Any2Point: Empowering Any-modality Large Models for Efficient 3D Understanding

Yiwen Tang[1,2*], Ray Zhang[3*], Jiaming Liu[4*], Zoey Guo[3*], Bin Zhao[1,2†], Zhigang Wang[1], Peng Gao[1], Hongsheng Li[3], Dong Wang[1†], and Xuelong Li[5]

[1]Shanghai AI Laboratory [2]Northwestern Polytechnical University [3]The Chinese University of Hong Kong [4]Peking University [5]TeleAI

Abstract. Large foundation models have recently emerged as a prominent focus of interest, attaining superior performance in widespread scenarios. Due to the scarcity of 3D data, many efforts have been made to adapt pre-trained transformers from vision to 3D domains. However, such 2D-to-3D approaches are still limited, due to the potential loss of spatial geometries and high computation cost. More importantly, their frameworks are mainly designed for 2D models, lacking a general any-to-3D paradigm. In this paper, we introduce **Any2Point**, a parameter-efficient method to empower any-modality large models (vision, language, audio) for 3D understanding. Given a frozen transformer from any source modality, we propose a 3D-to-any (1D or 2D) virtual projection strategy that correlates the input 3D points to the original 1D or 2D positions within the source modality. This mechanism enables us to assign each 3D token with a positional encoding paired with the pre-trained model, which avoids 3D geometry loss caused by the true projection and better motivates the transformer for 3D learning with 1D/2D positional priors. Then, within each transformer block, we insert an any-to-3D guided adapter module for parameter-efficient fine-tuning. The adapter incorporates prior spatial knowledge from the source modality to guide the local feature aggregation of 3D tokens, compelling the semantic adaption of any-modality transformers. We conduct extensive experiments to showcase the effectiveness and efficiency of our method. The code is released at `https://github.com/Ivan-Tang-3D/Any2Point`.

Keywords: Large Foundation Model · Cross-modality Transfer · Parameter Efficient Fine-Tuning

1 Introduction

Driven by the growing volume of model parameters and training data, large foundation models have gained unprecedented attention in a diverse array of domains and tasks. Numerous large models have been pre-trained for natural language

* Equal Contribution.
† Corresponding author.

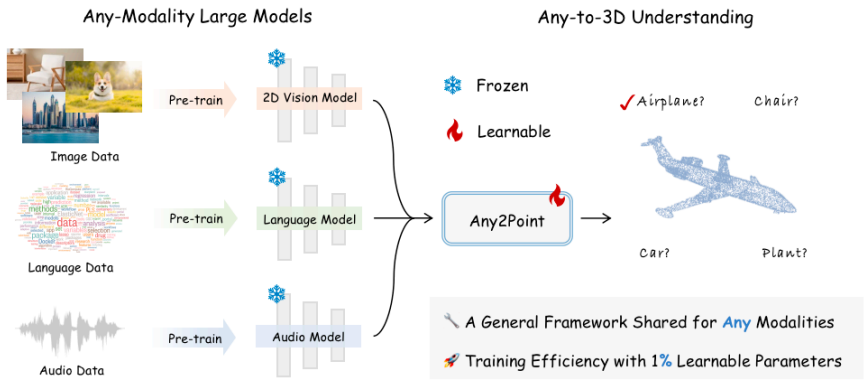

Fig. 1: Overview of Any2Point. We propose a general framework for any-to-3D learning, which is shared for any modalities with parameter-efficient fine-tuning.

process, including BERT [4], T5 [33], GPT series [6, 23], and LLaMA [16, 49], as well as visual understanding like DINOV2 [24], MAE [11, 39, 41], and ViT-22B [3]. Existing works [1, 13, 14, 18, 50] also explore efficient fine-tuning techniques to transfer pre-trained large models to a variety of downstream tasks, consistently achieving excellent performance. Meanwhile, 3D visual understanding [2, 10, 26, 46, 48, 53] is also a significant topic, with its rich geometric representation contributing to the development of many applications (e.g., robotics [17, 21, 30] and autonomous driving [15, 31, 43]). Unfortunately, due to a lack of large-scale 3D data, the efforts towards 3D foundational model are significantly lagging compared to language and 2D vision. Specifically, the acquisition and annotation of high-quality 3D data requires expensive resources and human labor, while synthetic 3D data training falls short of distribution diversity and real-world applications.

Therefore, some previous works have transferred pre-trained models from other modalities (mainly 2D vision) to 3D modality, leveraging sufficient pre-trained knowledge from diverse sources. We categorize existing 2D-to-3D works into two groups. **1) Data modality transformation.** This type of approach involves projecting 3D point clouds into 2D images [38, 48, 54], which are subsequently fed into 2D pre-trained models. Despite the promising performance on downstream tasks, the process of modality transformation inevitably causes the loss of spatial information in 3D data, hindering the full potential for 3D understanding. **2) Cross-modality knowledge distillation.** These approaches involve the pre-training knowledge transfer from 2D or vision-language models to a newly trained 3D model [5, 42, 51]. They are not only required to forward propagate both the 2D and 3D models during training, but also highly rely on the large-scale paired 2D-3D data. This leads to substantial computation costs and data engineering, limiting their ability for efficient implementation. Besides the aforementioned issues, more importantly, current methods mostly focus on the model adaption from 2D vision to 3D point clouds, rather than a shared

methodology for other modalities. Therefore, we pose a question: *can we develop a general any-to-3D paradigm that empowers any-modality large models for efficient and effective point cloud understanding?*

To address this issue, we propose Any2Point, a unified any-to-3D framework that transfers any 1D (language) or 2D (image/audio) large models to 3D domains with Parameter-Efficient Fine-Tuning (PEFT), as shown in Fig. 1. Different from prior methods, our Any2Point avoids the point cloud projection, thereby mitigating the 3D information loss, and directly fine-tunes pre-trained models from source modalities, which saves resources by knowledge distillation. Specifically, given an any-modality pre-trained transformer, we first introduce a 3D-to-any (1D or 2D) virtual projection mechanism. This mechanism establishes a positional mapping between the input 3D points and their virtually projected 1D lines or 2D planes. This enables us to encode 3D coordinates using the original positional embeddings of the source modality of pre-trained large models. In this way, we no longer need to conduct a true projection losing 3D geometries, while better promoting the pre-trained transformer to acquire 3D features with their original 1D/2D positional priors. Then, for each transformer block, we insert an any-to-3D guided adapter module for PEFT. This adapter leverages the 1D/2D spatial guidance to aggregate the local semantics of 3D tokens, facilitating fine-grained feature interaction. Afterward, we perform an adaptive ensemble for the 3D features guided by different 1D/2D priors, which attains superior 3D representations.

Extensive experiments across various tasks demonstrate that our Any2Point framework achieves superior performance compared to current 3D pre-trained models, while utilizing only 1.0% of the trainable parameters. Using the pre-trained CLIP Text Encoder [32], Any2Point fine-tunes only 0.8M parameters and attains 91.9% on ScanObjectNN [36], outperforming the previous state-of-the-art (SOTA) 3D pre-trained model by +1.3%, and 94.3% on ModelNet40 [40]. Furthermore, Any2Point also achieves comparable results and efficiency by utilizing other pre-trained models [7,8,19,24,35] of different modalities, including 2D vision, language, and audio, validating the robustness of our approach. The contributions of our paper are as follows:

- To enable a general any-to-3D transferring framework, we propose Any2Point, which empowers any-modality pre-trained large models (e.g., 2D vision, language, and audio) for efficient 3D understanding.
- We introduce two techniques, i.e., 3D-to-any virtual projection and any-to-3D guided adapter, to effectively overcome the issues within current methods, such as 3D geometry loss and excessive resource cost.
- Any2Point achieves superior performance compared to previous SOTA 3D pre-trained models across various tasks. Notably, these competitive results remain consistent regardless of leveraging pre-trained models from different modalities, such as 2D vision, language, and audio.

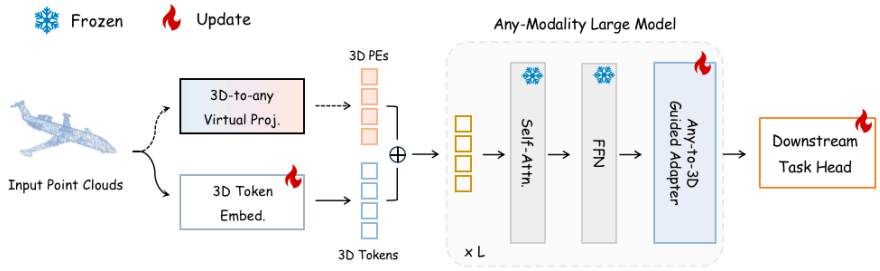

Fig. 2: Overall Pipeline of Any2Point. For efficiently fine-tuning Any-modality pre-trained models, our Any2Point framework contains two components: a 3D-to-any Virtual Projection, which pairs the pre-trained positional encodings with 3D tokens to avoid the 3D geometric information loss, and a Any-to-3D Guided Adapter to effectively grasp local structures.

2 Any2Point

In Sec. 2.1, we first provide a paradigm overview of Any2Point, including the problem definition and network architecture. Then, in Sec. 2.2 and Sec. 2.3, we respectively elaborate on the methodologies of our proposed two techniques for adapting any-modality large models for 3D domains.

2.1 Method Overview

Problem Definition. Given a pre-trained transformer from any modality, e.g., vision, language, and audio, our objective is to empower it with 3D understanding capabilities in an effective and efficient manner. Instead of employing full fine-tuning on 3D data, we seek a parameter-efficient solution with the source transformers frozen, since their large-scale parameters might cause high computation cost and over-fitting issues on the limited 3D dataset. We generally divide the source models into two categories according to their pre-training data dimension, denoted as 1D and 2D transformers. The 1D transformers are specialized in processing sequential data, exemplified by language models like RoBERTa [19], T5 [33], and CLIP's text encoder [32]. The 2D transformers are expert at 2D spatial data, including vision models, e.g., DINOv2 [24] and DeiT [35], and audio models, e.g., ImageBind Audio Encoder [7] and SSAST [8].

Model Pipeline. The overall paradigm of Any2Point is depicted in Fig. 2. To encode the input point cloud, we discard the original embedding modules in source transformers, e.g., tokenizers in 1D language models and convolutions in 2D vision/audio models, and employ a 3D mini-network for point cloud tokenization. On top of this, the encoded 3D tokens are fed first into a 3D-to-any virtual projection module for positional encoding, and then into the frozen 1D/2D transformer with any-to-3D guided adapters. The former mechanism aims to assign each 3D token with positional information within the source modality, and the

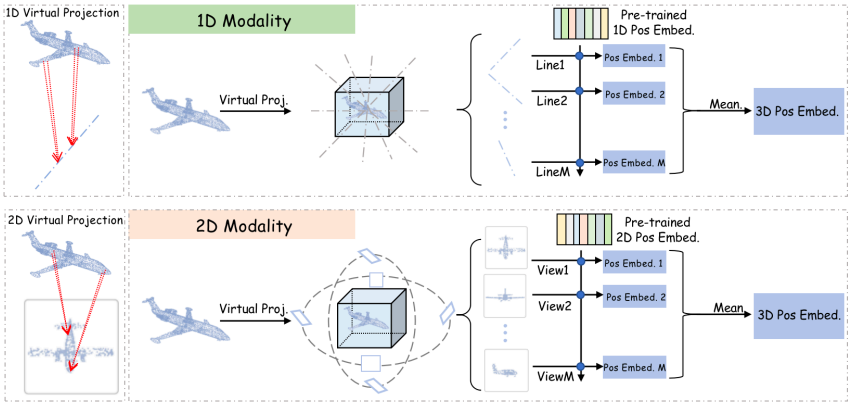

Fig. 3: 3D-to-any Virtual Projection. To prevent the loss of 3D geometric information, the module assigns 3D tokens with the positional encodings that are paired with the pre-trained model.

latter is designed for adaptive 1D/2D-guided 3D representation learning, which we will detail in the following sections. Note that, as the source transformers are kept frozen, only the initial tokenization network and the inserted adapters are learnable for parameter-efficient fine-tuning.

2.2 3D-to-any Virtual Projection

Many current 2D-to-3D methods [38, 48, 54] project 3D point clouds into multi-view images to meet the input modality of pre-trained 2D models. This dimension reduction process potentially leads to the information loss of 3D geometries and deep measurements, enabling insufficient 3D feature encoding. In addition, these approaches are merely validated on the large models within 2D images, without considering other modalities like language and audio. Therefore, we propose a 3D-to-any virtual projection strategy that mitigates the geometric loss, and is generalizable to any 1D/2D pre-trained models, as shown in Fig. 3.

Tokenization in 3D Space. To avoid any information degradation, we directly tokenize the input point cloud within the 3D space for the subsequent 1D/2D transformer. Specifically, we employ a 3D mini-network containing small-scale parameters, which is a lighter-weight variant of Point-PN [52, 53] . The tokenization process involves Farthest Point Sampling (FPS) [26] for point number downsampling, k-Nearest Neighbor (k-NN) algorithm for local aggregation, and learnable linear layers for feature encoding. After this, we transform the raw point clouds into high-dimensional vectors, obtaining N 3D tokens as $\{T_i\}_{i=1}^{N}$, with $\{p_i^{3D}\}_{i=1}^{N}$ denoting their 3D coordinates.

Motivations for Virtual Projection. Positional encodings (PEs) serve as the only indicator for positional information to the transformer model, since

the inner attention mechanism is permutation-invariant, treating every token at different orders all the same. Therefore, a straightforward way for 1D/2D transformers to comprehend 3D positional information is to integrate new 3D PEs with 3D tokens. However, the source transformers are pre-trained paired with their original PEs in 1D/2D space, which leads to semantic discrepancy between the frozen 1D/2D weights and newly learned 3D PEs. To address this issue, we virtually project 3D tokens into the source modality, and obtain the corresponding 1D/2D PEs for better aligning with the transformers.

3D-to-2D Virtual Projection. For 2D transformers in 2D vision and audio modalities, we virtually project each 3D coordinate, e.g., p_i^{3D}, into M views, deriving the corresponding 2D coordinates as $\{p_{ij}^{2D}\}_{j=1}^{M}$. The M different perspectives are capable of providing diverse positional relations within 2D space. We adopt a simple projection in PointCLIP [48] without learnable parameters. Importantly, we do not truly produce the projected multi-view images, but only aim to obtain the virtual 2D positions. Then, according to the original 2D PEs within pre-trained transformers, we assign each 3D token, e.g., T_i, with M different PEs, denoted as $\{\mathrm{PE}^{2D}(p_{ij}^{2D})\}_{j=1}^{M}$.

3D-to-1D Virtual Projection. Similarly, for 1D transformers in language modality, we virtually project the 3D coordinates into different 1D lines. To align the number with 2D modality, we also select M lines passing through the center of the point cloud with M uniform rotation angles. For simplicity, we suppose the point cloud center as the origin, the unit direction vectors of M lines as $\{\boldsymbol{v}_j^{1D}\}_{j=1}^{M}$, and the point coordinate, p_i^{3D}, vectorized as $\boldsymbol{p}_i^{3D}$. Then, the 1D coordinate of point i in line j is formulated by the dot production of

$$p_{ij}^{1D} = \boldsymbol{v}_j^{1D} \cdot \boldsymbol{p}_i^{3D}, \tag{1}$$

denoting the projected length. In this way, we refer to the original 1D PEs, and assign each 3D token, e.g., T_i, with M different PEs as $\{\mathrm{PE}^{1D}(p_{ij}^{1D})\}_{j=1}^{M}$.

Encoding 3D Positions in 1D/2D PEs. After acquiring the corresponding 1D/2D PEs, we average them as an overall positional indicator, and incorporate it with the 3D token, e.g., T_i, by

$$T_i^{in} = T_i + \frac{1}{M}\sum_{j=1}^{M} \mathrm{PE}^{1D/2D}(p_{ij}^{1D/2D}). \tag{2}$$

With this approach, we inject sufficient positional information of the source modality into 3D tokens to better collaborate with the frozen transformer, while mitigating the information loss of the true projection.

2.3 Any-to-3D Guided Adapter

Different from existing distillation-based methods [9, 51] training a new 3D network, we directly feed the encoded 3D tokens $\{T_{ij}^{in}\}_{i=1}^{N}$ to the pre-trained

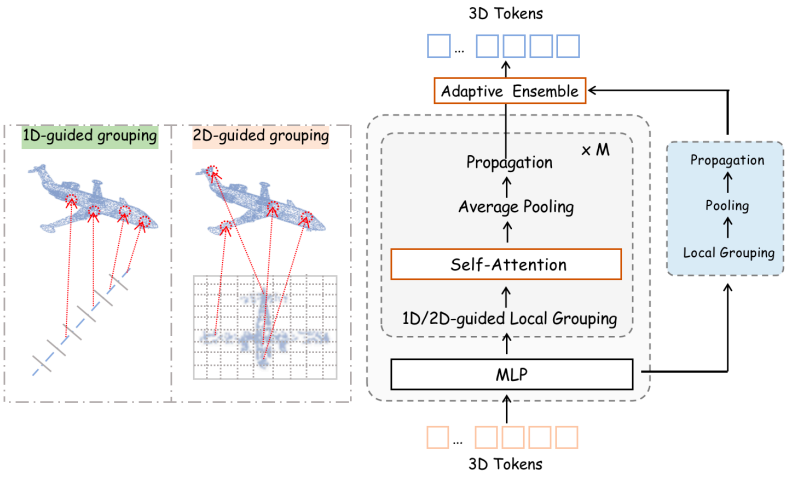

Fig. 4: Any-to-3D Guided Adapter. Inserted into every transformer block, the adapter leverages the 1D/2D-guided Local Aggregation module to capture 3D local semantics and utilizes the Adaptive Any-to-3D Ensemble to obtain high-quality features.

1D/2D transformer. Although the PEs of 3D tokens have been aligned with the source model, the entirely frozen weights pre-trained by other modalities are still restricted to learning superior 3D representations. Considering this, we introduce a learnable any-to-3D guided adapter within each transformer block, as shown in Fig. 4. The adapters are inserted after the Feed-Forward Networks (FFNs), and further incorporate 1D/2D-prior knowledge for parameter-efficient fine-tuning.

Motivations for Inserting Adapters. The self-attention mechanisms within source transformers normally focus on long-range token interaction in global contexts, which lacks local feature extraction. However, the detailed spatial geometries are also significant for the fine-grained understanding of 3D shapes. To complement the gap, we utilize the proposed adapter layers for specifically capturing 3D semantics within local neighborhoods. In addition, as the source transformers are powered by 1D/2D PEs as discussed above, the naive FPS and k-NN for 3D local grouping might cause positional discrepancy. Therefore, we further design a 1D/2D-guided aggregation strategy and an adaptive any-to-3D ensemble approach for robust 3D fine-grained encoding.

1D/2D-guided Local Aggregation. Within the adapter, we first group 3D tokens into different local neighborhoods guided by 1D/2D positional priors, which better align the adopted 1D/2D PEs. For M different views/lines, we conduct M concurrent local aggregation process to make the best of different projection perspectives. Specifically, for 2D transformers, we divide each virtually projected image, e.g., the j-th view, into uniform local 2D patches, and group the 3D to-

kens within the same patch into a neighborhood, according to their 2D positions $\{p_{ij}^{2D}\}_{i=1}^{N}$. For 1D transformers, we similarly divide each virtually projected line, e.g., the j-th direction, into uniform local 1D segments, and group the 3D tokens within different segments referring to their 1D positions $\{p_{ij}^{1D}\}_{i=1}^{N}$. On top of this, we adopt a self-attention layer for 3D tokens within each 1D/2D neighborhoods, performing local feature interaction guided by 1D/2D priors. Then we employ the operations of pooling and propagation to propagate the local aggregated feature to every points within the same neighborhood.

Adaptive Any-to-3D Ensemble. After the parallel local aggregation, we obtain M sets of 3D tokens, each representing a 2D view or 1D line. As different projection perspectives normally showcase different significance for 3D representations, we propose an adaptive any-to-3D ensemble approach to aggregate the M features for each token. We denote the i-th 3D token with M sets of features at this stage as $\{F_{ij}\}_{j=1}^{M}$. To properly indicate the relative importance of each view/line, we additionally employ a 3D feature transformation branch independent of the M 2D-guided local aggregation. This non-parametric branch only contains the local grouping in 3D space, feature average pooling within local groups, and propagation operations, converting the 3D token before the adapter into a feature baseline for adaptive ensemble, denoted as B_i. Then, we calculate the relative weights for different views/lines by the cosine similarity, and finally aggregate their features to obtain the final output as

$$T_i^{out} = \frac{1}{M} \sum_{j=1}^{M} \mathrm{Sim}(B_i, F_{ij}). \tag{3}$$

With the ensemble strategy, we integrate M different features with dynamic weights, enabling the adapter to adaptively determine which view/line is more critical, contributing to high-quality adapted features.

3 Experiments

3.1 Experimental Settings

ScanObjectNN. The ScanObjectNN dataset [36] consists of real-world 3D object scans, categorized into 15 distinct classes. We select the most challenging PB-T50-RS split to test the performance of the Any2Point framework without the voting strategy. For all models, we employ the AdamW optimizer [20] and the CosineAnnealing scheduler. The initial learning rate is set to 5e-4, with a weight decay factor of 0.05. We fine-tune the model for 300 epochs with a batch size of 32. For data augmentation, we use Random scaling, translation, and rotation. For language, 2D vision, and audio modalities, we respectively select the CLIP Text Encoder [32], DINO V2 [24], and ImageBind Audio Encoder [7] as pre-trained models. For these three models, the transformer architecture is the same: a 12-block encoder with 768 feature channels and 1,024 input point number. The hyperparameter M in the 3D-to-any Virtual Projection is set to 6 with

Table 1: Comparisons on accuracy with previous methods on 3D classification datasets. We report the pre-training modality (Pre-train), the number of learnable parameters (#Param) on the "PB-T50-RS" split of ScanObjectNN (SCAN.) and ModelNet40 (MN.).† indicates utilizing the voting strategy.

Method	Pre-train	#Param(M)	SCAN.(%)	MN.(%)
Point-NN [52]	N/A	0.0	64.9	81.8
PointNet [26]	N/A	3.5	68.0	89.2
PointNet++ [27]	N/A	1.5	77.9	90.7
DGCNN [37]	N/A	1.8	78.1	92.9
PointMLP [22]	N/A	12.6	85.4	94.1
Point-PN [52]	N/A	0.8	87.1	93.8
PointNeXt [29]	N/A	1.4	87.7	94.0
Point-BERT [44]	3D	22.1	83.1	92.7
w/ Point-PEFT [34]	3D	0.6	85.0	93.4
Point-MAE [25]	3D	22.1	85.2	93.2
Point-M2AE [47]	3D	15.3	86.4	93.4
P2P-HorNet [38]	2D	1.2	89.3	94.0†
ACT [5]	3D+2D	22.1	88.2	93.7
w/ IDPT [45]	3D+2D	1.7	87.7	94.0†
I2P-MAE [51]	3D+2D	12.9	90.1	93.7
ReCon [28]	3D+2D+Language	43.6	90.6	94.1
Any2Point	Audio	**0.8**	**87.0**	**92.7**
	2D	**0.8**	**87.7**	**93.2**
	Language	**0.9**	**91.9**	**94.3**

identical angles for the Any-Modality Transformers. To match the shape of the original PEs within pre-trained models, we virtually project 3D points into a 1D line segment of length 77 with a line size of 2 in the language modality; a 2D plane measuring 512x512 with a patch size of 26 in the 2D vision modality; and a 2D plane sized 192x304 with a patch size of 16 in the audio modality.

ModelNet40. The ModelNet40 dataset [40] consists of 40 categories of synthesized 3D CAD models, with 9,843 training samples and 2,468 test samples. In our experiments on ModelNet40, we adopt the same fine-tuning settings and the same pre-trained models as in ScanObjectNN. For data augmentation, we utilize default random scaling and translation. Notably, during the testing process, we do not employ the voting strategy.

In this section, we conduct extensive experiments on the ScanObjectNN [36] and ModelNet40 [40] datasets. We first introduce the fine-tuning settings and implementation details in Sec. 3.1. Then, in Sec. 3.2, we present the main experiment of transferring any-modality large models (language, 2D image and audio) to 3D classification tasks. Finally, in Sec. 3.3, we conduct ablation studies to evaluate each component within our proposed Any2Point framework.

3.2 Quantitative Analysis

The results are shown in Tab. 1. It is observed that: (i) On the 3D real-world object dataset ScanObjectNN, the Any2Point framework achieves 91.9%, 87.7%,

Table 2: Ablation Study on Different PEFT Methods. We report the number of learnable parameters (#P) and classification accuracy(%) of CLIP-Text (1D.) and DINO V2 (2D.) on the "PB-T50-RS" split of the ScanObjectNN dataset.

Method	#P(M)	1D.(%)	2D.(%)
Full Fine-Tuning	86.3	79.9	85.3
Prompt Tuning [14]	0.4	89.1	86.4
Adapter Tuning [12]	0.4	89.6	85.9
LoRA [13]	0.9	86.3	85.1
Any2Point	**0.8**	**91.9**	**87.7**

Table 3: Ablation Study on Main Components. To validate the effectiveness of 3D-to-any Virtual Projection (V.P.) and Any-to-3D Guided Adapter (G.A.).

3D-to-any V.P.	Any-to-3D G.A.	#P(M)	1D.(%)	2D.(%)
-	-	0.3	88.7	86.1
✓	-	0.3	89.3	86.6
-	✓	0.8	90.9	87.6
✓	✓	**0.8**	**91.9**	**87.7**

and 87.0% accuracy based on Language (CLIP-Text), 2D Vision (DINO V2-B), and Audio (ImageBind-Audio) modalities, respectively. Compared to the previous SOTA method (ReCon), 1D language pre-trained Any2Point achieves a 1.3% improvement with only 0.9M learnable parameters. For the 2D (Vision/Audio) modalities, Any2Point significantly outperforms Point-M2AE, which is the SOTA method pre-trained only on 3D datasets, by 0.6% and 1.3%, respectively. This reveals that our framework is capable of fully exploiting pre-trained knowledge from other modalities to solve 3D recognition tasks. (ii) On the 3D synthetic object dataset ModelNet40, across the Language, 2D Vision, and Audio modalities, our Any2Point framework attains 94.3%, 93.2%, and 92.7%. Our framework exclusively utilizes one pre-trained model in the 1D language modality, achieving a 0.2% improvement over the previous SOTA method (ReCon), and reducing 42.7M learnable parameters. For 2D modalities, Any2Point demonstrates performance on par with models pre-trained exclusively on 3D datasets. (iii) Surprisingly, whether on the ScanObjectNN or the ModelNet40 dataset, the Any2Point framework maintains a performance trend where 1D modality (language) outperforms 2D modalities (image and audio). Large language models provide abundant spatial and semantic information in low-dimensional spaces to assist in 3D learning. This trend is further validated in the upcoming Sec. 3.3.

3.3 Ablation Study

In this section, we conduct extensive ablation studies to explore the effectiveness of different components within our Any2Point framework. We adopt CLIP-Text (1D) and DINO V2 (2D) as the pre-trained transformer, and report the classification accuracy (%) on the "PB-T50-RS" split of the ScanObjectNN dataset.

Table 4: Ablation Study on 3D-to-any Virtual Projection. Sinusoidal, Learnable and 3D-to-any V.P. refer to sinusoidal positional encoding, learnable positional encoding and 3D-to-any Virtual Projection.

Sinusoidal	Learnable	3D-to-any V.P.	**1D.(%)**	**2D.(%)**
-	-	-	90.9	87.6
✓	-	-	87.4	86.0
-	✓	-	90.5	86.5
-	-	✓	**91.9**	**87.7**

Table 5: Ablation Study on Any-to-3D Guided Adapter. To validate the effectiveness of 1D/2D-guided Local Aggregation (L.A.) and Adaptive (Ada.) Any-to-3D Ensemble (Ens.).

1D/2D-guided L.A.	Ada. Any-to-3D Ens.	**#P(M)**	**1D.(%)**	**2D.(%)**
-	-	0.25	89.3	86.6
✓	-	0.8	90.2	86.8
✓	✓	**0.8**	**91.9**	**87.7**

Table 6: More Results on ScanObjectNN.

Method	**Pre-train**	**Model**	**#Param(M)**	**SCAN.(%)**
Any2Point	Audio	SSAST [8]	**0.8**	**87.1**
	2D	DeiT [35]	**0.8**	**87.3**
	Language	RoBERTa [19]	**0.9**	**89.7**

Comparison with traditional PEFT methods. As demonstrated in Tab. 2, our Any-to-3D Guided Adapter significantly outperforms traditional PEFT techniques when utilizing pre-trained models from 1D or 2D modalities. In comparison to Prompt Tuning [14], it achieves improvements of 2.8% and 1.3%; compared to Adapter Tuning [12], it achieves improvements of 2.3% and 1.8%; and in contrast to Low-Rank Adaptation (LoRA) [13], it achieves improvements of 5.6% and 2.6%, respectively. The experimental results demonstrate that our Any-to-3D Guided Adapter can efficiently mine and integrate pre-trained knowledge from other modalities to understand the semantics of 3D objects. Unlike other methods, our framework leverages 1D/2D spatial guidance to aggregate the local semantics of 3D tokens, capturing the local fine-grained information of 3D objects.

Effectiveness of Main Components. As shown in Tab. 3, to substantiate the efficacy of our proposed methods, we conducted ablation experiments by progressively incorporating each component into the baseline. The first row indicates the baseline configuration, which consists of the 3D tokenizer, the pre-trained transformer, and the task head, with updates applied only to the tokenizer and head. Introducing the 3D-to-any Virtual Projection resulted in performance improvements to 89.3% in the 1D modality and 86.6% in the 2D modality. This

suggests that using virtual projection, rather than true projection, helps mitigate the loss of 3D spatial information caused by modality conversion. Following the inclusion of the Any-to-3D Guided Adapter, performance in the 1D modality surged to 90.9%, while in the 2D modality, it rose to 87.6%, with a focus on local structures leading to greater improvements. Introducing both aforementioned methods simultaneously led to a surge in performance to 91.9% in the 1D modality and a rise to 87.7% in the 2D modality, effectively showcasing the effectiveness of our comprehensive framework.

Effects of 3D-to-any Virtual Projection. In Tab. 4, we investigated the effects of employing different positional encoding methods on the Any2Point framework. The first row indicates the absence of any positional encoding. Introducing sinusoidal positional encoding or learnable positional encoding led to a certain degree of performance degradation. This is due to the conflict between the newly introduced positional information and the inherent semantics within the source modality transformer. On the other hand, employing 3D-to-any Virtual Projection resulted in respective improvements of 1.0% and 0.1% accuracy. The results demonstrate that using original 1D/2D positional priors can promote the pre-trained transformer to acquire 3D features.

Components of Any-to-3D Guided Adapter. As shown in Tab. 5, we conduct ablation experiments by incrementally adding components to the Any-to-3D Guided Adapter. The first row signifies the baseline adapter, consisting of only an MLP with bottleneck layers. By incorporating 1D/2D-guided Local Aggregation, composed of local aggregation in 1D/2D spaces, self-attention interactions, pooling, and propagation, our approach achieves performance gains of 0.9% and 0.2%. Leveraging the positional priors from the pre-trained model facilitates mining fine-grained 3D structural information from different perspectives. The Adaptive Any-to-3D Ensemble brings further improvements of 1.7% and 0.9% for 1D and 2D modalities, effectively integrating parallel features in accordance with 3D structural features. The experiments demonstrate the effectiveness of each component in our Any-to-3D Guided Adapter to gather 3D local geometric information, complementing the global attention in the pre-trained model.

More Results on Performance Trend. To further validate our previous findings that the Any2Point framework, based on 1D Language pre-trained models, significantly outperforms those based on 2D modalities (Vision/Audio) in the 3D object recognition task, we conduct additional experiments in Tab. 6. On the "PB-T50-RS" split of ScanObjectNN dataset, we select RoBERTa (1D), DeiT (2D Vision), and SSAST (Audio) as the pre-trained models, with fine-tuning settings consistent with our previous experiments. These models achieve performance of 89.7%, 87.3%, and 87.1%, respectively. The performance trend across modalities is observed: 1D language $>$ 2D Vision $>$ 2D Audio. We suspect that due to the pre-training data, large language models possess stronger semantic

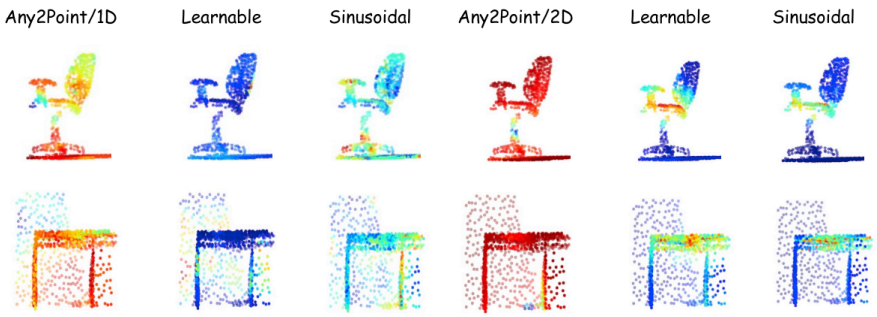

Fig. 5: Visualization of Different Positional Encoding Methods. For the 1D/2D modalities, we visualize the attention scores of the **[CLS]** token to other point cloud tokens, utilizing sinusoidal positional encoding, learnable positional encoding, and 3D-to-any Virtual Projection. **The red color indicates higher values.**

information compared to other modalities, which is beneficial for the deep understanding of different 3D objects.

4 Visualization

In this section, we opt to validate the efficacy of the proposed 3D-to-any Virtual Projection and the Any-to-3D Guided Adapter by visualizing on the ScanObjectNN test set, utilizing the CLIP-Text Encoder (1D) [32] and DINO V2 (2D) [24].

4.1 Different Positional Encoding Methods

Our 3D-to-any Virtual Projection fully exploits the positional encoding paired with the pre-trained model, injecting the source modality spatial knowledge into the 3D tokens during fine-tuning. In Fig. 5, when using sinusoidal positional encodings, learnable positional encodings, and our 3D-to-any Virtual Projection respectively, we visualize the attention scores of the [CLS] token to other point cloud tokens. As illustrated, for the 1D language modality, learnable positional encodings grasp useless information. After applying the commonly used sinusoidal positional encodings in Large Language Models, they fail to capture the critical 3D semantics. However, our method focuses more on the salient object parts, such as the armrests and wheels of chairs, and the legs of tables. For the 2D visual modality, learnable encodings are slightly better than sinusoidal positional encodings, as 2D pre-trained models mainly adopt the learnable encoding method. Meanwhile, our method directly recognizes the whole object and its key parts, for example, giving high weights to the chair's backrest.

4.2 Effects of Any-to-3D Guided Adapter

The Any-to-3D Guided Adapter captures the 3D fine-grained information through interactions within the local regions of the source modality. In Fig. 6, we visualize

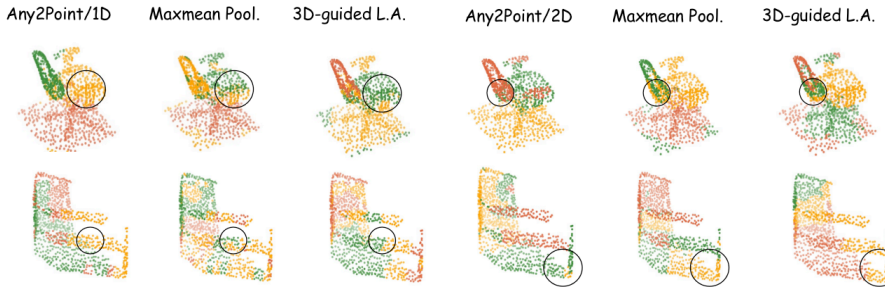

Fig. 6: Visualization of Effects of Any-to-3D Guided Adapter. For 1D/2D modalities, we visualize the clusters of the similarities between the **[CLS]** token and other point features, with the number of clusters set to 3. It is conducted for the complete Any-to-3D Guided Adapter, for replacing the Adaptive Any-to-3D Ensemble with maxmean pooling (Maxmean Pool.), and for performing local aggregation solely based on 3D positions (3D-guided L.A.).

the clustering results of the similarities between the [CLS] token and other point token features, utilizing the complete Any-to-3D Guided Adapter, replacing the Adaptive Any-to-3D Ensemble with maxmean pooling, and further only using 3D positional information. As shown, for simple objects like chairs (1^{st} row), our method effectively distinguishes between the chair's backrest, armrests, seat, and wheels, whereas removing components fails to capture the differences between key parts. For more challenging objects like shelves (2^{nd} row), removing any components leads to semantic confusion of the object, while our approach clearly differentiates the shelf's base, middle layer, and backrest. These experiments indicate that each component within the Adapter effectively utilizes the positional information from different modalities to promote the 3D structure extraction.

5 Conclusion

In conclusion, our paper proposes Any2Point to enable a general any-to-3D transferring framework, empowering any-modality pre-trained large models (e.g., 2D vision, language, and audio) for efficient 3D understanding. Within Any2Point framework, we introduce two techniques, named 3D-to-any virtual projection and any-to-3D guided adapter, to extract 3D structure knowledge while efficiently fine-tuning pre-trained models. This enables us to overcome issues within current methods, such as 3D geometry loss and excessive resource cost. Our extensive experiments across various tasks demonstrate the superior performance and efficiency of Any2Point compared to previous SOTA 3D pre-trained models, achieving remarkable results with only a fraction of the trainable parameters.

Acknowledgements

This work is partially supported by the Shanghai AI Laboratory, National Key R&D Program of China (2022ZD0160101), the National Natural Science Foundation of China (62376222), and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

1. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems **35**, 16664–16678 (2022)
2. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
3. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. In: International Conference on Machine Learning. pp. 7480–7512. PMLR (2023)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Dong, R., Qi, Z., Zhang, L., Zhang, J., Sun, J., Ge, Z., Yi, L., Ma, K.: Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? arXiv preprint arXiv:2212.08320 (2022)
6. Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines **30**, 681–694 (2020)
7. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2023)
8. Gong, Y., Lai, C.I., Chung, Y.A., Glass, J.: Ssast: Self-supervised audio spectrogram transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 10699–10709 (2022)
9. Guo, Z., Zhang, R., Qiu, L., Li, X., Heng, P.A.: Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. arXiv preprint arXiv:2302.14007 (2023)
10. Guo, Z., Tang, Y., Zhang, R., Wang, D., Wang, Z., Zhao, B., Li, X.: Viewrefer: Grasp the multi-view knowledge for 3d visual grounding with gpt and prototype guidance. arXiv preprint arXiv:2303.16894 (2023)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
12. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

14. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
15. Jing, L., Xue, Y., Yan, X., Zheng, C., Wang, D., Zhang, R., Wang, Z., Fang, H., Zhao, B., Li, Z.: X4d-sceneformer: Enhanced scene understanding on 4d point cloud videos through cross-modal knowledge transfer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2670–2678 (2024)
16. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
17. Li, X., Zhang, M., Geng, Y., Geng, H., Long, Y., Shen, Y., Zhang, R., Liu, J., Dong, H.: Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. arXiv preprint arXiv:2312.16217 (2023)
18. Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602 (2021)
19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Ma, C., Liu, Y.L., Wang, Z., Liu, W., Liu, X., Wang, Z.: Humannerf-se: A simple yet effective approach to animate humannerf with diverse poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1460–1470 (2024)
22. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual mlp framework. arXiv preprint arXiv:2202.07123 (2022)
23. OpenAI: GPT-4 technical report (2023)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
25. Pang, Y., Wang, W., Tay, F.E., Liu, W., Tian, Y., Yuan, L.: Masked autoencoders for point cloud self-supervised learning. In: European conference on computer vision. pp. 604–621. Springer (2022)
26. Qi, C.R., Su, H., Mo, K., 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 (2017)
27. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
28. Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., Yi, L.: Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. arXiv preprint arXiv:2302.02318 (2023)
29. Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., Ghanem, B.: Pointnext: Revisiting pointnet++ with improved training and scaling strategies. Advances in Neural Information Processing Systems **35**, 23192–23204 (2022)
30. Qu, D., Chen, Q., Zhang, P., Gao, X., Zhao, B., Wang, D., Li, X.: Livescene: Language embedding interactive radiance fields for physical scene rendering and control. arXiv preprint arXiv:2406.16038 (2024)

31. Qu, D., Yan, C., Wang, D., Yin, J., Chen, Q., Xu, D., Zhang, Y., Zhao, B., Li, X.: Implicit event-rgbd neural slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19584–19594 (2024)
32. 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)
33. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
34. Tang, Y., Zhang, R., Guo, Z., Ma, X., Zhao, B., Wang, Z., Wang, D., Li, X.: Point-peft: Parameter-efficient fine-tuning for 3d pre-trained models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5171–5179 (2024)
35. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
36. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., 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)
37. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (tog) **38**(5), 1–12 (2019)
38. Wang, Z., Yu, X., Rao, Y., Zhou, J., Lu, J.: P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. Advances in neural information processing systems **35**, 14388–14402 (2022)
39. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14668–14678 (2022)
40. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., 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)
41. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)
42. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., 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)
43. Yang, S., Liu, J., Zhang, R., Pan, M., Guo, Z., Li, X., Chen, Z., Gao, P., Guo, Y., Zhang, S.: Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. arXiv preprint arXiv:2312.14074 (2023)
44. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., 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)
45. Zha, Y., Wang, J., Dai, T., Chen, B., Wang, Z., Xia, S.T.: Instance-aware dynamic prompt tuning for pre-trained point cloud models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)

46. Zhang, D., Li, C., Zhang, R., Xie, S., Xue, W., Xie, X., Zhang, S.: Fm-ov3d: Foundation model-based cross-modal knowledge blending for open-vocabulary 3d detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 16723–16731 (2024)
47. Zhang, R., Guo, Z., Gao, P., Fang, R., Zhao, B., Wang, D., Qiao, Y., Li, H.: Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. Advances in neural information processing systems **35**, 27061–27074 (2022)
48. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., 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)
49. Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
50. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. arXiv preprint arXiv:2305.03048 (2023)
51. Zhang, R., Wang, L., Qiao, Y., Gao, P., Li, H.: Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21769–21780 (2023)
52. Zhang, R., Wang, L., Wang, Y., Gao, P., Li, H., Shi, J.: Starting from non-parametric networks for 3d point cloud analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5344–5353 (2023)
53. Zhu, X., Zhang, R., He, B., Guo, Z., Liu, J., Xiao, H., Fu, C., Dong, H., Gao, P.: No time to train: Empowering non-parametric networks for few-shot 3d scene segmentation. arXiv preprint arXiv:2404.04050 (2024)
54. Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., 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)