Dense Multimodal Alignment for Open-Vocabulary 3D Scene Understanding

Ruihuang Li[1], Zhengqiang Zhang[1], Chenhang He[1], Zhiyuan Ma[1], Vishal M. Patel[2], and Lei Zhang[1(✉)]

[1] Hong Kong Polytechnic University
[2] Joins Hopkins University
{csrhli, cslzhang}@comp.polyu.edu.hk, vpatel36@jhu.edu
https://github.com/lslrh/DMA

Abstract. Recent vision-language pre-training models have exhibited remarkable generalization ability in zero-shot recognition tasks. Previous open-vocabulary 3D scene understanding methods mostly focus on training 3D models using either image or text supervision while neglecting the collective strength of all modalities. In this work, we propose a Dense Multimodal Alignment (DMA) framework to densely co-embed different modalities into a common space for maximizing their synergistic benefits. Instead of extracting coarse view- or region-level text prompts, we leverage large vision-language models to extract complete category information and scalable scene descriptions to build the text modality, and take image modality as the bridge to build dense point-pixel-text associations. Besides, in order to enhance the generalization ability of the 2D model for downstream 3D tasks without compromising the open-vocabulary capability, we employ a dual-path integration approach to combine frozen CLIP visual features and learnable mask features. Extensive experiments show that our DMA method produces highly competitive open-vocabulary segmentation performance on various indoor and outdoor tasks.

Keywords: 3D Scene understanding · Open-vocabulary · Multimodal alignment

1 Introduction

3D scene understanding, which aims to achieve accurate comprehension of objects as well as their attributes and relationships within a scene, has gained significant attention in recent years due to its popular applications in autonomous driving [32], virtual reality (VR) [2,40,50] and robot navigation [3], *etc*. However, the annotation of large-scale 3D data is very costly [7,11], impeding the training of generalizable models for open-vocabulary scene understanding. Though many existing methods [9,10,20,29–31,41,46,58] have achieved significant advancements in recognizing closed-set categories for specific tasks, they fail to identify novel categories and other types of queries [42] without 3D supervision, hindering the application of existing 3D scene understanding methods to real-world settings, where the number of possible classes is unlimited.

In contrast to the limited 3D data, modalities such as images and texts are more abundantly available. Existing pre-trained multimodal models, such as CLIP [43] and ALIGN [24], have shown impressive zero-shot recognition ability by training on large-scale noisy image-text pairs, and have been successfully adapted for open-vocabulary classification [53, 54], detection [5, 38] and segmentation tasks [33, 47, 52]. Based on these observations, researchers have attempted to use image or natural language modalities to provide supervisory signals for learning 3D representations [13, 36, 42, 55]. Some methods use fixed 2D features as supervision and distill the knowledge from either the pre-trained 2D encoder of CLIP [36] or 2D open-vocabulary segmentation (OVSeg) models [42] into 3D representations (NeRF or point clouds). However, they overlook the fact that 3D models can in turn enhance 2D models by leveraging the strong 3D structural information. Besides, the 2D OVSeg models compromise their open-vocabulary ability since they are primarily fine-tuned on in-vocabulary datasets. There are also some methods that directly align 3D features to semantic captions [13, 45, 55]. However, they only capture coarse image- or region-level descriptions without establishing dense point-to-text correspondences or exploiting image features that involve rich semantic contexts and more variations. Though some methods [53, 54] simultaneously leverage visual and textual supervisions, they only conduct coarse multimodal alignment for object-level point cloud classification.

In order to leverage the synergistic benefits of multiple modalities for dense prediction tasks, we propose a dense multimodal alignment (DMA) strategy to co-embed 3D points, image pixels, and text strings into a shared latent space. To build dense associations across different modalities, the primary bottleneck is *how to obtain rich and reliable text descriptions without relying on manual labeling*. To this end, we generate two types of prompts using large Vision-Language Models (VLMs). Firstly, we employ the tagging model such as RAM [57] to detect as many categories as possible from an image, ensuring alignment with **complete** semantic patterns. Considering that category names might not provide sufficient details and contextual information, we incorporate Multimodal Large Language Models (MLLM) such as LLaVA [35] to generate linguistically expressible scene descriptions, thereby enhancing the **scalability** of text queries. In addition, we use the GPT to filter out the noise in the generated texts for improving the **reliability**. As a result, we establish a highly scalable and informative text modality, enhancing the overall understanding of 3D scenes.

As for the image modality, we adopt a **dual-path integration** strategy to extract robust 2D features as supervision. Specifically, we employ the FC-CLIP [56] as the feature extractor. On one hand, we fix its CLIP visual encoder to maintain the open-world recognition ability. On the other hand, by fine-tuning its mask head, we *incorporate 3D structural priors into 2D features*, better adapting the model to 3D dense tasks. Then we build triplets of points, pixels, and their corresponding texts by taking image modality as the bridge. Given the generated triplets of different modalities and their dense correspondences, we finally adopt the **mutually inclusive loss** function to align multiple modalities. In this

way, we can effectively unleash the potential of existing foundation VLMs and maximize the complementary effects of multiple modalities.

In summary, (1) we first present a dense multimodal alignment framework, which establishes dense correspondences among points, pixels and texts, to learn robust 3D representations for open-vocabulary 3D scene understanding. (2) To generate complete and scalable language modality without relying on manual annotations, we leverage a tagging model and an MLLM to extract category information and scene descriptions, respectively. (3) Finally, to improve the segmentation ability without compromising the open-vocabulary ability, we integrate 3D priors into 2D features by fine-tuning the 2D mask head with the backbone frozen. Extensive experiments demonstrate the outstanding open-world 3D segmentation ability of our DMA model on various indoor and outdoor tasks.

2 Related Work

Open-Vocabulary 3D Scene Understanding. 3D scene understanding is a popular research topic in computer vision. Most previous methods [10, 15, 19, 20, 58] focus on training models on manually labeled close-set categories, and have yielded promising performance on popular 3D benchmarks [4, 11]. However, most of these methods are designed for a specific task, such as object classification [51], detection [8], semantic/instance segmentation [10, 15, 58], and they cannot identify novel categories, restricting their applications to real-world settings. To overcome this limitation, recent works have been focused on the open-vocabulary scene understanding problem [36]. Rozenberszki *et al.* [45] proposed a language-driven pre-training method to enforce 3D feature to be close to text embeddings, and finetune the 3D encoder with ground-truth annotations. PLA [13] and RegionPLC [55] explicitly associate 3D points with image- and region-level image captions, respectively. However, existing image captioning models can only identify sparse and salient objects while missing other important categories. Besides, textual signals lack variations and contexts, making them insufficient for dense prediction tasks. Some methods [36, 42] distill knowledge from large-scale pre-trained 2D models, such as image-text contrastive learning models [43] and open-vocabulary segmentation models [14, 33, 52, 56]. However, the performance of pre-trained models drops a lot on the downstream datasets due to the large domain shift. These methods also overlook the fact that 3D models can in turn enhance 2D models by leveraging the strong structural information inherent in 3D data.

Vision-Language Foundation Models. Recent vision-language foundation models have exhibited remarkable generalization ability on zero-shot prediction tasks. Segment Anything Model (SAM) [26] leads a new trend of universal image segmentation and exhibits promising results on diverse downstream tasks. Recognize Anything Model (RAM) [57] presents a novel paradigm for image tagging (multi-label classification) by leveraging large-scale image-text pairs for training without manual annotations. The recent success of ChatGPT and GPT4 have stimulated tremendous interests in developing multimodal large

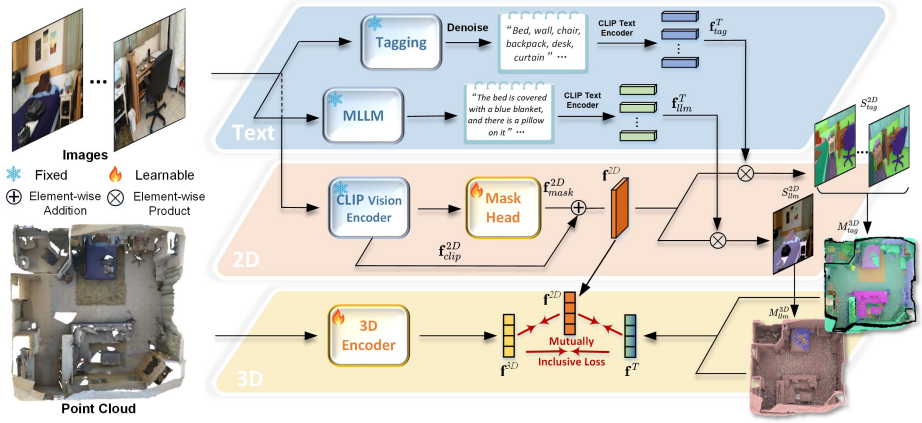

Fig. 1: Framework of our proposed Dense Multimodal Alignment (DMA) method. We generate comprehensive language modality data by leveraging a tagging model and an MLLM. As for 2D modality, we fix the CLIP visual backbone $\mathbf{f}^{2D}_{clip}$ but finetune the mask head $\mathbf{f}^{2D}_{mask}$ for better adaptation to downstream 3D tasks without compromising the open-vocabulary ability. Then the dense correspondences between pixels $\mathbf{f}^{2D}$ and texts $\mathbf{f}^{T}_{tag}/\mathbf{f}^{T}_{llm}$ can be built by computing their feature similarities, resulting in semantic score maps $S^{2D}_{tag}/S^{2D}_{llm}$. By taking image modality as the bridge, we back-project text labels to each point and obtain the 3D label maps $M^{3D}_{tag}/M^{3D}_{llm}$. Finally, we co-embed point $\mathbf{f}^{3D}$, pixel $\mathbf{f}^{2D}$, and text embeddings $\mathbf{f}^{T}$ into a common space to learn a robust 3D representation by optimizing the mutually inclusive loss function.

language models (LLMs). LLaVA [35] is an early exploration to apply LLMs to the multimodal fields by connecting a vision encoder to LLM for general-purpose visual and language understanding. The recent open-vocabulary methods [33, 52, 56] shed lights on the direct use of pre-trained foundation models for handling different visual tasks. ODISE [52] explores the potential ability of pretrained text-to-image diffusion models [44] for open-vocabulary panoptic segmentation. FC-CLIP [56] utilizes a shared frozen convolutional CLIP backbone to maintain the ability of open-vocabulary classification without compromising accuracy.

3 Method

As illustrated in Fig. 1, we propose a dense multimodal alignment (DMA) framework for open-vocabulary 3D scene understanding, where we construct dense correspondences across 2D image pixels, 3D points and 1D texts, and embed them into a common latent space. In this section, we will elaborate the construction of text and image modalities, and explain how we associate and align them in a dense manner.

3.1 Comprehensive Text Modality Generation

Learning a robust 3D model that is generalizable to open vocabularies is challenging since it is unclear how to acquire the dense text labels for point clouds.

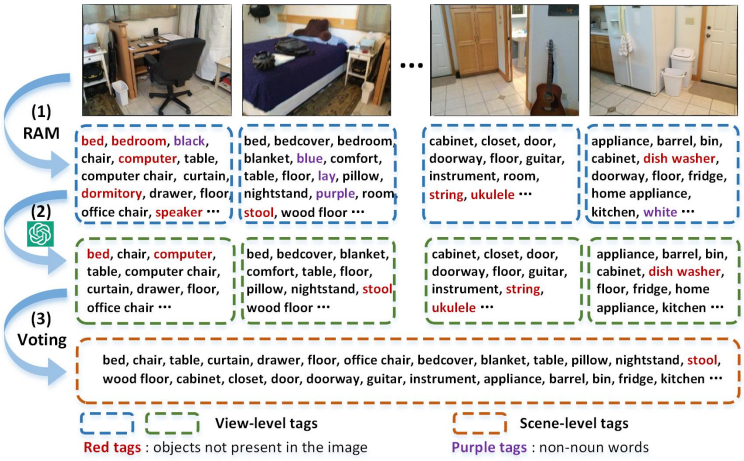

Fig. 2: Scene tagging generation. (1) We first employ RAM [57] to generate view-level tags, and then (2) reduce the tag noise with GPT. Finally, scene-level tags are generated by (3) multi-view voting.

Although well-trained human annotators could potentially provide detailed language descriptions of 3D scenes, such a method is costly and lacks scalability. To overcome this limitation, we leverage a tagging model and an MLLM to extract **complete** category information and **scalable** scene descriptions, respectively.

Complete Category Information. The scene tagging process is illustrated in Fig. 2. Firstly, we use the image tagging foundation model such as RAM [23] to extract all possible categories from an image, and utilize category names and short descriptions derived from the metadata as the text query, referred to as T_{tag}, such as "`There is a {category name} in the scene`", "`A photo of a {category name}`", *etc*. Unlike image captioning models [1, 28] that can only identify sparse and salient objects in a scene, RAM can recognize as many tags as possible without missing important parts, *ensuring a high recall rate and alignment with complete semantic patterns*. The more complete and accurate the detected categories are, the easier we can establish precise dense correspondences between text and 3D modalities, and hence the open-vocabulary capability of the 3D model can be enhanced.

Reliable GPT-based Denoising. While there are many tags recognized by RAM, some redundant or irrelevant tags are also included, such as non-noun words ("**purple**", "**blue**", "**lay**", *etc*.) and objects that do not exist in the image ("**bed**", "**ukulele**", *etc*.), as shown in Fig. 2. We address this issue in two steps. Firstly, we utilize GPT to filter potential noisy vocabulary. Given the input list, we instruct GPT to examine the words one by one and perform reasoning according to the chain of thought, outputting a boolean list indicating whether a word is an outlier. Please refer to *Fig. 1 of supplemental material* for the detailed instructions and examples to reduce the noisy tags. Secondly, to decrease the non-existent categories in a scene, we conduct multi-view voting and neglect categories that appear in fewer than five views. Please refer to *Fig. 2*

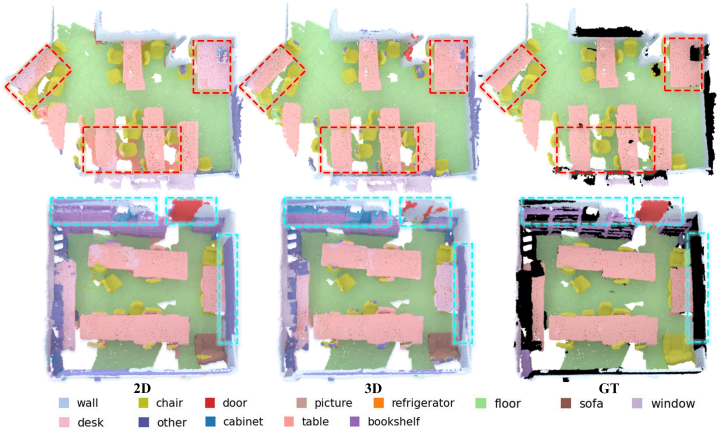

Fig. 3: Segmentation results using 2D and 3D models. 2D model has advantages in segmenting background objects (in **blue** boxes), while 3D model is more favorable for foreground objects with distinct structures (in **red** boxes).

of the supplemental material for the denoised scene tagging results and the corresponding visualizations of 3D label maps.

Scalable Scene Description. Although scene-level tags have already covered most of the categories, the limited scalability and variation of category names hinders their provision of rich contexts and details. To address this limitation and enable arbitrary queries for 3D networks, we additionally leverage MLLMs such as LLaVA [35] to generate diverse and linguistically expressible descriptions, denoted by T_{llm}. Owing to the exposure to a diverse range of linguistic patterns and contextual nuances, the MLLMs can generate comprehensive and in-depth descriptions based on input images. Consequently, these LLMs can enhance the richness and granularity of the generated textual representations, thereby facilitating a more comprehensive understanding of the 3D scenes. Please refer to *Fig. 3 of supplemental material* for the examples of scene-level captions and corresponding visualizations.

Finally, we generate the text embeddings $\mathbf{f}_{tag}^{T}$ and $\mathbf{f}_{llm}^{T}$ using CLIP text encoder based on the generated tags T_{tag} and scene descriptions T_{llm}, respectively, which are utilized to supervise the training of 3D networks subsequently.

3.2 Structure-aware Image Feature Extraction

Compared to language modality, the image modality offers a wealth of contextual information and exhibits significant variations among different pixels, which could provide more effective supervision. Inspired by this observation, OpenScene [42] distills the knowledge from frozen open-vocabulary 2D segmentation models, such as LSeg [27] and OpenSeg [14]. However, these methods suffer from two major limitations. Firstly, they are fine-tuned on in-vocabulary datasets, which leads to a misalignment between image and text features and consequently results in poor performance on open-vocabulary categories. Secondly, all of these methods

freeze 2D networks, failing to perceive the 3D structure of objects and leading to inaccurate supervision. As shown in Fig. 3, we visualize the segmentation results using 2D and 3D features. One can observe that although 2D features are more advantageous in segmenting background objects with ambiguous geometry, such as "bookshelf", "door" and "blackboard", they are less effective in segmenting objects with distinct shapes, such as "table" and "chair". Therefore, it is necessary to distill the structural priors of 3D networks into 2D ones as well in order to facilitate fine-grained scene understanding.

In this paper, we adopt FC-CLIP [56] as the backbone to extract image features. On one hand, we use the frozen CLIP visual encoder to ensure the intactness of image-text alignment, obtaining CLIP features $\mathbf{f}^{2D}_{clip}$. On the other hand, to facilitate the synergistic benefits of both 2D and 3D modalities, we fine-tune the mask head and attain the mask features $\mathbf{f}^{2D}_{mask}$. In contrast to previous methods that rely on potentially noisy fixed image features for supervision, the fine-tuned mask features enhance the adaptability to downstream 3D tasks. We explore different fine-tuning strategies, such as LoRA [18], Adapter [17], and full parameter fine-tuning and compare them in experiments.

3.3 Dense Associations across Modalities

Once the text and image modalities are constructed, the subsequent step is to associate each point to its corresponding pixel and text. We utilize the image modality as a bridge to establish separate associations between pixels and other modalities. Firstly, we construct the associations between image and language modalities by taking C different text embeddings $\mathbf{f}^T = \{\mathbf{f}^T_1, \cdots, \mathbf{f}^T_C\}$ as classifier to assign text labels to each pixel, obtaining a 2D semantic score map, denoted by $S^{2D} \in \mathbb{R}^{H\times W\times C}$. This process can be formulated as follows:

$$S^{2D}_c(u,v) = \sigma(< \mathbf{f}^{2D}(u,v), \mathbf{f}^T_c > /\tau_1), \tag{1}$$

where $S^{2D}_c(u,v)$ denotes the probability that the pixel at location (u,v) belongs to the c-th text label, and $< \cdot >$ represents the cosine similarity between two ℓ_2-normalized feature vectors. τ_1 is a temperature parameter. Then we establish the associations between images and point clouds by back-propagating 3D points $\mathbf{p} = (x,y,z)$ onto 2D positions (u,v) using a projection matrix $T \in \mathbb{R}^{3\times 4}$, *i.e.*, $[u,v,w] = T[x,y,z,1]$, where w is a scaling factor.

Finally, we associate the text and 3D modalities by taking image as the bridge. Given K different projection views for one point $\mathbf{p}$, we compute its average semantic score, denoted as $\bar{S}^{3D}$, across K views $[S^{2D}_1(u_1,v_1), \cdots, S^{2D}_K(u_K,v_K)]$. Based on the aggregated 3D semantic scores, the final text-to-3D label map, denoted as $M^{3D} \in \mathbb{R}^{N\times C}$, can be derived by:

$$M^{3D}_c = \begin{cases} 1, \text{ if } \ \bar{S}^{3D}_c > \text{threshold} \\ 0, \qquad\quad else \end{cases}, \tag{2}$$

where N denotes the number of points and M^{3D}_c indicates whether the point belongs to the c-th text label or not. M^{3D} can be regarded as the pseudo label map for point cloud, serving as the supervision signal for training 3D models.

It is noteworthy that instead of generating one-hot label through the `argmax` operation, we select all confident text labels whose scores exceed the threshold. This is because the generated text categories may exhibit similarities in semantics (like 'suitcase' and 'luggage') or inclusion relationships (such as 'kitchen' and 'stove'). As a result, it is highly possible that *one single point corresponds to multiple text labels simultaneously.*

3.4 Dense Multimodal Alignment

After obtaining the triplets of different modalities and their dense correspondences, the subsequent objective is to align the 3D points with their corresponding text and pixel embeddings. This alignment process involves several steps. Firstly, we extract 3D features for the point cloud by utilizing a 3D network, denoted as ε_{3D}. These features are then projected to match the dimension of the CLIP features.

Next, we assign text labels to different 3D points by computing the cosine similarities between point and text embeddings $\mathbf{f}^T$, yielding a 3D segmentation probability map P^{3D}:

$$P^{3D}_{i,c} = \sigma(< \mathbf{f}^{3D}_i, \mathbf{f}^T_c > /\tau_2), \tag{3}$$

where $\mathbf{f}^{3D}_i$ denotes the feature of the i-th point, and $P^{3D}_{i,c}$ denotes the probability that the i-th point belongs to the c-th text label. Here we employ the Sigmoid activation function $\sigma(\cdot)$ since it will not lead to mutually exclusive relationships among different categories.

Text-to-3D Supervision. We use the text-to-3D label map M^{3D} as the pseudo label to facilitate the alignment of point and text features. Different loss functions are employed for aligning the point embeddings with the tag $\mathbf{f}^T_{tag}$ and scene description $\mathbf{f}^T_{llm}$ embeddings. As can be seen in Fig. 1, we build dense associations between $\mathbf{f}^T_{tag}$ and the **entire point cloud**, resulting in M^{3D}_{tag}. Consequently, The Binary Cross Entropy (BCE) loss is used to effectively penalize both positive and negative samples:

$$\mathcal{L}_{3d-text(tag)} = \mathcal{L}_{BCE}(P^{3D}, M^{3D}_{tag}). \tag{4}$$

As for $\mathbf{f}^T_{llm}$, since it corresponds only to salient objects, we can only obtain the mask for **partial points**, denoted as M^{3D}_{llm}. (Visualizations of M^{3D}_{tag} and M^{3D}_{llm} are given in *Fig. 2 and Fig. 3 of the supplementary material*, respectively.) We utilize the cosine similarity loss to supervise only the positive samples:

$$\mathcal{L}_{3d-text(llm)} = 1 - \cos(\mathbf{f}^T_{llm}, \mathbf{f}^{3D}). \tag{5}$$

Mutually Inclusive Loss. In this work, we do not employ the Cross-Entropy loss because it would result in a mutually exclusive relationship between different classes, meaning that each point is assigned to only one class of interest. However, in text-to-3D alignment, one point may simultaneously associate with multiple text prompts, such as 'bed' and 'bedroom', 'chair' and 'office chair', 'curtain' and 'drape', *etc*. To handle this issue, we employ mutually inclusive losses (**MIL**),

such as BCE loss and cosine similarity loss, to ensure that each point is aligned with all its corresponding tags/descriptions simultaneously, avoiding the potential conflicts between categories with overlapping or similar semantics.

2D-to-3D Supervision. For 3D-2D pairs, we follow the previous work [42] to fuse pixel embeddings across K different views, represented as $[\mathbf{f}_1^{2D}, \cdots, \mathbf{f}_K^{2D}]$, into a single feature vector $\bar{\mathbf{f}}^{2D}$, and align 2D and 3D features by minimizing the cosine similarity loss:

$$\mathcal{L}_{3d-2d} = 1 - \cos(\bar{\mathbf{f}}^{2D}, \mathbf{f}^{3D}). \tag{6}$$

Since 2D mask head is also trainable, we additionally add the text-to-2D supervision and compute the BCE loss between 2D predictions and 2D masks, obtaining $\mathcal{L}_{text-2d}$. Finally, the overall objective function to perform dense multimodal alignment is defined as:

$$\mathcal{L}_{3D} = \mathcal{L}_{3d-text(tag)} + \mathcal{L}_{3d-text(llm)} + \mathcal{L}_{3d-2d} + \mathcal{L}_{text-2d}, \tag{7}$$

where the language modality provides comprehensive textual descriptions, and the image modality gives precise supervision on object edges and contextual information. Additionally, the 3D modality reveals crucial structural information of objects. By densely aligning these modalities in a shared space, our method can maximize the synergistic benefits among them and achieve outstanding segmentation performance without compromising the open-vocabulary classification ability of the model.

4 Experiments

4.1 Setups

Datasets. To demonstrate the effectiveness of our proposed method, we employ three popular datasets, *i.e.*, ScanNet [11], Matterport3D [6], and nuScenes [4]. The first two datasets are indoor ones, comprising RGBD images and 3D meshes. The third one is an outdoor dataset, consisting of data collected from two sensors, *i.e.*, LiDAR and camera. We conduct comparisons with state-of-the-art methods on each of these datasets. The mean Intersection-of-Union (mIoU), mean Accuracy (mACC), Precision, and Recall are employed as the evaluation metrics.

Implementation Details. In this work, MinkowskiNet [10] is employed as the 3D backbone, whose voxel size is set to 2cm for ScanNet and Matterport3D and 5cm for nuScenes. As for the 2D backbone, we use OpenSeg [14] and FC-CLIP [56] that perform mask-wise classification. The parameters τ_1 in Eq. 1 and τ_2 in Eq. 3 are both set to 0.1. We use Adam [25] as the optimizer and the initial learning rate is set to $1e-4$. The model is trained for 100 epochs. We set the batch size as 8 for indoor datasets and use one single NVIDIA RTX A6000 for training. As for nuScenes dataset, we use 8 GPUs for training and set the batch size as 16.

	Methods	mIoU	mACC	mIoU(F)	mIoU(B)	Latency
fully-supervised	TangentConv [49]	40.9	–	–	–	–
	TextureNet [21]	54.8	–	–	–	–
	ScanComplete [12]	56.6	–	–	–	–
	Mix3D [41]	**73.6**	–	–	–	–
	VMNet [20]	73.2	–	–	–	–
	MinkowskiNet [20]	69.0	–	–	–	–
Zero-shot	PLA [13]	17.7	33.5	–	–	0.07s
	RegionPLC [55]	43.8	65.6	–	–	0.07s
	OpenScene [42](LSeg)-3D	52.9	63.2	–	–	0.07s
	OpenScene(LSeg)-2D3D	54.2	66.6	–	–	102.6s
	OpenScene†(OpenSeg)-3D	46.6	66.5	50.0	47.1	0.07s
	OpenScene†(OpenSeg)-2D3D	47.9	**71.7**	49.5	51.0	89.4s
	DMA(OpenSeg)-text only	50.5	63.7	56.7	48.0	0.07s
	DMA(OpenSeg)-3D	53.3	70.3	58.3	51.5	0.07s
	DMA(LSeg)-3D	**54.8**	66.9	**59.9**	**51.9**	0.07s
	DMA(FC-CLIP)-3D	51.8	68.7	56.0	51.4	0.07s

Table 1: Comparison on the ScanNet [11] validation set. "F" and "B" denote foreground and background classes, respectively. † denotes our reproduced results.

	Methods	Anno.	mIoU	mIoU (Base)	mIoU (Long-Tail)
fully-sup.	RangeNet++ [39]	100%	65.5	76.4	56.4
	Cylinder3D [58]		75.4	**84.1**	69.4
	SPVNAS [48]		74.8	82.3	67.2
	AMVNet [34]		**77.0**	83.9	**70.8**
weakly-sup.	ContrastiveSC [16]	0.9%	64.5	79.7	53.8
	LESS [37]		**74.8**	**81.6**	**68.7**
	ContrastiveSC	0.2%	63.5	78.4	51.6
	LESS		73.5	81.1	66.6
Zero-shot	OpenScene [42](LSeg)-2D3D	No	36.7	55.0	22.3
	OpenScene(OpenSeg)-2D3D		42.1	52.6	33.8
	DMA(OpenSeg)-3D	No	45.1	59.3	33.9
	DMA(FC-CLIP)-3D		**47.4**	**61.4**	**35.3**

Table 2: Comparison on the nuScenes [4] validation set. We partition all categories into base and long-tail classes according to their frequencies.

4.2 Comparison with State-of-the-Arts

We compare the proposed DMA with fully-/weakly-supervised and zero-shot methods [13,42,55]. Tab. 1 presents the segmentation results on the ***ScanNet*** [11] dataset. To facilitate comparison, we measure the results of OpenScene by using 3D and 2D-3D integrated features as supervision. As can be seen in Tab. 1, although OpenScene(LSeg) attains better results (54.2% mIoU) by using both 2D and 3D encoders, it results in significantly **increased inference latency**. This is because the parameter size of 2D encoder is much larger than 3D encoder, and the 2D encoder needs to perform inference on multi-view images of the scene. Our DMA(OpenSeg) using only 3D model for prediction outperforms OpenScene(OpenSeg)-2D3D by 5.4% mIoU at a significantly lower latency, wherein the mIoU (F) and mIoU (B) are improved by 8.8% and 0.5%, respectively. This is because we perform additional alignment with text modality, thereby compensating for the decreased open-vocabulary ability of 2D model. When using text supervision only, our method outperforms the text-supervised approach RegionPLC [55] by 9.5%, and even surpasses OpenScene(OpenSeg)-2D3D by 2.6% in terms of mIoU. This indicates that, compared to previous methods that generate image- and region-level captions, our method establishes dense and precise correspondences between text and 3D points by taking 2D modality as the bridge, achieving more precise supervision. The suboptimal performance of our method using FC-CLIP as the 2D encoder may be attributed to the low resolution of the images (320×240), which limits the capabilities of FC-CLIP.

Outdoor Scenes. To validate the effectiveness of our method on outdoor point clouds, we evaluate the performance of DMA on the ***nuScenes*** dataset [4]. Due to the highly imbalanced class distribution of outdoor scenes, we additionally measure the performance on base and long-tail categories. As shown in Tab. 2, by densely aligning with the tagging information and the detailed description extracted from each scene, our DMA(OpenSeg) using only 3D encoder significantly improves the performance over OpenScene(OpenSeg)-2D3D by 3.0% mIoU. Additionally, the final performance is further improved by 2.3% and attains

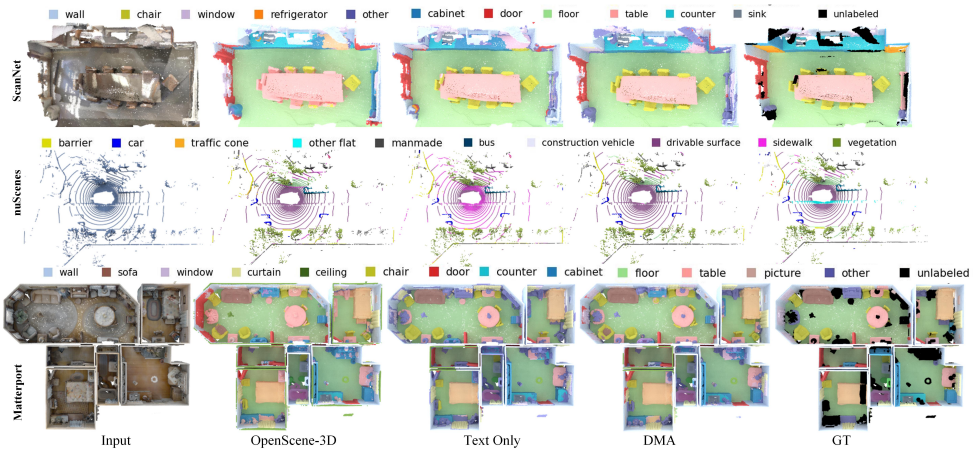

Fig. 4: Qualitative results of different methods on both indoor and outdoor datasets.

	mIoU				mACC				Precision				Recall			
	Head	Common	Tail	All	Head	Common	Tail	All	Head	Common	Tail	All	Head	Common	Tail	All
OpenScene† [42]-2D3D	21.2	8.4	4.0	6.2	35.5	15.7	9.4	12.0	43.9	18.6	7.8	14.5	34.5	15.7	9.4	12.0
PLA [13]	–	–	–	1.8	–	–	–	3.1	–	–	–	–	–	–	–	–
RegionPLC [55]	–	–	–	6.5	–	–	–	15.9	–	–	–	–	–	–	–	–
DMA-text only	23.2	7.6	2.0	6.9	32.6	13.4	5.9	11.3	40.0	15.1	5.8	13.7	32.6	13.4	5.9	11.3
DMA(OpenSeg)-3D	25.3	10.8	5.5	7.6	36.7	18.2	10.7	14.6	44.5	23.6	10.5	14.9	36.7	18.2	10.7	13.2
DMA(FC-CLIP)-3D	27.2	11.5	5.8	7.9	37.4	19.2	11.2	15.2	46.2	24.9	11.3	15.7	38.2	20.4	11.1	14.0
Fully-Sup	45.5	13.6	3.4	20.8	–	–	–	–	66.8	55.7	23.3	34.4	57.6	19.1	5.8	27.8

Table 3: Comparison on ScanNet200 [45] validation set. † means our reproduced results.

47.4% mIoU by employing FC-CLIP [56] to extract 2D features. This is because FC-CLIP could achieve more precise segmentation while maintaining outstanding open-vocabulary recognition ability of CLIP. Besides, by fine-tuning the mask head, FC-CLIP could incorporate the 3D structural priors into mask features and produce better results.

Long-Tail Datasets. As shown in Tab. 3, we validate the open-vocabulary methods on the more challenging long-tail 3D scene understanding datasets, *i.e.*, ***ScanNet200*** [45]. Following [45], we partition the 200 categories into three splits, *i.e.*, `head`, `common`, and `tail` sets, facilitating a more comprehensive comparison across categories with different frequencies. When training on `head` classes (ceiling, curtain, window, *etc.*), the fully-supervised method performs much better than zero-shot methods due to the sufficient 3D labels for supervision. However, on the `common` and `tail` splits, our DMA method approaches to or even surpasses the fully-supervised competitors. This is because there are only a few instances available on these long-tail categories, which is not sufficient to train a robust model from scratch. Our method does not rely on ground truth 3D labels but instead distill knowledge from pretrained vision-language models, thus it is more robust to rare objects.

To further validate the robustness of our method on rare objects/classes, we evaluate on the most frequent K classes of ***Matterport3D*** [6], where $K = 21, 40, 80, 160$. We train a 3D model by taking our generated textual descriptions

		mIoU				mACC			
	# of classes K	21	40	80	160	21	40	80	160
fully-sup.	TangentConv [49]	–	–	–	–	46.8	–	–	–
	TextureNet [21]	–	–	–	–	63.0	–	–	–
	ScanComplete [12]	–	–	–	–	44.9	–	–	–
	DCM-Net	–	–	–	–	66.2	–	–	–
	VM-Net [20]	–	–	–	–	**67.2**	–	–	–
	MinkowskiNet [20]	54.2	–	–	–	64.6	–	–	–
zero-shot.	OpenScene [42](LSeg)-3D	41.9	25.4	12.0	5.9	51.2	30.7	15.2	7.5
	OpenScene(LSeg)-2D3D	43.4	26.8	13.1	6.4	53.5	33.0	17.4	8.6
	OpenScene(OpenSeg)-3D	41.3	33.4	18.1	8.2	55.1	46.7	26.2	13.9
	OpenScene(OpenSeg)-2D3D	42.6	34.2	18.8	8.4	**59.2**	47.5	**27.1**	14.5
	DMA-text only	39.8	25.4	11.7	6.2	49.5	31.6	16.1	8.0
	DMA(OpenSeg)-3D	45.1	37.9	19.7	9.4	57.6	47.7	26.7	14.1
	DMA(FC-CLIP)-3D	**46.2**	**38.4**	**20.1**	**9.8**	58.4	**48.3**	26.5	**15.2**

Table 4: Comparison on the Matterport [6] test set.

and image features as supervision, and perform inference on different K categories. As shown in Tab. 4, when employing the same 2D network, *i.e.*, OpenSeg, our method demonstrates superior zero-shot segmentation capability on both common and rare categories. Specifically, our DMA(OpenSeg)-3D surpasses OpenScene(OpenSeg)-3D by 3.8%, 4.5%, 1.6%, and 1.2% in terms of mIoU at different K. This can be attributed to that OpenScene heavily relies on 2D model for supervision without aligning with text prompts, which limits its open-vocabulary ability. Our method, however, directly aligns with the textual modality, overcoming the limitations of 2D models.

Qualitative Comparison. Fig. 4 visualizes the segmentation results of different methods. We can observe that OpenScene [42] with only 3D encoder exhibits poor performance in segmenting objects that lack spatial structures, such as "door", "window", "counter", *etc*. In contrast, text supervision offers more refined guidance by establishing dense correspondences between the texts and points, thereby enabling more precise alignment. Our approach leverages the advantages of both language and 2D modalities, and achieves excellent segmentation results for both foreground and background classes using only the 3D model.

4.3 Ablation Study

2D Features *vs*. 3D Features. In Fig 5, we compare the segmentation performance on ScanNet by using different features. 'F' and 'B' denote foreground and background classes, respectively. For OpenScene [42], we observe that its 2D features are more advantageous for segmenting background categories with ambiguous geometry than 3D ones, *i.e.*, 49.1% *vs*. 47.1% mIoU(B), while 3D features excel at segmenting the foreground objects with distinct shapes, *i.e.*, 50.0% *vs*. 42.5% mIoU(F). Although the 2D-3D hybrid feature can leverage the strengths of both features simultaneously, utilizing 2D models for inference introduces significant computational overhead (please refer to the latency in Tab. 1). By additionally aligning with our generated text modality, our method can achieve outstanding performance on both foreground (58.3%) and background (51.5%) categories using only 3D features. Besides, DMA achieves comparable

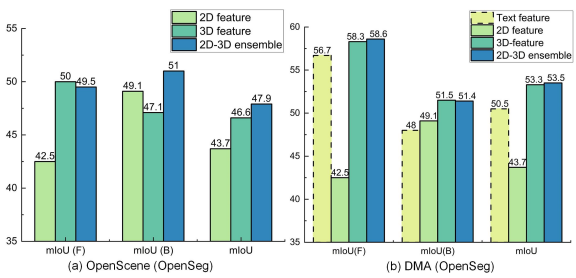

Fig. 5: Comparisons of text, 2D, and 3D features. "F" and "B" denote foreground and background classes.

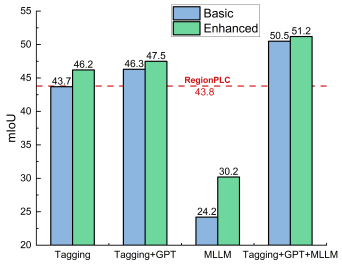

Fig. 6: Comparisons of tagging models and MLLMs.

Method	mIoU	mACC	mIoU (In)	mIoU (Out)
CLIP feature	35.2	51.3	36.7	31.8
Mask feature (w/o FT)	40.1	55.4	48.3	21.3
Mask feature (w/ FT)	42.0	57.4	50.5	24.1
CLIP+Mask	44.8	59.7	51.7	28.5

Table 5: Comparisons of CLIP and Mask features of FC-CLIP on ScanNet. "FT" denotes fine-tuning.

Method	2D Mask	3D Mask
w/o FT	36.6	40.1
Full Parameter	40.4	42.0
LoRA [18]	39.0	41.3
Adapter [17]	37.9	40.9

Table 6: Comparisons of different fine-tuning methods.

performance to using both 2D and 3D encoders by solely utilizing the 3D encoder, *i.e.*, 53.3% *vs*. 53.5% mIoU(F), and hence significantly reducing inference time.

Tagging Models *vs*. MLLMs. In Tab. 6, we compare the results of using different tagging models and MLLMs on ScanNet. For the enhanced version, we replace RAM with RAM++ [22], and LLaVA-7B with LLaVA-13B. We can observe that our method outperforms RegionPLC [55] by a large margin (about 6.7%) by building dense point-to-text correspondences. The tagging model plays a key role for performance improvement since it encompasses extensive semantic patterns, while MLLM further enhances the final performance by incorporating rich contextual information. By filtering out noisy tags with GPT, the performance can be improved by 2.6% and 1.3% for the basic and the enhanced versions, respectively. The final performance can be further improved when stronger tagging models/MLLMs are employed.

CLIP Features *vs*. Mask Features. In addition to OpenSeg [14], we employ FC-CLIP [56] to extract 2D features due to its effectiveness. As show in Tab. 5, we compare the performance by using CLIP and Mask features as supervision. 'In' and 'Out' denote in-vocabulary and out-vocabulary classes, respectively. FC-CLIP contains an in-vocabulary classifier and an out-vocabulary classifier, which correspond to the seen and unseen categories in the training process, respectively. As can be seen, the fixed CLIP feature is more advantageous in segmenting unseen categories, which outperforms mask feature by 10.5% in terms of mIoU(Out). This demonstrates that the fixed CLIP visual encoder could maintain the strong generalization ability on novel classes. While for in-vocabulary classes, mask features outperform the CLIP feature by 11.6%. By combining these features, we can simultaneously achieve competitive results on in- and out-vocabulary categories, attaining 44.8% mIoU over all classes.

Comparisons of Different Fine-Tuning Methods. We fine-tune the mask head of FC-CLIP with different strategies for incorporating 3D structural

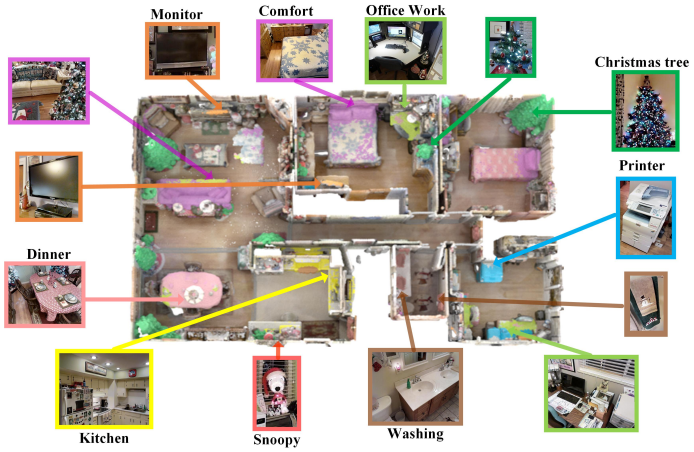

Fig. 7: Open-vocabulary segmentation results on rare categories and different forms of queries. The same color corresponds to the same query/category.

priors into mask features. As can be seen in Tab. 6, by fully fine-tuning the mask head, the performances of 2D and 3D masks are improved by 3.8% and 1.9%, respectively. LoRA [18] and Adapter [17] can also achieve obvious improvements by tuning a small amount of parameters.

Open-Vocabulary Segmentation for Different Text Queries. We finally investigate the ability of our method to segment rare categories. As shown in Fig. 7, our method can accurately segment the corresponding regions for the given texts/queries in 3D scenes, even for unseen categories. For instance, our well-trained model can quickly locate the position of new categories such as "Snoopy", or functional areas such as "kitchen", *etc*. On one hand, we align with 2D CLIP features that have been trained with a vast corpus of text. On the other hand, we construct a comprehensive and scalable textual modality by using VLMs, further enhancing the understanding ability.

5 Conclusion

We presented a dense multimodal alignment (DMA) framework for open-vocabulary 3D scene understanding by establishing dense correspondences between 3D points, 2D images and 1D texts, and leveraging their synergistic benefits to learn robust and generalizable 3D representations. To build a scalable language modality, we utilized powerful vision-language models to extract comprehensive scene descriptions and category information. Furthermore, we preserved the open-vocabulary recognition ability of the image modality by combining frozen CLIP features with trainable mask features. Extensive experiments demonstrate the promising performance of our method in open-vocabulary segmentation tasks across various indoor and outdoor scenarios.

Limitations. Our method relies on the quality of generated text descriptions and image features. In addition, collecting a larger 3D scene dataset is crucial for improving the generalization ability to unseen categories and variations.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1534–1543 (2016)
3. Cadena, C., Dick, A.R., Reid, I.D.: Multi-modal auto-encoders as joint estimators for robotics scene understanding. In: Robotics: Science and systems. vol. 5 (2016)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
5. Cao, Y., Yihan, Z., Xu, H., Xu, D.: Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. Advances in Neural Information Processing Systems **36** (2024)
6. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 2017 International Conference on 3D Vision (3DV). pp. 667–676. IEEE Computer Society (2017)
7. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
8. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: European conference on computer vision. pp. 202–221. Springer (2020)
9. Chen, L., Lei, C., Li, R., Li, S., Zhang, Z., Zhang, L.: Fpr: False positive rectification for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1108–1118 (2023)
10. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3075–3084 (2019)
11. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
12. Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., Nießner, M.: Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2018)
13. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Pla: Language-driven open-vocabulary 3d scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7010–7019 (2023)
14. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European Conference on Computer Vision. pp. 540–557. Springer (2022)
15. Han, L., Zheng, T., Xu, L., Fang, L.: Occuseg: Occupancy-aware 3d instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2940–2949 (2020)

16. Hou, J., Graham, B., Nießner, M., Xie, S.: Exploring data-efficient 3d scene understanding with contrastive scene contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15587–15597 (2021)
17. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
19. Hu, W., Zhao, H., Jiang, L., Jia, J., Wong, T.T.: Bidirectional projection network for cross dimension scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14373–14382 (2021)
20. Hu, Z., Bai, X., Shang, J., Zhang, R., Dong, J., Wang, X., Sun, G., Fu, H., Tai, C.L.: Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15488–15498 (2021)
21. Huang, J., Zhang, H., Yi, L., Funkhouser, T., Nießner, M., Guibas, L.J.: Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4440–4449 (2019)
22. Huang, X., Huang, Y.J., Zhang, Y., Tian, W., Feng, R., Zhang, Y., Xie, Y., Li, Y., Zhang, L.: Open-set image tagging with multi-grained text supervision. arXiv e-prints pp. arXiv–2310 (2023)
23. Huang, X., Zhang, Y., Ma, J., Tian, W., Feng, R., Zhang, Y., Li, Y., Guo, Y., Zhang, L.: Tag2text: Guiding vision-language model via image tagging. arXiv preprint arXiv:2303.05657 (2023)
24. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
27. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=RriDjddCLN
28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
29. Li, R., He, C., Li, S., Zhang, Y., Zhang, L.: Dynamask: dynamic mask selection for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11279–11288 (2023)
30. Li, R., He, C., Zhang, Y., Li, S., Chen, L., Zhang, L.: Sim: Semantic-aware instance mask generation for box-supervised instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7193–7203 (2023)

31. Li, R., Li, S., He, C., Zhang, Y., Jia, X., Zhang, L.: Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11593–11603 (2022)
32. Li, Y., Yu, A.W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q.V., et al.: Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17182–17191 (2022)
33. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023)
34. Liong, V.E., Nguyen, T.N.T., Widjaja, S., Sharma, D., Chong, Z.J.: Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. arXiv preprint arXiv:2012.04934 (2020)
35. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
36. Liu, K., Zhan, F., Zhang, J., Xu, M., Yu, Y., Saddik, A.E., Theobalt, C., Xing, E., Lu, S.: Weakly supervised 3d open-vocabulary segmentation. arXiv preprint arXiv:2305.14093 (2023)
37. Liu, M., Zhou, Y., Qi, C.R., Gong, B., Su, H., Anguelov, D.: Less: Label-efficient semantic segmentation for lidar point clouds. In: European Conference on Computer Vision. pp. 70–89. Springer (2022)
38. Lu, Y., Xu, C., Wei, X., Xie, X., Tomizuka, M., Keutzer, K., Zhang, S.: Open-vocabulary point-cloud object detection without 3d annotation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1190–1199 (2023)
39. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: Rangenet++: Fast and accurate lidar semantic segmentation. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 4213–4220. IEEE (2019)
40. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2917 (2021)
41. Nekrasov, A., Schult, J., Litany, O., Leibe, B., Engelmann, F.: Mix3d: Out-of-context data augmentation for 3d scenes. In: 2021 International Conference on 3D Vision (3DV). pp. 116–125. IEEE (2021)
42. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 815–824 (2023)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
45. Rozenberszki, D., Litany, O., Dai, A.: Language-grounded indoor 3d semantic segmentation in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)

46. Schult, J., Engelmann, F., Kontogianni, T., Leibe, B.: Dualconvmesh-net: Joint geodesic and euclidean convolutions on 3d meshes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8612–8622 (2020)
47. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. arXiv preprint arXiv:2306.13631 (2023)
48. Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S.: Searching efficient 3d architectures with sparse point-voxel convolution. In: European conference on computer vision. pp. 685–702. Springer (2020)
49. Tatarchenko, M., Park, J., Koltun, V., Zhou, Q.Y.: Tangent convolutions for dense prediction in 3d. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3887–3896 (2018)
50. Vu, T., Kim, K., Luu, T.M., Nguyen, T., Yoo, C.D.: Softgroup for 3d instance segmentation on point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2708–2717 (2022)
51. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
52. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
53. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1179–1189 (2023)
54. Xue, L., Yu, N., Zhang, S., Li, J., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip-2: Towards scalable multimodal pre-training for 3d understanding. arXiv preprint arXiv:2305.08275 (2023)
55. Yang, J., Ding, R., Wang, Z., Qi, X.: Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. arXiv preprint arXiv:2304.00962 (2023)
56. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In: NeurIPS (2023)
57. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al.: Recognize anything: A strong image tagging model. arXiv preprint arXiv:2306.03514 (2023)
58. Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D.: Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9939–9948 (2021)