Joint-MAE: 2D-3D Joint Masked Autoencoders for 3D Point Cloud Pre-training

Ziyu Guo[1], Renrui Zhang[2], Longtian Qiu[5], Xianzhi Li[3], Pheng-Ann Heng[1, 4]
[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2] CUHK MMLab
[3]Huazhong University of Science and Technology
[4]Institute of Medical Intelligence and XR, The Chinese University of Hong Kong
[5]ShanghaiTech University
{zyguo, pheng}@cse.cuhk.edu.hk, xzli@hust.edu.cn

Abstract

Masked Autoencoders (MAE) have shown promising performance in self-supervised learning for both 2D and 3D computer vision. However, existing MAE-style methods can only learn from the data of a single modality, i.e., either images or point clouds, which neglect the implicit semantic and geometric correlation between 2D and 3D. In this paper, we explore how the 2D modality can benefit 3D masked autoencoding, and propose **Joint-MAE**, a 2D-3D joint MAE framework for self-supervised 3D point cloud pre-training. Joint-MAE randomly masks an input 3D point cloud and its projected 2D images, and then reconstructs the masked information of the two modalities. For better cross-modal interaction, we construct our JointMAE by two hierarchical 2D-3D embedding modules, a joint encoder, and a joint decoder with modal-shared and model-specific decoders. On top of this, we further introduce two cross-modal strategies to boost the 3D representation learning, which are local-aligned attention mechanisms for 2D-3D semantic cues, and a cross-reconstruction loss for 2D-3D geometric constraints. By our pre-training paradigm, Joint-MAE achieves superior performance on multiple downstream tasks, e.g., 92.4% accuracy for linear SVM on ModelNet40 and 86.07% accuracy on the hardest split of ScanObjectNN.

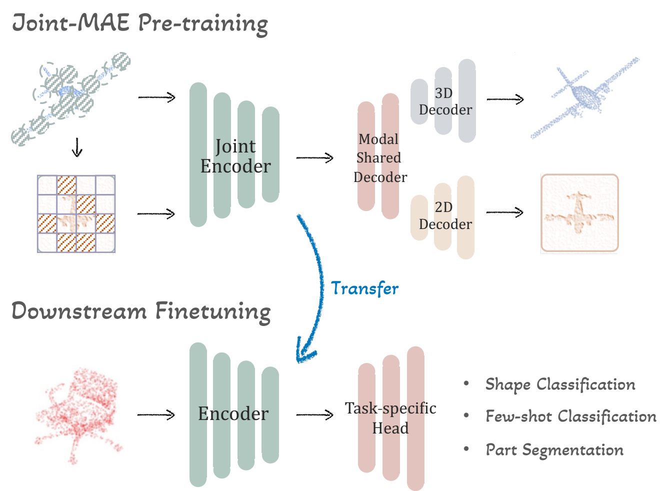

Figure 1: **The Pre-training and Fine-tuning Paradigms of Joint-MAE.** By joint masked autoencoding on both 3D and 2D, our Joint-MAE learns strong point cloud representations and exerts superior transferable capacity on 3D downstream tasks.

1 Introduction

Self-supervised pre-training aims to learn more universal representations from unlabeled data that work across a wider variety of downstream tasks and datasets. It has achieved great success in both 2D images and 3D point clouds with the main schemes including contrastive learning [Afham *et al.*, 2022] and masked autoencoding [He *et al.*, 2021; Gao *et al.*, 2022; Pang *et al.*, 2022; Zhang *et al.*, 2022a]. Recently, as the newly proposed paradigm, masked autoencoders have outperformed contrastive learning on multiple downstream tasks. It randomly masks a portion of input data and adopts a transformer encoder to extract the unmasked features. Then, a lightweight transformer decoder is utilized to reconstruct the information of the masked positions.

In 2D modality, MAE [He *et al.*, 2021] and its follow-up work efficiently conduct 2D masked autoencoding with multi-scale convolution stages [Gao *et al.*, 2022] and visible feature mimicking [Gao *et al.*, 2023]. In 3D modality, Point-MAE [Pang *et al.*, 2022], Point-M2AE [Zhang *et al.*, 2022a] and I2P-MAE [Zhang *et al.*, 2022b] divide the point clouds into groups and directly reconstruct the masked coordinates in 3D space. At this point, existing methods have verified the effectiveness of MAE respectively in 2D and 3D. However, they can only process a single modality independently without utilizing their implicit correlations. Compared to the irregular and sparse point clouds, 2D images can provide dense and fine-grained visual signals concerning both geometries and semantics, which might assist the 3D representation learning during MAE pre-training. Therefore, we ask the question: *Can we design a 2D-3D joint MAE frame-*

work to boost 3D point cloud pre-training by leveraging the information from 2D images?

To address this issue, we propose **Joint-MAE**, a unified masked autoencoding framework that leverages cross-modal 2D knowledge to boost 3D representation learning. As shown in Figure 1, by joint pre-training on both 2D images and 3D point clouds, our Joint-MAE can produce high-quality point cloud representation and present strong transferable capacity on 3D downstream tasks. Specifically, to obtain the multi-modal input without using extra 2D data, we efficiently project the 3D point cloud into a 2D depth map from a random view. On top of the 2D-3D input data, we introduce two hierarchical modules to respectively embed 2D and 3D tokens and randomly mask them with a large ratio. Then, we concatenate the visible tokens of two modalities and feed them into a joint encoder, which interacts 2D-3D semantics via stacked transformer blocks. After that, we utilize a joint decoder to simultaneously reconstruct the masked point cloud coordinates and the image pixels, which consists of a modal-shared decoder and two modal-specific decoders.

Upon this unified paradigm, we further benefit 3D representation learning via two cross-modal strategies. First, in every transformer block of the joint encoder, we design a local-aligned attention mechanism for better interacting 2D-3D features. Only geometrically correlated 2D-3D tokens are valid for attention calculation, which can boost the point cloud representation learning with 2D fine-grained semantics. Second, besides the independent reconstruction losses of 2D and 3D, we propose an extra cross-reconstruction loss that provides additional 2D-3D geometric constraints.

The contributions of our paper are as follows:

- We propose Joint-MAE, a 2D-3D joint MAE framework for self-supervised 3D point cloud pre-training, including two hierarchical 2D-3D embedding modules, a joint encoder, and a joint decoder with modal-shared and model-specific components.
- To better interact 2D-3D semantics and geometries, we introduce local-aligned attention mechanisms in the joint encoder and a cross-reconstruction loss for self-supervision.
- We evaluate Joint-MAE on various 3D downstream tasks, e.g., shape classification and part segmentation, where our approach shows leading performance compared with existing methods.

2 Related Work

Pre-training in 3D Vision. Recently, supervised learning in 3D vision [Qi *et al.*, 2017a; Qi *et al.*, 2017b; Guo *et al.*, 2021; Wang *et al.*, 2019; Zhang *et al.*, 2023] has achieved promising performance with delicate network designs. However, they are confined to limited data domains [Wu *et al.*, 2015a; Uy *et al.*, 2019] that they are trained on, without satisfactory generalization ability. Therefore, self-supervised learning emerges in this context, and has significantly enhanced the 3D transfer learning. That is, big models with abundant pre-trained knowledge show promising results after fine-tuning. Meanwhile, the development of generative models for 3D point cloud [Xie *et al.*, 2016; Xie *et al.*, 2018; Xie *et al.*, 2020; Xie *et al.*, 2021; Nichol *et al.*, 2022] promotes exploration of the pre-training paradigm on large-scale data. Most self-supervised pre-training methods [Poursaeed *et al.*, 2020] include an encoder and a decoder to firstly transform the input point clouds into latent representations, and then recover them into the original data form. Other works [Rao *et al.*, 2020] conducts self-supervised pre-training via contrastive learning. Also, inspired by masked image modeling [He *et al.*, 2021], a series of works for masked point modeling [Liu *et al.*, 2022; Yu *et al.*, 2021] are put forward. Therein, Point-MAE [Pang *et al.*, 2022] conduct masked autoencoding for 3D point clouds directly. Point-M2AE [Zhang *et al.*, 2022a] and I2P-MAE [Zhang *et al.*, 2022b] further adopt hierarchical MAE transformers. Our Joint-MAE is also based on the masked autoencoding framework, and further utilizes the correlated information between 2D and 3D modalities to build a more powerful 3D self-supervised learner.

Multi-Modality Learning. Multi-modality learning, namely, cross-modal learning, aims at learning from signals of multiple modalities at the same time, which achieves more robust performance than the single-modality learning method. Most works [Desai and Johnson, 2021] have proved the powerful capability of multi-modality pre-training. CLIP [Radford *et al.*, 2021] bridges the gap between textual space and image space through contrastive learning. Based on CLIP, some works [Gao *et al.*, 2021; Zhang *et al.*, 2021a; Guo *et al.*, 2022] further propose more advanced structures to promote the performance, and adopt CLIP's pre-trained knowledge on different downstream tasks, e.g., PointCLIP V1 [Zhang *et al.*, 2021b], V2 [Zhu *et al.*, 2022] for 3D learning, and DepthCLIP [Zhang *et al.*, 2022c] for depth estimation. [Zhang *et al.*, 2021c] conducts a point-voxel joint pre-training design, and CrossPoint [Afham *et al.*, 2022] proposes an image-point contrastive learning network. Also, via filter inflation, [Xu *et al.*, 2021] utilizes pre-trained 2D knowledge on 3D point cloud understanding. Inspired by the success of MAE pre-training independently in 2D and 3D vision [He *et al.*, 2021; Gao *et al.*, 2022; Pang *et al.*, 2022; Zhang *et al.*, 2022a], we expect to fully incorporate MAE pre-training and multi-modality learning to unleash their potientials for 3D representation learning.

3 Method

The overall pre-training pipeline of Joint-MAE is shown in Figure 2. Given an input 3D point cloud, we first generate 2D inputs via efficient depth map projection, and adopt 2D-3D embedding modules for initial tokenization (Section 3.1). Then, we feed the visible tokens into a 2D-3D transformer, which consists of a joint encoder and a joint decoder (Section 3.2). Finally, we further introduce two cross-modal strategies, i.e., the local-aligned attention mechanism (Section 3.3) and the cross-reconstruction loss (Section 3.4).

3.1 2D-3D Data Embedding

2D Depth Projection from 3D. To conduct efficient multi-modal learning for 3D point cloud pre-training, we require

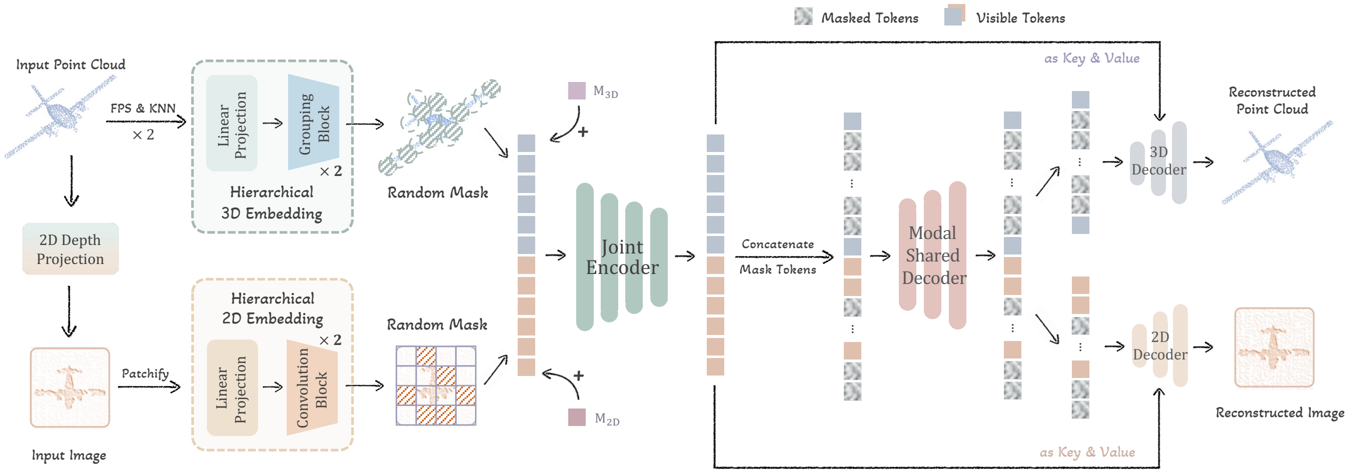

Figure 2: **The Pipeline of Joint-MAE,** which takes as input the 3D point cloud with its projected 2D image to conduct cross-modal masked autoencoding. To boost 3D representation learning by 2D-3D interaction, Joint-MAE consists of two hierarchical 2D-3D embedding modules, a joint encoder, and a joint decoder with modal-shared and modal specific components.

no additional real-world 2D images and simply project 3D point clouds into depth maps as the 2D input. For an input point cloud $P \in R^{N\times 3}$ with N points, we select a random view at every iteration for augmentation and perspectively project [Goyal *et al.*, 2021; Zhang *et al.*, 2021b] 3D points onto a pre-defined image plane as I, formulated as

$$I = \text{Project}(P), \quad \text{where } I \in R^{H\times W\times 1}. \tag{1}$$

In this way, we not only discard the time-consuming rendering, but also preserve the 2D-3D geometric correlations.

Hierarchical 2D-3D Tokenization. Unlike MAE [He *et al.*, 2021] and Point-MAE [Pang *et al.*, 2022] that tokenize the 2D and 3D data with a constant resolution, we propose hierarchical 2D-3D embedding modules to extract multi-scale features independently for the two modalities. For the 3D branch, we first adopt two cascaded Farthest Point Sampling (FPS) with K Nearest Neighbors (k-NN) to acquire the multi-scale representations of point cloud, which contain $N/8$ and $N/32$ points respectively. After transforming the raw points via a linear projection layer, we utilize two grouping blocks to hierarchically tokenize the point cloud according to the obtained multi-scale representations, which produces the C-dimensional 3D tokens, $T_{3D} \in R^{N/32\times C}$. The grouping block consists of a mini-PointNet [Qi *et al.*, 2017a] for aggregating local features within each k neighbors. For the 2D branch, we first patchify the depth map with a patch size 4×4 and utilize a linear projection layer for high-dimensional embedding. Then, two 3×3 convolution blocks are adopted to progressively tokenize the 2D image as the C-dimensional 2D tokens with a downsample ratio $1/16$, denoted as $T_{2D} \in R^{HW/256\times C}$. We formulate the 2D-3D embedding modules as

$$T_{3D} = \text{Embed}_{3D}(P), \quad T_{2D} = \text{Embed}_{2D}(I). \tag{2}$$

2D-3D Masking. On top of this, we randomly mask the 2D and 3D tokens respectively with a large ratio, e.g., 75%. This builds a challenging pre-text task for self-supervised learning by masked auto-encoding. We denote the visible tokens as T^v_{2D} and T^v_{3D}, which are fed into a proposed 2D-3D joint transformer for multi-modal reconstruction.

3.2 2D-3D Joint Transformer

The transformer consists of a joint encoder and a joint decoder. The former encodes cross-modal features upon the visible 2D-3D tokens. The latter reconstructs the masked 2D-3D information with a modal-shared decoder and two modal-specific decoders.

Joint Encoder

We concatenate the visible tokens of two modalities along the token dimension as the input for joint encoder, which is composed of 12 transformer blocks with self-attention layers for 2D-3D feature interaction. To distinguish the modality characters during attention mechanisms, we introduce the modality encodings by using two learnable tokens and randomly initializing them before pre-training. The two modality tokens, $M_{2D}, M_{3D} \in \mathbb{R}^{1\times C}$, are respectively added to all the 2D and 3D tokens, which can provide modal-specific cues for attention layers. We formulate it as

$$E^v_{2D\text{-}3D} = \text{JointEnc}\left(\text{Concat}(T^v_{3D} + M_{3D}, T^v_{2D} + M_{2D})\right),$$

where $E^v_{2D\text{-}3D}$ denotes the visible 2D-3D features of the input point cloud. By the multiple attention layers, the 3D tokens can adaptively aggregate informative 2D semantics for 2D-3D interaction, and obtain a cross-modal understanding of the point cloud.

Joint Decoder

After the encoding of visible 2D-3D tokens, we propose a joint decoder for multi-modal reconstruction of the masked information, which consists of a modal-shared decoder and two modal-specific decoders.

Modal-shared Decoder. We adopt two learnable C-dimensional mask tokens in the decoder respectively for 2D and 3D, which are shared for all masked positions, denoted

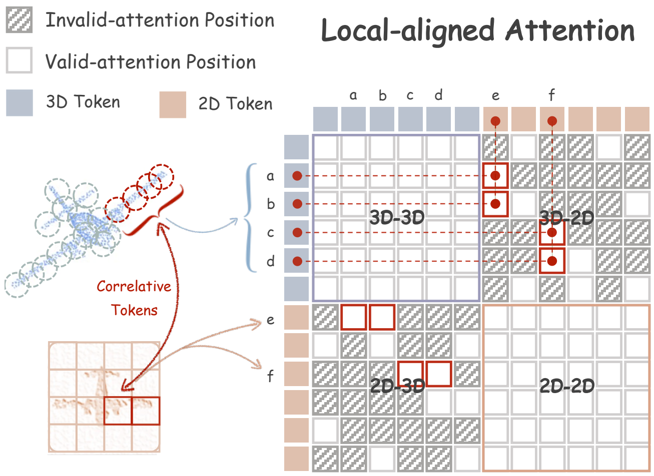

Figure 3: **Local-aligned Attention.** For self-attention mechanisms of the joint encoder, we only interact cross-modal features between the geometrically correlative 2D-3D tokens (highlighted in red). This contributes to more fine-grained and intensive feature learning.

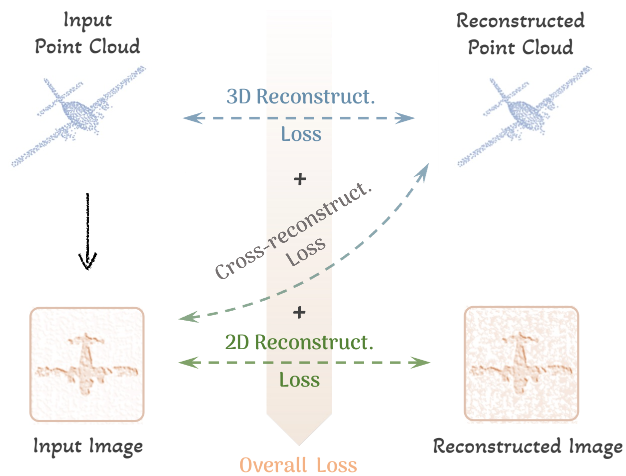

Figure 4: **2D-3D Reconstruction Loss.** We apply three losses for Joint-MAE pre-training, which are 2D and 3D reconstruction losses, along with a cross-reconstruction loss that provides additional 2D-3D geometric constrains.

as T^m_{3D} and T^m_{2D}. We concatenate them with the encoded 2D-3D tokens $E^v_{2D\text{-}3D}$ along the token dimension, and feed them into the model-shared decoder, which contains 2 transformer blocks with self-attention layers. We formulate it as

$$D_{2D\text{-}3D} = \mathrm{ShareDec}\big(\mathrm{Concat}(E^v_{2D\text{-}3D}, T^m_{3D}, T^m_{2D})\big), \quad (3)$$

where $D_{2D\text{-}3D}$ denotes the decoded 2D-3D features. By the model-shared decoder, the 3D mask tokens T^m_{3D} can simultaneously capture information from both 2D and 3D visible tokens. In this way, the reconstruction of masked 3D point clouds can be benefited from both the visible 3D geometries and the additional 2D patterns, which contributes to better 3D representation learning and vice versa.

Modal-specific Decoders. After the modal-shared decoder, we divide $D_{2D\text{-}3D}$ along the token dimension as the 2D and 3D tokens, denoted as D_{2D} and D_{3D}. Then, they are fed into two concurrent decoders for specific 2D and 3D decoding. This enables the network to respectively focus on the unique characteristics of 2D and 3D for their reconstruction. Importantly, we adopt cross-attention layer in the 1-block specific decoders, where D_{2D}, D_{3D} serve as queries and E^v_{2D}, E^v_{3D} serve as keys/values. E^v_{2D}, E^v_{3D} are the token-wise division of $E^v_{2D\text{-}3D}$ respectively for 2D and 3D modalities, which can complement more model-specific cues for the mask tokens during the cross-attention mechanism. We formulate them as

$$\begin{aligned} D'_{3D} &= \mathrm{SpeDec_{3D}}(D_{3D}, E^v_{3D}) \\ D'_{2D} &= \mathrm{SpeDec_{2D}}(D_{2D}, E^v_{2D}) \end{aligned} \quad (4)$$

where D'_{2D}, D'_{3D} represent the modal-specific decoded features for 2D and 3D.

3.3 Local-aligned Attention

To further enhance the 2D-3D feature interaction, we propose a local-aligned attention mechanism in every transformer block of the joint encoder. Figure 3 illustrates its methodology. During the attention layer, all tokens are mutually visible within their own modalities: every 3D token conducts attention calculation with each other, and 2D likewise. However, only the geometrically correlated 2D and 3D tokens would be calculated the attention scores for feature interaction (Valid-attention Position), and others would not (Invalid-attention Position). We regard a 3D token is geometrically correlated with a 2D token if its 3D center point is projected onto the pixels within the 2D token on the image plane. As shown in the figure, the 3D tokens 'a, b, c, d' are correlated with the 2D tokens 'e, f, g', both of which represent the same part-wise features of the object, i.e., the wing of an airplane. We only enable the attention scores between these correlated 2D-3D tokens to be valid. By such local-aligned attention, the cross-modal interaction is effectively constrained into different local regions, which guides the tokens to intensively aggregate fine-grained features from the other modality and avoids the long-range distractions of non-relevant regions.

3.4 2D-3D Reconstruction Loss

As shown in Figure 4, we adopt two reconstruct heads of simple linear layers to respectively recover the masked 2D pixels and 3D coordinates following MAE [He *et al.*, 2021] and Point-MAE [Pang *et al.*, 2022]. We adopt Mean Squared Error (MSE) and l_2 Chamfer Distance (CD) [Fan *et al.*, 2017] as two loss functions of 2D and 3D reconstruction, i.e.,

$$\begin{aligned} L_{3D} &= \mathrm{CD}\left(\mathrm{RecHead_{3D}}(D'_{3D}),\ P\right), \\ L_{2D} &= \mathrm{MSE}\left(\mathrm{RecHead_{2D}}(D'_{2D}),\ I\right). \end{aligned} \quad (5)$$

Besides, for better 2D-3D geometric alignment, we propose an extra cross-reconstruction loss between the reconstructed point cloud and the 2D input I, the initially projected depth map. Specifically, we project the reconstructed point cloud by the same view of I and then calculate the MSE loss as

$$L_{3D\text{-}2D} = \mathrm{MSE}\Big(\mathrm{Project}\big(\mathrm{RecHead_{3D}}(D'_{3D})\big),\ I\Big). \quad (6)$$

Method	Accuracy (%)
SO-Net [Li *et al.*, 2018]	87.3
FoldingNet [Yang *et al.*, 2018]	88.4
VIP-GAN [Han *et al.*, 2019]	90.2
DGCNN + Jiasaw [Sauder and Sievers, 2019]	90.6
DGCNN + OcCo [Wang *et al.*, 2021]	90.7
DGCNN + CrossPoint [Afham *et al.*, 2022]	91.2
Transformer + OcCo [Wang *et al.*, 2021]	89.6
Point-BERT [Yu *et al.*, 2021]	87.4
Point-MAE [Pang *et al.*, 2022]	91.0
Joint-MAE	**92.4**

Table 1: **Linear SVM Performance.** We report shape classification results of different self-supervised pre-training methods on ModelNet40 [Wu *et al.*, 2015a] via linear SVM.

Method	Accuracy (%)
PointNet [Qi *et al.*, 2017a]	89.2
PointNet++ [Qi *et al.*, 2017b]	90.5
DGCNN [Wang *et al.*, 2019]	92.9
PCT [Guo *et al.*, 2021]	93.2
Point Transformer [Zhao *et al.*, 2021]	93.7
Transformer + OcCo [Wang *et al.*, 2021]	92.1
Point-BERT [Yu *et al.*, 2021]	93.2
Point-MAE [Pang *et al.*, 2022]	93.8
MaskPoint [Liu *et al.*, 2022]	93.8
Joint-MAE	**94.0**

Table 2: **Shape Classification on ModelNet40 [Wu *et al.*, 2015a].** We report the shape classification accuracy (%) of Joint-MAE fine-tuned on ModelNet40 dataset.

Such cross-reconstruction loss can better self-supervise the spatial structure of reconstructed point clouds and enforce the 3D-2D geometric alignment in the 2D space. Then, the overall loss for Joint-MAE pre-training is formulated as

$$L_{overall} = L_{2D} + L_{3D} + L_{3D\text{-}2D}. \tag{7}$$

The three self-supervisory signals can well interact 2D-3D features from both semantics and geometries, which effectively boosts the representation learning for 3D poitn clouds.

4 Experiments

In this section, we evaluate the performance of Joint-MAE on various downstream tasks, i.e., shape classification, few-shot classification, and part segmentation.

4.1 Pre-training.

Settings. Following existing works [Pang *et al.*, 2022; Zhang *et al.*, 2022a], we pre-train our Joint-MAE on ShapeNet [Chang *et al.*, 2015], which covers 57,448 3D shapes of 55 categories. The input point number N is set as 2,048 and the depth map size $H \times W$ is set as 224×224. We adopt a feature dimension C as 384. Please refer to the Supplementary Material for detailed implementation details.

Method	OBJ-BG	OBJ-ONLY	PB-T50-RS
PointNet	73.3	79.2	68.0
PointNet++	82.3	84.3	77.9
DGCNN	82.8	86.2	78.1
PointCNN	86.1	85.5	78.5
GBNet	-	-	80.5
SimpleView	-	-	80.5
PRANet	-	-	81.0
MVTN	-	-	82.8
PointMLP	-	-	85.2
Transformer	79.86	80.55	77.24
Transformer + OcCo	84.85	85.54	78.79
Point-BERT	87.43	88.12	83.07
MaskPoint	89.30	88.10	84.30
Point-MAE	90.02	88.29	85.18
Joint-MAE	**90.94**	**88.86**	**86.07**
Improvement	+0.92	+0.57	+0.89

Table 3: **Shape Classification on ScanObjectNN [Uy *et al.*, 2019].** We report the shape classification accuracy (%) on the three splits of ScanObjectNN.

After Joint-MAE pre-training, we utilize the pre-trained encoder along with the hierarchical 3D embedding module for 3D downstream tasks, where we only preserve the 3D branch, i.e., only 3D point clouds as input without projecting into depth maps.

Performance of Linear SVM. We first evaluate the Linear SVM performance of Joint-MAE on the well-known ModelNet40 dataset [Wu *et al.*, 2015b]. Following CrossPoint [Afham *et al.*, 2022], we utilize the pre-trained encoder and freezes its weights to extract the 3D point cloud features. We report the shape classification accuracy in Table 1. As shown, our Joint-MAE outperforms all existing self-supervised pre-training methods and surpasses the second-top Point-MAE [Pang *et al.*, 2022] with **+1.4%** improvement, indicating the superior 3D representation assested by the 2D modality.

4.2 Downstream Tasks

After pre-training, we fine-tune the 3D branch of Joint-MAE on multiple 3D downstream tasks, i.e., shape classification, few-shot classification, and part segmentation. In each task, we discard the 2D branch with the decoder, and equip the pre-trained encoder by the task-specific heads.

Shape Classification. We utilize a simple a classification head of linear layers and evaluate the accuracy on ModelNet40 [Wu *et al.*, 2015a] and ScanObjectNN [Uy *et al.*, 2019] datasets, which contain synthetic objects and real-world instances, respectively. Table 2 demonstrates the classification results on ModelNet40, where we train on 9,843 instances and test on 2,468 instances with 40 categories. Joint-MAE achieves the highest accuracy compared with other methods. Note that we randomly sample 1,024 points from each object with only the coordinate information as the input data. Besides, we report in Table 3 the performance on the real-world ScanObjectNN dataset, which consists of 2,304 ob-

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
PointNet [Qi *et al.*, 2017a]	52.0 ± 3.8	57.8 ± 4.9	46.6 ± 4.3	35.2 ± 4.8
PointNet + OcCo [Wang *et al.*, 2021]	89.7 ± 1.9	92.4 ± 1.6	83.9 ± 1.8	89.7 ± 1.5
PointNet + CrossPoint [Afham *et al.*, 2022]	90.9 ± 4.8	93.5 ± 4.4	84.6 ± 4.7	90.2 ± 2.2
DGCNN [Wang *et al.*, 2019]	31.6 ± 2.8	40.8 ± 4.6	19.9 ± 2.1	16.9 ± 1.5
DGCNN + OcCo [Wang *et al.*, 2021]	90.6 ± 2.8	92.5 ± 1.9	82.9 ± 1.3	86.5 ± 2.2
DGCNN + CrossPoint [Afham *et al.*, 2022]	92.5 ± 3.0	94.9 ± 2.1	83.6 ± 5.3	87.9 ± 4.2
Transformer [Vaswani *et al.*, 2017]	87.8 ± 5.2	93.3 ± 4.3	84.6 ± 5.5	89.4 ± 6.3
Transformer + OcCo [Wang *et al.*, 2021]	94.0 ± 3.6	95.9 ± 2.3	89.4 ± 5.1	92.4 ± 4.6
Point-BERT [Yu *et al.*, 2021]	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1
MaskPoint [Liu *et al.*, 2022]	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
Joint-MAE	**96.7 ± 2.2**	**97.9 ± 1.8**	**92.6 ± 3.7**	**95.1 ± 2.6**

Table 4: **Few-shot Classification on ModelNet40[Wu *et al.*, 2015a]**. We report the average accuracy (%) and standard deviation (%) of 10 independent experiments.

Method	mIoU$_C$	mIoU$_I$
PointNet [Qi *et al.*, 2017a]	80.39	83.70
PointNet++ [Qi *et al.*, 2017b]	81.85	85.10
DGCNN [Wang *et al.*, 2019]	82.33	85.20
Transformer [Vaswani *et al.*, 2017]	83.42	85.10
Transformer + OcCo [Wang *et al.*, 2021]	83.42	85.10
Point-BERT [Yu *et al.*, 2021]	84.11	85.60
MaskPoint [Liu *et al.*, 2022]	84.40	86.00
Point-MAE [Pang *et al.*, 2022]	-	86.10
Joint-MAE	**85.41**	**86.28**

Table 5: **Part Segmentation on ShapeNetPart [Yi *et al.*, 2016]**. 'mIoU$_C$' (%) and 'mIoU$_I$' (%) denote the mean IoU across all part categories and all instances in the dataset, respectively.

jects for training and 576 objects for testing. Joint-MAE exceeds Point-MAE by an improvement of +0.89% on the most challenging part, i.e., 'PB-T50-RS' in Table 3. As the pre-training dataset, ShapeNet, is composed of synthetic point clouds, this results indicate the superiority of Joint-MAE for out-of-distribution data.

Few-shot Classification. To evaluate the representation ability of our Joint-MAE on limited training data, we further conduct few-shot classification task on ModelNet40 [Wu *et al.*, 2015a] as well. Following the common routine, we perform N-way K-shot experiments on Joint-MAE for 10 times, which means we randomly select N classes from ModelNet40, and sample K objects from each class. Table 4 reports the comparison results, where we exhibit the mean and standard deviation over 10 runs. As reported, our Joint-MAE outperforms prior works in almost all few-shot settings, indicating strong capacity under low-data regimes.

Part Segmentation. To evaluate the understanding ability of Joint-MAE on the fine-grained dense 3D data, we also conduct the part segmentation experiment on ShapeNetPart dataset [Yi *et al.*, 2016], which contains 16,881 instances of 16 categories. Following Point-MAE [Pang *et al.*, 2022], we sample 2,048 points from each input instance, and adopt the

Attention Schemes		Acc. (%)
2D-3D	3D-2D	
Global	Global	91.7
Local	Global	91.8
Global	Local	92.0
Local	Local	**92.4**

Table 6: **Ablation Results on Local-aligned Attention**. We utilize 'Global' and 'Local' to denote the standard global attention and local-aligned attention mechanism respectively. '2D-3D' and '3D-2D' denote the cross-modal positions during attention calculation.

same segmentation head for fair comparison, which concatenates the output features from different transformer blocks of the encoder. Table 5 shows the comparison results of our Joint-MAE with existing supervised methods and self-supervised methods. We here report the mean IoU across all part categories (denoted as mIoU$_C$) and all instances (denoted as mIoU$_I$), respectively. Clearly, our Joint-MAE achieves the highest scores on both metrics. This well illustrates the superior dense 3D representation capacity of Joint-MAE learned from our 2D-3D cross-modal pre-training.

4.3 Ablation Study

To explore the contributions of each major component in Joint-MAE, we conduct extensive ablations on ModelNet40 and evaluate the performance by the Linear SVM classification accuracy.

2D-3D Embedding Modules and Joint Decoder. On top of our final design, we conduct experiments by removing the hierarchical 2D-3D embedding modules and modal-shared decoder, respectively. As reported in Table 7, the absence (denoted by '-') of either modal-specific embedding module hurts the performance. The results show the effectiveness of both our initial embedding modules for multi-scale learning and the modal-shared decoder for 2D-3D feature interaction for mask tokens.

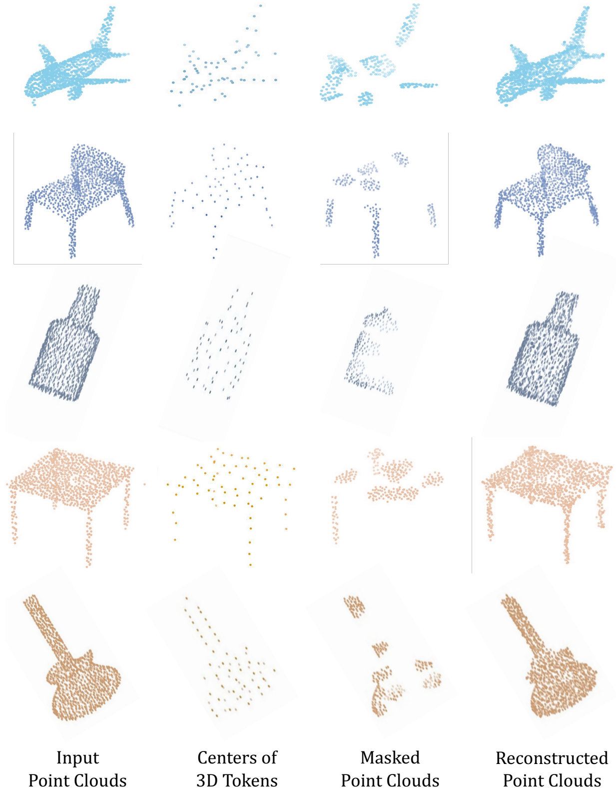

Figure 5: **Visualization of Point Clouds from Joint-MAE.** In each row, we visualize the input point clouds, the centers of 3D tokens, the masked point clouds, and the reconstructed coordinates.

Embed. Modules		Joint Decoder		Acc. (%)
2D	3D	Modal-shared	Modal-specific	
✓	✓	✓	✓	**92.4**
✓	-	✓	✓	91.2
-	✓	✓	✓	91.4
✓	✓	-	✓	91.8

Table 7: **Ablation Results on Hierarchical Embedding Modules and Joint Decoder**. 'Acc.' denotes the accuracy performance.

Local-aligned Attention. In Table 6, we experiment different attention schemes for cross-modal positions shown in Figure 3, i.e., 2D-3D and 3D-2D correlative tokens. We denote the local-aligned attention and the original attention as 'Local' and 'Global', respectively. The last row represents our Joint-MAE with local-aligned attention for all cross-modal positions. Compared with others, the local-aligned attention mechanism contributes to +0.5% classification accuracy, indicating that the proposed method can boost point cloud representation learning with 2D fine-grained semantic cues.

Cross-reconstruction Loss. In Table 8, we explore the projection settings of the cross-reconstruction loss. As reported, the proposed cross-reconstruction loss contributes to at least +0.5% on the classification scores. The comparison of the middle two rows reveals that projecting from multiple perspectives, which enforces more geometric constraints, can

Cross-Reconstruct.	Proj. Settings		Acc. (%)
	View Num.	Render	
-	-	-	91.1
✓	1	-	91.6
✓	4	-	**92.4**
✓	4	✓	91.9

Table 8: **Ablation Results on Loss Functions.** 'View Num' denotes the view number of projections in the cross-reconstruction loss. 'Render' denotes whether we conduct color rendering on the projected 2D depth maps.

well benefit 3D point cloud pre-training. Also, the performance decay in the last row indicates the rendering operation might introduce extra noises that disturb the effective supervision on the reconstructed 3D point clouds.

5 Visualization

In Figure 5, to intuitively verify the constructed performance of Joint-MAE, we visualize the input point clouds, the centers of 3D tokens ($N/32$ point number), the masked point clouds, and the reconstructed coordinates, respectively in each row. As shown, with the proposed 2D-3D joint framework and cross-modal learning strategies, the network can well generate masked point clouds with fine-grained local structures.

6 Conclusion

We propose Joint-MAE, a multi-modal masked autoencoding framework for 3D point cloud pre-training. With the hierarchical embedding modules and the joint 2D-3D transformer, Joint-MAE can well fuse the informative characters between 2D and 3D modalities within a unified MAE architecture. By equipping the local-aligned attention and cross-reconstruction loss, Joint-MAE further learns strong cross-modal knowledge for 3D point cloud learning. Extensive experiments are conducted to demonstrate the superiority of Joint-MAE on 3D point cloud representation capability. We expect Joint-MAE can inspire more works to further combine multi-modality learning with 3D point cloud MAE using a unified framework. As for limitations, although Joint-MAE has verified the 2D modality can effectively benefit 3D point cloud pre-training, we will further explore how 3D modality improves 2D MAE as the future work. We do not foresee a negative social impact from the proposed work.

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant No. 2022YFE0200700), the Hong Kong Innovation and Technology Commission (Project No. MHP/086/21), the China National Natural Science Foundation (Project No. 62202182), and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: T45-401/22-N).

Contribution Statement

Ziyu Guo and Renrui Zhang equally contribute to this work. Xianzhi Li is the corresponding author.

References

[Afham *et al.*, 2022] 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. *arXiv preprint arXiv:2203.00680*, 2022.

[Chang *et al.*, 2015] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[Desai and Johnson, 2021] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021.

[Fan *et al.*, 2017] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[Gao *et al.*, 2021] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[Gao *et al.*, 2022] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022.

[Gao *et al.*, 2023] Peng Gao, Renrui Zhang, Rongyao Fang, Ziyi Lin, Hongyang Li, Hongsheng Li, and Qiao Yu. Mimic before reconstruct: Enhancing masked autoencoders with feature mimicking. *arXiv preprint arXiv:2303.05475*, 2023.

[Goyal *et al.*, 2021] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. *arXiv preprint arXiv:2106.05304*, 2021.

[Guo *et al.*, 2021] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021.

[Guo *et al.*, 2022] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022.

[Han *et al.*, 2019] Zhizhong Han, Mingyang Shang, Yu-Shen Liu, and Matthias Zwicker. View inter-prediction gan: Unsupervised representation learning for 3d shapes by learning global shape memories to support local view predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8376–8384, 2019.

[He *et al.*, 2021] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[Li *et al.*, 2018] Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9397–9406, 2018.

[Liu *et al.*, 2022] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. *arXiv preprint arXiv:2203.11183*, 2022.

[Nichol *et al.*, 2022] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

[Pang *et al.*, 2022] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. *arXiv preprint arXiv:2203.06604*, 2022.

[Poursaeed *et al.*, 2020] Omid Poursaeed, Tianxing Jiang, Han Qiao, Nayun Xu, and Vladimir G Kim. Self-supervised learning of point clouds via orientation estimation. In *2020 International Conference on 3D Vision (3DV)*, pages 1018–1028. IEEE, 2020.

[Qi *et al.*, 2017a] 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*, pages 652–660, 2017.

[Qi *et al.*, 2017b] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[Rao *et al.*, 2020] Yongming Rao, Jiwen Lu, and Jie Zhou. Global-local bidirectional reasoning for unsupervised representation learning of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5376–5385, 2020.

[Sauder and Sievers, 2019] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32, 2019.

[Uy *et al.*, 2019] 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*, pages 1588–1597, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Wang *et al.*, 2019] 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.

[Wang *et al.*, 2021] 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*, pages 9782–9792, 2021.

[Wu *et al.*, 2015a] 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*, pages 1912–1920, 2015.

[Wu *et al.*, 2015b] 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*, pages 1912–1920, 2015.

[Xie *et al.*, 2016] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644. PMLR, 2016.

[Xie *et al.*, 2018] Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Learning descriptor networks for 3d shape synthesis and analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8629–8638, 2018.

[Xie *et al.*, 2020] Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Generative voxelnet: learning energy-based models for 3d shape synthesis and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2468–2484, 2020.

[Xie *et al.*, 2021] Jianwen Xie, Yifei Xu, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Generative pointnet: Deep energy-based learning on unordered point sets for 3d generation, reconstruction and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14976–14985, 2021.

[Xu *et al.*, 2021] Chenfeng Xu, Shijia Yang, Bohan Zhai, Bichen Wu, Xiangyu Yue, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2point: 3d point-cloud understanding with pretrained 2d convnets. *arXiv preprint arXiv:2106.04180*, 2021.

[Yang *et al.*, 2018] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018.

[Yi *et al.*, 2016] 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.

[Yu *et al.*, 2021] 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. *arXiv preprint arXiv:2111.14819*, 2021.

[Zhang *et al.*, 2021a] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[Zhang *et al.*, 2021b] 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. *arXiv preprint arXiv:2112.02413*, 2021.

[Zhang *et al.*, 2021c] 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*, pages 10252–10263, 2021.

[Zhang *et al.*, 2022a] 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. *arXiv preprint arXiv:2205.14401*, 2022.

[Zhang *et al.*, 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. *arXiv preprint arXiv:2212.06785*, 2022.

[Zhang *et al.*, 2022c] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022.

[Zhang *et al.*, 2023] Renrui Zhang, Liuhui Wang, Ziyu Guo, and Jianbo Shi. Nearest neighbors meet deep neural networks for point cloud analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1246–1255, 2023.

[Zhao *et al.*, 2021] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021.

[Zhu *et al.*, 2022] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022.