompt proves most effective. This highlights the importance of prompt capacity: sufficient to capture robust knowledge, yet compact enough to avoid overfitting. We adopt the (10, 3) configuration for all experiments.

# Conclusion

In this paper, we introduce Multimodal Robust Prompt Distillation, a novel framework that enhances 3D model robustness by distilling knowledge from image, text, and 3D teachers into lightweight prompts. The core of our MRPD framework is a novel distillation architecture that translates robust knowledge from powerful image, text, and 3D teachers into complementary point and text prompts. To manage the potential for conflicting advice arising from these distinct modalities, our Confidence-Gated Contrastive loss acts as a dynamic arbiter, selectively filtering inconsistent signals to stabilize the learning process and distill a truly robust representation. Experiments show MRPD significantly outperforms existing methods against a wide range of attacks, especially on real-world data, while critically adding zero inference overhead. This work establishes an effective and practical paradigm for robust 3D vision. Future directions include applying this prompt-distillation technique to other 3D tasks like detection and segmentation.

# Acknowledgments

Xiang Gu, Liming Lu and Shuchao Pang are supported by the National Natural Science Foundation of China (Grant No.62206128), National Key Research and Development Program of China under (Grant No.2023YFB2703900) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No.KYCX24_0723). Anan Du is supported by the Start-up Fund for New Talented Researchers of Nanjing University of Industry Technology (Grant No.YK24-05-04). Yongbin Zhou is supported by the National Natural Science Foundation of China (Grant No.U2336205).

# References

Hamdi, A.; Rojas, S.; Thabet, A.; and Ghanem, B. 2020. Advpc: Transferable adversarial perturbations on 3d point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 241–257. Springer.

Huang, T.; Dong, B.; Yang, Y.; Huang, X.; Lau, R. W.; Ouyang, W.; and Zuo, W. 2023. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22157–22167.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7482–7491.

Liu, D.; Yu, R.; and Su, H. 2019. Extending adversarial attacks and defenses to deep 3d point cloud classifiers. In *2019 IEEE International Conference on Image Processing (ICIP)*, 2279–2283. IEEE.

Liu, H.; Jia, J.; and Gong, N. Z. 2021. Pointguard: Provably robust 3d point cloud classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6186–6195.

Liu, M.; Zhu, Y.; Cai, H.; Han, S.; Ling, Z.; Porikli, F.; and Su, H. 2023. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21736–21746.

Lu, L.; Chen, Z.; Lu, X.; Rao, Y.; Li, L.; and Pang, S. 2024. Uniads: Universal architecture-distiller search for distillation gap. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14167–14174.

Lu, L.; Pang, S.; Zheng, X.; Gu, X.; Du, A.; Liu, Y.; and Zhou, Y. 2025. CIARD: Cyclic Iterative Adversarial Robustness Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 350–359.

Pang, S.; Chen, Z.; Zhang, S.; Lu, L.; Liang, S.; Du, A.; and Zhou, Y. 2025a. Towards a 3D Transfer-based Blackbox Attack via Critical Feature Guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 26912–26922.

Pang, S.; Rao, Y.; Lu, Z.; Wang, H.; Zhou, Y.; and Xue, M. 2025b. PriDM: Effective and Universal Private Data Recovery via Diffusion Models. *IEEE Transactions on Dependable and Secure Computing*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rusu, R. B.; Marton, Z. C.; Blodow, N.; Dolha, M.; and Beetz, M. 2008. Towards 3D point cloud based object maps for household environments. *Robotics and Autonomous Systems*, 56(11): 927–941.

Sohail, S. S.; Himeur, Y.; Kheddar, H.; Amira, A.; Fadli, F.; Atalla, S.; Copiaco, A.; and Mansoor, W. 2024. Advancing 3D point cloud understanding through deep transfer learning: A comprehensive survey. *Information Fusion*, 102601.

Uy, M. A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, T.; and Yeung, S.-K. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1588–1597.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, H.; Pang, S.; Lu, Z.; Rao, Y.; Zhou, Y.; and Xue, M. 2024. dp-promise: Differentially private diffusion probabilistic models for image synthesis. In *33rd USENIX Security Symposium (USENIX Security 24)*, 1063–1080.

Wang, X.; Cai, M.; Sohel, F.; Sang, N.; and Chang, Z. 2021. Adversarial point cloud perturbations against 3D object detection in autonomous driving systems. *Neurocomputing*, 466: 27–36.

Wen, Y.; Lin, J.; Chen, K.; Chen, C. P.; and Jia, K. 2020. Geometry-aware generation of adversarial point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 2984–2999.

Wicker, M.; and Kwiatkowska, M. 2019. Robustness of 3d deep learning in an adversarial setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11767–11775.

Wu, Z.; Duan, Y.; Wang, H.; Fan, Q.; and Guibas, L. J. 2020. If-defense: 3d adversarial point cloud defense via implicit function based restoration. *arXiv preprint arXiv:2010.05272*.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920.

Xia, J.; Zhu, H.; Pang, S.; Lu, Z.; Li, B.; Zhou, Y.; and Xue, J. 2025. One Head to Rule Them All: Amplifying LVLM Safety through a Single Critical Attention Head. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xiao, C.; Li, B.; Zhu, J.-Y.; He, W.; Liu, M.; and Song, D. 2018. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*.

Xue, L.; Gao, M.; Xing, C.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J. C.; and Savarese, S. 2023.

Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1179–1189.

Zhang, J.; Chen, L.; Liu, B.; Ouyang, B.; Xie, Q.; Zhu, J.; Li, W.; and Meng, Y. 2023a. 3d adversarial attacks beyond point cloud. *Information Sciences*, 633: 491–503.

Zhang, J.; Chen, L.; Ouyang, B.; Liu, B.; Zhu, J.; Chen, Y.; Meng, Y.; and Wu, D. 2022a. Pointcutmix: Regularization strategy for point cloud classification. *Neurocomputing*, 505: 58–67.

Zhang, J.; Dong, Y.; Kuang, M.; Liu, B.; Ouyang, B.; Zhu, J.; Wang, H.; and Meng, Y. 2023b. The art of defense: Letting networks fool the attacker. *IEEE Transactions on Information Forensics and Security*, 18: 3267–3276.

Zhang, K.; Zhou, H.; Zhang, J.; Huang, Q.; Zhang, W.; and Yu, N. 2023c. Ada3diff: Defending against 3d adversarial point clouds via adaptive diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8849–8859.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2022b. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8552–8562.

Zhang, S.; Du, A.; Zhang, J.; Gao, Y.; and Pang, S. 2024. MAFFN-SAT: 3D Point Cloud Defense via Multi-view Adaptive Feature Fusion and Smooth Adversarial Training. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhou, H.; Chen, K.; Zhang, W.; Fang, H.; Zhou, W.; and Yu, N. 2019. Dup-net: Denoiser and upsampler network for 3d adversarial point clouds defense. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1961–1970.

Zhou, J.; Wang, J.; Ma, B.; Liu, Y.-S.; Huang, T.; and Wang, X. 2023. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*.

Zhu, X.; Zhang, R.; He, B.; Guo, Z.; Zeng, Z.; Qin, Z.; Zhang, S.; and Gao, P. 2023. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2639–2650.