Learning Robust 3D Representation from CLIP via Dual Denoising

Shuqing Luo [1] **Bowen Qu** [1] **Wei Gao** [1]

Abstract

In this paper, we explore a critical yet under-investigated challenge: *whether pre-trained vision language models like CLIP can be adapted for zero-shot adversarial-robust point cloud recognition*. Since point clouds are a crucial data format for representing a 3D world, particularly in safety-critical applications, there is a pressing need to develop adversarially robust 3D recognition algorithms due to the inherent vulnerability of deep models to adversarial attacks. Recent advances in vision-language pre-training have endowed point cloud recognition models with powerful zero-shot generalization capacity, leading to a new paradigm for large-scale 3D recognition. This is usually achieved via cross-modal distillation, a scalable approach for multi-modal aware 3D learning. However, current methods primarily rely on direct alignment to map point cloud features to a shared multi-modal feature space, providing no improvement in 3D robustness. This raises a critical question: can both high-performing zero-shot 3D recognition and zero-shot 3D adversarial robustness be achieved in large-scale 3D learning? Our answer is affirmative. In this paper, we propose a novel distillation algorithm designed to learn robust 3D representations from CLIP. It is capable of simultaneously enhancing both zero-shot 3D recognition performance and zero-shot 3D adversarial robustness compared to baseline models. Our approach is built upon two key components, namely robust 3D pre-training and parallel feature denoising. This enables robust and high-performing 3D zero-shot generalization without the dependence on adversarial training, which is often inefficient and prone to overfit. Experiments indicate that our model achieves a 7% improvement with clean input and varying degrees of enhancement with perturbed input, outperforming other models of similar scale on zero-shot 3D recognition benchmarks.

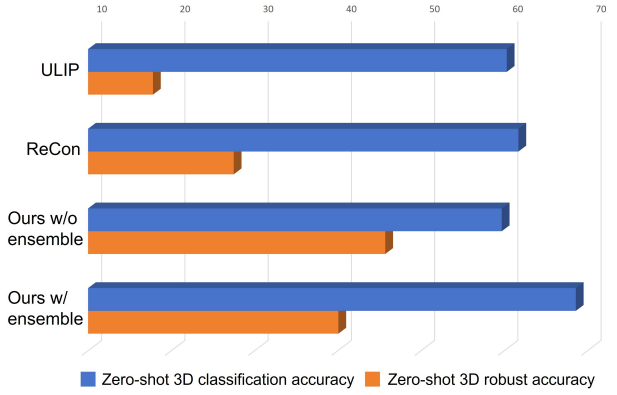


Figure 1. **Comparison on model performance and robustness.** Models are pre-trained on ShapeNet with ViT-B level scale, using CLIP-B as the teacher model. Zero-shot accuracy is obtained on the ModelNet40 testset, and robust accuracy is tested using IFGM with ℓ_∞ norm $\epsilon = 0.01$ for 50 steps. With our proposed robust 3D pre-training algorithm, zero-shot robustness is effectively boosted, while zero-shot accuracy has no significant change. Combined with our proposed ensemble approach, both robustness and performance can be boosted under a zero-shot setting.

1. Introduction

As one of the most common representations of the 3D world, point cloud plays a vital part in computer vision. Point cloud is composed of unordered points with *xyz*-coordinates or other attributes like RGB, sampled from various sources including LiDAR, stereo cameras, or computer-aided design (CAD) models. The unique *xyz*-coordinates enable point clouds to represent spatial geometry, providing complementary information from other modalities like images. In recent years, research on deep learning has also been extended to point clouds (Qi et al., 2017a;b), facilitating massive 3D-based real-life applications such as autonomous driving. Although deep learning has significantly improved performance on 3D recognition tasks, extensive studies have demonstrated that point cloud learning models are vulnerable to adversarial attacks (Xiang et al., 2019). In the adversarial settings, some carefully crafted imperceptible perturbations are added to the input point cloud, which can lead to

[1]School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, China. Correspondence to: Wei Gao <gaowei262@pku.edu.cn>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models.

wrong recognition results while the input still maintains its semantics. Since point cloud learning models are broadly applied to safety-critical areas, it is imperative to enhance their adversarial robustness.

Given the unstructured nature of point cloud data, crafting 3D adversarial examples differs from that on images. Current approaches include shifting or removing a subset of the whole point cloud (Xiang et al., 2019), creating additional point clusters (Sun et al., 2021), and generating new point clouds as adversarial examples (Zhou et al., 2020). To make adversarial point clouds visually imperceptible, optimizations are employed to restrict the number or the amplitude of perturbed points. To defend against the attacks and train adversarially robust 3D recognition models, previous research has proposed some strategies, including pre-processing (Zhou et al., 2019), adversarial training, and adversarial purification (Sun et al., 2023a). However, current research is mainly conducted on small-scale datasets with a limited number of classes, and models are trained with traditional supervised learning, which is less practical or convincing. Besides, although adversarial training can effectively improve robustness, it has been proven to have many limitations, including overfitting, inefficient training, and invalidity on kNN-based point cloud learning (Sun et al., 2023a). This calls for the exploration of robust and efficient point cloud recognition on larger-scale datasets.

Recent advances in large-scale vision-language pre-training (Radford et al., 2021; Zhai et al., 2023) have also inspired 3D recognition, *i.e.*, large-scale multi-modal aware pre-training can significantly improve generalization (Xue et al., 2023). Previous research has explored different strategies, including leveraging pre-trained models like CLIP on point cloud rendered images directly for 3D perception (Zhang et al., 2022; Zhu et al., 2023), or aligning 3D representations to the pre-trained shared feature space using correspondence among text, image, and point cloud (Xue et al., 2023; 2024). Through cross-modal distillation, high-performing zero-shot 3D recognition can be achieved by scaling the dataset and 3D backbone. While this marks a new trend in 3D learning, the robustness of zero-shot 3D recognition has not yet been fully studied. In this paper, we extend 3D adversarial attack algorithms to the multi-modal settings and reveal the vulnerability of simply aligning 3D features with a shared multi-modal feature space. This motivates us to explore *whether pre-trained vision language models can be adapted for zero-shot adversarial-robust 3D recognition*.

Our research starts by examining the adversarial robustness of current methods that distill 3D knowledge from CLIP-like models. We find that incorporating self-supervised learning on point clouds in the pre-training stage can help improve the adversarial robustness, compared to direct alignment. Building on this insight, we develop a robust 3D pre-training approach named Point Denoising AutoEncoder (PointDAE), inspired by the principles of denoise diffusion models (DDM) and the effectiveness of diffusion-based adversarial purification (Nie et al., 2022). While the combination of PointDAE and cross-modal distillation improves zero-shot 3D adversarial robustness, it offers limited gains for zero-shot 3D recognition performance. Inspired by the recent in-depth study on knowledge ensemble (Allen-Zhu & Li, 2023), we come up with the idea of implementing an efficient ensemble approach via the proposed parallel feature denoising. Different from introducing randomness via training models with different random seeds, we attach randomness to the training process via feature denoising, *i.e.*, initializing CLS token with Gaussian noise and training the model to perform feature denoising. This will introduce no additional computation overhead during training, but can effectively facilitate test-time augmentation, thereby improving zero-shot generalization. A comparison with our method and others can explain the effectiveness of our design, as shown in Figure 1, where experiments are conducted under the same settings. ULIP (Xue et al., 2023), ReCon (Qi et al., 2023), and ours represent direct alignment, distillation with mask reconstruction, and distillation with robust pre-training, respectively.

Our contributions can be outlined as follows:

(1) We propose to learn robust and multi-modal aware 3D representation from the perspective of adversarial machine learning. Based on the experimental insights, we design a simple yet effective approach named *Dual Denoising*. Our method can surpass previous methods with similar scale and configuration, both in terms of model and training dataset, on standard zero-shot point cloud recognition benchmarks.

(2) We extend conventional 3D adversarial attack algorithms on supervised learning to multi-modal learning, so as to evaluate the adversarial robustness of different large-scale models and examine the effectiveness of our approach under zero-shot settings.

(3) We find that incorporating self-supervised learning with cross-modal distillation can improve the adversarial robustness of point cloud learning under zero-shot settings. Based on this insight, we propose a robust 3D pre-training approach named PointDAE inspired by the principles of denoising diffusion models and diffusion-based adversarial purification algorithms. Experimental results demonstrate the effectiveness of our design.

(4) We propose an efficient approach for knowledge ensemble, since only employing robust pre-training in cross-modal distillation offers minimal improvement

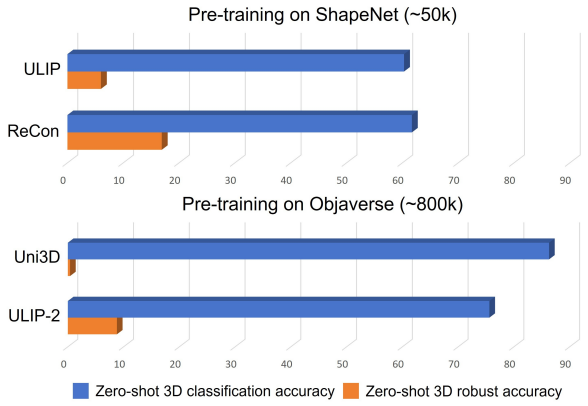


Figure 2. **Comparison on model performance and robustness with current methods.** For fair comparison, experiments are conducted on ShapeNet (∼50k) and Objaverse (∼800k), and zero-shot accuracy is evaluated on ModelNet40 testset. Robust accuracy is tested using IFGM with ℓ_∞ norm $\epsilon = 0.025$.

for zero-shot recognition. By making slight refinements to the training pipeline and incorporating corresponding test-time augmentation during inference, we can significantly enhance zero-shot performance while ensuring adversarial robustness.

2. Related Work

Adversarial attacks and defenses for point cloud learning. Xiang et al. (2019) first demonstrated that point cloud recognition models are vulnerable to adversarial attacks. Hamdi et al. (2020) designed transformable black-box attacks on point cloud learning models. Zhou et al. (2020) proposed to generate adversarial point clouds via label-guided GAN. To enhance robustness for point cloud learning, Zhou et al. (2019) proposed to purify the input point cloud via preprocessing. Li et al. (2022b) proposed constrained optimization to defend against adversarial attacks through implicit gradients. (Sun et al., 2023a) proposed to use 3D diffusion model for adversarial purification.

Contrastive Language Image Pre-training. The pioneering CLIP (Radford et al., 2021) proposes to project images and text into a shared feature space during pre-training, where downstream tasks are conducted via retrieval based on the cosine similarity of the extracted feature. FLIP (Li et al., 2023) accelerates and scales the pre-training via masking. SLIP (Mu et al., 2022) combines self-supervised learning and CLIP pre-training for better zero-shot generalization. EVA-CLIP (Sun et al., 2023b) incorporates novel designs for representation learning, optimization, and augmentation to achieve better performance than CLIP with an equal number of parameters and smaller training costs. SigLIP (Zhai et al., 2023) replaces *Softmax* in CLIP with *Sigmoid* operation for large-scale distributed pre-training with better efficiency, since *Sigmoid* loss does not need to access the whole mini-batch.

Multi-modal aware point cloud pre-training. ULIP (Xue et al., 2023) aligns point cloud features to CLIP feature space using the ternary of point cloud, image, and text. ULIP-2 (Xue et al., 2024) scales the model using multi-view rendered images and more detailed text captions. ReCon (Qi et al., 2023) combines cross-modal distillation with mask reconstruction-based self-supervised learning. Uni3D (Zhou et al., 2023) and OpenShape (Liu et al., 2024) scale both the training dataset and model size and achieve extremely high performance on zero-shot 3D recognition benchmarks.

Adversarial attack on point cloud. We use IFGM, PGD, and C&W attack in this paper, where C&W attack is L_2 norm-based and the others are L_∞ norm-based. For the C&W attack, we set the loss function as:

$$\mathcal{L} = (\max_{i \neq t'} \mathcal{Z}(\boldsymbol{X}')_i - \mathcal{Z}(\boldsymbol{X}')_{t'})^+ + \lambda \cdot \|\boldsymbol{X} - \boldsymbol{X}'\|_2, \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{n\times 3}$ is the clean point cloud, $\mathbf{X}' \in \mathbb{R}^{n\times 3}$ is the optimized adversarial point cloud, $\mathcal{Z}(\boldsymbol{X})_i$ is the i-th element of the output logits, and t' is the target class. Here, logits are computed with the dot product between the point cloud feature and the set of text CLIP features. We leverage a 10-step binary search to find the appropriate hyperparameter λ from [10, 80]. We use the whole test set of ModelNet40 and ScanObjectNN (OBJ_ONLY) for evaluation. The step size of the adversarial optimization is 0.01, and we allow at most 500 iterations of optimization in each binary search to find the adversarial examples. For the L_∞ norm-based PGD attack, we adopt the formulation as:

$$\boldsymbol{X}_{t+1} = \Pi_{\boldsymbol{X}+\mathcal{S}}(\boldsymbol{X}_t + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{X}_t}\mathcal{L}(\boldsymbol{X}_t, \boldsymbol{\theta}, \boldsymbol{y}))), \quad (2)$$

where $\boldsymbol{X}_t$ is the adversarial point cloud in the t-th iteration during attack, Π is the projection function to project the adversarial point cloud to a pre-defined space $\boldsymbol{X} + \mathcal{S}$, the L_∞ norm ball. α is the step size. We use the sign function to normalize the gradient into the L_∞ norm ball at each iteration. We set the boundary of allowed perturbations as $\epsilon = \{0.01, 0.025, 0.05, 0.075\}$ for space $\mathcal{S}$. Since point cloud data is continuous within the range of $[-1, 1]$, we set the step size as $\alpha = \epsilon/10$. IFGM is basically similar to PGD, with a difference in perturbation initialization.

3. Our Method

We first present our motivation by extending traditional adversarial attacks to the setting of multi-modal pre-training, and evaluate current large-scale multi-modal aware point cloud learning models on a zero-shot 3D recognition benchmark. Based on the insights from experimental results, we

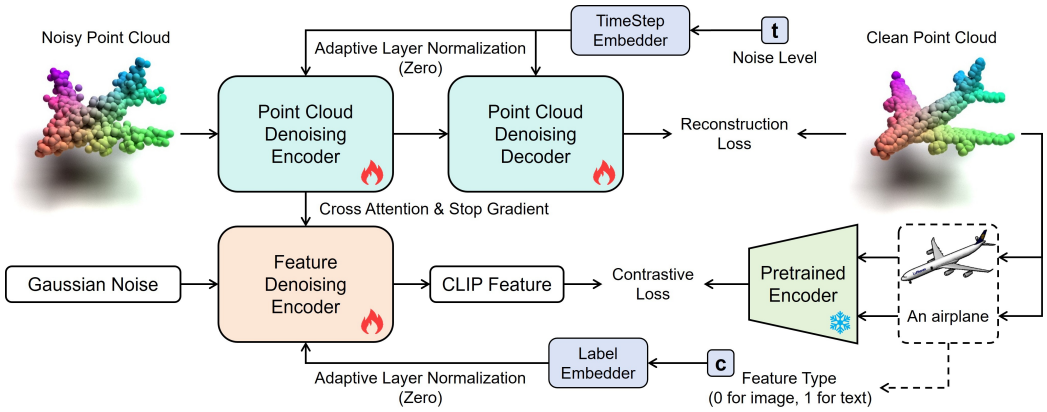


Figure 3. **Pipeline of our proposed *Dual Denoising* framework**. It is built on two key components, namely the point cloud denoising autoencoder (PointDAE) and feature denoising module. The two components are coupled with cross-attention and stop gradient operation, aiming at aligning the point cloud feature with the multi-modal feature space in a robust manner.

present our robust point cloud pre-training algorithm. Considering that only employing robust pre-training to cross-modal distillation offers minimal increase on zero-shot recognition performance, we present the design of an efficient knowledge ensemble approach, which can effectively balance adversarial robustness and recognition performance under zero-shot settings for multi-modal aware point cloud learning. We present the overall pipeline in Figure 3.

3.1. Adversarial attacks under multi-modal setting

The basic idea of an adversarial attack is optimizing the input so that the transformed data does not differ too much from the initial input, while the output of the model can be changed. We take the commonly used C&W attack to illustrate the extension from traditional supervised learning to multi-modal settings. For the adversarial attack on traditional supervised learning, the loss function is usually defined as:

$$\mathcal{L} = (\max_{i \neq t'} \mathcal{Z}(\boldsymbol{X}')_i - \mathcal{Z}(\boldsymbol{X}')_{t'})^+ + \lambda \cdot \|\boldsymbol{X} - \boldsymbol{X}'\|_2, \quad (3)$$

where $\mathbf{X} \in \mathbb{R}^{n\times 3}$ is the clean point cloud, $\mathbf{X}' \in \mathbb{R}^{n\times 3}$ is the optimized adversarial point cloud, $\mathcal{Z}(\boldsymbol{X})_i$ is the i-th element of the output *logits*, and t' is the target class. Here we replace *logits* as the dot-product between the point cloud feature and the set of text CLIP features, instead of the raw output of the model. We keep the other parts consistent with traditional 3D adversarial attacks. We evaluate current models with both zero-shot classification accuracy and zero-shot robust accuracy, as shown in Figure 2. In the upper sub-figure, ReCon (Qi et al., 2023) outperforms ULIP (Xue et al., 2023) on robust accuracy, since it integrates self-supervised learning in the cross-modal distillation stage. In the lower sub-figure, ULIP-2 (Xue et al., 2024) surpasses Uni3D (Zhou et al., 2023) on robust accuracy. This is because the text captions used in ULIP-2 are much more detailed, since it is obtained from BLIP (Li et al., 2022a) inference. Zero-shot accuracy of Uni3D is significantly better, since it scaled the model size to an extremely large level. Zero-shot accuracy of the lower sub-figure also surpasses the upper part because of the scaled dataset for training. However, zero-shot robust accuracy of the lower two models still falls behind ReCon, although both the training dataset and model size have been scaled. I believe that it benefits from the incorporation of self-supervised learning. Based on these experimental observations, it is promising to explore whether incorporating a more powerful self-supervised learning approach can further improve zero-shot adversarial robustness.

3.2. Diffusion-based robust point cloud pre-training

Inspired by the adversarial purification ability of point cloud diffusion model (Sun et al., 2023a), we propose to explore *whether a diffusion-based point cloud pre-training method can improve adversarial robustness*, since a diffusion process can disturb the adversarial property within the input point cloud, and a following denoising process can restore the clean input with high probability (Nie et al., 2022). Although research on diffusion-based point cloud generation has been fully studied (Zhou et al., 2021; Mo et al., 2023), the representation capacity of the point cloud diffusion model has not been fully examined. The commonly used data format of point voxel (Liu et al., 2019) is not appropriate for self-supervised learning due to its inherent sparsity. Therefore, a new pipeline for diffusion-based point cloud pre-training is required for robust 3D perception.

Current self-supervised learning methods mainly take a vanilla ViT architecture as a point cloud backbone for mask reconstruction (Pang et al., 2022; Yu et al., 2022). However, it is not practicable to adapt denoise reconstruction on these pipelines, since point tokens are usually overlapped with each other. To tackle it, we re-design the diffusing process on the point cloud, as shown in Figure 4. The raw

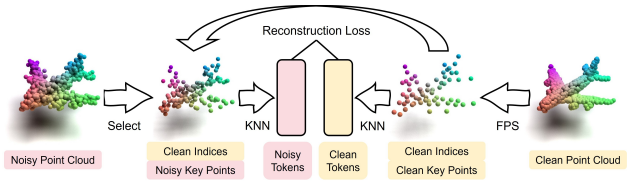


Figure 4. **Redesigned point cloud diffusion and reconstruction process.** FPS is only applied on the clean point cloud, while the resultant support point indices are applied to both diffused and clean ones to ensure the correspondence for reconstruction.

Stage \ Method	Baseline	Vanilla ensemble	Efficient ensemble
Training	F	$E \cdot F$	F
Inference	F	$E \cdot F$	$F \cdot (1 + E/T)$

Table 1. **Theoretical analysis of computation overhead**. The computation cost of our method during training is the same as non-ensemble methods, while during inference, the extra computation is minimal, since in most cases $T >> E$.

point cloud is diffused on the whole point set, while farthest point sampling (FPS) is only conducted on the clean point cloud, and the resultant support point indices are applied to the diffused one. k nearest neighbor (kNN) is applied on both point clouds, respectively, based on the support points. Reconstruction loss is later computed between the correlated point tokens. We take this redesign to ensure the correspondence between diffused and raw point tokens for reconstruction.

The diffusion process follows the conventional design of DDM. We denote the raw data as z_0, diffused data at step t as z_t, which is derived from:

$$z_t = \gamma_t z_0 + \sigma_t \epsilon, \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian noise. We set $\gamma_t \equiv 1$ and σ_t as a simple linear schedule from 0 to s, as they are unessential for representation learning (Chen et al., 2024). Although conventional DDMs for generation usually set the total time step T as a large value such as 1000, we find that a merged time step is better for representation learning. Denoting the merging interval as Δ, the reduced time step for t is $t' = \lfloor t/\Delta \rfloor$, which is then embedded and modulated to the transformer backbones using AdaLN-Zero (Peebles & Xie, 2023).

3.3. Efficient knowledge ensemble via parallel feature denoising

Although the diffusion-based point cloud pre-training can effectively improve zero-shot adversarial robustness when incorporated with cross-modal distillation, the recognition performance could not be enhanced and even degraded, as shown in Figure 1. To tackle it, we need to refine the pipeline so that both zero-shot recognition performance and zero-shot adversarial robustness can be ensured. Inspired by the recent research on knowledge ensemble (Allen-Zhu & Li, 2023), we propose a simple and efficient ensemble approach. As shown in Figure 3, the point cloud feature is obtained in the lower branch, transforming the input token to a CLIP feature via the cross-attention operation, where the local point cloud features obtained in the upper branch serve as the cross-attention context. Unlike previous work (Qi et al., 2023), which takes a learnable input token as the initialized feature, if we take Gaussian noise as the initial feature and train the model to learn feature denoising, then we can conduct efficient knowledge ensemble via initializing multiple input features and averaging the output. We present a theoretical analysis of the computation overhead of different ensemble strategies in Table 1. Assuming that the FLOPs of a sample is F for the baseline model, the point tokens of a sample is T, and the ensemble times is E. Our method does not require training multiple models for ensemble, and the inference cost is only slightly higher than the baseline. Figure 1 demonstrates the validity of our model, in which both zero-shot recognition performance and zero-shot adversarial robustness are largely improved compared to the baselines.

3.4. Dual Denoising

Built on the above motivations and insights, the overall pipeline of our approach is shown in Figure 3. Loss function is formulated as a weighted combination of denoise reconstruction loss $\mathcal{L}_r$ and denoise contrastive loss $\mathcal{L}_c$:

$$\mathcal{L}_{train} = \lambda_t \cdot \mathcal{L}_r^t + \alpha \cdot \mathcal{L}_c, \tag{5}$$

where t is randomly sampled from 0 to $T-1$, T is the total time steps before merging, and λ_t is empirically set as $\lambda_t = 1/(1+\sigma_t^2)$. We train the model to fit different types of CLIP features, since there is a gap between image and language features. The feature type is also embedded and modulated using AdaLN-Zero, similar to the diffusion time step. The difference is that the time step is modulated to the PointDAE encoders, while the feature type is modulated to the feature denoising blocks. Please see Appendix A for a more detailed illustration of our network architecture. Since we align the diffused point cloud with the CLIP feature in the pre-training stage, zero-shot performance can be boosted by applying the same diffusion process on the input. We can denote the model at inference stage as $y_i^t = f_i(x, t, \epsilon_1^t, \epsilon_2)$, where y is the predicted CLIP feature, x is the input point cloud, $t \in \{0, \ldots, T-1\}$ is the diffuse time step, i is the feature type and ϵ_1^t, ϵ_2 are the standard Gaussian noise with the same shape as x, y. Knowledge ensemble can be formulated as:

$$y_i^t = \frac{1}{E} \cdot \sum_{j=1}^{E} f_i(x, t, \epsilon_1^t, \epsilon_{j,2}), \tag{6}$$

Method	Pre-train Dataset	Teacher Model	ModelNet10	ModelNet40	S-OBJ_ONLY	S-OBJ_BG	S-PB_T50_RS
CLIP2Point (Huang et al., 2023)	ShapeNet	CLIP	66.6	49.4	35.5	30.5	23.3
PointCLIP (Zhang et al., 2022)	-	-	30.2	23.8	21.3	19.3	15.4
PointCLIP V2 (Zhu et al., 2023)	-	-	73.1	64.2	50.1	41.2	35.4
ULIP (Xue et al., 2023)	ShapeNet	SLIP	72.8	60.4	49.9	44.2	27.2
ReCon (Qi et al., 2023)	ShapeNet	CLIP	75.6	61.7	43.7	38.6	28.6
Ours	ShapeNet	CLIP	**79.5**	**69.0**	**52.7**	**48.7**	**39.8**
Improvement			+3.9	+4.8	+2.6	+4.5	+4.4

Table 2. **Zero-shot 3D classification accuracy (%) on ModelNet10, ModelNet40 and ScanObjectNN.** We report the performance of other methods with their ***best-performing settings, e.g., visual encoder, projected view number, and textual input***.

where E is the ensemble times. CLIP features with different types can also be ensembled for better performance.

4. Network architecture details

We present the architecture details of *Dual Denoising* in Figure 5. The left part is a trainable basic block of the PointDAE encoder, right is a basic block of the feature denoising module, which is also trainable. The whole PointDAE encoder and feature denoising module are cascaded with these basic blocks, connected with cross-attention within each pair of basic blocks. The PointDAE decoder is also built upon left blocks without a cross-attention connection. For the left blocks, we use AdaLN-Zero (Peebles & Xie, 2023) to regress the scale factor of inputs, the shift of inputs, and the scale factor of outputs for both self-attention and feed-forward networks, while for the right blocks, we only regress the variations of the feed-forward network.

5. Experiments

5.1. Implementation

Currently, there are many solutions for learning 3D representations from pre-trained vision-language models (VLMs) with various configurations. For example, the pre-training datasets include ShapeNet (∼50k+) (Chang et al., 2015) and Objaverse (∼800k+) (Deitke et al., 2023). Prompt templates can be hand-crafted (Zhang et al., 2022; Qi et al., 2023) or synthesized using LLMs (Zhu et al., 2023) or multi-modal LLMs (Xue et al., 2023; Qi et al., 2024). Model scale also varies from millions to billions. To make a fair comparison with existing methods, we use basic experiment settings similar to ReCon (Qi et al., 2023), including: **(1) Dataset**: we take ShapeNet, the most commonly used dataset for 3D pre-training, **(2) Prompt**: we follow PointCLIP (Zhang et al., 2022) to use hand-crafted templates, and **(3) Model**: we use vanilla transformer (Vaswani et al., 2017) encoder blocks with dimension 384 and a tiny PointNet patch embedding module to learn 3D tokens. The PointDAE encoder contains 12 blocks, and the decoder contains 4 blocks. We use Vision Transformer (ViT-B) (Dosovitskiy et al., 2020) and text encoder from CLIP (Radford et al., 2021) as the vision and language teacher, respectively. The image and text encoders are frozen during pre-training, using Smooth l_1-based positive-only distillation loss (Chen & He, 2021). PointDAE uses a reconstruction loss based on l_2 Chamfer-Distance. All the experiments are conducted on a single NVIDIA GeForce RTX 3090.

5.2. Zero-shot point cloud recognition

We take multiple datasets for zero-shot evaluation, following the previous benchmark (Zhu et al., 2023). The evaluation datasets include the real-world object recognition dataset ScanObjectNN and the synthetic object dataset ModelNet. ScanObjectNN (Uy et al., 2019) is one of the most common and challenging 3D datasets containing $\sim$ 15 K real-world objects from 15 categories. We take 3 splits of it, including OBJ_ONLY, OBJ_BG, and PB_T50_RS. ModelNet (Wu et al., 2015) is also a commonly used 3D dataset, containing $\sim$ 12 K CAD objects of 40 (ModelNet40) or 10 (ModelNet10) categories. We use both for evaluation. Following the zero-shot principle, we directly test the classification performance on the full test set without learning from the training set. We compare existing methods under their best settings to fully achieve their performance, following PointCLIP v2 (Zhu et al., 2023). For fair comparison, we do not compare with models that scaled up on model size or pre-training dataset, like ULIP-2(Xue et al., 2024) and Uni3D (Zhou et al., 2023), or methods that adopt external knowledge from LLMs like ShapeLLM (Qi et al., 2024). Zero-shot 3D object classification results are shown in Figure 2. We surpass almost all of the previous methods with a similar configuration and scale. For ModelNet, we achieve 79.5% accuracy on ModelNet10 and 69.0% accuracy on ModelNet40, with an improvement of 3.9% and 4.8%. For ScanObjectNN, we achieve 52.7 accuracy on OBJ_ONLY, 48.7 accuracy on OBJ_BG, and 39.8 accuracy on PB_T50_RS, with an improvement of 2.6%, 4.5%, and 4.4%, respectively. The best performance of our method is obtained with $E = 8, t = 600$ on ModelNet and $E = 8, t = 100$ on ScanObjectNN.

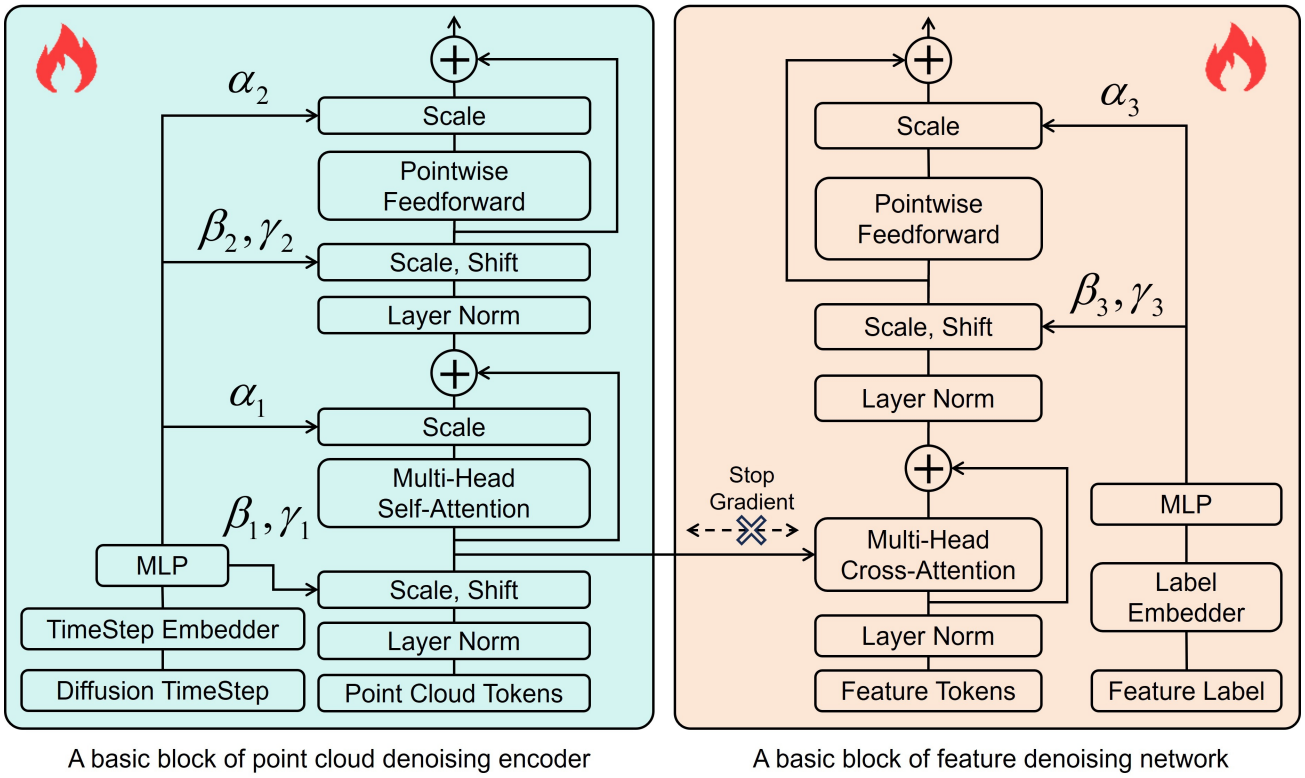


Figure 5. **A basic block of *Dual Denoising*.** PointDAE encoder is built upon the cascade of left blocks, while the feature denoising module is built with right blocks. The PointDAE decoder is also composed of the left blocks without a cross-attention connection.

5.3. Adversarial Robustness under Zero-Shot Settings

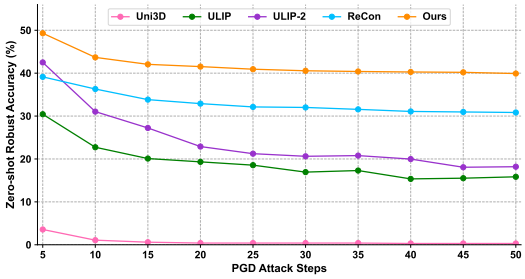
We extend 3D adversarial attack algorithms on a standard classification task to a zero-shot classification task, and evaluate existing cross-modal distillation algorithms for point clouds under the same settings. The gradient-based 3D adversarial attack algorithms can largely reduce the performance of the 3D learning model by adding a slight perturbation to the input. Generally speaking, the 3D adversarial example is computed by raising the *logit* value of a target while minimizing the perturbation on the input through optimization (Xiang et al., 2019). In this way, we can change the model output while keeping the input point cloud almost unchanged. To extend previous methods to CLIP-like zero-shot classification task, we only need to change the *logits* into cosine similarity in Eq 3. We choose the first candidate (the element with the 2nd highest similarity value) as the target, and conduct a targeted adversarial attack (Xiang et al., 2019). We use iterative attack algorithms, including IFGM (Ding et al., 2023), PGD (Sun et al., 2021) and C&W Perturb (Xiang et al., 2019), as they have stronger attack capacity. Please see Appendix B for more details on these algorithms. Experiment results are shown in Table 3. We investigate current methods under these attacks, including ReCon (Qi et al., 2023), ULIP (Xue et al., 2023), ULIP-2 (Xue et al., 2024), and Uni3D (Zhou et al., 2023). ϵ is used for gradient-based optimization, where a larger one means a larger degree of perturbation. We implement IFGM and PGD for 50 steps, while conducting C&W Perturb attack for 10-step binary search and 500 iterations of optimization in each binary search to find the adversarial examples. From Table 3 we can find that our method is more robust under adversarial attacks. We also visualize these methods under PGD attack under different optimization steps with $\epsilon = 0.01$, shown in Figure 6. This also shows the robustness of our method. Notice that we also compare with the methods that scaled up like Uni3D and ULIP-2. When taking adversarial examples as input, our model performs better than them.

5.4. Ablation Study

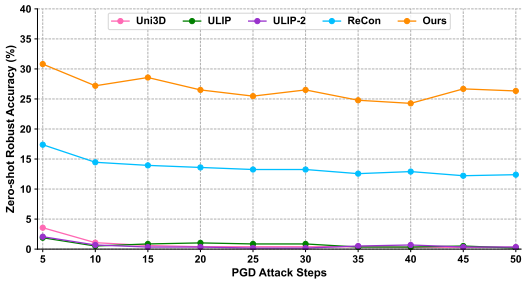
Noise scale. Noise scale plays an important role in our method. If the noise scaling factor s is too small, we can hardly learn a robust 3D representation. If the factor is set too large, the pre-training would be difficult to converge. We first present Figure 7 as a visualization for noise scale and diffusion time step on the point cloud. In Figure 7(a), we change the noise scale and fix the time step to the max steps (999/1000). When s is set to be too large, like 0.10 or 0.12,

Method	Adversarial robustness on zero-shot 3D classification task on ModelNet40 test dataset							
	ReCon (Qi et al., 2023)	ULIP (Xue et al., 2023)	ULIP-2 (Xue et al., 2024)	Uni3D (Zhou et al., 2023)	Ours (E=8, t=100)	Ours (E=8, t=500)	Ours (E=8, t=900)	Ours (E=16, t=500)
Clean point cloud	61.7	60.3	75.6	86.3	68.4	68.8	68.3	68.6
IFGM ($\epsilon = 0.01$)	27.5	17.8	29.1	0.4	33.5	38.5	40.6	**42.3**
IFGM ($\epsilon = 0.025$)	16.9	6.0	8.8	0.2	9.7	**20.7**	20.4	18.8
IFGM ($\epsilon = 0.05$)	7.1	3.0	3.1	0.0	3.5	8.9	**10.1**	8.3
IFGM ($\epsilon = 0.075$)	3.8	2.3	2.8	0.0	2.6	**4.7**	4.4	4.1
PGD ($\epsilon = 0.01$)	31.7	18.0	18.2	0.3	43.8	**48.1**	46.3	47.9
PGD ($\epsilon = 0.025$)	13.3	5.0	5.3	0.0	24.9	31.8	**31.9**	31.8
PGD ($\epsilon = 0.05$)	3.5	2.7	2.1	0.0	13.0	20.2	19.0	**20.7**
PGD ($\epsilon = 0.075$)	0.9	1.1	2.0	0.0	9.7	15.3	**15.6**	15.1
C&W Perturb	6.6	0.0	0.0	0.0	8.1	14.6	**15.0**	14.5

Table 3. **Comparison of adversarial robustness in zero-shot 3D classification task**. The best scores are in bold.



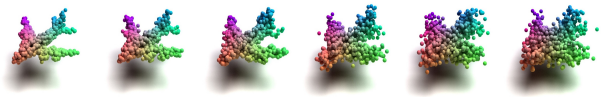
(a) Adversarial Robustness on ModelNet40.



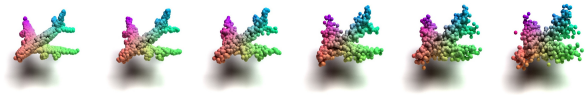
(b) Adversarial Robustness on ScanObjectNN.

Figure 6. **Visualization of adversarial robustness under PGD attack on ModelNet40 and ScanObjectNN.** We use ModelNet40 test set and OBJ_ONLY test set, respectively.

even humans can hardly recognize it. When s is set to be too small, like 0.02, there is little difference from the clean data, as the point cloud itself has already contained certain perturbations. In Figure 7(b), we visualize the point cloud at different time steps with $s = 0.08$. We can generally recognize the shape in most cases. We ablate s in the zero-shot classification task, as shown in Table 4(a). We find that even setting $s = 0$ can still have a decent performance, *i.e.*, forcing the network to predict the input point tokens themselves in the pre-training stage. This indicates that the most critical point for 3D self-supervised learning may not lie in the proxy task. This result is also similar to ReCon (Qi et al., 2023) (61.7%), which also reveals the weak effect of mask reconstruction in this knowledge distillation setting.



(a) Diffused point cloud with scale 0.02, 0.04, 0.06, 0.08, 0.10, and 0.12 from left to right (full time step).



(b) Diffused point cloud at time step 0, 200, 400, 600, 800, and 999 from left to right with scale 0.08.

Figure 7. **Visualization of different noise scaling factor (up) and diffuse time step (down).**

(a) Ablation study on noise scaling factor s.

s	0	0.02	0.04	0.06	0.08	0.1	0.12
Acc	60.4	65.6	67.3	68.1	**69.0**	68.3	67.9

(b) Ablation study on time step merging factor Δ.

Δ	1	10	20	50	100	200	1000
Acc	66.9	67.5	67.3	68.1	68.7	**69.0**	65.4

Table 4. **Ablation study on scaling factor s and merging factor Δ in zero-shot 3D classification task on ModelNet40 test set.** We report the best performance for each case.

Time step merging. We ablate the time step merging interval parameter Δ in zero-shot 3D classification task. Results are shown in Table 4(b). Notice that $\Delta = 1$ means do not conduct time step merging, and $\Delta = 1000$ means do not use AdaLN-Zero in PointDAE. We can see that the time step embedding module plays a relatively important role in representation learning, and time step merging can slightly improve the performance of pre-training.

Stop gradient. Since the two main components of our model (shown in Figure 3) use different types of loss functions, the training process would collapse if parameters are updated by gradients from both losses. A stop gradient can isolate the two components to avoid representation collapse.

stop-grad	ModelNet40	ScanObjNN
✗	56.3	43.2
✔	69.0	52.7

Table 5. **Ablation study on *stop-gradient* operation of *Dual Denoising***. Overall accuracy (%) is reported under the same configuration.

Dual denoising. We ablate the two denoising designs in our model, *i.e.*, PointDAE and feature denoising. We set removing PointDAE as setting $s = 0$, and set removing feature denoising as using learnable tokens to replace the standard Gaussian noise for the input of the feature branch. Results are shown in Figure 8. We can see that PointDAE and feature denoising both play a part in adversarial robustness, while PointDAE is more critical for both representation learning and adversarial robustness.

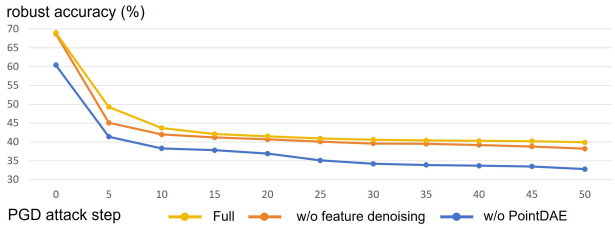


Figure 8. **Ablation study on dual denoising**. We evaluate the robust accuracy under PGD attack with $\epsilon = 0.01$, and report the performance for each model under the best configuration of E and t.

Inference. We conduct comprehensive experiments on ensemble times and diffuse time steps for zero-shot classification and adversarial robustness. Results are shown in Figure 9. The distribution difference for ModelNet40 and ScanObjectNN makes the best configuration vary with each other. We can generally conclude that using a knowledge ensemble and input diffusion can improve the performance and robustness under zero-shot settings. When $t = 0$ and E changes from 1 to 16, performance can be significantly promoted, showing the effectiveness of the feature denoising module.

6. Conclusion

In this paper, we propose to learn a robust 3D representation from pre-trained VLMs like CLIP. Our method is composed of a robust point cloud pre-training algorithm and an efficient ensemble algorithm named parallel feature denoising. Experiments on zero-shot recognition benchmark show that our method can generalize better than others with similar scale and settings, while experiments on zero-shot recognition under adversarial attack show that our method can learn more adversarial robust 3D representations. Ablation studies show the effectiveness of the two modules we proposed, *i.e.*, PointDAE and parallel feature denoising.

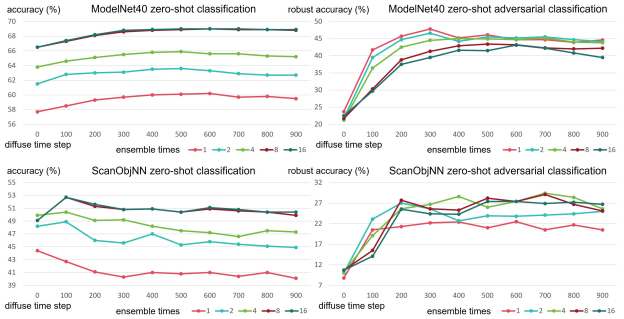


Figure 9. **Ablation study on ensemble times and diffuse time steps in zero-shot classification task on ModelNet40 and ScanObjNN test set**. Adversarial robustness is evaluated using a PGD attack for 10 steps with $\epsilon = 0.01$.

References

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Chen, X., Liu, Z., Xie, S., and He, K. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

Ding, D., Jiang, E., Huang, Y., Zhang, M., Li, W., and Yang, M. Cap: Robust point cloud classification via semantic and structural modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12260–12270, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Hamdi, A., Rojas, S., Thabet, A., and Ghanem, B. 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*, pp. 241–257. Springer, 2020.

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

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022a.

Li, K., Zhang, Z., Zhong, C., and Wang, G. Robust structured declarative classifiers for 3d point clouds: Defending adversarial attacks with implicit gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15294–15304, 2022b.

Li, Y., Fan, H., Hu, R., Feichtenhofer, C., and He, K. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23390–23400, June 2023.

Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., and Su, H. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36, 2024.

Liu, Z., Tang, H., Lin, Y., and Han, S. Point-voxel cnn for efficient 3d deep learning. *Advances in neural information processing systems*, 32, 2019.

Mo, S., Xie, E., Chu, R., Hong, L., Niessner, M., and Li, Z. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. *Advances in neural information processing systems*, 36:67960–67971, 2023.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pp. 529–544. Springer, 2022.

Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., and Anandkumar, A. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, pp. 16805–16827. PMLR, 2022.

Pang, Y., Wang, W., Tay, F. E., Liu, W., Tian, Y., and Yuan, L. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pp. 604–621. Springer, 2022.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

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

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

Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., and Yi, L. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, pp. 28223–28243. PMLR, 2023.

Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Wang, H., Yi, L., and Ma, K. Shapellm: Universal 3d object understanding for embodied interaction. *arXiv preprint arXiv:2402.17766*, 2024.

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

Sun, J., Cao, Y., Choy, C. B., Yu, Z., Anandkumar, A., Mao, Z. M., and Xiao, C. Adversarially robust 3d point cloud recognition using self-supervisions. *Advances in Neural Information Processing Systems*, 34:15498–15512, 2021.

Sun, J., Wang, J., Nie, W., Yu, Z., Mao, Z., and Xiao, C. A critical revisit of adversarial robustness in 3d point cloud recognition with diffusion-driven purification. In *International Conference on Machine Learning*, pp. 33100–33114. PMLR, 2023a.

Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023b.

Uy, M. A., Pham, Q.-H., Hua, B.-S., Nguyen, T., and Yeung, S.-K. 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*, pp. 1588–1597, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Xiang, C., Qi, C. R., and Li, B. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J. C., and Savarese, S. 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*, pp. 1179–1189, 2023.

Xue, L., Yu, N., Zhang, S., Panagopoulou, A., Li, J., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J. C., et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27091–27101, 2024.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., and Lu, J. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19313–19322, 2022.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

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

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

Zhou, H., Chen, D., Liao, J., Chen, K., Dong, X., Liu, K., Zhang, W., Hua, G., and Yu, N. 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*, pp. 10356–10365, 2020.

Zhou, J., Wang, J., Ma, B., Liu, Y.-S., Huang, T., and Wang, X. Uni3d: Exploring unified 3d representation at scale. In *The Twelfth International Conference on Learning Representations*, 2023.

Zhou, L., Du, Y., and Wu, J. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5826–5835, 2021.

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