

---

# Bridging the Domain Gap: Self-Supervised 3D Scene Understanding with Foundation Models

---

**Zhimin Chen**<sup>1</sup>  
Clemson University  
zhiminc@clemson.edu

**Longlong Jing**<sup>2</sup>  
The City University of New York  
ljing@gradcenter.cuny.edu

**Yingwei Li**<sup>3</sup>  
Johns Hopkins University  
yingwei.li@jhu.edu

**Bing Li**<sup>✉1</sup>  
Clemson University  
bli4@clemson.edu

## Abstract

Foundation models have achieved remarkable results in 2D and language tasks like image segmentation, object detection, and visual-language understanding. However, their potential to enrich 3D scene representation learning is largely untapped due to the existence of the domain gap. In this work, we propose an innovative methodology called Bridge3D to address this gap by pre-training 3D models using features, semantic masks, and captions sourced from foundation models. Specifically, our method employs semantic masks from foundation models to guide the masking and reconstruction process for the masked autoencoder, enabling more focused attention on foreground representations. Moreover, we bridge the 3D-text gap at the scene level using image captioning foundation models, thereby facilitating scene-level knowledge distillation. We further extend this bridging effort by introducing an innovative object-level knowledge distillation method that harnesses highly accurate object-level masks and semantic text data from foundation models. Our methodology significantly surpasses the performance of existing state-of-the-art methods in 3D object detection and semantic segmentation tasks. For instance, on the ScanNet dataset, Bridge3D improves the baseline by a notable margin of 6.3%. Code will be available at: <https://github.com/Zhimin-C/Bridge3D>

## 1 Introduction

In recent years, task-agnostic pre-trained representations have fundamentally reshaped the landscape of Natural Language Processing (NLP), driven by the success of foundation models such as GPT-3 [55], PALM [13], T-NLG [31], and BERT [7]. Parallel advancements have been observed in the realm of computer vision, where foundation models like CLIP [54], Grounding DINO [43], DINOv2 [49], BLIP [39], and SAM [37] have established new benchmarks in 2D vision tasks, emerging as the leading approach for achieving state-of-the-art performance. However, the potential of these powerful models in advancing 3D scene understanding is yet to be fully realized, primarily due to the limited availability of large-scale 3D-text pair datasets and the considerable cost associated with procuring high-quality 3D annotations. Despite recent studies demonstrating the potential of individual foundation models like CLIP [54] or MOCO [27] in enhancing 3D scene understanding, a comprehensive exploration of the utility of other foundation models and their synergistic combinations remains a largely uncharted territory.

To address this challenge, we propose a novel framework **Bridge3D** that harnesses the strengths of multiple foundation models to advance 3D representation learning through a self-supervised learning

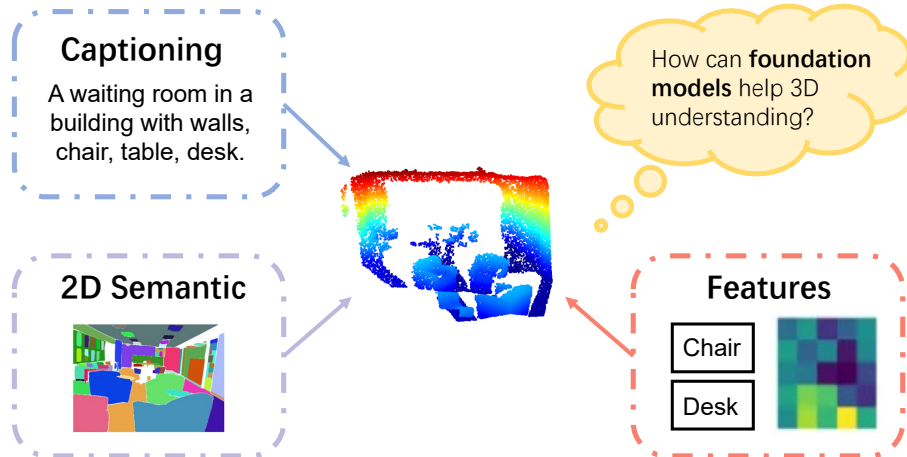


Figure 1: **The motivation of Bridge3D.** The driving force behind Bridge3D is to create a method that bridges the gap between 3D models and text/image foundation models via self-supervised learning.

approach. Specifically, the Bridge3D leverages image captioning outputs from foundation models to generate 3D scene prompts, establishing a bridge between 3D and text domains for scene-level knowledge distillation. These generated prompts are further utilized to produce instance segmentation results with the assistance of Grounding DINO and SAM. Subsequently, we employ a 3D network trained on a self-supervised task that distillates the knowledge from text and 2D to point clouds at the scene level. This is followed by the distillation of these multidimensional features into 3D features at the object level. Additionally, to optimize point reconstruction, we propose an inventive masking and patch-dropping strategy that redirects the model’s attention toward foreground object representation learning.

Our proposed framework effectively circumvents three significant hurdles in 3D self-supervised learning. **Firstly**, the adoption of foundation models in our novel masking and dropping strategy allows the network to concentrate more on foreground object representation learning, thereby enhancing the performance of the 3D model. This is a distinct shift from traditional 3D masked autoencoder methods that rely on a random masking strategy and reconstruct all point clouds, which impairs representation learning due to the imbalance between foreground and background points. **Secondly**, the lack of datasets incorporating both 3D and text description pairs significantly hampers the potential for large-language models to contribute to 3D understanding. To overcome this obstacle, Our method first employs image captioning to generate text descriptions from paired images of point clouds, effectively bridging the 3D-text gap at the scene level. This novel integration of 3D and text modalities presents a compelling new frontier for improving self-supervised 3D scene understanding. **Lastly**, our approach stands in contrast to previous methodologies [9; 59] that facilitated either 2D to 3D or text to 3D distillation in isolation due to inherent limitations in mask generation. Instead, our strategy leverages foundation models to generate highly precise object-level masks and semantic text information. This approach seamlessly integrates object-level 3D, visual, and textual features, thereby significantly enhancing the quality of 3D scene representation learning.

We evaluate our method on multiple datasets, including SUN RGB-D [75] and ScanNet [14] for 3D object detection and S3DIS [5] for 3D semantic segmentation. Our approach outperforms state-of-the-art self-supervised learning methods in both tasks, demonstrating the effectiveness of our proposed framework. The contributions of Bridge3D can be summarized as follows:

1. We propose a novel masking and patch-dropping strategy based on foundation models to refine the focus of the network on foreground representation learning for 3D masked autoencoders.
2. We propose a novel scene-level and object-level multi-modality knowledge distillation method that pre-trains a 3D network via features, semantic masks, and captions obtained from foundation models

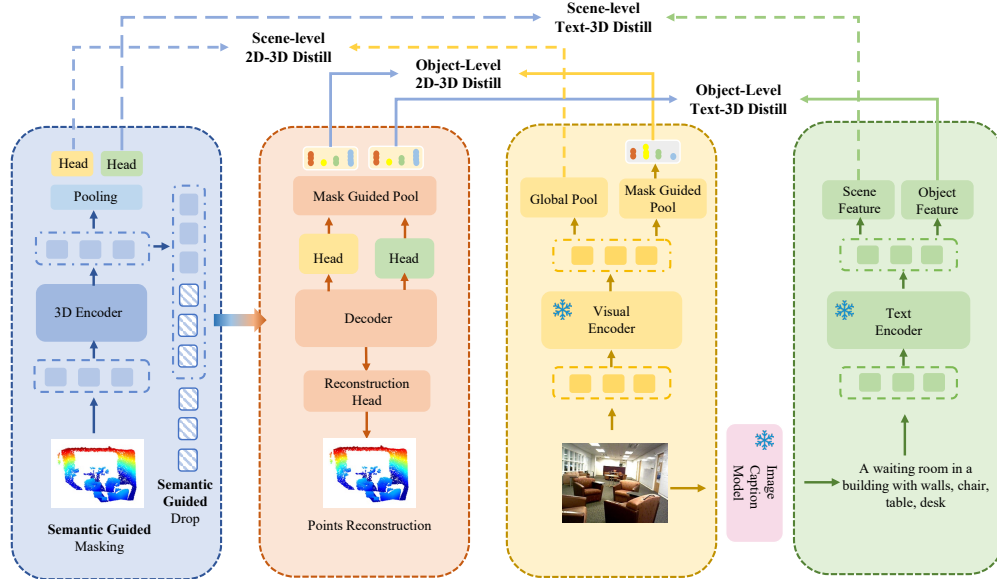


Figure 2: **Overview of Bridge3D.** Our method employs features, semantic masks, and captions derived from foundation models to improve 3D representation learning. We use semantic masks to guide the masking and reconstruction phases in the masked autoencoder, which intensifies the network’s attention on foreground objects. At the scene level, we use image captioning foundation models to bridge the scene-level 3D-text gap. Additionally, we facilitate the distillation of well-learned 2D and text representations to the 3D model at the object level by leveraging foundation models to generate accurate object-level masks and semantic text information.

3. To the best of our knowledge, this is the first research to harness multiple foundation models for self-supervised 3D scene understanding. This pioneering approach has been shown to outperform state-of-the-art methods in various downstream tasks.

## 2 Related Work

**3D Self-Supervised Representation Learning.** Recently, self-supervised pre-training on unlabelled point clouds [2; 24; 12; 66; 73; 35; 22] has shown promising transferable ability, providing a good network initialization for downstream fine-tuning. Several methods have been proposed for pre-training point cloud features, including learning relative position [58], multiple pretext tasks [26], and contrastive learning [17; 33; 57; 65; 72; 3; 35; 23; 21]. Info3D [57] extends the InfoMax and contrastive learning principles to 3D shapes. PointContrast [65] conducts point-level contrast on two transformed views of the same point cloud. Zhang [72] contrasts instance-level representations obtained from the same scenario but processed by different model architectures. CrossPoint [3] introduces an auxiliary multi-modal contrastive objective that captures 3D-2D correspondence, leveraging the complementary attributes of point clouds and images. Point-BERT [67] uses pre-trained tokenizers to indicate discrete point tokens, while Point-MAE [50] applies Masked Autoencoders (MAE) to directly reconstruct the 3D coordinates of masked tokens. Our proposed method uses Point-MAE as the baseline, but leverages foundation models to guide the masking and reconstruction stages. Additionally, we leverage image and text knowledge from foundation models to enhance 3D self-supervised learning.

**Foundation Models.** The field of AI research has experienced a paradigm shift with the emergence of models trained on massive amounts of data at scale, commonly referred to as foundation models [19; 4; 56; 62; 25; 60]. These models have demonstrated remarkable performance in various language and visual-related tasks. The use of large-scale text pre-training on attention-based models [15; 71] has led to the increasing popularity of vision-language models (VLM) due to their impressive performance in visual understanding tasks [54; 46; 40]. Recent advancements in contrastive learning have enabled CLIP [54] to perform multimodal learning with 400M data crawled from the web. CLIP has been extended for high-efficiency model training and cycle consistency through various methods

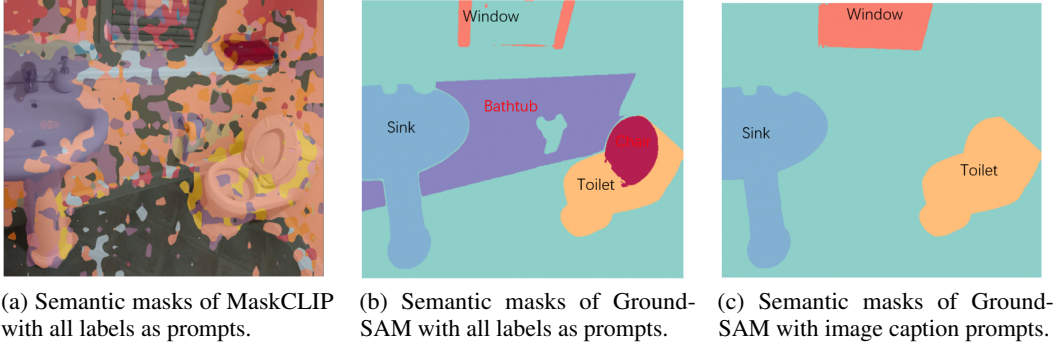


Figure 3: **The comparison of zero-shot semantic results.** (a) MaskCLIP failed to perform semantic segmentation accurately, yielding low accuracy overall (b) when all possible labels were used as prompts, the instance segmentation from foundation models results were prone to false positives (as seen in the Chair and Bathtub in this example) (c) by leveraging image captioning model to generates text prompts, the performance of the instance segmentation from Ground-SAM is further improved. This approach is beneficial for 3D scene understanding.

[40; 39; 38]. BLIP [39] includes text-to-image generation as an auxiliary task, which results in better performance by utilizing synthetic data as a bonus. More recently, the success of the foundation models has been achieved in the pure computer vision area. Segment Anything (SAM) [37] has been proposed to act as a generic image segmentation model trained on the large visual corpus. To overcome the drawback of CLIP which overlooks the visual local information, DINOv2 [49] is proposed, which is trained with self-supervised learning and achieves results that match or surpasses the standard approach used in task-specific fields. Grounding DINO [43] extends a closed-set detector DINO to open-set object detection by performing vision-language modality fusion at multiple phases.

**Self-supervised 3D Understanding with Foundation Models.** A number of studies have proposed strategies for knowledge transfer from pre-trained 2D foundation models to 3D representations at the object level [28; 30; 69; 68; 51]. For comprehensive scene understanding, recent efforts have improved 3D point representations by exploiting pixel-point alignments for distillation or contrastive learning [59; 9]. The I2P-MAE [70] approach takes advantage of 2D semantic saliency maps from CLIP [54] to guide masking and facilitate knowledge distillation from 2D to 3D at the instance level. Despite these advancements, most current 3D understanding methodologies employing foundation models focus predominantly on CLIP and distill knowledge through feature-level consistency or contrastive learning [9; 59]. The potential to utilize the capabilities of other foundation models, such as BLIP [39] for image-to-text captioning, SAM [37] for mask generation, Grounding DINO [43] for zero-shot detection, and DINOv2 [49] for high-performance features with detailed localized information, remains largely unexplored. Therefore, we propose to advance self-supervised 3D scene understanding by newly incorporating features, semantic masks, and captions obtained from various foundation models, achieving superior performance compared to state-of-the-art methodologies.

### 3 Methodology

The pipeline of Bridge3D is illustrated in Fig. 2. Our approach employs semantic information and features extracted from well-established foundation models to enhance 3D scene representation learning using self-supervised learning. The proposed method consists of three components: a semantic-guided masked autoencoder, multi-modal scene-level knowledge distillation, and multi-modal object-level knowledge distillation.

#### 3.1 Mask Generation by Foundation Models

Our proposed methodology leverages existing foundation models to produce instance segmentation masks. Initially, we use Tag2text [34], based on BLIP [39], to create image captions. We leverage ChatGPT to filter captions objects neither in ScanNet [14] nor SUN RGB-D [61] dataset to generate text prompts for the 3D scene. Subsequently, we employ SAM [37] to generate masks for the image.

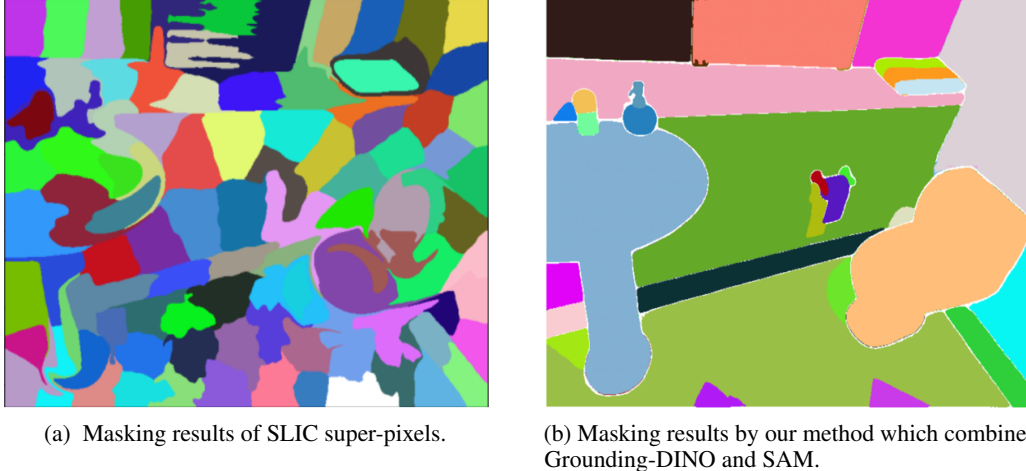


Figure 4: **The comparison of masks.** (a) Masks generated from super-pixel and (b) Masks generated from our method. Our method can provide much more accurate object-level masks compared to super-pixel and thus benefits the multi-modality knowledge distillation.

Lastly, we use the text labels as prompts for Grounding Dino [43] to create the corresponding bounding box, which is then inputted into SAM to produce both zero-shot instance segmentation labels and segmentation masks  $\mathcal{O}_1, \dots, \mathcal{O}_N$ . To establish the dense visual token-point token correspondence  $x_i, p_i$ , we calibrate the point cloud with the respective images, where  $x_i$  and  $p_i$  signify the  $i$  paired image feature and point feature. This procedure is completed offline and saved locally, with the generated labels being used directly during the self-supervised training stage. As shown in Fig.3 and Fig.4, the instance segmentation masks generated from foundation models outperform previous methods in terms of semantic results and object-level masks. Furthermore, Fig. 3 demonstrates that the performance of the foundation model is further improved when using caption methods as prompts and filtering out 3D-unrelated text.

### 3.2 Semantic Guided 3D Masked Autoencoder

To let the 3D model understand 3D-specific representations, we leverage Point-MAE [50] as the baseline for pre-training, which learns a meaningful representation with the pretext task of recovering the original inputs from visible ones. The Point-MAE [50] method utilizes standard Transformers as the backbone of its architecture, with an encoder-decoder structure that is asymmetric. The encoder takes visible tokens  $T^v$  as input and generates encoded tokens  $T^e$ , while the decoder contains fewer Transformer blocks and is responsible for generating the reconstructed masked 3D coordinates. Positional embeddings are incorporated into each Transformer block to provide location-based information. The encoded tokens  $T^e$  are padded with learnable mask tokens  $T^m$  and sent to the decoder. A set of positional embeddings is added to each Transformer block in the decoder to provide location information to all tokens. The output of the decoder  $H^m$  is then passed through a simple fully connected layer to reconstruct the masked 3D coordinates  $P^{pre}$ . After that, it restores the coordinates of the points in each masked point patch, and to evaluate the accuracy of the predicted coordinates, it computes the reconstruction loss using  $l_2$  Chamfer Distance [18], which is formulated as:

$$\mathcal{L}_{mae} = \frac{1}{M_{mask}} \text{Chamfer}(P^{pre}, P^{mask}) \quad (1)$$

where  $P_{mask}$  represents the ground truth of masked points.

**Foreground-aware Masking and Patch Dropping.** The Point-MAE [50] relies on a random masking strategy, resulting in dispersed attention across the entire image and insufficient focus on foreground objects. This dispersion of attention can lead to the model wasting computational resources on unimportant background elements, thereby leading to weaker learned representations. Moreover, the Point-MAE’s approach to reconstructing all masked tokens, including those from the background, may further weaken representation learning as the model overly concentrates on the

background due to the foreground-background imbalance. Although MAE [70] has suggested using 2D semantic saliency maps to guide the masking of point tokens for 3D instances in classification tasks, the problem of effectively guiding 3D scene masking still remains a key research challenge.

To address those problems, we propose a semantic-guided masking strategy based on segmentation results obtained from foundation models. Specifically, for foreground objects obtained from the segmentation mentioned before, we mask a higher percentage  $r_f$  of foreground points compared to the whole masking ratio  $r_w$ . This generates a more challenging reconstruction task and forces the model to focus more on foreground objects. In addition, instead of inputting all patches  $\{P_i\}_{i=1}^M$  as in Point-MAE, we randomly drop a percentage  $r_d$  of background patches to obtain  $\{P_i\}_{i=1}^N$ . Where  $N = (1 - r_d) \times M$ . The transformer decoder reconstructs masked point patches by using features from visible tokens and the positional information of both visible patches and mask patches. The patches that are dropped will not be reconstructed by the decoder. With enough background patches dropped, the decoder sees fewer points to perform the trivial up-sampling, which further improves the 3D representation learning and accelerates the pre-training by reducing the input data.

### 3.3 Scene-level Multi-modal Knowledge Distillation

Although some works have investigated scene-level multimodal learning using image and point clouds [36; 11; 41; 70], exploring scene-level multimodal learning with text and point clouds remains a challenge due to the lack of corresponding text descriptions for current 3D scene datasets. To address this problem, we propose to leverage image captioning methods to generate corresponding captions and filter out other 3D irrelevant objects to obtain 3D scene description texts  $t_s$ . Then, we train the 3D network to align 3D point clouds with their corresponding scene-level images  $i_s$  and texts  $t_s$ . We formulate the proposed method below. Consider a set of  $N$  point cloud-image-text pairs  $\{F^{3D}, F^{2D}, F^{text}\}$ , where  $F^{3D}$  represents scene-level point cloud features from the encoder,  $F^{2D}$  is the corresponding image features obtained from pre-trained foundation model based on  $i_s$ , and  $F^{text}$  is the text features from pre-trained foundation model based on  $t_s$ . Our model maps scene-level 3D features  $F^{3D}$  to the hidden representation  $\hat{F}^m$  for each modality  $m$  with a projection head  $E_m$ . The mapping process can be formulated as:

$$\hat{F}^m = E_m(F^{3D}), \quad (2)$$

Due to the attributes to avoid representation collapsing, previous methods [9; 59] utilize InfoNCE loss [48] to conduct multi-modality knowledge distillation. However, in this work, we find that leveraging positive only  $L_1$  smooth loss generates better results. We think this is because those foundation models have learned discriminative features during the pre-training stage, and thus negative pairs are not necessary for the distillation stage. The scene-level distillation between 3D-image features; and 3D-text features are defined by:

$$\mathcal{L}_{scene} = L_1(\hat{F}^{2D}, F^{2D}) + L_1(\hat{F}^{text}, F^{text}) \quad (3)$$

where  $L_1$  represents the  $L_1$  smooth loss.

### 3.4 Object-level Multi-modal Knowledge Distillation

While the value of object-wise feature representations in downstream tasks like semantic segmentation and detection is well-proved [74; 29; 64; 6], the generation of unsupervised masks presents a substantial challenge. Traditional techniques in computer vision, such as Felzenszwalb-Huttenlocher [20] and super-pixel [1], have been employed in earlier methods [59]. However, these techniques yield subpar masking results and cannot generate semantic labels, which hinders their ability to bridge the 3D-2D-text gap and leaves the potential of powerful language foundation models unrealized. The recent CLIP2Scene method [9] uses the MaskClip [76] model to generate dense semantic predictions, but it falls short of generating instance semantic results, which prevents object-level visual representations from being distilled into 3D models. The inferior quality outputs of MaskClip, Felzenszwalb-Huttenlocher, and super-pixel impede 3D representation learning. In contrast, our proposed Bridge3D method leverages high-quality masks obtained from foundation models to guide object-level knowledge distillation, thereby enhancing 3D scene understanding.

In the object-level knowledge distillation phase of Bridge3D, we propose a distinctive approach to multi-modality distillation through reconstructed tokens from the decoder, as opposed to previous

methods [9; 59] that directly distill the knowledge from other modalities to 3D post-encoder. The Point-MAE uses the decoder to reconstruct these masked tokens and learn specific features. In our method, we not only task the decoder with reconstructing masked point clouds but also reconstruct text and visual features that correlate with visible point tokens to distill the multi-modal knowledge. We do not reconstruct text and visual features of masked tokens that are dropped in the point reconstruction part, as mentioned in Sec. 3.2. To be specific,  $I_i$  represents the visual features obtained from the pre-trained visual model and belonging to the mask  $O_i$ . We use the mapping function to group the points into corresponding masks:  $\mathcal{G}_1, \dots, \mathcal{G}_k$ . For each scene, we compute mask visual features and mask point features by average pooling:

$$f_{l,i}^{2D} = \frac{1}{\mathcal{O}_i} \sum_{j \in \mathcal{O}_i} (I_j) \quad (4)$$

$$\hat{f}_{l,i}^{2D} = \frac{1}{\mathcal{G}_i} \sum_{j \in \mathcal{G}_i} (\mathcal{F}_{2D}(H_j)) \quad (5)$$

Where  $H_i$  represents visible tokens from the decoder,  $\mathcal{F}_{2D}$  is the projection head. For the text-to-point cloud knowledge distillation, as only foreground objects have the text information from the semantic labels, we choose corresponding foreground masks  $\mathcal{J}_1, \dots, \mathcal{J}_s$  from  $\mathcal{G}$ . Text features are obtained following:

$$f_{l,i}^{text} = \phi_{text}(t_{s,i}) \quad (6)$$

$$\hat{f}_{l,i}^{text} = \frac{1}{\mathcal{J}} \sum_{j \in \mathcal{J}_i} (\mathcal{F}_{text}(H_j)) \quad (7)$$

Where  $t_{s,i}$  is the corresponding mask semantic label and  $\phi_{text}$  is the pre-trained text encoder, and  $\mathcal{F}_{text}$  is the projection head. We then transfer visual-text pairs to point-text pairs  $(f_{l,i}^{3D}, f_{l,i}^{text})$  and distill the knowledge from text to the point cloud in the object-level. The objective function is as follows:

$$\mathcal{L}_{object} = \frac{1}{K} \sum_i L_1(\hat{f}_{l,i}^{2D}, f_{l,i}^{2D}) + \frac{1}{S} \sum_i L_1(\hat{f}_{l,i}^{text}, f_{l,i}^{text}) \quad (8)$$

The  $L_1$  is the smooth  $L_1$  loss. Our final loss is the sum of previous loss terms.

$$L_{final} = L_{mae} + L_{scene} + L_{object} \quad (9)$$

## 4 Experiments

In this section, we first introduce the pre-training setting of Bridge3D. Then, we show the effectiveness of our method on several popular downstream tasks, including 3D object detection and 3D semantic segmentation. Finally, we conduct extensive ablation studies to show the effectiveness of each design. We put more details into the *supplementary materials*.

### 4.1 Self-supervised Pre-training

**Network architectures.** For the 3D backbone encoder, we utilize the same architecture as Point-MAE [50]. For the image branch, we follow DINOv2 ViT-B [49] to divide 518x518 images into regular patches with a size of  $37 \times 37$ , before the ViT backbone. For the image branch, we directly utilize the CLIP ViT-B [54] to extract text features. For image captioning, we leverage Tag2text [34].

**Pre-training.** During this stage, we perform training of the model for 120 epochs by employing the ScanNet dataset [14] consisting of point clouds and their corresponding images. For the text prompts, we only utilize all class names of ScanNet and SUN RGB-D as the prompts and filter other classes generated from image captioning. We use AdamW [45] optimizer with a base learning rate of  $5e-4$  and weight decay of  $5e-2$ , along with a batch size of 64. The whole masking ratio  $r_w$  is set to 70% and the drop ratio  $r_d$  is set to 40%. The cosine learning rate scheduler is applied, with a drop path rate and warm-up epochs set to 0.1 and 10, respectively. The encoder depth is set to 6, and we utilize the same decoder as Point-MAE [50], with the decoder depth set to 2.

Methods	Pre-trained	SUN RGB-D		ScanNetV2	
		$AP_{25}$	$AP_{50}$	$AP_{25}$	$AP_{50}$
VoteNet [52]	<i>None</i>	57.7	32.9	58.6	33.5
PointContrast [65]	✓	57.5	34.5	59.2	38.0
Hou et al. [32]	✓	-	36.4	-	39.3
4DContrast [10]	✓	-	38.2	-	40.0
DepthContrast [72]	✓	61.6	35.5	64.0	42.9
DPCo [41]	✓	60.2	35.5	64.2	41.5
3DETR [47]	<i>None</i>	58.0	30.3	62.1	37.9
+Bridge3D(from scratch)	<i>None</i>	57.6	31.9	61.1	38.6
+Point-BERT[67]	-	-	-	61.0	38.3
+Point-MAE [50]	✓	-	-	63.4	40.6
+MaskPoint [42]	✓	-	-	63.4	40.6
+ACT [16]	✓	-	-	63.5	41.0
+PiMAE [8]	✓	59.9	33.7	63.0	40.2
+Bridge3D	✓	<b>61.8(+3.8)</b>	<b>35.9(+5.6)</b>	<b>65.3(+3.2)</b>	<b>44.2(+6.3)</b>
GroupFree3D [44]	<i>None</i>	63.0	45.2	67.3	48.9
+Bridge3D(from scratch)	<i>None</i>	62.2	45.0	66.1	48.3
+Point-MAE [50]	✓	63.9	46.1	67.4	49.8
+PiMAE [8]	✓	65.0	46.8	67.9	50.5
+Bridge3D	✓	<b>67.9(+4.9)</b>	<b>48.5(+3.3)</b>	<b>69.1(+1.8)</b>	<b>51.9(+3.0)</b>

Table 1: **3D object detection results on ScanNet and SUN RGB-D dataset.** We adopt the average precision with 3D IoU thresholds of 0.25 ( $AP_{25}$ ) and 0.5 ( $AP_{50}$ ) for the evaluation metrics.

Methods	Pre-trained	S3DIS		ScanNetV2	
		$mIoU$	$mAcc$	$mIoU$	$mAcc$
SR-UNet [65]	<i>None</i>	68.2	75.5	72.1	80.7
PointContrast [65]	✓	70.9	77.0	74.1	81.6
DepthContrast [72]	✓	70.6	-	73.1	-
Hou et al. [32]	✓	72.2	-	73.8	-
Standard Transformer [67]	<i>None</i>	60.0	68.6	-	-
PointBert [67]	✓	60.8	69.9	-	-
PViT [53]	<i>None</i>	64.4	69.9	-	-
PViT+Pix4Point [53]	✓	69.6	75.2	-	-
Ours(from scratch)	<i>None</i>	61.1	67.2	67.3	73.1
+Point-MAE [50]	✓	64.8	70.2	-	-
+Bridge3D	✓	<b>70.2(+9.1)</b>	<b>76.1(+8.9)</b>	<b>73.9(+6.6)</b>	<b>80.2(+7.1)</b>

Table 2: **3D semantic segmentation results on S3DIS dataset.** We adopt the mean accuracy (mAcc) and mean IoU (mIoU) for the evaluation metrics.

## 4.2 Results on Downstream Tasks

For fine-tuning downstream tasks, we discard decoders in pre-training and append task-specific decoders onto the encoder for different tasks.

**Object Detection.** To demonstrate the generality of the proposed method, we also pre-train it on the indoor ScanNetV2 dataset [14] and subsequently fine-tune our method on the object detection task in ScanNetV2 dataset and SUN RGBD [75]. We report our performance on indoor 3D detection based on SOTA methods 3DETR [47] and GroupFree3D [44]. The Table 1 indicates that Our method achieves 66.3  $AP_{25}$  (+4.2) and 45.5  $AP_{50}$  (+7.6) compared to the baseline 3DETR on the ScanNetV2 dataset and also brings significant improvements to both models, surpassing previous baselines consistently in all other datasets and criteria. These experiments’ results showcase our method’s



Point-MAE	Semantic Guided Reconstruction	Scene-level Distillation	Object-level Distillation	ScanNetV2		S3DIS	
				$AP_{25}$	$AP_{50}$	$mIoU$	$mAcc$
✓				62.3	39.9	64.8	70.2
✓	✓			63.2	41.1	66.2	71.1
✓	✓	✓		64.4	43.0	68.4	73.7
✓	✓	✓	✓	<b>65.3</b>	<b>44.2</b>	<b>70.2</b>	<b>76.1</b>

Table 3: **The effectiveness of each component.** Ablation study on the effectiveness of each component on 3D object detection and semantic segmentation tasks.

Text	Image	ScanNetV2		S3DIS	
		$AP_{25}$	$AP_{50}$	$mIoU$	$mAcc$
		62.1	37.9	61.1	67.2
✓		64.2	42.5	67.8	74.1
	✓	64.7	43.3	68.3	74.5
✓	✓	<b>65.3</b>	<b>44.2</b>	<b>70.2</b>	<b>76.1</b>

Table 4: **The effectiveness of each modality.** Ablation study on the effectiveness of each modality on 3D object detection and semantic segmentation tasks.

effectiveness in learning superior 3D representations for object detection, highlighting its potential to benefit a wide range of 3D applications.

**Semantic Segmentation.** In Tab. 2, we present the semantic segmentation results on the S3DIS dataset. Despite their efforts, prior 3D self-supervised methods such as PointContrast[65] and [72] only achieved marginal improvements post-pre-training (+2.7 and +2.4 in  $mIoU$ ). Conversely, Pix4Point [53], utilizing a pre-trained 2D model, demonstrated significant progress compared to training from scratch. Most notably, our proposed method incorporates multiple foundation models during pre-training, elevating the metrics by 10.0 and 10.3 respectively, markedly surpassing other state-of-the-art 3D self-supervised methods. These results substantiate the effectiveness of utilizing multiple foundation models for enhancing 3D representation learning for semantic segmentation.

### 4.3 Ablation Studies

**The effectiveness of Each Component.** As shown in Table 3, the results indicate that each component, including the foreground-aware masking strategy, multi-modal scene-level knowledge distillation, and multi-modal object-level knowledge distillation, contributes to better results. Moreover, when we combined all components, our proposed method achieved the best performance. The foreground-aware masking strategy proved to be important as it enhanced the learning of foreground object representations in the 3D masked autoencoder. The multi-modal scene-level knowledge distillation, which leverages an image captioning model to generate text descriptions from paired images of point clouds, helped bridge the gap between 3D and text at the scene level. The multi-modal object-level knowledge distillation, which uses foundation models to generate accurate object-level masks and semantic text information, bridged the gap between 3D, 2D, and text object-level features. Overall, our ablation study demonstrates the effectiveness of each component in our proposed framework and highlights the importance of leveraging foundation models to improve 3D scene representation learning.

**The effectiveness of Each Modality.** Table 4 provides a clear insight into the significant contribution of each modality to the overall performance of our method. By integrating all modalities, our method realizes its full potential, showcasing the best results. However, it’s worth noting the inherent resilience our method exhibits to modality variations. This adaptability implies that even when the modality mix is altered or reduced, the system remains relatively unaffected in terms of its output quality. This inherent resilience not only underscores the robust architecture of our model but also offers users the freedom to customize the framework based on their specific requirements. The adaptable nature of modality inclusion thus ensures that our method remains both versatile and efficient, enabling users to balance computational overhead with optimal performance.

Semantic Mask Method	ScanNetV2		S3DIS	
	$AP_{25}$	$AP_{50}$	$mIoU$	$mAcc$
Point-MAE [50]	62.3	39.9	64.8	70.2
MaskCLIP [76] + Superpixels [1]	64.1	43.0	67.1	73.4
Bridge3D	<b>65.3</b>	<b>44.2</b>	<b>70.2</b>	<b>76.1</b>

Table 5: **The effectiveness of mask.** Ablation study on unsupervised mask methods on 3D object detection and semantic segmentation tasks.

Methods	Pre-trained	SUN RGB-D		ScanNetV2	
		$AP_{25}$	$AP_{50}$	$AP_{25}$	$AP_{50}$
CAGroup3D [63]	<i>None</i>	66.8	50.2	75.1	61.3
+Bridge3D (scene & object level distillation)	✓	<b>68.7</b>	<b>52.1</b>	<b>76.3</b>	<b>62.2</b>

Table 6: **The performance on SOTA method.** 3D object detection results on ScanNet and SUN RGB-D dataset based on CAGroup3D.

**The Effectiveness of Masks from Foundation Models.** In Table 5, we conduct a comparative study of mask generation strategies employed by our method and those utilized by prior works. For instance, CLIP2Scene[9] employs MaskCLIP [76] to produce semantic masks. However, this approach fails to provide instance segmentation masks and is solely capable of guiding text-to-3D knowledge distillation. Alternatively, SLiDR [59] leverages superpixels [1] for generating object-level masks. Despite this, the superpixels lack semantic information, limiting their use to guiding 2D-to-3D knowledge distillation. In contrast, our method generates instance semantic information, enhancing the functionality and accuracy of the masks. For a comparative analysis, we combine superpixels and MaskCLIP to direct both 2D-to-3D and text-to-3D distillation, essentially mimicking the fusion of CLIP2scene and SLiDR. The experimental results reveal that masks generated through our method, which leverages foundation models, yield substantial improvements compared to the combined use of MaskCLIP and superpixels.

**Apply Bridge3D in SOTA Method.** Our method’s adaptability extends to its successful application to the recent state-of-the-art (SOTA) work, CAGroup3D [63]. While the network structure made a direct application of the plain transformer challenging, we devised a unique adaptation, pre-training the CAGroup3D’s backbone with our scene and object-level distillation, excluding the reconstruction part. Table 2 illustrates how our pre-training approach can enhance and benefit current SOTA 3D detection methodologies, showcasing the potential reach of our technique. Importantly, it should be noted that our framework is optimized for plain transformer-based backbones, and thus the application of Bridge3D to alternative backbones may lead to a reduction in performance.

## 5 Conclusion

In this work, we introduce a pioneering method Bridge3D that capitalizes on foundation models to overcome the hurdles in self-supervised 3D learning. This innovative approach not only enables more focused learning on foreground object representation but crucially bridges the domain gap between 3D and text at the scene level. Furthermore, it enriches the quality of 3D scene representation learning by generating highly accurate object-level masks and semantic textual information, effectively bridging the gap between 3D, 2D, and text object-level features. Our comprehensive experimental results corroborate the superior effectiveness of our approach in amplifying 3D scene understanding. However, the current work primarily focuses on indoor 3D scene understanding, which constitutes a limitation. Looking ahead, we plan to broaden the applicability of Bridge3D to encompass a more diverse set of 3D tasks, including outdoor scene understanding and open-vocabulary 3D tasks. Our work is expected to have no negative societal implications.

## References

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.
- [2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2017.
- [3] 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*, pages 9902–9912, 2022.
- [4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [5] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016.
- [6] Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C Berg. Point-level region contrast for object detection pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16061–16070, 2022.
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [8] Anthony Chen, Kevin Zhang, Renrui Zhang, Zihan Wang, Yuheng Lu, Yandong Guo, and Shanghang Zhang. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. *arXiv preprint arXiv:2303.08129*, 2023.
- [9] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. *arXiv preprint arXiv:2301.04926*, 2023.
- [10] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 543–560. Springer, 2022.
- [11] Zhimin Chen, Longlong Jing, Yang Liang, YingLi Tian, and Bing Li. Multimodal semi-supervised learning for 3d objects. *arXiv preprint arXiv:2110.11601*, 2021.
- [12] Zhimin Chen, Longlong Jing, Liang Yang, Yingwei Li, and Bing Li. Class-level confidence based 3d semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 633–642, 2023.
- [13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [14] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [16] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022.
- [17] Bi’an Du, Xiang Gao, Wei Hu, and Xin Li. Self-contrastive learning with hard negative sampling for self-supervised point cloud learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3133–3142, 2021.
- [18] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.
- [19] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

- [20] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004.
- [21] Ziyue Feng, Longlong Jing, Peng Yin, Yingli Tian, and Bing Li. Advancing self-supervised monocular depth learning with sparse lidar. In *Conference on Robot Learning*, pages 685–694. PMLR, 2022.
- [22] Ziyue Feng, Liang Yang, Pengsheng Guo, and Bing Li. Cvrecon: Rethinking 3d geometric feature learning for neural reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17750–17760, 2023.
- [23] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *European Conference on Computer Vision*, pages 228–244. Springer, 2022.
- [24] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3d point cloud processing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.
- [25] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all, 2023.
- [26] Kaveh Hassani and Mike Haley. Unsupervised multi-task feature learning on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8160–8171, 2019.
- [27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [28] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal M Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. *arXiv preprint arXiv:2303.11313*, 2023.
- [29] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and Joao Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021.
- [30] Georg Hess, Adam Tonderski, Christoffer Petersson, Lennart Svensson, and Kalle Åström. Lidarclip or: How i learned to talk to point clouds. *arXiv preprint arXiv:2212.06858*, 2022.
- [31] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [32] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021.
- [33] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021.
- [34] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023.
- [35] Longlong Jing, Yucheng Chen, Ling Zhang, Mingyi He, and Yingli Tian. Self-supervised modal and view invariant feature learning. *arXiv preprint arXiv:2005.14169*, 2020.
- [36] Longlong Jing, Zhimin Chen, Bing Li, and Yingli Tian. Self-supervised modality-invariant and modality-specific feature learning for 3d objects, 2022.
- [37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
- [38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [39] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- [40] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.
- [41] Lanxiao Li and Michael Heizmann. A closer look at invariances in self-supervised pre-training for 3d vision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 656–673. Springer, 2022.

- [42] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 657–675. Springer, 2022.
- [43] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [44] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021.
- [45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [46] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [47] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021.
- [48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [50] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 604–621. Springer, 2022.
- [51] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. *arXiv preprint arXiv:2211.15654*, 2022.
- [52] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019.
- [53] Guocheng Qian, Xingdi Zhang, Abdullah Hamdi, and Bernard Ghanem. Pix4point: Image pretrained transformers for 3d point cloud understanding. *arXiv preprint arXiv:2208.12259*, 2022.
- [54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [55] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [56] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [57] Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *European Conference on Computer Vision*, pages 626–642. Springer, 2020.
- [58] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32, 2019.
- [59] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022.
- [60] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. *arXiv preprint arXiv:2303.13496*, 2023.
- [61] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

- [62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [63] Haiyang Wang, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, Liwei Wang, et al. Cagroup3d: Class-aware grouping for 3d object detection on point clouds. *Advances in Neural Information Processing Systems*, 35:29975–29988, 2022.
- [64] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021.
- [65] 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.
- [66] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 206–215, 2018.
- [67] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322, 2022.
- [68] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. *arXiv preprint arXiv:2303.04748*, 2023.
- [69] 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.
- [70] Renrui Zhang, Lihui 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.
- [71] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [72] 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.
- [73] Yongheng Zhao, Tolga Birdal, Haowen Deng, and Federico Tombari. 3d point capsule networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1009–1018, 2019.
- [74] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luwei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.
- [75] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.
- [76] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021.