Multimodal Masked Point Distillation for 3D Representation Learning

Anonymous authors
Paper under double-blind review

Abstract

We propose a two-stage pre-training approach using point clouds for a diverse set of 3D understanding tasks. In the first stage, we pre-train the 3D encoder to acquire knowledge from the other modalities such as vision and language. This stage aligns 3D representations with multiple modalities by leveraging several pre-trained foundation models, unlike the current cross-modal paradigm that typically uses only a single pre-trained model. In the second stage, the pre-training approach is improved upon masked point modeling by global-local feature distillation of semantic 3D embeddings and token shuffling approach. These techniques enable the model to focus on the 3D modality while leveraging the multimodal information associated with the point clouds. This pre-training approach is model-agnostic and can be applied to any 3D transformer encoder. We conduct extensive experiments on a wide range of 3D understanding tasks, from synthetic and real-world object recognition to indoor semantic segmentation and object detection, achieving state-of-the-art results. For instance, on the ScanObjectNN variants, our approach achieves **96.1%**, **94.2%** and **91.2%** accuracy using multi-scale 3D encoder proposed in Point-M2AE.

1 Introduction

Self-supervised learning (SSL) approaches have paved the way for the creation of foundation models (Bommasani et al., 2022) that leverage abundant unlabeled data, enabling adaptation to various downstream tasks with only a small amount of labeled data. This paradigm has led to remarkable success across multiple domains, including NLP (Radford & Narasimhan, 2018; Devlin et al., 2019; Wei et al., 2023; Ouyang et al., 2022), 2D vision (He et al., 2022; 2020; Chen et al., 2020), and vision-language (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023). Nevertheless, the success of these models fundamentally depends on the availability of large-scale pre-training data.

Recently, many successful pre-training strategies have been adapted for 3D point cloud understanding. However, compared to NLP and vision, the 3D domain suffers from a lack of data at scale, as collecting and annotating point cloud data is challenging. This phenomenon, often referred to as a *data desert* (Dong et al., 2022), can hinder the scaling and development of 3D foundation models. Contrastive learning based approaches (Afham et al., 2022; Xie et al., 2020; Sanghi, 2020) that learn instance-discriminative representations by maximizing the similarity of different views of the point cloud, fail to generalize when there is a lack of pre-training data (Qi et al., 2023). On the other hand, generative pre-training (Pang et al., 2022; Zhang et al., 2022a; Zha et al., 2023a) which follows the masked autoencoding (MAE) in 2D, can bring significant improvement with less amount of data. However, the representations learned by these approaches may not fully capture the holistic representations due to the intricate geometric and spatial relationships in 3D data. Since point clouds are unstructured and unordered set of coordinates, there is no straightforward or exact supervision method like one-to-one mean squared error (MSE) loss between the ground truth and the reconstructed one. MAE-based approaches typically use Chamfer distance as a pre-training loss to compute an approximate matching between two sets of points. This allows for an initial alignment of point clouds, though it still has limitations in capturing fine-grained details or handling unordered nature of point clouds effectively. Additionally, there are several shortcomings to using Chamfer distance as a loss function for point cloud completion (Wu et al., 2021; Liu et al., 2019; Huang et al., 2023a).

Another line of work is cross-modal learning (Xue et al., 2023; Dong et al., 2022; Zhang et al., 2022b; Zhu et al., 2022; Xu et al., 2022; Tang et al., 2024; Qian et al., 2024) which leverages other modalities like image, depth maps and text associated with point cloud data, and has shown impressive performance. In this context, most approaches utilize pre-trained foundation models to align multimodal features with point cloud features, resulting in enhanced 3D representations complemented by multimodal capabilities. However, these methods primarily rely on a single vision or language foundation model, which can limit the feature representation of the multimodal data. CLIP (Radford et al., 2021), which has been mostly used in cross-modal learning, has demonstrated remarkable performance in a zero-shot setting, but it under-performs compared to vision transformer (Dosovitskiy et al., 2021) as reported by ReCon (Qi et al., 2023). CLIP focuses more on global representations, which may not be optimal for dense tasks such as semantic segmentation, as noted by (Li et al., 2022). We hypothesize that by leveraging unique properties from multiple foundation models can further enhance the multimodal representations. For example, compared to CLIP, DINOv2 (Oquab et al., 2023) has emerged with strong spatial features for dense tasks. Segment Anything (SAM) (Kirillov et al., 2023) has shown remarkable performance on segmentation due to dense feature representations. Furthermore, existing cross-modal methods (Huang et al., 2023b; Qi et al., 2023; Xue et al., 2023) rely on limited prompt templates based on category names, which both constrains scalability and reduces the diversity of text descriptions available for pre-training.

To address these issues, we propose a two-stage pre-training approach that can learn holistic 3D representations using rich and well-aligned multimodal data. In the first stage, we pre-train a 3D encoder using multimodal data by leveraging multiple vision and language foundation models using contrastive objective. We follow ULIP-2 (Xue et al., 2024) to generate diverse language descriptions of point cloud data using BLIP-2 (Li et al., 2023), which encapsulates all expressible information about 3D data. We also render 2D images and depth from a fixed set of viewpoints to generate 2D modality input. This stage harnesses insights from other multimodal data and learns enhanced 3D representations by aligning the features of the 3D modality with those of other multimodal data. Next, we continue training the 3D encoder using only point cloud data with improved masked point modeling. The model is tasked with predicting both global embeddings and token-wise embeddings from the first stage based on unmasked input point clouds. We also apply random shuffling to the encoder's output tokens before passing them to the decoder to avoid learning shortcuts. By focusing solely on point cloud data, this stage emphasizes local geometric and spatial relationships while retaining the semantic knowledge learned in the first stage. Our approach is model-agnostic and can be used with any MAE based point cloud architecture such as Point-MAE (Pang et al., 2022), Point-M2AE (Zhang et al., 2022a) etc. We conduct extensive experimental study on a broader range of 3D understanding tasks across five benchmark datasets that includes 3D object recognition, 3D semantic and part segmentation and 3D object detection. We also provide comprehensive ablation study to verify the effectiveness of different components of our approach. The main contributions of our paper are as follows:

- We propose a two-stage pre-training approach that learns semantically enriched 3D representations with high generalization capability. The first stage leverages knowledge from multiple foundation models using multimodal data, while the second stage enhances masked point modeling through local-global feature distillation.

- Our approach consistently outperforms competing methods on wide variety of 3D tasks. For example, in zero-shot setting, our first-stage pre-training surpasses ULIP2 (Xue et al., 2024) by **+1.2%** on ModelNet40 dataset (Wu et al., 2015). More notably, on the real-world ScanObjectNN dataset Uy et al. (2019), our approach significantly outperforms the Point-MAE baseline by **+5.18%**, **+5.44%**, and **+4.3%** on all three variants, respectively.

- We also demonstrate that our approach is compatible with any MAE-based architecture. When applied to Point-M2AE (Zhang et al., 2022a), a multi-scale hierarchical point cloud encoding framework, our method achieves state-of-the-art (SOTA) performance across a range of downstream 3D understanding tasks.

2 Related Work

Pre-training for Point Clouds. There has been significant effort in developing pre-training methods for point clouds, following the recent success of self-supervised approaches in 2D vision and the NLP domain. Generally, 3D pre-training approaches can be divided into two categories: contrastive learning (Afham et al., 2022; Xie et al., 2020; Sanghi, 2020; Huang et al., 2021; Liu et al., 2022a) and generative learning (Yu et al., 2022; Pang et al., 2022; Zhang et al., 2022a; Zha et al., 2023a; Zhang et al., 2023). Contrastive learning aims to learn instance-discriminative representations by maximizing the similarity of different views of the same sample while minimizing the similarity of views from different samples. These views are typically generated using data augmentations. In 3D, PointContrast (Xie et al., 2020) uses geometric transformations to generate multiple views for learning discriminative representations. In contrast, generative pre-training focuses on learning representations by masking portions of the input and reconstructing them either in the input space or the latent space. Point-BERT (Yu et al., 2022), inspired by BERT (Devlin et al., 2019), pre-trains a transformer model by predicting the masked point tokens using tokenizer with discrete Variational AutoEncoder (dVAE). Point-MAE (Pang et al., 2022) follows the MAE (He et al., 2022) paradigm to reconstruct masked point patches in the point space. Point-M2AE (Zhang et al., 2022a) introduces a hierarchical multi-scale transformer to pre-train the model using masked point modeling. Point-FEMAE (Zha et al., 2023a) proposes local enhancement modules and incorporate global random and local block masking to learn compact 3D representations. Pix4Point (Qian et al., 2024) differs from cross-modal contrastive/generative methods; it adapts image-pretrained transformers to point clouds via weight transfer. Mamba3D (Han et al., 2024) proposes a state-space backbone with local pooling to better capture local geometry, achieving high accuracy with improved efficiency. In contrast, our work focuses on transformer-based architectures, and our two-stage framework is orthogonal and can be combined with alternative backbones such as Mamba3D.

Cross-Modal Pre-training. In addition to contrastive and generative pre-training, there have been efforts to learn representations by acquiring knowledge from other modalities such image, text etc. In the realm of 3D, CrossPoint (Afham et al., 2022) employs contrastive learning to align point clouds with their rendered images while ensuring invariance to spatial transformations. PointCLIP (Zhang et al., 2022b) projects point clouds into depth maps and leverages CLIP (Radford et al., 2021) to bridge 2D image–text representations with 3D data. ACT (Dong et al., 2022) introduces masked point modeling with feature distillation. The target features are generated by pre-training a cross-modal autoencoder that acquires knowledge from other modalities by leveraging a pre-trained vision transformer. OpenShape (Liu et al., 2023)introduces large-scale multimodal 3D pretraining with millions of shapes, leveraging image–text–point correspondences to improve generalization across 3D recognition tasks. Uni3D (Zhou et al., 2024) introduces a scalable 3D foundation model by initializing a ViT from 2D pretraining and aligning point cloud features with image–text representations. By scaling up to one billion parameters, Uni3D achieves strong performance on zero-shot, few-shot, and open-world 3D tasks. ReCon (Qi et al., 2023) combines both contrastive and generative learning to incorporate knowledge from other modalities. ReCon++ (Qi et al., 2024) extends ReCon with stronger objectives and larger-scale training. I2P-MAE (Zhang et al., 2023) integrates 2D guided masking and 2D visual features, in addition to 3D coordinate reconstruction, by leveraging 2D pre-trained models. Point-Bind (Guo et al., 2023b) constructs a joint embedding space by aligning point clouds with image, text, video and audio via contrastive learning under ImageBind. It shows strong zero-shot, any-to-3D generation, and open-world understanding capabilities, further illustrating the power of multi-modal supervision in 3D learning. Multi-View masked learner (Chen et al., 2025) introduces a masked learner that projects point clouds into multi-view 2D features (using pose), with a two-stage teacher-student scheme and MSMH attention.

Knowledge Distillation. The concept of training a smaller network (student) from a large network (teacher) was first proposed for model compression (Bucilu et al., 2006). The goal is to transfer the "dark knowledge" from the teacher model to the student model. (Hinton et al., 2015) extends this idea for deep neural network by using the logits of the teacher model to distill knowledge in the student model. Following this breakthrough, a plethora of work (Furlanello et al., 2018; Cho & Hariharan, 2019; Mirzadeh et al., 2019; Yang et al., 2018) have expanded on the use of logits or soft labels of the teacher model to guide the student model. Another line of work in this area is the feature distillation (Heo et al., 2019b; Huang & Wang, 2019; Heo et al., 2019a; Park et al., 2019; Kim et al., 2018; Peng et al., 2019), which focuses on transferring

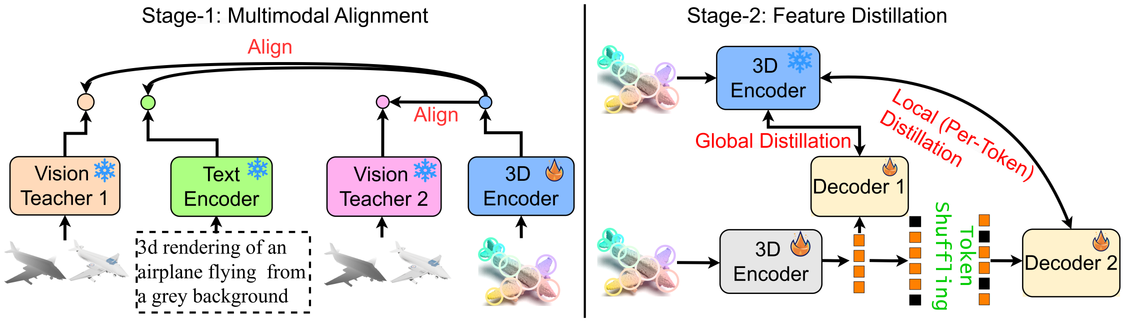
Stage-1: Multimodal Alignment
Align
Align
Vision Teacher 1
Text Encoder
Vision Teacher 2
3D Encoder
3d rendering of an airplane flying from a grey background
Stage-2: Feature Distillation
3D Encoder
Global Distillation
Local (Per-Token) Distillation
Decoder 1
3D Encoder
Token shuffling
Decoder 2

Figure 1: **Overview of our two-stage pre-training.** In the first stage, we pre-train the 3D encoder by aligning multimodal features with point cloud features using various vision and language foundation models. In the second stage, we further pre-train the 3D encoder with enhanced masked point modeling. The Stage-1 3D encoder generates global and token-level embeddings, which then serve as targets for the second-stage distillation.

the features of the teacher model directly into the backbone of the student model. In 3D understanding, ACT (Dong et al., 2022) uses a cross-modal teacher as the target for masked point modeling in the 3D student model. The cross-modal teacher is trained to acquire knowledge from other modalities through self-supervised prompt tuning.

3 Approach

Our goal is to learn semantically and geometrically enriched 3D representations from multiple modalities, such as images, depth maps, and text, by leveraging pre-trained vision and language models. We propose a simple yet effective two-stage pre-training approach that captures both multimodal 3D representations and local geometric features. Given a batch of multimodal data, we first pre-train the 3D encoder by aligning these modalities with point clouds using contrastive learning with vision and language pre-trained models. In the second stage, we fine-tune the 3D encoder solely on point clouds with enhanced masked point modeling, effectively distilling the knowledge acquired in the first stage. The 3D encoder from the first stage serves as a cross-modal teacher during this process. Figure 1 illustrates the pre-training pipeline of our approach.

3.1 Stage-1: Multimodal Representation Alignment

In the first stage, we conduct contrastive learning to align images, depth and text with point clouds. Specifically, given a point cloud $\mathcal{P} = \{\mathbf{p}_i | i = 1, 2, \dots, N\} \in \mathbb{R}^{N\times3}$ with N coordinates, a randomly sampled 2D rendered image or associated depth $\mathcal{I}$, and a language description $\mathcal{T}$ (e.g., a description generated by BLIP-2), we extract modality-specific features using the modality-specific encoders. We pass the point clouds $\mathcal{P}$ to 3D point cloud encoder E_{P} to extract 3D representations $\mathbf{f}^{\mathrm{P}} = E_{\mathrm{P}}(\mathcal{P})$. Next, we extract text features $\mathbf{f}^{\mathrm{T}} = E_{\mathrm{T}}(\mathcal{T})$ using text encoder E_{T} and image features $\mathbf{f}^{\mathrm{I}} = \{\mathbf{f}^{(\mathrm{I,m})} | m = 1, 2, \dots, N\}$ where $\mathbf{f}^{\mathrm{I}}_m = E^{\mathrm{I}}_m(\mathcal{I})$ using N different vision foundation models E^{I}_m. In experiments, we use two vision foundation models i.e., $N = 2$. To match the feature dimension of the teacher model, we employ a teacher-specific projection layer. Following prior works (Xue et al., 2023; 2024; Radford et al., 2021), we apply a symmetric cross-entropy loss over the similarity score for multimodal alignment. More specifically, 3D-to-text alignment can be formulated as:

$$\mathcal{L}^{P\leftrightarrow T} = -\frac{1}{2}\sum_i \Big(\log \frac{\exp(\mathbf{f}^{\mathrm{P}}_i \mathbf{f}^{\mathrm{T}}_i/\tau)}{\sum_j \exp(\mathbf{f}^{\mathrm{P}}_i \mathbf{f}^{\mathrm{T}}_j/\tau)} + \log \frac{\exp(\mathbf{f}^{\mathrm{T}}_i \mathbf{f}^{\mathrm{P}}_i/\tau)}{\sum_j \exp(\mathbf{f}^{\mathrm{T}}_i \mathbf{f}^{\mathrm{P}}_j/\tau)}\Big). \tag{1}$$

Similarly, the 3D-to-image alignment for a vision teacher m can be written as:

$$\mathcal{L}_m^{P\leftrightarrow I} = -\frac{1}{2}\sum_i \Big(\log \frac{\exp(\mathbf{f}_i^{\mathrm{P}}\mathbf{f}_i^{(\mathrm{I,m})}/\tau)}{\sum_j \exp(\mathbf{f}_i^{\mathrm{P}}\mathbf{f}_j^{(\mathrm{I,m})}/\tau)} + \log \frac{\exp(\mathbf{f}_i^{(\mathrm{I,m})}\mathbf{f}_i^{\mathrm{P}}/\tau)}{\sum_j \exp(\mathbf{f}_i^{(\mathrm{I,m})}\mathbf{f}_j^{\mathrm{P}}/\tau)}\Big), \tag{2}$$

Overall, the pre-training objective for stage-1 is to minimize the sum of two contrastive alignment losses, which is:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}^{P\leftrightarrow T} + \sum_m \mathcal{L}_m^{P\leftrightarrow I}. \tag{3}$$

Although Stage-1 follows the multi-modal alignment paradigm proposed in recent works (Xue et al., 2023; 2024; Guo et al., 2023b), our key contribution lies in leveraging multiple vision foundation models analogous to multi-teacher distillation within this stage and integrating them into a novel two-stage framework, where Stage-1 establishes a strong foundation for Stage-2's masked point modeling.

3.2 Stage-2: Masked Point Modeling

While the first stage captures semantic relationships through multimodal alignment, pre-training on multimodal data using contrastive learning may not fully capture the complex nature of the 3D data, as it tends to overlook local geometric features and spatial relationships, focusing primarily only on the global similarities between modalities. To address this, our second-stage pre-training focuses on learning representations solely from point cloud data. We build upon the masked point modeling techniques (Dong et al., 2022; Zhang et al., 2022a), while ensuring that the model retains the semantics and other representations learned in the first stage.

In this stage, we initialize the second-stage 3D encoder with the first-stage model's weights and then pre-train it on point cloud data using an enhanced masked point modeling approach. To prevent the model from learning shortcuts, we incorporate a token shuffling scheme. Additionally, we introduce both global and local token-wise feature distillation strategies, which allow the second-stage model to effectively leverage the knowledge acquired during the first stage. Specifically, the second-stage model learns to predict all tokens from the first-stage model based on the masked point cloud.

Local Distillation. Since we initialized the model from the first stage, it may begin to develop shortcuts for decoding the unmasked tokens while predicting the masked ones. This could simplify the task, making it easier than predicting all the tokens simultaneously. To mitigate this issue, we first concatenate the visible and masked tokens, randomly shuffle the token sequence, and then pass it to the decoder, where positional embeddings are added after the shuffling. This operation encourages the model to avoid learning shortcuts for unmasked tokens. The approach is analogous to the Jigsaw pretext task (Noroozi & Favaro, 2017) in SSL, which the decoder tries to solve for the unmasked tokens while predicting the masked ones. This masked point modeling, referred to as local feature distillation, minimizes the negative cosine similarity between the output tokens of the first-stage and second-stage models.

$$\mathcal{L}_{local} = -\sum_{i=1}^{N_t} \mathbf{1} - \frac{\mathbf{z}_i^1 \cdot \mathbf{z}_i^2}{\|\mathbf{z}_i^1\| \cdot \|\mathbf{z}_i^2\|} \tag{4}$$

where N_t is the total number of tokens, and $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$ represents the i^{th} token from the first-stage model and the decoder-2 of the second-stage model, respectively.

Global Distillation. We find that our second-stage model slightly under-performs compared to the first-stage model on synthetic data when using only local distillation. This can likely be attributed that to

Table 1: **Few-shot classification results on ModelNet40.** Overall mean accuracy (%) with standard deviation (without voting) is reported.

Methods	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
Supervised Learning Only				
PointNet Qi et al. (2017a)	52.0±3.8	57.8±4.9	46.6±4.3	35.2±4.8
DGCNN Wang et al. (2019)	31.6±2.8	40.8±4.6	19.9±2.1	16.9±1.5
OcCo Wang et al. (2021)	90.6±2.8	92.5±1.9	82.9±1.3	86.5±2.2
with Single-Modal Self-Supervised Representation Learning				
Point-BERT Yu et al. (2022)	94.6±3.1	96.3±2.7	91.0±5.4	92.7±5.1
MaskPoint Liu et al. (2022b)	95.0±3.7	97.2±1.7	91.4±4.0	93.4±3.5
Point-MAE Pang et al. (2022)	96.3±2.5	97.8±1.8	92.6±4.1	95.0±3.0
Point-M2AE Zhang et al. (2022a)	96.8±1.8	98.3±1.4	92.3±4.5	95.0±3.0
PointGPT Chen et al. (2024)	96.8±2.0	98.6±1.1	92.6±4.6	95.2±3.4
Point-FEMAE Zha et al. (2023a)	97.2±1.9	98.6±1.3	94.0±3.3	95.8±2.8
with Cross-Modal Self-Supervised Representation Learning				
ACT Dong et al. (2022)	96.8±2.3	98.0±1.4	93.3±4.0	95.6±2.8
Joint-MAE Guo et al. (2023a)	96.7±2.2	97.9±1.9	92.6±3.7	95.1±2.6
I2P-MAE Zhang et al. (2023)	97.0±1.8	98.3±1.3	92.6±5.0	95.5±3.0
TAP Wang et al. (2023)	97.3±1.8	97.8±1.9	93.1±2.6	95.8±1.0
ReCon Qi et al. (2023)	97.3±1.9	98.9±1.2	93.3±3.9	95.8±3.0
Ours (Point-M2AE)	**97.6±1.7**	**99.0±0.6**	**94.1±3.0**	**96.2±2.6**
Improve (over Point-M2AE)	+0.8	+0.7	+1.8	+1.2
Mamba3D Han et al. (2024)	96.4±2.2	98.2±1.2	92.4±4.1	95.2±2.9

catastrophic forgetting (McCloskey & Cohen, 1989) during the two-stage pre-training. To address this issue, we distill the global 3D representations from the first stage using the full point cloud. Given the visible tokens from the second-stage 3D encoder, we employ a decoder network (e.g., transformer blocks or MLP) to output a global embedding. We do not apply token shuffling or add positional embedding for this decoder. We then align the second-stage global embedding with the global 3D embedding from the first stage by minimizing the negative cosine similarity.

$$\mathcal{L}_{global} = -\sum_{i} \mathbf{1} - \frac{\mathbf{f}^{\mathrm{P}}_{(1,i)} \cdot \mathbf{f}^{\mathrm{P}}_{(2,i)}}{\|\mathbf{f}^{\mathrm{P}}_{(1,i)}\| \cdot \|\mathbf{f}^{\mathrm{P}}_{(2,i)}\|} \tag{5}$$

where $\mathbf{f}^{\mathrm{P}}_{(1,i)}$ and $\mathbf{f}^{\mathrm{P}}_{(2,i)}$ represents the global embeddings from the first-stage model and the second-stage model respectively.

Overall, the pre-training objective of the second stage is the sum of two feature distillation losses, given by:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{local} + \mathcal{L}_{global}. \tag{6}$$

4 Experiments

We conduct experiments across a broader range of 3D understanding tasks to evaluate the generalizability of our pre-trained model. Specifically, we evaluate the model on 3D object recognition, indoor segmentation, part segmentation and 3D object detection. Please see supplementary materials for implementation details.

Table 2: **Zero-shot 3D classification on ModelNet40.** We report top-1 and top-5 accuracy. A "✓" in the Manual captions means that the models use text prompt with category names to generate language description for the 3D data while "✗" means the opposite.

Model	**Pre-train dataset**	**Pre-train method**	**Manual captions?**	**ModelNet40 top-1**	**top-5**
PointCLIP (Zhang et al., 2022b)	–	–	–	19.3	34.8
PointCLIPv2 (Zhu et al., 2022)	–	–	–	63.6	85.0
ReCon (Qi et al., 2023)	ShapeNet	ReCon (Qi et al., 2023)	✓	61.2	78.1
CLIP2Point (Huang et al., 2023b)	ShapeNet	CLIP2Point (Huang et al., 2023b)	✗	49.5	81.2
Point-BERT (Yu et al., 2022)	ShapeNet	OpenShape (Liu et al., 2023)	✓	70.3	91.3
Point-BERT (Yu et al., 2022)	ShapeNet	ULIP Xue et al. (2023)	✓	60.4	84.0
		Ours	✓	**65.8**	**93.5**
		ULIP-2 Xue et al. (2024)	✗	70.0	89.7
		Ours	✗	**71.2**	**93.0**
	Objaverse(no LVIS) + ShapeNet	ULIP Xue et al. (2023)	✓	68.6	86.4
	Objaverse + ShapeNet	ULIP Xue et al. (2023)	✓	69.6	85.9
PointM2AE (Zhang et al., 2022a)	ShapeNet	Ours	✗	**73.4**	**94.5**

4.1 Results on 3D Object Classification

We first evaluate the performance of our approach on 3D object classification. Both the Point-MAE and Point-M2AE based 3D encoders are used for fine-tuning. Following previous works (Qi et al., 2023; Dong et al., 2022), we adopt the same transfer learning protocols: Full, MLP-3, and MLP.

3D Synthetic Object Recognition We evaluate our approach on ModelNet40 (Wu et al., 2015), which contains 12311 3D CAD objects spanning 40 categories. During fine-tuning, we apply the *ScaleandTranslate* data augmentation. The results are reported in Table 3, both with and without the voting strategy. It can be seen from the table that our approach sets the state-of-art for these strategies. Our pre-trained model, using multi-scale architecture (Zhang et al., 2022a), achieves **94.3%** accuracy under full protocol, improving by **+0.9** over Point-M2AE performance. Using the standard plain transformer, our approach achieves **94.0%** accuracy, improving by **+0.8** over Point-MAE baseline. Furthermore, compared to ACT (Dong et al., 2022), a two-stage pre-training approach, our method outperforms it by **+0.8%** and **+1.1%** on both strategies, respectively. Finally, for other transfer learning protocols, our approach consistently outperforms the other baselines, including ReCon (Qi et al., 2023).

3D Real-World Object Recognition Next, we evaluate our approach on ScanObjectNN (Uy et al., 2019) dataset, which consists of 15k 3D objects. Following previous works (Qi et al., 2023; Dong et al., 2022), we evaluate on three variants of this dataset: OBJ-BG, OBJ-ONLY, and PB-T50-RS. We only apply *rotation* augmentation during fine-tuning as done in (Qi et al., 2023; Dong et al., 2022). The results are reported in the Table 3 using both standard transformer and multi-scale transformer architectures. As shown in the table, our approach sets new state-of-the results across all variants of ScanObjectNN. More specifically, the standard transformer fine-tuned model improves by **5.18%**, **5.44%**, and **4.92%** over Point-MAE across all three variants, respectively. Similarly, compared to Point-M2AE, our approach shows improvements of **4.88%**, **5.41%**, and **4.77%** on these variants. Furthermore, compared to ReCon (Qi et al., 2023), our approach show improvement of **+0.9%** on OBJ_BG variant of ScanObjectNN. Furthermore, under the MLP-3 protocol, our approach with the standard transformer model outperforms ReCon by **+1.08%** on the ScanObjectNN hard variant, demonstrating that the frozen features are more representative and discriminative compared to other SSL methods. This observation holds consistent across the other two variants. Lastly, we can draw the same conclusion for MLP protocol.

Few-shot Classification We opt for ModelNet40 for few-shot classification experiments, following previous works (Dong et al., 2022; Qi et al., 2023; Pang et al., 2022) in 3D pre-training. We use the standard "n-way", "m-shot" configuration, where "n" represents the number of randomly sampled class and "m" represents the number of training examples for each class. We report the mean accuracy with standard deviation by

Table 3: **Classification accuracy (%) on ScanObjectNN and ModelNet40.** We report the overall accuracy (%) on three variants of ScanObjectNN. For ModelNet40, we report the overall accuracy (%) for both with and without voting. "#P" means the model's parameters.

Methods	#P	ScanObjectNN				ModelNet40		
		Input	OBJ_BG	OBJ_ONLY	PB_T50_RS	Input	w/o Vote	w/ Vote
Supervised Learning Only								
PointNet Qi et al. (2017a)	3.5	1k Points	73.3	79.2	68.0	1k Points	89.2	-
PointNet++ Qi et al. (2017b)	1.5	1k Points	82.3	84.3	77.9	1k Points	90.7	-
DGCNN Wang et al. (2019)	1.8	1k Points	82.8	86.2	78.1	1k Points	92.9	-
SimpleView Goyal et al. (2021)	-	6 Images	-	-	80.5±0.3	6 Images	93.9	-
MVTN Hamdi et al. (2021)	11.2	20 Images	92.6	92.3	82.8	12 Images	93.8	-
PointMLP Ma et al. (2022)	12.6	1k Points	-	-	85.4±0.3	1k Points	94.5	-
SFR Zha et al. (2023b)	-	20 Images	-	-	87.8	20 Images	93.9	-
P2P-HorNet Wang & Yoon (2021)	195.8	40 Images	-	-	89.3	40 Images	94.0	-
with Single-Modal Self-Supervised Learning								
Point-BERT Yu et al. (2022)	22.1	1k Points	87.43	88.12	83.07	1k Points	92.7	93.2
MaskPoint Liu et al. (2022b)	-	2k Points	89.30	88.10	84.30	1k Points	-	93.8
Point-MAE Pang et al. (2022)	22.1	2k Points	90.02	88.29	85.18	1k Points	93.2	93.8
Point-M2AE Zhang et al. (2022a)	15.3	2k Points	91.22	88.81	86.43	1k Points	93.4	94.0
PointGPT Chen et al. (2024)	19.5	2k Points	91.60	90.00	86.90	1k Points	-	94.0
Point-FEMAE Zha et al. (2023a)	27.4	2k Points	95.18	93.29	90.22	1k Points	94.0	94.5
with Cross-Modal Self-Supervised Learning								
ACT Dong et al. (2022)	22.1	2k Points	93.29	91.91	88.21	1k Points	93.2	93.7
Joint-MAE Guo et al. (2023a)	-	2k Points	90.94	88.86	86.07	1k Points	-	94.0
I2P-MAE Zhang et al. (2023)	15.3	2k Points	94.14	91.57	90.11	1k Points	93.7	94.1
TAP Wang et al. (2023)	22.1	2k Points	90.36	89.50	85.67	-	-	-
ReCon Qi et al. (2023)	43.6	2k Points	95.18	93.63	90.63	1k Points	94.1	94.5
PViT Qian et al. (2024)	-	-	-	-	85.7	-	-	-
PViT+Pix4Point Qian et al. (2024)	-	-	-	-	87.9	-	-	-
Multi-View ML (Point-MAE) Chen et al. (2025)	22.1	-	93.32	92.69	88.93	-	93.8	94.1
Multi-View ML (Point-M2AE) Chen et al. (2025)	22.1	-	95.10	93.56	90.37	-	94.0	94.4
Ours (Point-MAE)	22.1	2k Points	95.20	93.73	90.10	1k Points	94.0	94.5
Improve (over Point-MAE)			+5.18	+5.44	+4.92	-	+0.8	+0.7
Ours (Point-M2AE)	15.3	2k Points	**96.10**	**94.25**	**91.20**	1k Points	**94.3**	**94.6**
Improve (over Point-M2AE)	-	-	+4.88	+5.41	+4.77	-	+0.9	+0.6
with Self-Supervised Representation Learning (MLP-Linear)								
Point-MAE Pang et al. (2022)	22.1	2k Points	82.77±0.30	83.23±0.16	74.13±0.21	1k Points	91.22±0.26	-
ACT Dong et al. (2022)	22.1	2k Points	85.20±0.83	85.84±0.15	76.31±0.26	1k Points	91.36±0.17	-
ReCon Qi et al. (2023)	43.6	2k Points	89.50±0.20	89.72±0.17	81.36±0.14	1k Points	92.47±0.22	-
Ours (Point-MAE)	22.1	2k Points	**90.57±0.19**	**90.68±0.14**	**82.41±0.16**	1k Points	**92.68±0.10**	-
with Self-Supervised Representation Learning (MLP-3)								
Point-MAE Pang et al. (2022)	22.1	2k Points	84.29±0.55	85.24±0.67	77.34±0.12	1k Points	92.33±0.09	-
ACT Dong et al. (2022)	22.1	2k Points	87.14±0.22	87.90±0.40	81.52±0.19	1k Points	92.69±0.18	-
ReCon Qi et al. (2023)	43.6	2k Points	90.62±0.22	90.71±0.30	83.80±0.42	1k Points	93.00±0.10	**-**
Ours (Point-MAE)	22.1	2k Points	**91.56±0.33**	**91.38±0.40**	**84.88±0.48**	1k Points	**93.35±0.17**	-

Table 4: **Ablation study for the proposed pre-training strategies.** Overall accuracy (%) without voting is reported.

Global Distillation	Token Shuffling	Local Distillation	MN40	ScanObjNN OBJ_BG
✓	✓	✓	**94.3**	**96.1**
✗	✓	✓	94.1	95.8
✓	✗	✓	93.9	95.7
✗	✗	✓	93.6	95.5
✗	✗	✗	93.7	95.0

evaluating on 10 independent experiments. It can be seen from the Table 1 that our approach achieves the state-of-the-art (SOTA) performance on all four settings. More specifically, our method shows improvements of **+0.8%**, **+0.7%**, **+1.8%** and **+1.2%** over Point-M2AE, respectively. Furthermore, compared to ReCon

Table 5: **Training Costs**. We compare the training costs in terms pre-training GPU hours, fine-tuning GPU hours and number of parameters on NVIDIA A100.

Method	#Params	Stage-1 Pre-train GPU hours	Stage-2 Pre-train GPU hours	Total Pre-train GPU hours	Fine-tune GPU hours	ScanObjectNN
Point-MAE	22.1M	-	15.2h	15.2h	9h	85.2
Point-MAE + 3.5 x pre-training	22.1M	-	54h	54h	9h	85.1
Ours (Point-MAE)	22.1M	24h	30h	54h	9h	**90.1**

and ACT, our model shows smaller deviations, suggesting that the pre-trained model has learned more robust and discriminative features that can be better adapted to downstream tasks in a low-data regime.

Zero-shot 3D Object Classification Our stage-1 pre-training aligns different modalities with the point cloud, enabling us to assess whether our approach exhibits a strong zero-shot capability. Following previous works (Qi et al., 2023; Xue et al., 2023), we conduct experiments on ModelNet40 (Wu et al., 2015). The results are shown in Table 2. Following PointCLIP (Zhang et al., 2022b) and ULIP-2 (Xue et al., 2024), we use both prompt templates with category labels and the top-1 BLIP-2 (Li et al., 2023) captions as the textual descriptions of the 3D objects. From Table 2, we observe that: (i) Our approach significantly outperforms the other zero-shot approaches; (ii) Compared to ULIP (Xue et al., 2023) and ReCon (Qi et al., 2023), our approach with prompt templates surpasses them by a clear margin achieving **93.5%** top-5 accuracy; (iii) Point-M2AE based 3D encoder trained with top-1 BLIP-2 descriptions achieves **73.4%** top-1 accuracy, surpassing ULIP-2 by **3.4%**.

Table 6: **Segmentation Results on S3DIS Area 5.** We report mean accuracy and mean IoU across all categories, i.e., mAcc (%) and mIoU(%) respectively.

Methods	Semantic Seg.	
	mAcc	mIoU
PointNet Qi et al. (2017a)	49.0	41.1
PointNet++ Qi et al. (2017b)	67.1	53.5
with Single-Modal Self-Supervised Representation Learning		
Transformer Vaswani et al. (2017)	68.6	60.0
Point-MAE Pang et al. (2022)	69.9	60.8
with Cross-Modal Self-Supervised Representation Learning		
ACT Dong et al. (2022)	71.1	61.2
Ours (Point-MAE)	**72.3**	**62.9**

4.2 Results on 3D Scene Segmentation

Next, we evaluate the approach on scene segmentation, a considerably complex task that requires the model to understand local geometric relationships and contextual semantics. We conduct experiment using S3DIS (Armeni et al., 2016) and use Area5 for the evaluation. The results are reported in Table 6. Our approach outperforms both PointMAE and ACT by **+2.4%** and **+1.1%** in mAcc, respectively.

4.3 Results on 3D Object Detection

We further conduct experiments on 3D object detection using ScanNetv2 (Dai et al., 2017) dataset. Following ACT (Dong et al., 2022), we use 3DETR (Misra et al., 2021), which has a transformer encoder consisting of three blocks and a transformer decoder. We compare various methods using mean Average Precision (mAP) at two different IoU thresholds of 0.50 and 0.25 with the results reported in Table 7. Our approach improves the 3DETR baseline by **+5.5%** and **+2.4%** on AP_{50} and AP_{25}, respectively. Compared to

Table 7: **3D object detection on the ScanNetV2 dataset.** We report Average Precision (AP) at two different IoU thresholds of 0.50 and 0.25.

Method	SSL	Input	AP_{50}	AP_{25}
VoteNet (Qi et al., 2019)	×	*xyz*	33.5	58.6
PointContrast (Xie et al., 2020)	✓	*xyz*	38.0	59.2
STRL (Huang et al., 2021)	✓	*xyz*	38.4	59.5
RandomRooms (Rao et al., 2021)	✓	*xyz*	36.2	61.3
DepthContrast (Zhang et al., 2021)	✓	*xyz*	-	61.3
3DETR (Misra et al., 2021)	×	*xyz*	37.9	62.1
Point-BERT (Yu et al., 2022)	✓	*xyz*	38.3	61.0
MaskPoint (Liu et al., 2022b)	✓	*xyz*	40.6	63.4
ACT Dong et al. (2022)	✓	*xyz*	42.1	63.8
Multi-View ML Chen et al. (2025)	✓	*xyz*	43.3	63.9
Ours (3DETR)	✓	*xyz*	**43.4**	**64.5**
3DTER-m Misra et al. (2021)	×	*xyz*	47.0	65.0
Point-M2AE Zhang et al. (2022a)	✓	*xyz*	48.3	66.3
Ours (3DETR-m)	✓	*xyz*	**49.4**	**67.5**

other SSL methods including ACT, our approach outperforms them by a significant margin. In addition to conducting experiments with 3DETR, we also evaluate the approach using 3DTER-m, following Point-M2AE. We keep the implementation and architecture details consistent with those of Point-M2AE. Our two-stage pre-training approach boost the performance of 3DETR-m by **+2.4%** on AP_{50} and **+2.5%** on AP_{25}, clearly outperforming Point-M2AE, as reported in Table 7.

4.4 Ablation Study

Effect of second stage pre-training objectives. To verify the effectiveness of different components of our second stage pre-training, we conduct an ablation study comparing the performance of token shuffling and global distillation with full second-stage model on ModelNet40 and OBJ_BG variant of ScanObjectNN. The results are reported in Table 4. We observe that: i) Both global distillation and token shuffling improve the overall performance of the second-stage model; ii) Token shuffling, in particular, provides a greater performance gain compared to global distillation. We conjecture that the more challenging learning objective introduced by token shuffling improves the overall 3D representations by encouraging the second-stage model to better capture geometric features in the point clouds; iii) Global distillation also boosts the performance of the second-stage model and helps mitigate catastrophic forgetting, particularly on ModelNet40.

Table 8: **Comparison between local distillation loss and reconstruction in point cloud space on ModelNet40 dataset.** Left table report results using Point-MAE while right table reports results using Point-M2AE. Overall accuracy (%) without voting is reported.

Target	Acc (%)
Point clouds	93.5
Features	**94.0**

Target	Acc (%)
Point clouds	93.7
Features	**94.3**

Reconstruction in the point space? We conduct experiments to compare pre-training the model using masked point modeling in the point cloud space with proposed local distillation. In Table 8, we report the results on ModelNet40 using Point-MAE and Point-M2AE. We find that reconstructing in the point cloud space degrades the overall performance for both architectures. One plausible reason for this is that, given the initialization from the stage-1 weights, the reconstruction task becomes easier, allowing the model to quickly find local optima when combined with global distillation, which may consequently degrade downstream performance.

Training Costs. In Table 5, we compare the training costs of our approach with the Point-MAE baseline on ScanObjectNN, reporting the pre-training numbers on ShapeNet. It is observed that our two-stage approach takes 3.5x longer than Point-MAE, but it achieves a significant improvement of **+4.9%** in terms of accuracy. When we extend the pre-training time for Point-MAE to match ours, we observe a marginal drop in performance, likely due to overfitting to the self-supervised task, while the performance gap between our approach and the baseline remains substantial. While our approach requires more computation during pretraining, it remains framework-agnostic, ensuring that inference speed and FLOPs during fine-tuning are comparable to Point-MAE and other variants. Additionally, we compare the fine-tuning cost with ReCon in (see Table **??** in supplementary), demonstrating that leveraging Point-MAE and Point-M2AE allows us to keep the parameter count and GFLOPS while improving downstream performance.

Stage-1 v/s Stage-2 results. In the last row of Table 4, we present the performance of our stage-1 pre-training. We observe that the stage-1 pre-trained model performs competitively, outperforming both Point-MAE and Point-M2AE on ModelNet40 and ScanObjectNN. Furthermore, by incorporating the stage-2 pre-training components, our approach achieves the highest performance on both datasets.

Additional Baselines We compare our approach to two additional baselines. The first baseline follows a single-stage pre-training with a multi-task learning paradigm. The second baseline reverses the learning order, performing masked point modeling in Stage-1 and multi-modal alignment in Stage-2. The results are reported in Table 9. It can be seen that both the baselines perform worse as compared to our approach. Joint training can lead to conflicting gradients which degrades the overall performance when combining these objectives. Reversing the order would prioritize local geometry too early, making it harder to align noisy text and captions effectively. Morever, random initialization in Stage-2 degrades token-wise distillation and yields lower accuracy.

Table 9: **Results on additional baselines.** Overall accuracy (%) without voting is reported.

Method.	MN40	ScanOBJNN
Joint training (single stage)	92.2	82.78
Stage-2 -> Stage-1	93.6	87.1
Stage-2 (random init)	93.7	88.2
Ours (Point-MAE)	**94.0**	**90.1**

4.5 Results on Part Segmentation

We evaluate the performance of our pre-trained approach in Part Segmentation using ShapeNetPart (Yi et al., 2016). This settings requires the model to understand local patterns by predicting part labels for each point. We follow PointM2AE (Zhang et al., 2022a) for segmentation head and other experimental details. We compare the performance of our approach with other baselines in Table 10 and report both the mean IoU over all categories and the mean IoU over all instances. Our approach achieves the SOTA performance and improves Point-M2AE by **+0.6%** and **+0.7%** on both metrics respectively.

5 Conclusion

In this paper, we present a novel two-stage pre-training approach for point clouds, leveraging foundation models and feature distillation. Our method tackles the scalability challenge by incorporating diverse language descriptions generated by multimodal models such as BLIP-2, alongside images and depth maps associated with point clouds. In the first stage, multiple foundation models are used to learn rich multimodal representations.

Table 10: **Part Segmentation Results on ShapeNetPart dataset.** We report mean IoU across all categories (Cls.mIoU) and mean IoU across all instances (Inst.mIoU). † means the transformer is pre-trained with ImageNet dataset.

Methods	Part Seg.	
	Cls.mIoU	Inst.mIoU
PointNet Qi et al. (2017a)	80.4	83.7
PointNet++ Qi et al. (2017b)	81.9	85.1
DGCNN Wang et al. (2019)	82.3	85.2
PointMLP Ma et al. (2022)	84.6	86.1
with Single-Modal Self-Supervised Representation Learning		
Transformer Vaswani et al. (2017)	83.4	84.7
CrossPoint Afham et al. (2022)	-	85.5
Point-BERT Yu et al. (2022)	84.1	85.6
MaskPoint Liu et al. (2022b)	84.4	86.0
Point-MAE Pang et al. (2022)	84.2	86.1
Point-M2AE Zhang et al. (2022a)	84.9	86.5
PointGPT Chen et al. (2024)	84.1	86.2
Point-FEMAE Zha et al. (2023a)	84.9	86.3
PViT Qian et al. (2024)	83.7	85.7
with Cross-Modal Self-Supervised Representation Learning		
ACT Dong et al. (2022)	84.7	86.1
ReCon Qi et al. (2023)	84.8	86.4
Ours (Point-M2AE)	**85.5**	**87.2**
Improve (over Point-M2AE)	+0.6	+0.7
Mamba3D Han et al. (2024)	83.6	85.6
PViT+Pix4Point Qian et al. (2024)†	**85.6**	86.8

In the second stage, the model is further pre-trained by improved masked point modeling, incorporating feature distillation and token shuffling. Our approach learns rich semantic and geometric representations, achieving state-of-the-art performance across a wide range of 3D understanding tasks. Furthermore, it is model-agnostic and can be applied to any transformer-based 3D encoder.

6 Discussion and Limitations

Our two-stage pre-training approach provides rich semantic and geometric representations by leveraging multimodal data as well as foundation models, and has demonstrated effectiveness in various 3D downstream tasks. However, there are some limitations to our approach. (i) In stage 1, we rely on single-view images or depth maps, which may not fully capture the complex geometry of 3D objects. While we still pre-train on ShapeNet, which contains 55k objects, there is potential to enhance the model's performance by incorporating multi-view images to ensure geometric consistency. (ii) Another avenue for improvement is the development of a separate encoder for depth, allowing for a depth-aligned understanding of both point clouds and images, as well as language. We will explore all of these limitations in the future. We also want to reiterate and advocate for the evaluation of 3D pre-training approaches on more complex and challenging scenarios, as we have recently seen the saturation of performance on small datasets such as ModelNet40 and a couple of variants of ScanObjectNN. We believe there is a need to shift focus toward indoor scene understanding or to construct a better evaluation dataset. Additionally, exploring 3D representations from a scaling perspective, including datasets like Objaverse-XL (Deitke et al., 2023), as well as leveraging large language models (LLMs) for 3D scene understanding, could significantly enhance embodied AI and intelligent systems for robotics.

References

Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9902–9912, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1534–1543, 2016.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Zhimin Chen, Xuewei Chen, Xiao Guo, Yingwei Li, Longlong Jing, Liang Yang, and Bing Li. Point cloud self-supervised learning via 3d to multi-view masked learner, 2025.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation, 2019.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pp. 5828–5839, 2017.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks, 2018.

Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *International Conference on Machine Learning*, pp. 3809–3820. PMLR, 2021.

Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzhi Li, and Pheng-Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *arXiv preprint arXiv:2302.14007*, 2023a.

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following, 2023b.

Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2021.

Xu Han, Yuan Tang, Zhaoxuan Wang, and Xianzhi Li. Mamba3d: Enhancing local features for 3d point cloud analysis via state space model, 2024.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation, 2019a.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2019b.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6535–6545, 2021.

Tianxin Huang, Zhonggan Ding, Jiangning Zhang, Ying Tai, Zhenyu Zhang, Mingang Chen, Chengjie Wang, and Yong Liu. Learning to measure the point cloud reconstruction loss in a representation space. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023a.

Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023b.

Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer, 2019.

Jangho Kim, Seonguk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Advances in Neural Information Processing Systems*, 2018.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

Fayao Liu, Guosheng Lin, Chuan-Sheng Foo, Chaitanya K Joshi, and Jie Lin. Point discriminative learning for data-efficient 3d point cloud analysis. In *2022 International Conference on 3D Vision (3DV)*, pp. 42–51. IEEE, 2022a.

Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *European Conference on Computer Vision*, pp. 657–675. Springer, 2022b.

Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. *arXiv preprint arXiv:1912.00280*, 2019.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 2023.

Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation. 1989.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019.

Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2906–2917, 2021.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles, 2017.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pp. 604–621. Springer, 2022.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation, 2019.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Yu Liu, Dongsheng Li, and Zhaoning Zhang. Correlation congruence for knowledge distillation, 2019.

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

Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9277–9286, 2019.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. *arXiv preprint arXiv:2302.02318*, 2023.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction, 2024.

Guocheng Qian, Abdullah Hamdi, Xingdi Zhang, and Bernard Ghanem. Pix4point: Image pretrained standard transformers for 3d point cloud understanding, 2024.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Yongming Rao, Benlin Liu, Yi Wei, Jiwen Lu, Cho-Jui Hsieh, and Jie Zhou. Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In *ICCV*. IEEE, 2021.

Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pp. 626–642. Springer, 2020.

Yiwen Tang, Ray Zhang, Jiaming Liu, Zoey Guo, Dong Wang, Zhigang Wang, Bin Zhao, Shanghang Zhang, Peng Gao, Hongsheng Li, and Xuelong Li. Any2point: Empowering any-modality large models for efficient 3d understanding, 2024.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9782–9792, 2021.

Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068, 2021.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019.

Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. Take-a-photo: 3d-to-2d generative pre-training of point cloud models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5640–5650, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Tong Wu, Liang Pan, Junzhe Zhang, Tai WANG, Ziwei Liu, and Dahua Lin. Balanced chamfer distance as a comprehensive metric for point cloud completion. In *Advances in Neural Information Processing Systems*, 2021.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 574–591. Springer, 2020.

Chenfeng Xu, Shijia Yang, Tomer Galanti, Bichen Wu, Xiangyu Yue, Bohan Zhai, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2point: 3d point-cloud understanding with 2d image pretrained models, 2022.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. 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.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding, 2024.

Chenglin Yang, Lingxi Xie, Chi Su, and Alan L. Yuille. Snapshot distillation: Teacher-student optimization in one generation, 2018.

Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. 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.

Yaohua Zha, Huizhen Ji, Jinmin Li, Rongsheng Li, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. Towards compact 3d representations via point feature enhancement masked autoencoders. *arXiv preprint arXiv:2312.10726*, 2023a.

Yaohua Zha, Rongsheng Li, Tao Dai, Jianyu Xiong, Xin Wang, and Shu-Tao Xia. Sfr: Semantic-aware feature rendering of point cloud. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35:27061–27074, 2022a.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8552–8562, 2022b.

Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. 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.

Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10252–10263, 2021.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022.