PiMAE: Point Cloud and Image Interactive Masked Autoencoders for 3D Object Detection

Anthony Chen[1,*], Kevin Zhang[1,*], Renrui Zhang[2], Zihan Wang[1],
Yuheng Lu[1,3], Yandong Guo[4], Shanghang Zhang[1,†]
[1]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University [2]The Chinese University of Hong Kong
[3]Wukong Lab, iKingtec [4]Beijing University of Posts and Telecommunications
{antonchen, kevinzyz6, shanghang}@pku.edu.cn
zhangrenrui@pjlab.org.cn, rzhevcherkasy@outlook.com
luyuheng@ikingtec.com, yandong.guo@live.com

Abstract

Masked Autoencoders learn strong visual representations and achieve state-of-the-art results in several independent modalities, yet very few works have addressed their capabilities in multi-modality settings. In this work, we focus on point cloud and RGB image data, two modalities that are often presented together in the real world, and explore their meaningful interactions. To improve upon the cross-modal synergy in existing works, we propose PiMAE, a self-supervised pre-training framework that promotes 3D and 2D interaction through three aspects. Specifically, we first notice the importance of masking strategies between the two sources and utilize a projection module to complementarily align the mask and visible tokens of the two modalities. Then, we utilize a well-crafted two-branch MAE pipeline with a novel shared decoder to promote cross-modality interaction in the mask tokens. Finally, we design a unique cross-modal reconstruction module to enhance representation learning for both modalities. Through extensive experiments performed on large-scale RGB-D scene understanding benchmarks (SUN RGB-D and ScannetV2), we discover it is nontrivial to interactively learn point-image features, where we greatly improve multiple 3D detectors, 2D detectors, and few-shot classifiers by 2.9%, 6.7%, and 2.4%, respectively. Code is available at https://github.com/BLVLab/PiMAE.

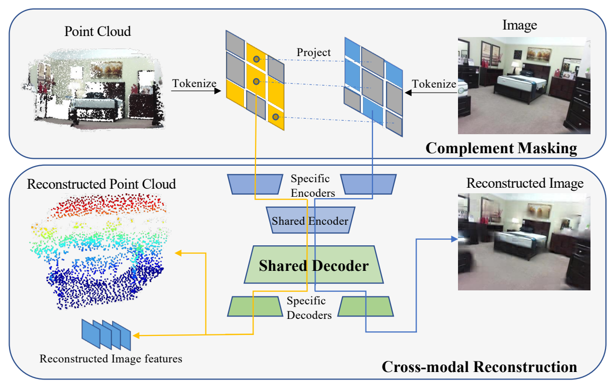

Figure 1. With our proposed design, PiMAE learns cross-modal representations by interactively dealing with multi-modal data and performing reconstruction.

1. Introduction

The advancements in deep learning-based technology have developed many significant real-world applications, such as robotics and autonomous driving. In these scenarios, 3D and 2D data in the form of point cloud and RGB images from a specific view are readily available. Therefore, many existing methods perform multi-modal visual learning, a popular approach leveraging both 3D and 2D information for better representational abilities.

Intuitively, the paired 2D pixels and 3D points present different perspectives of the same scene. They encode different degrees of information that, when combined, may become a source of performance improvement. Designing a model that interacts with both modalities, such as geometry and RGB, is a difficult task because directly feeding them to a model results in marginal, if not degraded, performance, as demonstrated by [34].

In this paper, we aim to answer the question: how to design a more interactive unsupervised multi-modal learn-

*Equal Contribution.
†Corresponding Author.

ing framework that is for better representation learning? To this end, we investigate the Masked Autoencoders (MAE) proposed by He *et al*. [20], which demonstrate a straightforward yet powerful pre-training framework for Vision Transformers [10] (ViTs) and show promising results for independent modalities of both 2D and 3D vision [2, 14, 17, 63, 64]. However, these existing MAE pre-training objectives are limited to only a single modality.

While much literature has impressively demonstrated MAE approaches' superiority in multiple modalities, existing methods have yet to show promising results in bridging 3D and 2D data. For 2D scene understanding among multiple modalities, MultiMAE [2] generates pseudo-modalities to promote synergy for extrapolating features. Unfortunately, these methods rely on an adjunct model for generating pseudo-modalities, which is sub-optimal and makes it hard to investigate cross-modality interaction. On the other hand, contrastive methods for self-supervised 3D and 2D representation learning, such as [1, 6, 7, 34, 58], suffer from sampling bias when generating negative samples and augmentation, making them impractical in real-world scenarios [8, 22, 73].

To address the fusion of multi-modal point cloud and image data, we propose PiMAE, a simple yet effective pipeline that learns strong 3D and 2D features by increasing their interaction. Specifically, we pre-train pairs of points and images as inputs, employing a two-branched MAE learning framework to individually learn embeddings for the two modalities. To further promote feature alignment, we design three main features.

First, we tokenize the image and point inputs, and to correlate the tokens from different modalities, we project point tokens to image patches, explicitly aligning the masking relationship between them. We believe a specialized masking strategy may help point cloud tokens embed information from the image, and vice versa. Next, we utilize a novel symmetrical autoencoder scheme that promotes strong feature fusion. The encoder draws inspiration from [31], consisting of both separate branches of modal-specific encoders and a shared-encoder. However, we notice that since MAE's mask tokens only pass through the decoder [20], a shared-decoder design is critical in our scheme for mask tokens to learn mutual information before performing reconstructions in separate modal-specific decoders. Finally, for learning stronger features inspired by [15,67], PiMAE's multi-modal reconstruction module tasks point cloud features to explicitly encode image-level understanding through enhanced learning from image features.

To evaluate the effectiveness of our pre-training scheme, we systematically evaluate PiMAE with different fine-tuning architectures and tasks, including 3D and 2D object detection and few-shot image classification, performed on the RGB-D scene dataset SUN RGB-D [51] and ScannetV2 [9] as well as multiple 2D detection and classification datasets. We find PiMAE to bring improvements over state-of-the-art methods in all evaluated downstream tasks.

Our main contributions are summarized as:

- To the best of our knowledge, we are the first to propose pre-training MAE with point cloud and RGB modalities interactively with three novel schemes.
- To promote more interactive multi-modal learning, we novelly introduce a complementary cross-modal masking strategy, a shared-decoder, and cross-modal reconstruction to PiMAE.
- Shown by extensive experiments, our pre-trained models boost performance of 2D & 3D detectors by a large margin, demonstrating PiMAE's effectiveness.

2. Related Work

3D Object Detectors. 3D object detection aims to predict oriented 3D bounding boxes of physical objects from 3D input data. Many CNN-based works propose two-stage methods for first generating region proposals and then classifying them into different object types. Prior 3D object detection methods adapt popular 2D detection approaches to 3D scenes, projecting point cloud data to 2D views [29, 33, 60] for 2D ConvNets to detect 3D bounding boxes. Other approaches adopt 3D ConvNets by grouping points into voxels [47, 75] and transposed convolutions for sparse detection [19]. Recently, the Transformer architecture [54] has demonstrated consistent and impressive performance in vision, specifically with object detectors [5, 35–37, 39, 46, 48, 49, 68, 72, 74]. Transformers are especially well-designed for 3D point clouds, needing not hand-crafted groupings and capable of having an invariant understanding. Specifically, [39] proposed an end-to-end Transformer-based object detection module using points cloud as input. Group-Free-3D [35] designed a novel attention stacking scheme and estimated detection results by fusing object features in different stages. In PiMAE, we draw inspiration from both projection-based and attention-based 3D object detectors; whereas the former projection mechanisms have been extensively utilized previously, the latter has shown better versatility and a more intuitive solution. Consequently, we design a MAE [20]-structured multi-modal learning framework that incorporates projection alignment for more interactive multi-modal learning.

Point Cloud and Image Joint Representation Learning. 3D point cloud and 2D image joint representation learning methods aim to explore the modal interaction between point clouds and images for feature fusion. Many recent studies have shown that cross-modal modules outperform single-modal methods on multiple tasks such as 3D object detection [25, 26, 56, 57, 62, 71], 3D semantic segmentation [23, 27, 42], and 3D open-world learning [18, 65, 69, 76]. In cross-modal self-supervised learning of point clouds and RGB images, several methods [24, 32, 34] based on contrastive learning propose to design specialized structures for learning from multiple modalities surpass single modalities when fine-tuned on downstream tasks including 3D object detection. As aforementioned, while contrastive methods have illustrated the significance of pairing RGB and point clouds, PiMAE has several advantages over contrastive methods, mainly requiring fewer augmentations.

Masked Autoencoders (MAE). Recently, inspired by advances in masked language modeling, masked image modeling (MIM) approaches [3, 20, 59] have shown superior performance, proposing a self-supervised training method based on masked image prediction. MAE [20], in particular, predicts pixels from highly masked images using a ViT decoder. Since MAE's success, several works [13, 21, 38, 41, 64, 70] have applied the framework to point cloud data, proposing to segment point cloud into tokens and perform reconstruction. Moreover, MultiMAE [2] investigates the alignment of various modalities with MAE among RGB images, depth images, and semantic segmentation. Recently, I2P-MAE [67] explores leveraging 2D pre-trained knowledge to guide 3D MAE pre-training. In this work, however, we demonstrate that earlier methods do not maximize the potential of point cloud and RGB scene datasets, because they cannot incorporate the RGB inputs with ease and bring only trivial performance gain. To the best of our knowledge, this is the pioneering work aligning RGB images with point cloud with MAE pre-training.

3. Methods

In this section, we first give an overview of our pipeline. Then, we introduce our novelly designed masking strategy, which aligns the semantic information between tokens from two modalities. Followingly, we present our cross-modal encoders and decoders design. Notably, the shared-decoder is a pioneering architecture. Finally, we finish with our cross-modal reconstruction module.

3.1. Pipeline Overview

As shown in Fig. 3, PiMAE learns cross-modal representations simultaneously by jointly learning features from point clouds and image modalities. In our proposed pipeline, we first embed point data into tokens by sampling and clustering algorithms and then perform random masking on point tokens. The mask pattern is transformed onto the 2D plane, where patches of images are complementarily masked and embedded into tokens.

Following this, we utilize a symmetrical joint-encoder-decoder scheme that promotes strong feature fusion. The encoder-decoder architecture consists of both separate

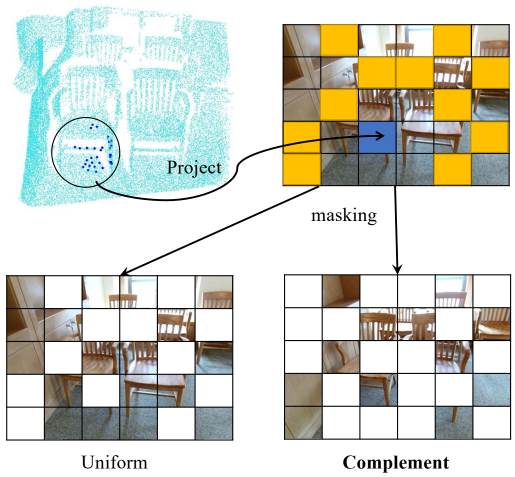

Figure 2. **Illustration of our projection operation and two different masking strategies.** A randomly sampled point cloud cluster (black circle) is projected onto the image patch (blue square), and the other clusters are done in a similar way (yellow squares). Under uniform masking, the yellow patches will be masked while other patches are visible. On the contrary, complement masking will result in a reversed masking pattern.

branches and shared modules, whereas the former protects modal-specific learning and the latter encourages cross-modal interaction for more robust features.

Finally, for learning stronger features from pre-training, PiMAE's cross-modal reconstruction module demands point cloud features to explicitly express image-level understanding.

3.2. Token Projection and Alignment

We follow MAE [20] and Point-M2AE [64] to generate input tokens from images and point clouds. An image is first divided into non-overlapping patches, before the embedding procedure that embeds patches by a linear projection layer with added Positional Embeddings (PE) and Modality Embeddings (ME). Correspondingly, a set of point clouds is processed into cluster tokens via Farthest Point Sampling (FPS) and K-Nearest Neighbour (KNN) algorithms and then embedded with a linear projection layer with added embeddings (i.e. PE, ME).

Projection. In order to achieve the alignment between multi-modality tokens, we build a link between the 3D point cloud and RGB image pixels by projecting the point cloud onto the camera's image plane. For 3D point $P \in \mathbb{R}^3$, a correlating 2D coordinate can be calculated using the projection function $Proj$ defined below,

$$\begin{bmatrix} u \\ v \\ z \end{bmatrix} = Proj(P) = K \cdot R_t \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{1}$$

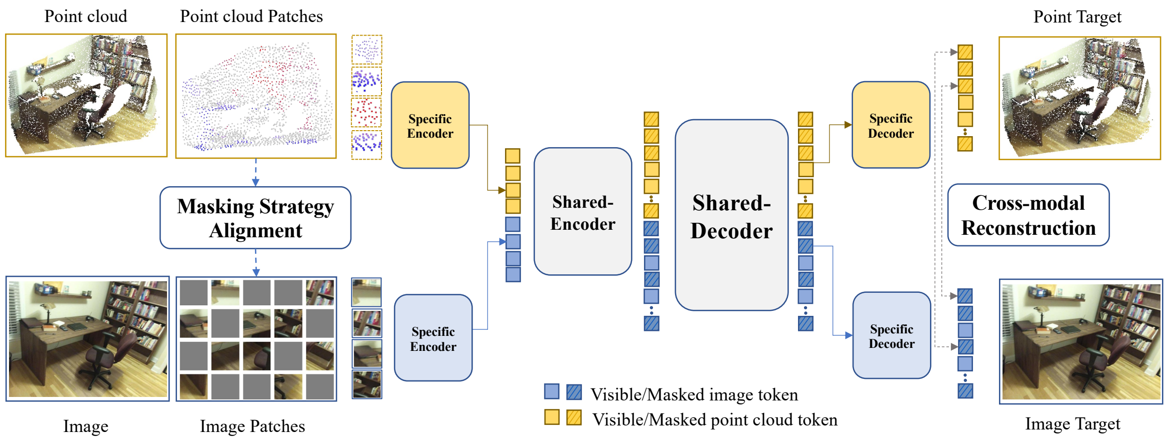

Figure 3. **Pre-training pipeline for PiMAE.** The point cloud branch samples and clusters point cloud data into tokens and randomly masks the input. Then the tokens pass through the masking alignment module to generate complement masks for image patches. After embedding, tokens go through a separate, then shared, and finally separated autoencoder structure. We lastly engage in a cross-modal reconstruction module to enhance point cloud representation learning. Point cloud is colored for better visualization.

where $K \in 3 \times 4$, $R_t \in 4 \times 4$ are the camera intrinsic and extrinsic matrices. (x, y, z), (u, v) are the original 3D coordinate and projected 2D coordinate of point P.

Masking with Alignment. Next, we generate masked tokens using the aforementioned projection function. Since point cloud tokens are organized by the cluster centers, we randomly select a subset of center points as well as their corresponding tokens, while keeping the rest masked. For the visible point cloud tokens T_p, we project their center point $P \in \mathbb{R}^3$ to the corresponding camera plane and attain its 2D coordinate $p \in \mathbb{R}^2$, which can naturally fall into an area of shape $H \times W$ (i.e. image shape), thus obtaining its related image patch index I_p by

$$I_p = \lfloor \lfloor v \rfloor / S \rfloor \times \lfloor W/S \rfloor + \lfloor \lfloor u \rfloor / S \rfloor, \tag{2}$$

where u and v denotes the x-axis value and y-axis value of 2D coordinate p, S is the image patch size.

After projecting and indexing each visible point cloud token, we obtain their corresponding image patches. Next, we explicitly mask these patches to reach a complement mask alignment. The rationale is that such masking pattern can make visible tokens more semantically abundant than under the uniform setting, and thus the model is able to extract rich cross-modal features. For visual demonstration of our projection and alignment, see Fig. 2.

3.3. Encoding Phase

Encoder. During this stage, we protect the integrity of different modalities. Inspired by AIST++ [31], our encoder consists of two modules: modal-specific encoders and a cross-modal encoder. The former is used to better extract modal-specific features, and the latter is used to perform interaction between cross-modal features.

The modality-specific encoder part contains two branches for two modalities, where each branch consists of a ViT backbone. First, for the encoders to learn modality differences through mapping inputs to feature spaces, we feed the aligned, visible tokens with their respective positional and modality embeddings to separate encoders.

Later, we promote feature fusion and cross-modality interactions of visible patches with a shared-encoder. The alignment of masks during this stage becomes critical, as aligned tokens reveal similar information reflected in both 3D and 2D data.

Formally, in the separate encoding phase, $E_I : T_I \mapsto L_I^1$ and $E_P : T_P \mapsto L_P^1$, where E_I and E_P are the image-specific and point-specific encoders, T_I and T_P are the visible image and point patch tokens, and L_I^1 and L_P^1 are the image and point latent spaces. Then, the shared-encoder performs fusion on the different latent representations $E_S : L_I^1, L_P^1 \mapsto L_S^2$.

3.4. Decoding Phase

Decoder. Generally, MAE encoders benefit from learning generalized encoders that capture high-dimensional data encoding representations for both image and point cloud data. Due to the differences between the two modalities, specialized decoders are needed to decode the high-level latent to the respective modality.

The purpose of additional shared-decoder layers is ultimately for the encoder to focus more on feature extraction and ignore the details of modality interactions. Because MAE uses an asymmetric autoencoder design where the mask tokens do not pass the shared-encoder, we complement the mask tokens to pass through a shared-decoder, along with the visible tokens. Without such a design, the entire decoder branches are segmented, and the mask tokens of different modalities do not engage in feature fusion. After the shared-decoder, we then design specialized decoders for the different modalities for better reconstructions.

Since the reconstructions of two modalities are involved, the losses of both the point cloud and the image modalities are obtained. For point clouds, we use ℓ_2 Chamfer Distance [11] for loss calculation, denoted as $\mathcal{L}_{pc}$, and for images, we use MSE to measure the loss, denoted as $\mathcal{L}_{img}$.

Formally, the input for our shared-decoder is $L_S^{2'}$, the full sets of tokens including encoded visible features and mask tokens of both modalities, and the shared-decoder performs cross-modal interaction on these latent representations $D_S : L_S^{2'} \mapsto L_I^3, L_I^3$. Then, in the separate decoder phase, the decoders map back to the image and point cloud space, $D_I : L_I^3 \mapsto T_I^{'}$ and $D_P : L_P^3 \mapsto T_P^{'}$, where D_I and D_P are the image-specific and point-specific decoders, $T_I^{'}$ and $T_P^{'}$ are the visible image and point cloud patches, and L_I^3 and L_P^3 are the image and point cloud latent spaces.

$$\mathcal{L}_{pc} = CD(D_P(l), P_{GT}), \tag{3}$$

where CD is ℓ_2 Chamfer Distance function [11], D_P represents the decoder reconstruction function, $l \in L_P^3$ is the point cloud latent representation, P_{GT} is the ground truth point cloud (i.e. point cloud input).

3.5. Cross-modal Reconstruction

We train PiMAE using three different losses: the point cloud reconstruction loss, the image reconstruction loss, and a cross-modal reconstruction loss that we design to further strengthen the interaction between the two modalities. In the final reconstruction phase, we utilize the previously aligned relationship to obtain the corresponding 2D coordinates of the masked point clouds. Then, we up-sample the reconstructed image features, such that each masked point cloud with a 2D coordinate can relate to a reconstructed image feature. Finally, the masked point cloud tokens go through a cross-modal prediction head of one linear projection layer to recover the corresponding visible image features. Note that we specifically avoid using visible point cloud tokens for this module, because they correspond to the masked image features (due to the complement masking strategy), which tend to have weaker representations and may harm representation learning. Formally, the cross-modal reconstruction loss is defined as

$$\mathcal{L}_{cross} = MSE(D_P(l_p^3), l_i^3), \tag{4}$$

where MSE denotes the Mean Squared Error loss function, D_P is the cross-modal reconstruction from the decoder, $l_p^3 \in L_P^3$ is the point cloud representation, $l_i^3 \in L_I^3$ is the image latent representation.

Our final loss is the sum of the previous loss terms, formulated in Eq. 5. By such design, PiMAE learns 3D and 2D features separately while maintaining strong interactions between the two modalities.

$$\mathcal{L} = \mathcal{L}_{pc} + \mathcal{L}_{img} + \mathcal{L}_{cross}. \tag{5}$$

4. Experiments

In this section, we provide extensive experiments to qualify the superiority of our methods. The following experiments are conducted. a) We pre-train PiMAE on the SUN RGB-D [51] training set. b) We evaluate PiMAE on various downstream tasks, including 3D object detection, 3D monocular detection, 2D detection, and classification. c) We ablate PiMAE with different multi-modal interaction strategies to show the effectiveness of our proposed design.

4.1. Implementation Details

Datasets and metrics. We pre-train our model on SUN RGB-D [51] and evaluate with different downstream tasks on several datasets including indoor 3D datasets (SUN RGB-D [51], ScanNetV2 [9]), outdoor 3D dataset (KITTI [16]), few-shot image classification datasets (CIFAR-FS [4], FC100 [40], miniImageNet [55]). Detailed descriptions of these datasets and evaluation metrics are in the Appendix.

Network architectures. Abiding by common practice [41, 64], we utilize a scaled-down PointNet [45] before a ViT [10] backbone in our point cloud branch. PointNet layers effectively reduce the sub-sampled points from 20,000 to 2,048. For the image branch, we follow [20] to divide images into regular patches with a size of 16×16, before the ViT backbone.

Pre-training. During this stage, we use the provided image and generated point cloud from SUN RGB-D [51] to train PiMAE for 400 epochs. AdamW [28] optimizer with a base learning rate of 1e-3 and weight decay of 0.05 is used, applied with a warm-up for 15 epochs. No augmentation is performed on both image and point cloud inputs, for the main goal of maintaining consistency between the patches. Experimentally, we find that a masking ratio of 60% is more appropriate. The reconstructed visualization results are in Fig. 4. Detailed configurations are in the Appendix.

Fine-tuning. With PiMAE's two multi-modal branches, we fine-tune and evaluate our learned features on both 3D and 2D tasks. For 3D tasks, we use the point cloud branch's specific encoder and the shared encoder as a 3D feature extractor. For 2D tasks, similarly, we utilize the image-specific encoder as well as the shared encoder as a 2D feature extractor. We fit our feature extractors into different baselines and keep the same training settings, except for the modifications on the backbone feature extractor. Detailed configurations are in the Appendix.

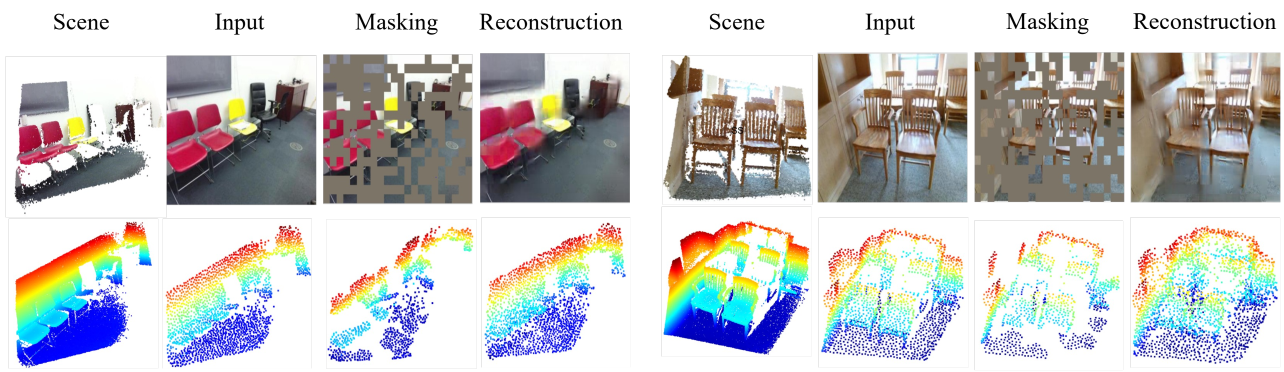

Figure 4. **Reconstruction results of images and point cloud from PiMAE.** Our model is able to perform image and point cloud reconstruction simultaneously, showing a firm understanding of the two modalities. Image results are in the first row, point cloud results in the second. The masking ratio for both branches is 60%. Point cloud is colored for better visualization.

Methods	Pre-trained	SUN RGB-D AP_{25}	SUN RGB-D AP_{50}	ScanNetV2 AP_{25}	ScanNetV2 AP_{50}
DSS [52]	*None*	42.1	-	15.2	6.8
PointFusion [61]	*None*	45.4	-	-	-
3D-SIS [23]	*None*	-	-	40.2	22.5
VoteNet [43]	*None*	57.7	32.9	58.6	33.5
3DETR [39]	*None*	58.0	30.3	62.1	37.9
+Ours(from scratch)	*None*	58.7	31.7	59.7	40.0
+Ours	SUN RGB-D	59.4(+1.4)	33.2(+2.9)	62.6(+0.5)	39.4(+1.5)
GroupFree3D [35]	*None*	63.0	45.2	67.3	48.9
+Ours(from scratch)	*None*	61.2	44.7	65.5	47.4
+Ours	SUN RGB-D	**64.6(+1.6)**	**46.2(+1.0)**	**67.6(+0.3)**	**49.7(+0.8)**

Table 1. **3D object detection results on ScanNetV2 [9] and SUN RGB-D [51].** We adopt the average precision with 3D IoU thresholds of 0.25 (AP_{25}) and 0.5 (AP_{50}) for the evaluation metrics.

4.2. Results on Downstream Tasks

In this work, we evaluate our method on four different downstream tasks dealing with different modalities , including 3D object detection, monocular 3D object detection, 2d object detection, and few-shot image classification.

Indoor 3D object detection. We apply our 3D feature extractor on 3D detectors by replacing or inserting the encoder into different backbones to strengthen feature extraction. We report our performance on indoor 3D detection based on SOTA methods 3DETR [39] and GroupFree3D [35]. As shown in Tab. 1, our model brings significant improvements to both models, surpassing previous baselines consistently in all datasets and criteria.

Furthermore, in the Appendix, we provide 3D object detection results with detailed per-class accuracy on SUN RGB-D, along with visualizations of detection results.

Outdoor monocular 3D object detection. To fully demonstrate the capacity of our approach, we report our performance in challenging outdoor scenarios, which have a large data distribution gap compared with our indoor pre-training data. As shown in Tab. 3, we brought substantial improvement to MonoDETR [66], validating that our pre-trained representations generalize well to both indoor and outdoor datasets.

2D object detection. Similarly, we apply our 2D branch's feature extractor to 2D detector DETR [5] by replacing its vanilla transformer backbone. We conduct experiments on both pre-trained and scratch backbones and report our performance on the ScanNetV2 2D detection dataset. As shown in Tab. 2, our model significantly improves the performance of DETR, demonstrating the strong generalization ability on 2D tasks of our model.

Few shot image classification. We conduct few-shot image classification experiments on three different benchmarks to explore the feature-extracting ability of PiMAE's

Methods	AP_{50}	AP_{75}	AP
*DETR [5]	39.8	26.2	25.3
+ PiMAE	**46.5(+6.7)**	**30.3(+4.1)**	**29.5(+4.2)**

Table 2. **2D object detection results on ScanNetV2 val set.** * denotes our implementation on ScanNetV2. We later load PiMAE pre-trained weights to the encoder.

Methods	Easy	Mod.	Hard
*MonoDETR [66]	23.1	17.3	14.5
+ PiMAE	**26.6(+3.5)**	**18.8(+1.5)**	**15.5(+1.0)**

Table 3. **Monocular 3D object detection results of car category on KITTI *val* set.** * denotes our implementation with adjusted depth encoder, to which we later load PiMAE pre-trained weights.

image encoder. To verify the effectiveness of PiMAE, we use no extra design for the classifier by only adding a linear layer to the feature encoder, predicting the class based on *[CLS]* token as input. Tab. 4 summarizes our results. We see significant improvements from PiMAE pre-training compared to models trained from scratch. Moreover, our performance surpasses previous SOTA self-supervised multi-modal learning method, CrossPoint [1].

4.3. Ablation Study

In this section, we investigate our methods, evaluating the quality of different PiMAE pre-training strategies both qualitatively and quantitatively. First, we attempt different alignment strategies between the masked tokens. Next, we pre-train with different reconstruction targets. Then, we ablate performance pre-trained with only a single branch. Finally, we examine our data efficiency by training on limited data. Additional ablation studies on model architecture and masking ratios are in our Appendix. All experiments are based on 3DETR and performed on SUN RGB-D unless otherwise stated.

Cross-modal masking. To better study the mask relationships between the two modalities, we design two masking strategies based on projection alignment: uniform masking and complement masking. Whereas the former masks both modalities in the same pattern that masked portions of one modality will correspond to the other when projected onto it, the latter is the opposite, i.e. a visible point cloud patch will be masked when projected on the image.

We pre-train both the uniform, complement as well as random masking strategies and evaluate their fine-tuning performance on 3D detection. Random masking acts as a baseline where the masking pattern of image and point patches are individually and randomly sampled. As shown in Tab. 5, masking the tokens from different modalities complementarily gains higher performance than randomly.

Compared to random masking, complement masking enables more cross-modal interactions between patches with diverse semantic information, thus helping our model to transfer 2D knowledge into the 3D feature extractor. However, with the uniform masking strategy, the extracted point cloud features and image features are semantically aligned, so the interaction does not help the model utilize 2D information better.

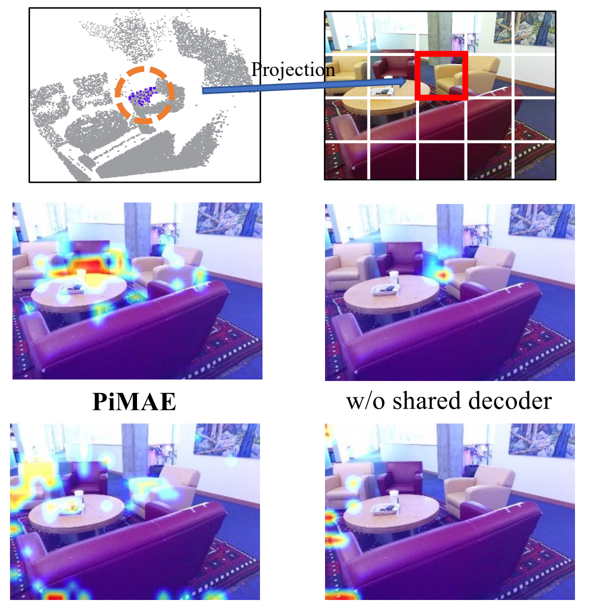

Figure 5. **Visualization of attention.** The encoder attention between the two modalities is visualized by computing self-attention from the query points (orange circle) to all the visible image tokens. We show the corresponding location (red square) of the query points after projection. See more examples in the Appendix.

Note that in Tab. 5, we purposely pre-train our model without the cross-modal reconstruction module. This is because the uniform masking strategy projects onto the masked image features, which are semantically weaker and may negatively influence our ablation study.

Effect of Cross-modal Reconstruction. Other than reconstructing the inputs, we promote cross-modal reconstruction by demanding point cloud features to reconstruct features or pixels of the corresponding image. We assess the significance of such design by ablating results on 3D detection. Shown in Tab. 6, the additional feature-level cross-modal reconstruction target brings additional performance gains.The promoted cross-modal reconstruction at the feature level encourages further interactions between modalities and encodes 2D knowledge into our feature extractor, improving model performance on downstream tasks.

Necessity of joint pre-training. To demonstrate the ef-

	CIFAR-FS 5-way		FC100 5-way		miniImageNet 5-way	
Method	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MAML [12]	58.9	71.5	-	-	48.7	63.1
Matching Networks [55]	-	-	-	-	43.6	55.3
Prototypical Network [50]	55.5	72.0	35.3	48.6	49.3	68.2
Relation Network [53]	55.0	69.3	-	-	50.4	65.3
CrossPoint [1]	64.5	80.1	-	-	-	-
PiMAE From Scratch	62.4	76.6	37.3	50.5	50.1	66.7
PiMAE Pre-trained	**66.9**	**80.7**	**39.0**	**53.3**	**55.3**	**70.2**

Table 4. **Few-shot image classification on CIFAR-FS, FC100 and miniImageNet test sets.** We report top-1 classification accuracy under 5-way 1-shot and 5-way 5-shot settings. Results of CrossPoint and previous methods are from [1, 4].

Masking Strategy	AP_{25}	AP_{50}
Random	58.0	32.9
Uniform	58.1	32.6
Complement	**59.0**	**33.0**

Table 5. **Comparisons of cross-modality masking strategies.**

Point Cloud			RGB	AP_{25}	AP_{50}
3D Geo	2D feat	2D pix	2D pix		
✓			✓	59.0	33.0
✓		✓	✓	58.0	31.6
✓	✓		✓	**59.4**	**33.2**

Table 6. **Ablation studies of cross-modal reconstruction targets.** Geo, feat, and pix refer to coordinates, features, and pixels, respectively.

fectiveness and the cruciality of double-branch pre-training in PiMAE, we provide the comparison of PiMAE pre-trained with one branch only. As shown in Tab. 7. A critical performance drop can be seen among double-branch PiMAE and single-branch PiMAE. Such evidence reveals the fact that PiMAE learns 3D and 2D features jointly and the cross-modal interactions that we propose help the model to utilize information from both modalities.

Furthermore, to demonstrate the effectiveness of PiMAE's cross-modal interaction design, we visualize the attention map in our shared encoder in Fig. 5. With our proposed design, PiMAE focuses on more foreground objects with higher attention values, showing a strong cross-modal understanding. See more examples in our Appendix.

Data efficiency. We train 3DETR and PiMAE-based 3DETR using limited annotated labels (varying from 1% to 100%) while testing on the full val set on SUN RGB-D. As shown in Tab. 6, PiMAE is able to outperform the baseline in every scenario, with the largest difference being 17.4% (AP_{25}) when 10% of labels are used.

	3D Object Detection		Few-shot image classification	
Input	AP_{25}	AP_{50}	5-way 1-shot	5-way 5-shot
RGB	-	-	66.3	79.5
Geo	58.4	32.3	-	-
RGB+Geo	**59.4**	**33.2**	**66.9**	**80.7**

Table 7. **Improvement from joint pre-training.** We compare results on 3D object detection (on SUN RGB-D) and few-shot image classification (on CIFAR-FS) tasks when pre-trained with a single-branch PiMAE.

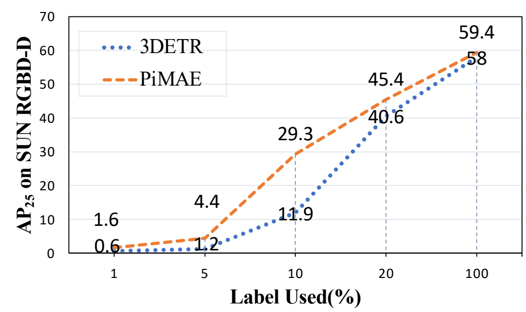

Figure 6. **Illustration of Data Efficiency.** Compared to 3DETR, PiMAE is able to ease the burden of data labeling and increase performance significantly.

5. Conclusion

In this work, we demonstrate PiMAE's simple framework is an effective and highly interactive multi-modal learning pipeline with strong feature extraction abilities on point cloud and image. We design three aspects for promoting cross-modality interaction. First, we explicitly align the mask patterns of both point cloud and image for better feature fusion. Next, we design a shared-decoder to accommodate mask tokens of both modalities. Finally, our cross-modality reconstruction enhances the learned semantics. In our extensive experiments and ablation studies performed on datasets of both modalities, we discover that PiMAE has great potential, improving multiple baselines and tasks.

References

[1] 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 (CVPR)*, pages 9902–9912, June 2022. 2, 7, 8

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022. 2, 3

[3] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. 2021. 3

[4] Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. 5, 8, 12

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020. 2, 6, 7, 12, 13

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2

[7] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 2

[8] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775. Curran Associates, Inc., 2020. 2

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2, 5, 6, 12

[10] 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. *ICLR*, 2021. 2, 5, 12

[11] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image, 2016. 5

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017. 8

[13] Kexue Fu, Peng Gao, ShaoLei Liu, Renrui Zhang, Yu Qiao, and Manning Wang. Pos-bert: Point cloud one-stage bert pre-training. *arXiv preprint arXiv:2204.00989*, 2022. 3

[14] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 2

[15] Peng Gao, Renrui Zhang, Hongyang Li, Hongsheng Li, and Yu Qiao. Mimic before reconstruct: Enhance masked autoencoders with feature mimicking. 2

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5, 12

[17] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. 2022. 2

[18] 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. 3

[19] JunYoung Gwak, Christopher B Choy, and Silvio Savarese. Generative sparse detection networks for 3d single-shot object detection. In *European conference on computer vision*, 2020. 2

[20] 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*, pages 16000–16009, 2022. 2, 3, 5

[21] Georg Hess, Johan Jaxing, Elias Svensson, David Hagerman, Christoffer Petersson, and Lennart Svensson. Masked autoencoders for self-supervised learning on automotive point clouds. *arXiv preprint arXiv:2207.00531*, 2022. 3

[22] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34:26449–26461, 2021. 2

[23] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans, 2018. 3, 6

[24] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts, 2020. 3

[25] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning?, 2021. 3

[26] Peixiang Huang, Li Liu, Renrui Zhang, Song Zhang, Xinli Xu, Baichao Wang, and Guoyi Liu. Tig-bev: Multi-view bev 3d object detection via target inner-geometry learning. *arXiv preprint arXiv:2212.13979*, 2022. 3

[27] Maximilian Jaritz, Tuan-Hung Vu, Raoul De Charette, Émilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. 5, 12

[29] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander. Joint 3d proposal generation and object detection from view aggregation, 2017. 2

[30] Jean Lahoud and Bernard Ghanem. 2d-driven 3d object detection in rgb-d images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4632–4640, 2017. 14

[31] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. 2, 4

[32] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, Junjun Jiang, Bolei Zhou, and Hang Zhao. Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1500–1508, 2022. 3

[33] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection, 2020. 2

[34] Yunze Liu, Li Yi, Shanghang Zhang, Qingnan Fan, Thomas A. Funkhouser, and Hao Dong. P4contrast: Contrastive learning with pairs of point-pixel pairs for RGB-D scene understanding, 2020. 2, 3

[35] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. *arXiv preprint arXiv:2104.00678*, 2021. 2, 6, 12, 14

[36] Zhe Liu, Xin Zhao, Tengteng Huang, Ruolan Hu, Yu Zhou, and Xiang Bai. Tanet: Robust 3d object detection from point clouds with triple attention. *ArXiv*, abs/1912.05163, 2020. 2

[37] Mingyuan Mao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Errui Ding, Baochang Zhang, Shumin Han, et al. Dual-stream network for visual recognition. *Advances in Neural Information Processing Systems*, 34:25346–25358, 2021. 2

[38] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxel-mae: Masked autoencoders for pre-training large-scale point clouds. *arXiv preprint arXiv:2206.09900*, 2022. 3

[39] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In *ICCV*, 2021. 2, 6, 12, 13, 14

[40] Boris N. Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. 2018. 5, 12

[41] 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. 3, 5

[42] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7108–7117, 2021. 3

[43] Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds, 2019. 6, 12, 14

[44] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum pointnets for 3d object detection from rgb-d data, 2017. 12, 14

[45] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016. 5, 12

[46] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. *arXiv preprint arXiv:2112.00322*, 2021. 2

[47] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. *CoRR*, abs/2106.01178, 2021. 2

[48] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020. 2

[49] S Shi, X Wang, H Pointrcnn Li, et al. 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE conference on computer vision and pattern recognition, Long Beach, CA, USA*, pages 15–20, 2019. 2, 12

[50] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning, 2017. 8

[51] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 2, 5, 6, 12

[52] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images, 2015. 6, 14

[53] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning, 2017. 8

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 2

[55] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 5, 8, 12

[56] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021. 3

[57] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual and language learning. *arXiv preprint arXiv:2209.14941*, 2022. 3

[58] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, pages 574–591. Springer, 2020. 2

[59] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[60] Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. In *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 2

[61] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation, 2017. 6, 14

[62] Hao Yang, Chen Shi, Yihong Chen, and Liwei Wang. Boosting 3d object detection via object-focused image fusion. *arXiv preprint arXiv:2207.10589*, 2022. 3

[63] Kevin Zhang and Zhiqiang Shen. i-mae: Are latent representations in masked autoencoders linearly separable? *arXiv preprint arXiv:2210.11470*, 2022. 2

[64] 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. 2, 3, 5, 12

[65] 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*, pages 8552–8562, 2022. 3

[66] Renrui Zhang, Han Qiu, Tai Wang, Xuanzhuo Xu, Ziyu Guo, Yu Qiao, Peng Gao, and Hongsheng Li. Monodetr: Depth-aware transformer for monocular 3d object detection. *arXiv preprint arXiv:2203.13310*, 2022. 6, 7, 12, 13

[67] 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. 2, 3

[68] Renrui Zhang, Liuhui Wang, Yali Wang, Peng Gao, Hongsheng Li, and Jianbo Shi. Parameter is not all you need: Starting from non-parametric networks for 3d point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[69] 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. 3

[70] 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. 3

[71] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives, 2020. 3

[72] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. *CoRR*, abs/2012.09164, 2020. 2

[73] Han Zhao, Xu Yang, Zhenru Wang, Erkun Yang, and Cheng Deng. Graph debiased contrastive learning with joint representation clustering. In *IJCAI*, pages 3434–3440, 2021. 2

[74] Minghang Zheng, Peng Gao, Renrui Zhang, Kunchang Li, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020. 2

[75] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection, 2017. 2

[76] 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. 3