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

Anthony Chen[1,*], Kevin Zhang[1,*], Zihan Wang[2],
Renrui Zhang[1,3], Yuheng Lu[4], Yandong Guo[5], Shanghang Zhang[1,†] , [*]
[1]Peking University [2]Beihang University
[3]Shanghai AI Laboratory [4]Wukong Lab, iKingtec [5]BUPT
{antonchen, kevinzyz6, shanghang}@pku.edu.cn
zhangrenrui@pjlab.org.cn, luyuheng@ikingtec.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), we discover it is nontrivial to interactively learn point-image features, where we greatly improve multiple 3D detectors and 2D few-shot classifiers by 2.9% and 2.4%, respectively.

1. Introduction

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

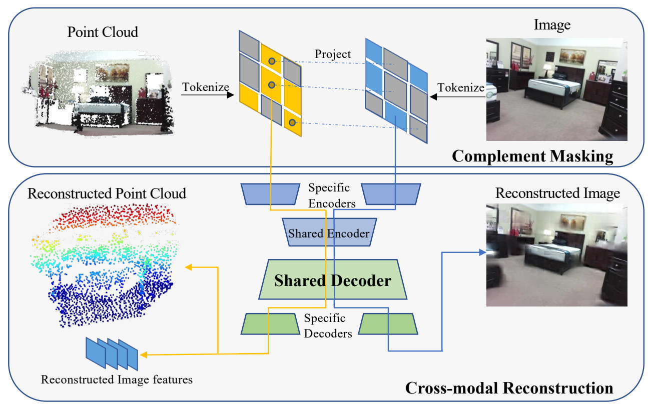

Figure 1. With our proposed design, PiMAE learns cross-modal representations by interactively dealing with multi-modal data and performing reconstruction.

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 be 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 both results in marginal, if not degraded, performance, as demonstrated by [26].

In this paper, we aim to answer the question: How to design a more interactive unsupervised multi-modal learning framework that is for better representation learning? To this end, we investigate the Masked Autoencoders (MAE) proposed by He et al. [16], 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 [13, 2, 14, 62, 61]. However, these existing MAE pre-training objectives are limited to a single modality.

While many literature have impressively demonstrated

** Equal Contribution, † Corresponding Author.

MAE approaches' superiority with multiple vision modalities, existing methods have yet to show promising results in bridging 3D and 2D data. For 2D scene understanding between multiple modalities, [2] adopts the approach of generating pseudo-modality to effectively promote synergy for extrapolating features. Unfortunately, these methods rely on an adjunct model for generating pseudo-modalities, which is suboptimal and makes it hard to investigate cross-modality interaction. On the other hand, contrastive methods for self-supervised 3D point cloud and 2D representation learning, such as [54, 7, 6, 26, 1], suffer from sampling bias when generating negative samples and augmentation, making them impractical in real-world scenarios. [8, 67, 18].

To address the problem of multi-modal attention-based point cloud and image fusion, we propose PiMAE, a simple yet effective pipeline that learns strong 3D and 2D features by increasing interaction. To promote cross-modal learning, we pre-train the 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 designed 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 the modalities. 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 [23], consisting of both separate branches of modal-specific encoders and a shared-encoder. However, we notice that MAE's mask tokens only pass through the decoder[16]; hence, for mask tokens of both modalities to learn mutual information, a shared-decoder design is critical in our scheme, before performing separate reconstructions in respective modal-specific decoders. Finally, for learning stronger features from pre-training, PiMAE's multi-modal reconstruction module demands point cloud features to explicitly express image-level understanding through enhanced learning from image information.

To evaluate the effectiveness of our pre-training scheme, we systematically evaluate PiMAE with different fine-tuning architectures and tasks, including 3D object detection and 2D few-shot image classification, performed on the RGB-D scene dataset SUN RGB-D [46] and multiple image classification datasets. We find PiMAE to bring improvements over state-of-the-art methods in all the 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 the point cloud and RGB modalities, developing a highly interactive pre-training scheme with three novel schemes.
- To promote better interactive multi-modal learning, we investigate different masking designs to propose a simple complement alignment strategy, novelly introduce a shared-decoder critical to MAE, and propose a cross-modal reconstruction module for PiMAE, creating a more difficult and diverse task for the point cloud branch.
- Through extensive experiments, we show it is nontrivial for PiMAE to pre-train with joint point cloud and image modalities. Specifically, fine-tuning our pre-trained models boosts performance by 2.9% and 2.4% over previous baselines in 3D detection and 2D few-shot classification, respectively. Moreover, we find our method to be more label-efficient.

2. Related Work

2.1. 3D Object Detection

3D object detection is an active research area aiming to predict oriented 3D bounding boxes of physical objects from 3D data. Many CNN-based works propose two-stage methods first generating regional proposals and then classifying them into different object classes [41, 43]. Recently, approaches based on Transformers [49] are also widely applied to 3D object detection [44, 42, 39]. 3DETR [31] proposes an end-to-end Transformer-based object detection module using points cloud as input. Group-Free-3D [27] designed a novel attention stacking scheme to fully take advantage of the attention mechanisms in the Transformers and estimate detection results by fusing object features in different stages. In this work, we perform multi-modal pre-training with PiMAE and fine-tune our encoder backbone on 3DETR and Group-Free-3D.

2.2. 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. [58, 52, 59, 51, 64], semantic segmentation [34, 19], 3D visual grounding [53, 68] and open-world 3D recognition [15, 63, 71]. In the field of cross-modal self-supervised learning of point clouds and RGB images, several methods [24, 26] based on contrastive learning propose to design specialized structures for learning from multiple modalities, which perform better than single modalities when fine-tuning downstream tasks including 3D object detection. As aforementioned, while contrastive methods have illustrated the significance of pairing RGB and point cloud inputs, our PiMAE has several advantages over contrastive-based approaches, mainly requiring fewer augmentations.

2.3. Masked Autoencoders (MAE)

Recently, inspired by advances in masked language modeling, masked image modeling (MIM) approaches [16, 55, 3] have shown superior performance, proposing a self-supervised training method based on a masked image prediction strategy. MAE [16], in particular, predicts pixels from highly masked images using a ViT decoder. Building off of MAE's success, several works [62, 33, 30, 65, 17] have applied the framework to point cloud data, proposing to segment the point cloud into tokens and perform reconstruction. Furthermore, MultiMAE[2] investigates the alignment of various modalities with MAE of RGB images, depth images, and semantic segmentation. In this work, however, we demonstrate that these methods both do not maximize the potential of point cloud and RGB scene datasets and that they cannot be easily adapted to image inputs with only a 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. Following, 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. 2, 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 into the 2D plane, where patches of images are complementarily masked and embedded into tokens.

Following, we utilize a symmetrical joint-encoder-decoder scheme that promotes strong feature fusion. The encoder-decoder architecture consists of both separate 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[16] and PointMAE[33] 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 the RGB image's pixels by projecting the 3D 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,

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

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

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 in order to reach a complement mask alignment. The rationale for this is that by promoting such a masking pattern, the visible point cloud tokens and image tokens are semantically more abundant than in the uniform setting, and thus the model is able to extract rich cross-modal features. For visual demonstration, see Fig. 3.

3.3. Encoding Phase

Encoder During this stage, we protect the integrity of different modalities. Inspired by AIST++ [23], our encoder

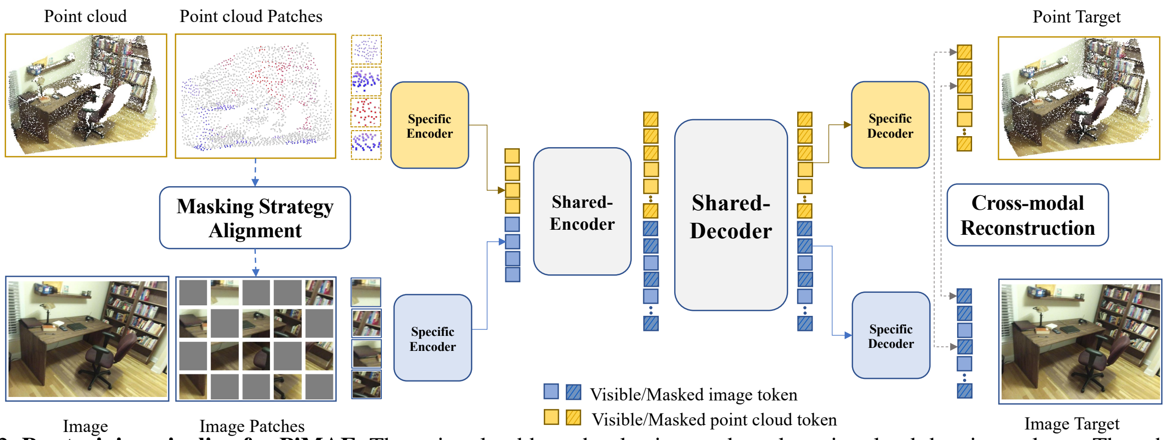

Figure 2. **Pre-training pipeline for PiMAE.** The point cloud branch tokenizes and masks point cloud data into tokens. Then the tokens are passed through a masking alignment module to generate complement masks for image patches. After embedding, tokens go through a first separate, then shared, and finally separated autoencoders structure. We last engage in a cross-modal reconstruction module to enhance point cloud representation learning. Point cloud colored for better visualization.

consists of two modules: modal-specific encoder and 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 encoder and the point-specific encoder, 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 shared-decoders, we then design specialized decoders for the different modalities for better reconstructions.

Since the reconstruction 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 shared-decoder performs fusion on the different 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, $D_I : L_I^3 \mapsto T_P$ and $D_P : L_P^3 \mapsto T_P$, where D_I and D_P are the image-specific decoder and the point-specific decoder, 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

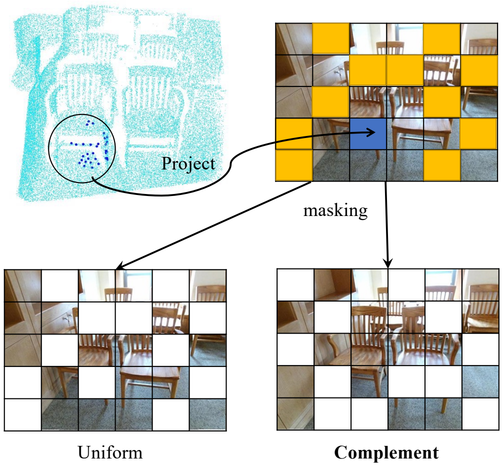

Figure 3. **Illustration of our projecting operation and two different masking strategies.** A randomly sampled point cloud cluster (black circle) is projected onto the image patch (blue square), and the other clusters are done in a similarly way (yellow squares). Under uniform masking, the yellow patches will be masked while other patches are sent into the encoders. On the contrary, complement masking will result in a reversed masking pattern.

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 are sent to a cross-modal prediction head of one linear projection layer to recover the corresponding masked image features and the cross-modal reconstruction loss is defined as

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

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

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 experimental results to qualify the superiority of our methods. The following experiments are conducted. a) We pre-train PiMAE on the SUN RGB-D [46] training set. b) We evaluate PiMAE on various downstream tasks, including 3D object detection and few-shot image classification. c) We ablate PiMAE with different cross-modal masking strategies as well as other modality-interaction strategies to show the effectiveness of our proposed design.

4.1. Datasets and metrics

SUN RGB-D[46] is a challenging large-scale 3D indoor dataset, consisting of 10,335 RGB-D images with labeled 3D bounding boxes for 37 categories. Depth images are converted to point clouds using provided camera poses, and we follow the standard 5285, 5050 splits for the training and testing stages, respectively. We report the accuracy on the test set of SUN RGB-D using the mean Average Precision at two different IoU thresholds, 0.25 and 0.5, respectively.

CIFAR-FS[4], FC100[32], miniImageNet[50], are widely used few-shot image classification datasets, and we use these datasets to evaluate PiMAE's image feature extractor. The top-1 accuracy under 5-way 1-shot and 5-way 5-shot settings on different datasets is adopted as the evaluation metric.

4.2. Implementation Details

Network architectures. Abiding by common practice [33, 62], we utilize a scaled-down PointNet [38] before a ViT [10] backbone in our point cloud branch. PointNet layers effectively reduce the sub-sampled points from 20,048 to 2,048. For the image branch, we follow [16] to divide image into regular patches with a size of 16×16, before the ViT backbone.

Pre-training. During the pre-training stage, we use the provided image and generated point cloud from SUN RGB-D [46] to train PiMAE for 400 epochs. AdamW [20] 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. The loss weights in Eq. 5 are equally set. 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 shown in Fig. 8.

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 only the image-specific encoder as well as the shared encoder as a 2D feature extractor, while discarding the rest and appending prediction heads for different scenarios. We follow the training settings same as the baselines we provide, except for the modifications on the backbone feature extractor.

4.3. Results on Downstream Tasks

In this work, we evaluate our method on two downstream tasks dealing with different modalities, 3D object detection

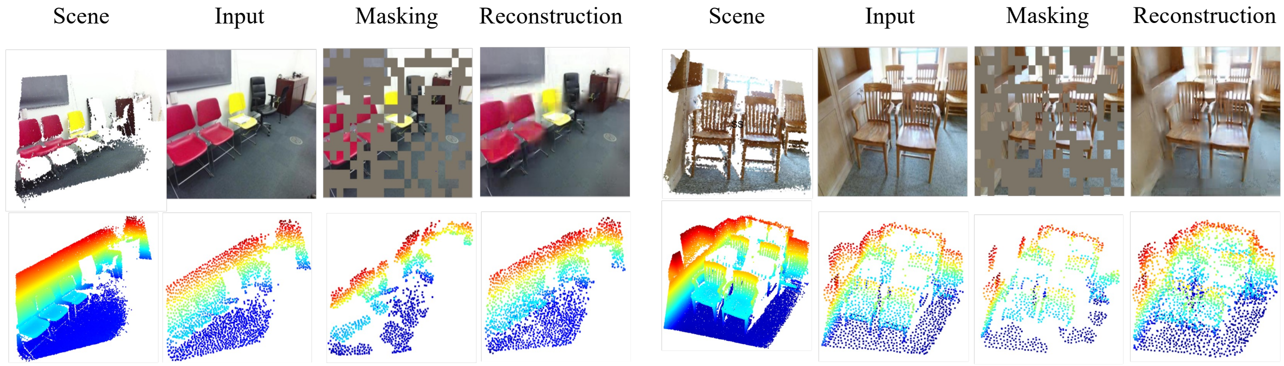

Figure 4. **Reconstruction results of images and point cloud from PiMAE.** Our model is able to perform image and point cloud reconstruction simultaneously, showing a firm understanding of the two modalities. Image results in first row, point clouds results in second row. The masking ratio for both branches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

Methods	Pre-trained	SUN RGB-D		ScanNetV2	
		AP_{25}	AP_{50}	AP_{25}	AP_{50}
DSS[47]	*None*	42.1	-	15.2	6.8
2D-driven[22]	*None*	45.1	-	-	-
PointFusion[57]	*None*	45.4	-	-	-
F-PointNet[37]	*None*	54.0	-	19.8	10.8
VoteNet[36]	*None*	57.7	32.9	58.6	33.5
ImVoteNet[35]	*None*	63.4	-	-	-
+P4Constrast[26]	ScanNetV2	63.5(+0.1)	-	-	-
3DETR[31]	*None*	58.0	30.3	62.1	37.9
+Point-BERT[60]	ScanNetV2	-	-	61.0(-1.1)	38.3(+0.4)
+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[27]	*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 SUNRGB-D[46].** We adopt the average precision with 3D IoU thresholds of 0.25 (AP_{25}) and 0.5 (AP_{50}) for the evaluation metrics.

and few-shot image classification.

3D object detection.

To demonstrate the effectiveness of PiMAE pre-training, 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 3D detection based on the SOTA methods, 3DETR[31] and GroupFree3D[27]. As shown in Tab. 1, our model brings significant improvements to both 3DETR and GroupFree3D, surpassing previous baselines consistently in all datasets and criteria. We use the mAP with 3D IoU thresholds of 0.25 (AP_{25}) and 0.5 (AP_{50}) as evaluation criteria. *Ours* refer to our implementation of the 3D detector with a ViT encoder backbone pre-trained with PiMAE. *Ours (from scratch)* refers to our implementation without pre-training. We compare both with and without PiMAE pre-training to illustrate that our performance boost comes from cross-modal interaction and not from larger backbones. Furthermore, in Tab. 2, we provide 3D object detection results with detailed per-class accuracy on SUN RGB-D.

Few shot image classification. On 2D tasks, we conduct few-shot image classification experiments on three different benchmarks to explore the feature-extracting ability of PiMAE's image encoder. Tab. 3 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], by 2.4% and 0.6% in terms of 5-way 1-shot and 5-way 5-shot results on CIFAR-FS, respectively. Such performance gain demonstrates that the cross-modal interactive pre-training of PiMAE not only helps the model to understand 3D scenes better but also brings a firm 2D understanding to the model.

Methods	bed	table	sofa	chair	toilet	desk	dresser	nightstd	bookshf	bathtub	AP_{25}
DSS[47]	78.8	50.3	53.5	61.2	78.9	20.5	6.4	15.4	11.9	44.2	42.1
2D-driven[22]	64.5	37.0	50.4	48.3	80.4	27.9	25.9	41.9	31.4	43.5	45.1
PointFusion[57]	68.6	31.0	53.8	55.1	83.8	17.2	23.9	32.3	**37.7**	37.3	45.4
F-PointNet[37]	81.1	51.1	61.1	64.2	90.9	24.7	32.0	58.1	33.3	43.3	54.0
VoteNet[36]	83.0	47.3	64.0	75.3	90.1	22.0	29.8	62.2	28.8	74.4	57.7
3DETR[31]	81.8	50.0	58.3	68.0	90.3	28.7	28.6	56.6	27.5	77.6	58.0
+Ours	85.4	48.9	62.5	69.0	93.8	28.2	33.0	62.8	30.4	80.3	59.4(+1.4)
GroupFree3D[27]	**87.8**	53.8	70.0	**79.4**	91.1	**32.6**	36.0	66.7	32.5	80.0	63.0
+Ours	85.4	**55.1**	**73.3**	78.1	**96.0**	31.5	**40.8**	**67.8**	28.4	**89.1**	**64.6**(+1.6)

Table 2. **3D objection Detection results on SUN RGB-D validation set.** Single-class precision is reported under average precision with 3D IoU threshold of 0.25. Results of VoteNet, GroupFree3D and 3DETR are taken from the original paper[31, 36, 27]. With PiMAE Pre-training, the network can improve performace by 1.4% and 2.9% .

	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[50]	-	-	-	-	43.6	55.3
Prototypical Ntwk[45]	55.5	72.0	35.3	48.6	49.3	68.2
Relation Network[48]	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 3. **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 taken from [1, 4].

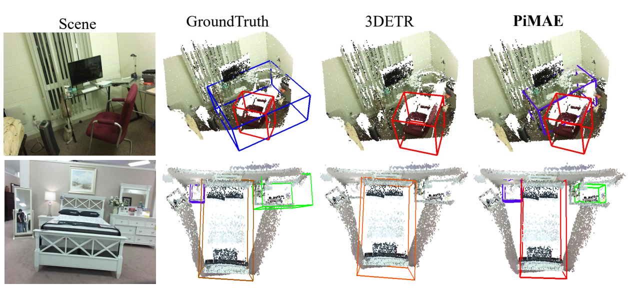

Figure 5. **Results visualization of 3D object detection.** With PiMAE pre-training, more 3D objects can be detected (e.g. dresser and nightstand in the second scene).

4.4. 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. Then, we compare various architectural settings. Following, we pre-train with different reconstruction targets. Finally, we ablate performance pre-trained with only a single branch.

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. As shown in Tab. 4, masking the tokens from different modalities complementarily gains higher performance than uniformly, with an improvement of 0.9% and 0.4% respectively on AP_{25} and AP_{50}. Complement masking enables more cross-modal interactions between patches with different semantic information, and thus helps our model to transfer the 2D knowledge into 3D feature extractor. While with the uniform masking strategy, the extracted point cloud features and image features are semantically aligned and the promoted interaction does not help the model to utilize 2D information better.

Note that we purposely fine-tune without cross-modal reconstruction, as uniform masking strategies project to masked image features, which are semantically weaker and negatively influence the evaluation of masking strategies.

Modality Interaction in Pipeline Architecture. During the reconstruction stage, as proposed in Sec. 3.4, a shared decoder architecture is adopted. The encoded features are

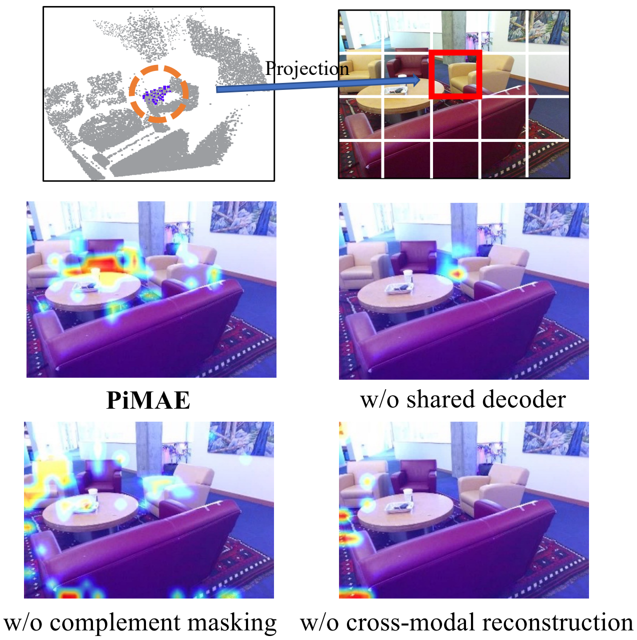

Figure 6. **Visualization of encoder attention.** The encoder attention between two modalities is visualized by computing self-attention from the query points (orange circle) to all the visible image tokens. We show the corresponding location (red square) of the query points after projection.

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

Table 4. **Comparisons of different cross-modality masking strategies.** We use specific encoders and a shared encoder of 3 layers respectively, and specific decoders of 3 layers as the default setting. Experiments are based on 3DETR[31] and performed on SUN RGB-D.

first disentangled by the cross-modality decoder, and reconstructions are completed afterward with task-specific decoders. From Tab. 5, we find the additional shared-decoder design performance-enhancing, as it considers the cross-modal influence of masked tokens. Specifically, shared-decoder is a novel contribution of PiMAE, and we find it non-trivial, because the interactions in the masked tokens improve feature extraction.

Effect of Cross-modal Reconstruction. Other than reconstructing the inputs, we promote cross-modal reconstruction by demanding point cloud features to reconstruct features and 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.

Encoder	Decoder	AP_{25}	AP_{50}
3+3	0+3	58.0	30.2
3+3	1+2	**59.4**	**33.2**
3+3	1+3	58.1	32.8
2+2	1+2	57.5	30.8

Table 5. **Effectiveness of our shared decoder.** a+b in encoder denotes specific encoders of a-layers ViT, and shared-encoder of b-layers ViT. c+d in decoder denotes c-layers ViT for shared-decoder and d-layers ViT for specific decoders. Experiments are based on 3DETR[31] and performed on SUN RGB-D.

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.** Note that Geo, feat and pix refers to coordinates, feature and pixel, respectively. Experiments are based on 3DETR and performed on SUN RGB-D. The cross-modal feature reconstruction target brings significant improvements to model performance on downstream tasks.

Necessity of joint pre-training. To demonstrate the effectiveness 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. 6. With our proposed design, PiMAE focuses on more foreground objects with higher attention values, showing strong cross-modal understanding.

4.5. Generalization Ability Study

As labeled data are rather expensive to acquire, the ability to make use of smaller labeled data while maintaining adequate performance is a key indicator to evaluate a model's generalization ability.

Data efficiency. We train 3DETR and 3DETR-based PiMAE using limited annotated labels (varying from 1% to 100%) while testing on the full val set on SUN RGB-D[46]. As shown in Tab. 7, PiMAE is able to outperform the base-

Input	3D Object Detection AP_{25}	AP_{50}	Few-shot image classification 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. **Necessity of joint pre-training.** We compare results on downstream tasks when pre-trained with a single-branch PiMAE. A performance drop occurs when only one modality is used, showing the necessity of our joint pre-training design. Experiments are performed on SUN RGB-D and CIFAR-FS.

line in each scenario. The largest gap can be seen when only 10% of labels are used and PiMAE can outperform baseline by 17.4% on AP_{25}.

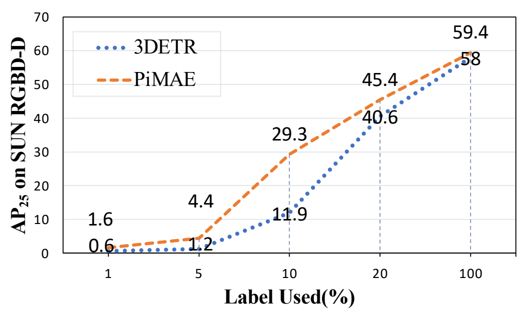

Figure 7. **Illustration of Data Efficiency.** Compared to 3DETR, PiMAE is able to ease the burden of data labeling and increases the performance significantly.

5. Conclusion

In this work, we demonstrate PiMAE's simple framework is an effectively and highly interactive multi-modal learning pipeline with strong feature extraction abilities on point cloud and image data. Concretely, we first design a projection module to complementarily align the mask portions of the different modalities. We design three aspects for promoting cross-modality interaction. First, explicitly align the mask patterns of both point cloud and image for better feature fusion. Ensuingly, we design a shared decoder to promote to accommodate for mask tokens of both modalities. Finally, our cross-modality reconstruction enhances the semantics learned by representations. In our extensive experiments and ablation studies performed on datasets of both modalities, we discover it is non-trivial for interactive multi-modal learning, increasing the performance of multiple baselines and tasks.

Appendix

A. Extended Related Work

A.1. 3D Object Detectors

Prior 3D object detection methods adapt popular 2D detection approaches to 3D scenes. Specifically, they project point cloud data to 2D views [21, 25, 56] for 2D ConvNets to detect 3D bounding boxes. Other approaches adopt 3D ConvNets for 3D object detection by grouping points into voxels [70, 40].

The Transformer architecture [49] has demonstrated consistent and impressive performance in vision, specifically with object detectors [5, 28, 66, 29, 69, 27, 31]. Transformers are well-designed for 3D point clouds, with the advantage of not needing hand-crafted groupings and an invariance understanding.

In PiMAE, we draw inspiration from both projection-based and attention-based 3D object detectors; Whereas the former projection mechanism has been extensively utilized by previous detectors, the latter has shown better versatility and a more intuitive solution. Consequently, we design a MAE [16] structured multi-modal learning framework that incorporates projection alignment for more interactive multi-modal learning.

B. Supplementary Implementation Details

In this section, we first introduce the descriptions of dataset we verified the effects of our pre-training scheme on. Then, we elaborate on the details of the baseline models we used in our empirical studies. Finally, we detail the specific configurations used in all of our experiments.

B.1. Datasets

ScanNetV2 [9] is a 3D interior scene dataset with rich annotations, consisting of 1513 indoor scenes and 18 object classes. The labels include semantic labels, per-point instances, and 3D bounding boxes. We use the common metrics for evaluation [36], measuring the mean Average Precision (mAP) under two IoU thresholds of 0.25 and 0.5.

B.2. Baseline Approaches

We evaluate our interactive multi-modal training pre-training scheme by fine-tuning on two state-of-the-art 3D object detectors, and our baseline implementations rigorously follow their publicly released codes.

3DETR[31] is a simple, end-to-end 3D detection pipeline that does not require a finely crafted 3D detection backbone. Instead, its versatile attention-based backbone maximally preserves the vanilla Transformer blocks to reach a comparable performance with CNN based detectors.

Group-Free 3D [27] is another approach implementing the Transformer models on 3D object detection task, using both a well-designed query locations for objects and an ensembling of detection results. Unlike PointNet-based networks [38, 37, 44] that create a local grouping scheme for each object candidate, Group-Free uses an attention mechanism on all the point cloud points.

Config	Value
optimizer	AdamW[20]
base lr	1e-3
weight decay	0.05
batch size	256
lr schedule	cosine decay
warmup epochs	15
epoch	400
augmentation	None

Table 8. **Pre-training configuration.**

B.3. Pre-training Details

The encoder and decoder architectures in PiMAE follow the standard ViT[10] design, which consists of several Transformer blocks. In our PiMAE, the number of Transformer blocks for specific encoders and shared-decoders is both set to 3, while the numbers for specific decoders and shared-decoders are set to 2 and 1, respectively. For encoders, each Transformer block has 256 hidden dimensions and 4 heads for the multi-head self-attention module. For decoders, the numbers are adjusted to 192 and 3.

For the point cloud branch, we sample 2048 points from each 3D scene in SUN RGB-D[46], following previous work [62]. For the image branch, we adjust the resolution of each image to 256×352, and we use a patch size of 16 to patchify images. The specific configuration for pre-training PiMAE is given in 8.

B.4. Fine-tuning Protocol on SUN RGB-D and ScanNetV2

For fine-tuning on GroupFree3D [27], we insert our 3D feature extractor (only the shared-encoder) into the pipeline. Compared to the original configuration, the only modification here is tuning the learning rate on the encoder lower to $lr = 3e-5$ to preserve the pre-trained prior. For detection on ScanNetV2, we lower the encoder learning rate to $lr = 6e-5$.

For experiments with 3DETR [31], our encoder consists of six Transformer blocks pre-trained with PiMAE. The exact setups as the original [31] are then used for fine-tuning, except that we apply a reduced learning rate of $lr = 1e-5$ to the encoder.

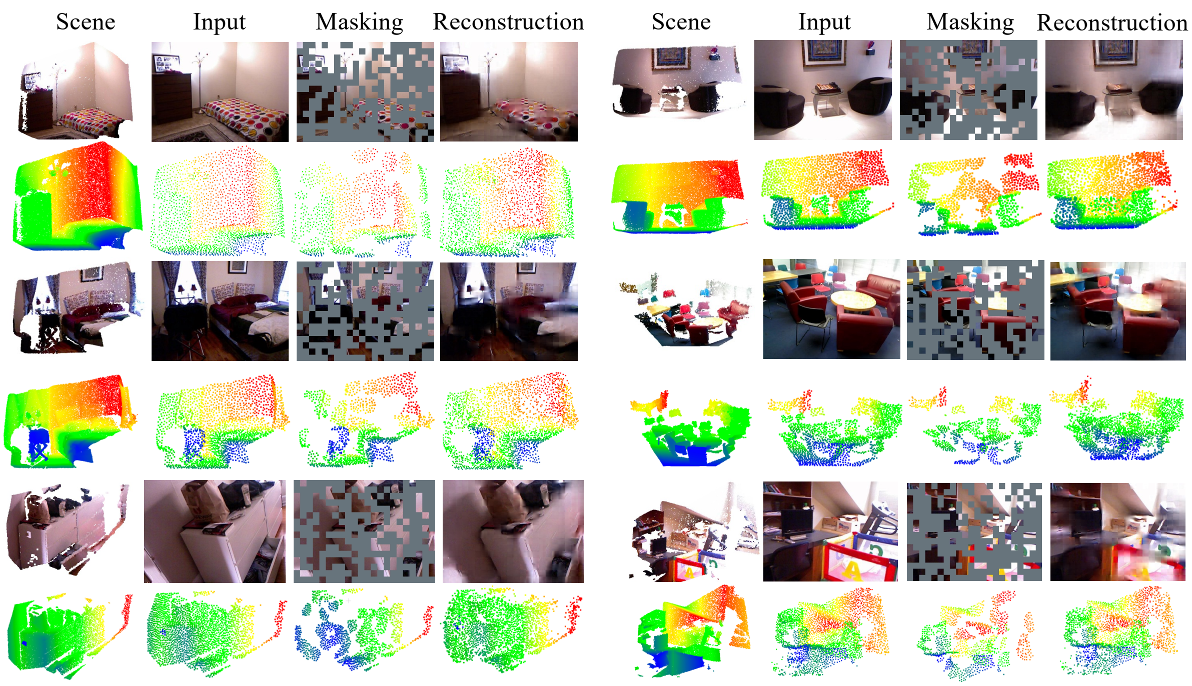

Figure 8. **Visualization of reconstruction results.** Our model is trained with 60% masking ratio. Point clouds are colored for better visualization purpose. PiMAE generalizes well for different scenes and reconstructs masked images (odd rows) and masked point clouds (even rows) simultaneously.

Mask Ratio	AP_{25}	AP_{50}
50%	58.7	33.1
60%	**59.4**	**33.2**
70%	58.4	33.0
80%	57.5	32.4

Table 9. **Ablation study on masking ratios.** Experiments with different masking ratio are conducted, and we report detection accuracy based of 3DETR on SUN RGB-D val set.

C. Additional Ablation Study

In this section, we give more ablation studies for further analysis of PiMAE.

Ablation Study on Masking Ratio. As reported in Tab. 9, we examined several masking ratios for PiMAE and find that the the model learns the best latent features when the masking ratio is set to 60%.

D. Additional Visualization

Reconstruction Results. In Fig. 8. We provide more examples of reconstruction visualizations. PiMAE simultaneously reconstructs masked point clouds and images with clear reconstructions reflecting semantic understanding.

Activation of Feature Map. This section provides more attention map examples generated by PiMAE's shared-encoder, where features from the two modalities first interact explicitly. By examining the self-attention weights, we can gain better insights on PiMAE's multi-modal interactions. We compute the self-attention from a point cloud token to all image tokens and show the attention values. In Fig. 9, PiMAE is able to attend to more foreground objects and with higher attention values, while other designs either attend to unrelated backgrounds(e.g. row 3, col 4), or have rather low attention values (e.g. row 2, col 4).

We also compute the self-attention from a image token to all point cloud tokens and display the attention weights. As shown in Fig. 10, given a image token as query, PiMAE accurately attends to the corresponding objects in the point cloud with highest values, showing a strong understanding of both 2D and 3D features.

Object Detection. We display qualitative results comparing PiMAE and baseline. On top of 3DETR [31], with PiMAE pre-training, we are able to detect more objects with more precise boxes.

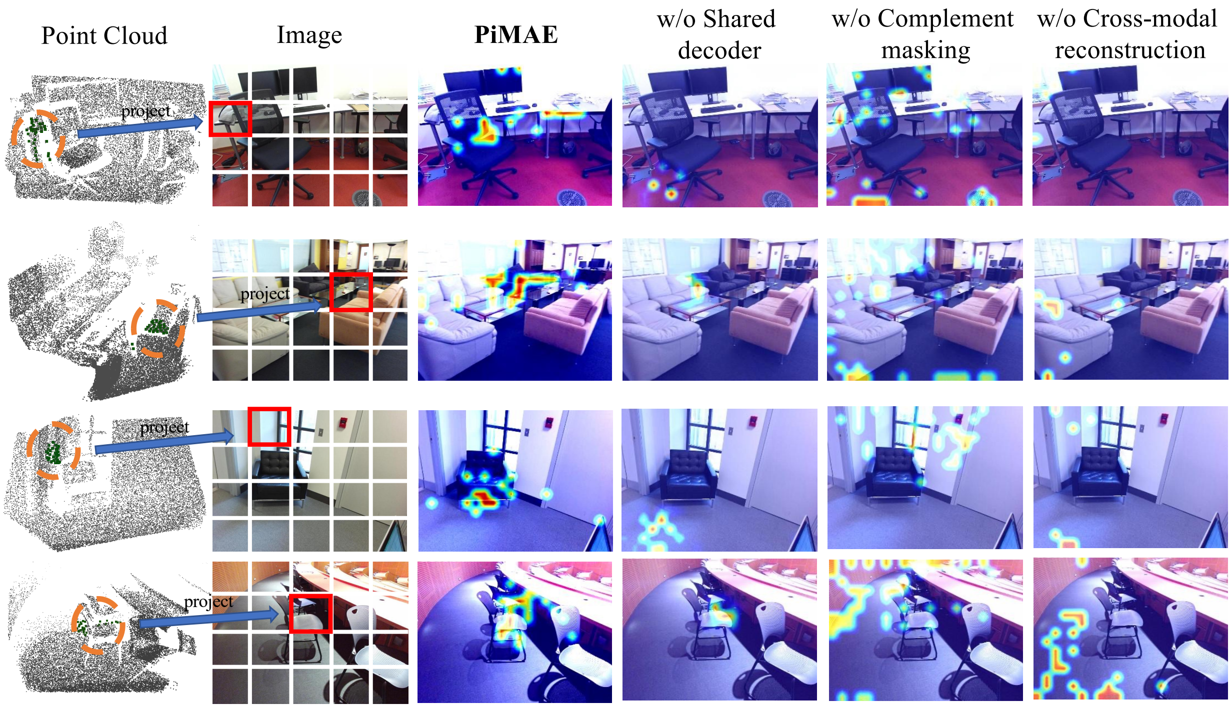

Figure 9. **Visualization of encoder attention.** The encoder's attention between two modalities is visualized by computing self-attention from the query points (orange circle) to all the visible image tokens. Highest values are shown in red. We show the corresponding location (red square) of the query points after projection. From left to right shows ablation of PiMAE with different designs, including our final proposal, and settings that exclude shared-decoder, complement masking strategy and cross-modal reconstruction, respectively.

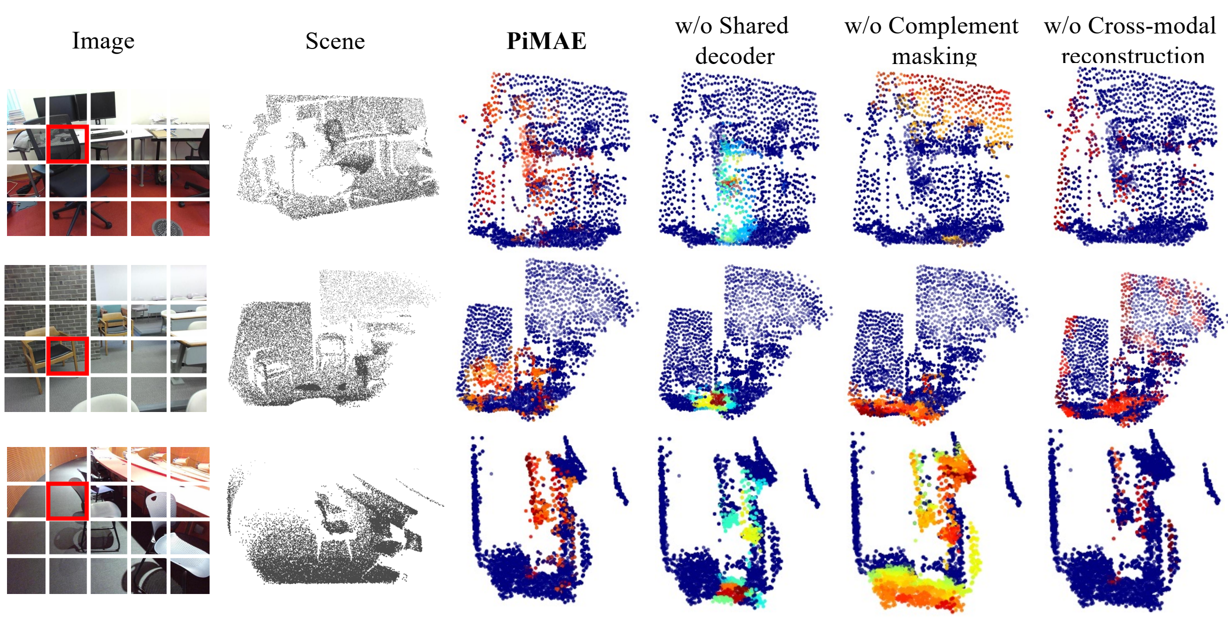

Figure 10. **Visualization of encoder attention.** The encoder attention between the two modalities is visualized by computing self-attention from the query of an image token (red square) to all the point cloud tokens. Highest values are shown in red. The attention intensity in the point cloud corresponds with the image patch query, showing the effectiveness of our cross-modal interactions during pre-training.

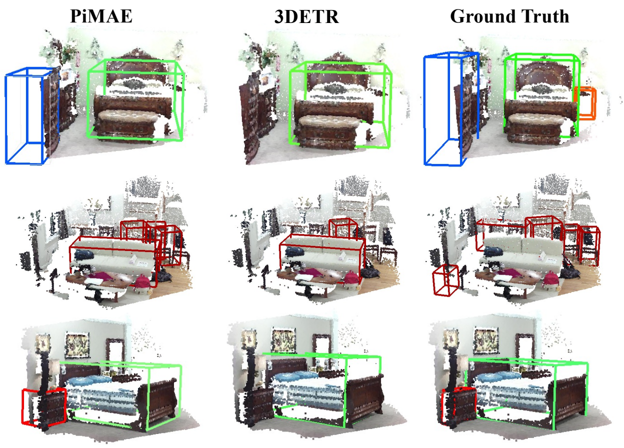

Figure 11. **Visualization of predictions on SUN RGB-D validation set.** Note that we correctly detect more objects (as in row 2), and predict bounding boxes with more precision (as in row 3).

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

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

[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] 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.

[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.

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

[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.

[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.

[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.

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

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

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

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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

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

[22] 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.

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

[24] 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.

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

[26] 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. abs/2012.13089, 2020.

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

[28] 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.

[29] 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.

[30] 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.

[31] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In *ICCV*, 2021.

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

[33] 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.

[34] 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.

[35] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

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

[38] 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.

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

[40] 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.

[41] 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, 2019.

[42] 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.

[43] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud, 2018.

[44] 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.

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

[46] 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.

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

[48] 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.

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

[50] 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.

[51] 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.

[52] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12186–12195, 2022.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

[58] Xiaobin Xu, Lei Zhang, Jian Yang, Chenfei Cao, Zhiying Tan, and Minzhou Luo. Object detection based on fusion of sparse point cloud and image information. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021.

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

[60] 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, 2021.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

[67] 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.

[68] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021.

[69] 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.

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

[71] 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.