Open-Vocabulary Octree-Graph for 3D Scene Understanding

Zhigang Wang[1,2*], Yifei Su[3,4,2*], Chenhui Li[2*],
Dong Wang[2], Yan Huang[3,4], Xuelong Li[5], Bin Zhao[1,2†]
[1]Northwestern Polytechnical University, [2]Shanghai AI Laboratory,
[3]University of Chinese Academy of Sciences, [4]CASIA, [5]TeleAI

Abstract

Open-vocabulary 3D scene understanding is indispensable for embodied agents. Recent works leverage pre-trained vision-language models (VLMs) for object segmentation and project them to point clouds to build 3D maps. Despite progress, a point cloud is a set of unordered coordinates that requires substantial storage space and does not directly convey occupancy information or spatial relation, making existing methods inefficient for downstream tasks, e.g., path planning and text-based object retrieval. To address these issues, we propose ***Octree-Graph****, a novel scene representation for open-vocabulary 3D scene understanding. Specifically, a Chronological Group-wise Segment Merging (CGSM) strategy and an Instance Feature Aggregation (IFA) algorithm are first designed to get 3D instances and corresponding semantic features. Subsequently, an adaptive-octree structure is developed that stores semantics and depicts the occupancy of an object adjustably according to its shape. Finally, the Octree-Graph is constructed where each adaptive-octree acts as a graph node, and edges describe the spatial relations among nodes. Extensive experiments on various tasks are conducted on several widely-used datasets, demonstrating the versatility and effectiveness of our method. Code is available here.*

1. Introduction

3D scene understanding is receiving increasing attention due to its widespread usage in robots [55] and VR/AR applications [17]. Previous works [19, 21, 34, 38, 42] trained models on particular 3D scene datasets to complete this task. Although significant progress has been achieved, they are limited to a closed-set category. Recently, we have witnessed the impressive generalization ability of foundation models (*e.g.*, SAM [18] and CLIP [33]) which can perceive various objects in unseen scenarios, inspiring a lot of open-vocabulary 3D scene understanding methods [4, 5, 7, 16, 25, 44, 49]. Given an RGB-D sequence with camera poses, mainstream methods leverage the off-the-shelf foundation models to generate 2D object masks and corresponding visual-language features, and then project them to point clouds to construct a semantic 3D map.

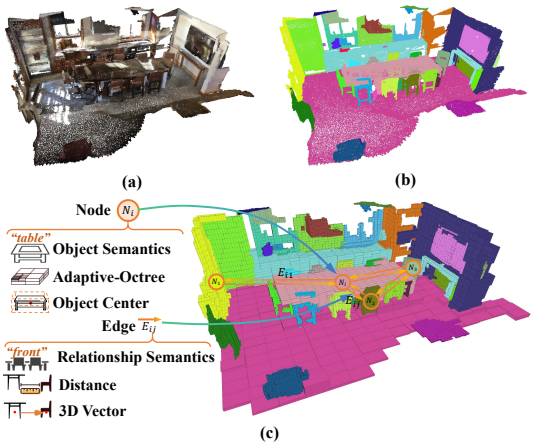

Figure 1. (a) A 3D scene. (b) The corresponding semantic 3D map based on point clouds (6.8M). (c) Our Octree-Graph where each object is represented by the proposed adaptive-octree and each edge contains rich spatial relations among objects. All adaptive-octrees occupy 42KB of storage space in total.

Despite the favorable open-vocabulary understanding capability, they have two drawbacks. 1) ***Inefficient space representation of 3D scenes***. Most mainstream methods [15, 31, 48] build the 3D map based on point clouds, as shown in Fig. 1 (b). Point clouds are unordered discrete coordinates that require considerable storage space, making existing methods inefficient to deploy on embodied agents with limited storage resources. Moreover, point clouds lack explicit representation of occupancy information and spatial connectivity which are critical for downstream tasks, *e.g.*, path planning and text-based object retrieval. 2) ***Inaccurate semantic object segmentation for 3D map construction***. Most methods overlook the inaccuracy of foundation models/vision-language models (VLMs) when conducting

*Equal contribution.
†Corresponding author.

object segmentation and feature extraction, inevitably causing imprecise 3D object segments and degraded semantics.

To alleviate these problems, we propose **Octree-Graph** as shown in Fig. 1 (c), a novel open-vocabulary scene representation designed to characterize the occupancy and semantics of each object, as well as the relations among them. Specifically, the adaptive-octree is first proposed to depict each object's occupancy, which inherits the advantages of the octree structure by hierarchically representing a 3D space with structured sub-regions. Compared to the point cloud without regional or hierarchical information, it can save significant storage space. Furthermore, our adaptive-octree initializes each object adaptively according to its shape, enabling a precise description of the occupancy within a limited octree depth. This is particularly suitable for objects with large aspect ratios, *e.g.*, walls and floors. Based on this, the Octree-Graph is constructed where each adaptive-octree acts as a graph node, and each edge encompasses rich relations among objects, *e.g.*, distances and relative orientations. The proposed Octree-Graph can be directly applied to downstream tasks such as object retrieval, occupancy queries, and path planning, thus providing significant convenience.

To obtain accurate semantic objects for Octree-Graph construction, we devise a **training-free** pipeline. First, given input images, 2D proposals are segmented via an off-the-shelf segmenter, and corresponding visual-language features are extracted by pretrained VLMs. Then, they are projected into 3D space as point cloud segments. Second, to correctly merge segments belonging to the same instance, a Chronological Group-wise Segment Merging (CGSM) strategy is proposed, where the segments are partitioned into several groups in time order. Each group is individually processed to leverage spatiotemporal details from the neighborhood while avoiding interference from global redundancy. Third, an Instance Feature Aggregation (IFA) method is proposed to obtain semantic representations for each object. Unlike existing works that directly average features as a result, we simultaneously consider the representativeness and distinctiveness of a feature during the fusion process. Our contributions are summarized as follows.

- We propose the Octree-Graph for open-vocabulary 3D scene understanding, which efficiently depicts objects' occupancies, semantics, and relations, benefiting several downstream tasks.
- We propose a Chronological Group-wise Segment Merging (CGSM) strategy and an Instance Feature Aggregation (IFA) method to obtain accurate semantic objects.
- We conduct extensive experiments, demonstrating the versatility, effectiveness, and efficiency of our method.

2. Related Work

Foundation Models. Recently, foundation models have exhibited impressive zero-shot perception ability. Here, we review several foundation models related to our work. CLIP [33] is a popular vision-language model that associates images and texts through contrastive learning, significantly promoting many vision-language tasks. SAM [18] is a class-agnostic 2D segmentation model trained with over 1 billion masks, demonstrating powerful zero-shot performance. OVSeg [22] finetunes CLIP to gain the ability of open-vocabulary semantic segmentation. CropFormer [32] fuses the full image and high-resolution image crops to improve segmentation performance. TAP [30] can simultaneously conduct recognition, segmentation, and caption generation. Additionally, many other methods [6, 8, 10, 14, 20, 24, 28, 35, 43, 45] are proposed for 2D open-vocabulary object detection and segmentation.

Open-Vocabulary 3D Scene Understanding. Based on the organization form of scene representation, we categorize these works into four types. 1) ***NeRF/Gaussian 3D mapping***. These methods perform 3D scene understanding and scene/object reconstruction simultaneously, *e.g.*, OpenObj [3]. Although achieving good performance, they need extra effort to train the NeRF or 3D Gaussian models. 2) ***point/grid-wise 3D mapping***. This branch involves directly projecting semantic features to each 3D point. OpenScene [31] and ConceptFusion [15] extract CLIP features from the images and densely project them to the point cloud. VLMaps [12] adopts a similar pipeline to project visual-language features to a grid-based BEV map. 3) ***instance-wise 3D mapping***. These works explicitly obtain each 3D instance and fuse its visual-language features for 3D mapping. OpenIns3D [13] is a 3D-input-only framework that gets objects by open-vocabulary 3D detection. OVIR-3D [25], SAM3D [51], and MaskClustering [48] follow a 2D-to-3D pipeline where 2D masks are projected into 3D space for instance merging based on semantic similarity and 3D overlap. SAI3D [53] uses both 2D proposals and 3D super points for instance segmentation. OpenMask3D [40], Open3DIS [29], and SA3DIP [50] leverage extra 3D instance detectors to get more accurate object proposals. However, the used 3D models cannot be considered purely zero-shot methods. 4) ***3D scene graph/octree***. A few works use a graph or octree to organize the scene. ConceptGraph [7] and Clio [27] cluster object segments and construct a scene graph to enhance spatial reasoning. HOV-SG [44] proposes a hierarchical 3D scene graph to enable scene representation of different granularities. OctreeOcc [26] and PlenOctrees [54] use the octree structure to store semantic class and rendering information, respectively. In contrast, our Octree-Graph represents each object using an adaptive-octree and models their relations using a graph, supporting efficient occupancy queries and spatial reasoning.

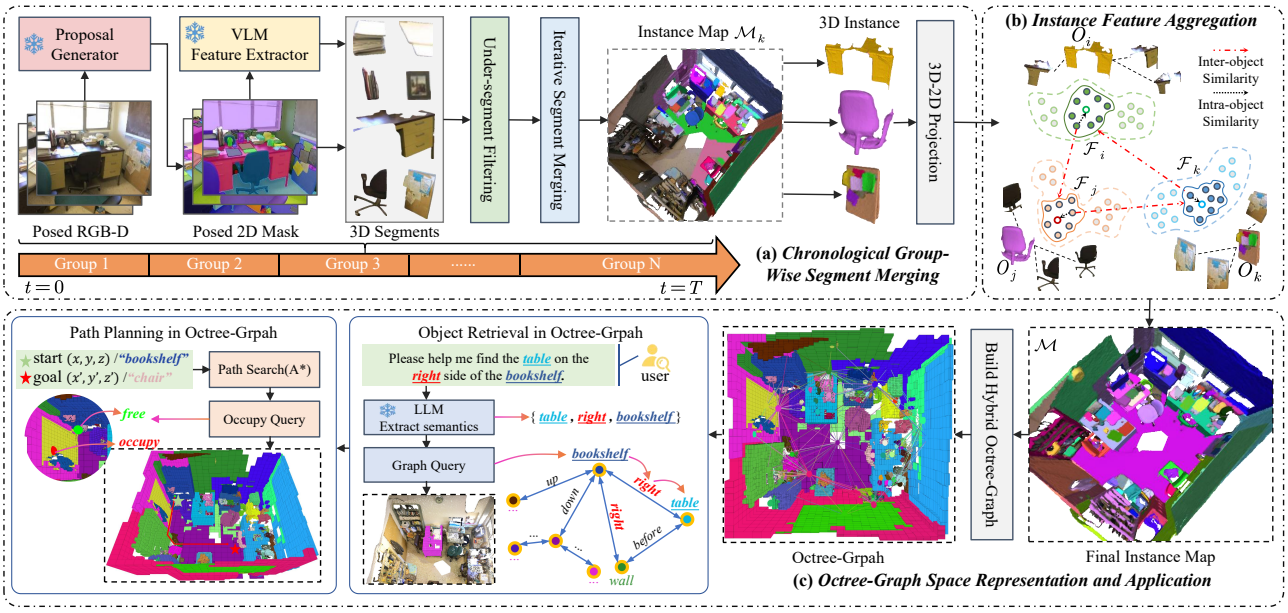

Figure 2. Overview of our Octree-Graph. (a) Chronological Group-wise Segment Merging (CGSM). Given posed RGB-D inputs, 2D masks with semantic features are first extracted and then projected into the 3D space, where CGSM is conducted to merge segments. (b) Instance Feature Aggregation (IFA). Feature aggregation is performed for each merged object, which considers both intra- and inter-object similarity. (c) The Octree-Graph is constructed to efficiently and accurately represent the scene, facilitating various downstream tasks.

3. Method

3.1. Framework Overview

As shown in Fig 2, given a sequence of RGB images $\mathcal{I}_c = \{\mathbf{I}_t^c\}_{t=1}^T$ and depth images $\mathcal{I}_d = \{\mathbf{I}_t^d\}_{t=1}^T$ scanned in a scene, we first leverage VLMs to extract segment proposals (§ 3.2). Next, we chronologically merge these segments into an instance map $\mathcal{M}$ via a group-wise merging strategy (§ 3.3). Then we dynamically aggregate the redundant semantics of each instance into a distinctive feature (§ 3.4). Finally, we build an Octree-Graph G to represent spatial relations among instances, with the adaptive-octree to detail instance occupancy. Based on this, we implemented LLM-based object retrieval and path planning algorithms (§ 3.5).

3.2. Segment Proposal and Comprehension

For each frame $\mathbf{I}_t^c$ at time t, we first adopt an off-the-shelf proposal generator, *e.g.*, CropFormer [32], to extract a set of 2D masks $\mathcal{P}_t^{2d} = \{\mathbf{m}_i\}_{i=1}^{n_t}$, where n_t is the mask number. We then filter out tiny and marginal masks to ensure the proposal quality. Next, each $\mathbf{m}_i$ is fed into the visual encoder and caption generator to obtain the visual feature $\mathbf{f}_i^v$ and caption feature $\mathbf{f}_i^c$. Finally, we project each mask $\mathbf{m}_i$ into the 3D space as a point cloud segment and perform DBSCAN [37] denoise, obtaining segments $\mathcal{P}_t^{3d} = \{\mathcal{S}_i\}_{i=1}^{n_t}$.

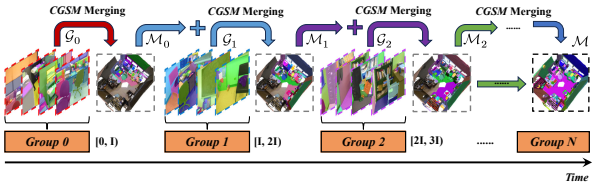

Figure 3. Illustration of group split and CGSM merging.

3.3. Chronological Group-wise Segment Merging

Existing segment merging strategies are typically categorized into two types: 1) *frame-wise*, which sequentially or hierarchically merges the adjacent frames [7, 25], integrating similar segments efficiently. 2) *graph-wise*, which merges segments across all frames [44, 48]. These methods have achieved great success, while the former solely relying on a single frame can be easily affected by proposal noises, *e.g.*, associating unrelated instances once an under-segment is merged. The latter, which processes all segments together, may introduce redundant computations and be affected by irrelevant segments. To this end, we propose a Chronological Group-wise Segment Merging (CGSM) strategy with semantic-guided under-segment filtering and a dynamic threshold decay strategy.

Chronological Group-Wise Split. Given the prior that *an instance often appears in multiple consecutive frames*, as shown in Fig. 3, CGSM first partitions all frames into several groups in time order with interval I, obtaining the set of segments $\mathcal{G}_i$ for each group. In this way, a group can retain

the adjacent segments while avoiding interference caused by global segments. Based on these groups, we perform iterations of merging to integrate separate segments into an instance map $\mathcal{M}$. Concretely, we start by merging $\mathcal{G}_0$ into an intermediate instance map $\mathcal{M}_0$. Subsequently, we iteratively take the union $\{\mathcal{M}_{k-1}, \mathcal{G}_k\}$ as input for the kth merging, until the final instance map $\mathcal{M}$ is constructed. Next, we elaborate on the details of a single merging step.

Segment Group Merging. For two segments $\{\mathcal{S}_m, \mathcal{S}_n\}$, we define $\phi^v_{\text{sem}}(m,n)$ as the cosine similarity between their visual features, and $\phi^c_{\text{sem}}(m,n)$ the cosine similarity between caption features. Regarding geometric similarity, we compute $\phi^{\text{iou}}_{\text{geo}}(m,n)$ as the intersection over union of two segments. Additionally, we calculate the ratio of $\mathcal{S}_n$ contained within $\mathcal{S}_m$ as $\phi^{\text{ior}}_{\text{geo}}(m,n) = |\mathcal{S}_m \cap \mathcal{S}_n| \,/\, |\mathcal{S}_n|$. $|\cdot|$ denotes the amount of points in a 3D segment. Intuitively, assuming $\mathcal{S}_m$ is an under-segment containing a correct segment $\mathcal{S}_n$, $\phi^{\text{ior}}_{\text{geo}}(m,n)$ will be relatively large. Based on this, we can collect all segments contained in $\mathcal{S}_m$ as $\{\mathcal{S}_j \mid \phi^{\text{ior}}_{\text{geo}}(m,j) \geq 0.8\}$. If the semantic feature variance of these contained segments exceeds a threshold τ_u, it indicates that $\mathcal{S}_m$ is probably an under-segment containing different objects, and $\mathcal{S}_m$ will be filtered out. We term this process as semantic-guided under-segment filtering. To merge the left segments, we compute an overall similarity $\phi = \phi^{\text{iou}}_{\text{geo}} + \phi^{\text{ior}}_{\text{geo}} + \phi^v_{\text{sem}} + \phi^c_{\text{sem}}$, and iteratively merge segments within group $\mathcal{G}_i$. At each iteration, we merge highly similar segments satisfying $\phi(m,n) \geq \theta_i$. However, simply doing so struggles to merge partially observed segments or over-segments sharing low spatial similarity. To this end, we linearly decay θ_i at each step inspired by [44, 53].

3.4. Instance Feature Aggregation

After obtaining the instance map $\mathcal{M}$, each 3D instance $\mathcal{O}_i$ in $\mathcal{M}$ is associated with multiple segment features $\mathcal{F}_i = \{\mathbf{f}^v_{i,j} \mid \mathbf{f}^v_{i,j} \in \mathcal{O}_i\}$ based on 2D-3D relations (for simplicity, we omit caption features here). To aggregate these features, previous methods either perform averaging [7, 25] or select the dominant feature via clustering [40, 44]. However, they overlook the distinction between different instance features. Hence, we propose a weighted average method to fuse an instance's features for an optimal feature both representative and distinctive, as shown in Fig. 2 (b). Specifically, taking the visual modality for illustration, we average $\mathcal{F}_i$ to a central feature $\bar{\mathbf{f}}^v_i$ for each instance, and the neighboring instances of $\mathcal{O}_i$ are then defined by $\mathcal{N}_i = \{\mathcal{O}_k \mid \cos\left(\bar{\mathbf{f}}^v_i, \bar{\mathbf{f}}^v_k\right) \geq \tau_d\}$. Based on this, we aggregate the features $\mathcal{F}_i$ into an optimal $\mathbf{f}^{v*}_i$ via assigning a dynamic fusion weight $a^v_{i,j}$ to each $\mathbf{f}^v_{i,j}$ in $\mathcal{F}_i$:

$$a^v_{i,j} = \cos\left(\mathbf{f}^v_{i,j}, \bar{\mathbf{f}}^v_i\right) - \sum_{O_k \in \mathcal{N}_i} \cos\left(\mathbf{f}^v_{i,j}, \bar{\mathbf{f}}^v_k\right), \qquad (1)$$

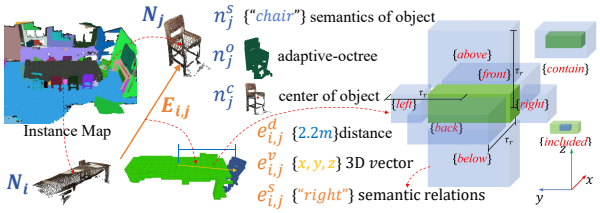

Figure 4. Illustration of the nodes and edges in Octree-Graph.

where $\cos(\cdot)$ denotes cosine similarity, and $a^v_{i,j}$ is normalized via softmax. Intuitively, a feature gets a larger weight if it is closer to its own cluster center and farther from neighboring instances. The caption feature $\mathbf{f}^{c*}_i$ can be formulated by replacing $\mathbf{f}^v_{i,j}$ with $\mathbf{f}^c_{i,j}$ in the above process. The final instance feature $\mathbf{f}^*_i$ is the average of $\mathbf{f}^{v*}_i$ and $\mathbf{f}^{c*}_i$.

3.5. Octree-Graph Construction and Applications

To efficiently and accurately represent a scene, we design a hybrid structure, termed Octree-Graph. This structure utilizes a graph as the high-level architecture to organize objects and their spatial relations. Furthermore, we propose an adaptive-octree to depict the occupancy information of each object, which acts as a node of the Octree-Graph.

Graph Construction. An Octree-Graph can be defined as G with nodes $\mathbf{N}_*$ and edges $\mathbf{E}_*$. The node $\mathbf{N}_i$ consists of correlated semantics n^s_i (*e.g.*, captions and features), center n^c_i, and adaptive-octree n^o_i. While the edge $\mathbf{E}_{i,j}$ comprises the semantic relation $e^s_{i,j}$, spatial distance $e^d_{i,j}$ and the 3D vector $\mathbf{e}^v_{i,j}$ between node i and node j. Notably, the semantic relations between nodes are aligned with the world coordinate system of the corresponding point cloud. As shown in Fig. 4, the semantic relation $e^s_{i,j}$ between node $\mathbf{N}_i$ and $\mathbf{N}_j$ is characterized as "right".

Adaptive-Octree Construction. The classical octree [11] is a tree-based structure capable of efficiently representing a 3D space with much less storage requirements than point clouds. During octree construction, the root node is defined by the minimal bounding box containing the point cloud. This box, centered at $c \in \mathbb{R}^3$ with a side length of d, is divided into eight sub-regions of side length $d/2$ using axis-aligned planes. Each sub-region serves as a child node, and the process continues recursively for each node until the desired octree depth $L_{\max}$ is reached or no point clouds are present within the node. We recommend referring to [11] for more details about the octree.

However, the traditional octree is proposed to represent an entire 3D space, which uses cubic voxels as units to depict occupancy details. This leads to dilemmas of redundant representation when depicting an object, *e.g.*, an object with a large aspect ratio requires a very deep octree to approximate its shape. To this end, we propose the adaptive-octree with varying voxels that adaptively adjust their sizes and shapes according to the object's shape. As shown in Fig.

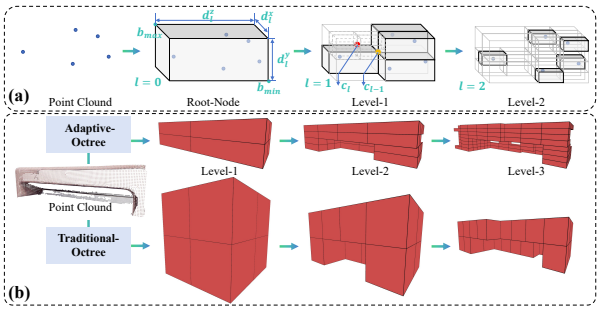

Figure 5. Illustration of the construction of the adaptive-octree. The above displays the process, and the below shows an example.

5, an adaptive-octree is constructed from an instance-level point cloud P. The size of each node in this adaptive-octree can be computed as follows:

$$\mathbf{d}_l = (\mathbf{b}_{\max} - \mathbf{b}_{\min}) / 2^l, \tag{2}$$

where $\mathbf{b}_{\max}$ and $\mathbf{b}_{\min}$ are the coordinates of the lower left corner and the upper right corner of P's bounding box. $l \in \{1, 2, \cdots, L_{\max}\}$ denotes the depth of the adaptive-octree. As shown in Fig. 5, the center $\mathbf{c}_l \in \mathbb{R}^3$ of the l-th layer node can be determined by the center $\mathbf{c}_{l-1}$ of the parent node and the edge length $\mathbf{d}_l$ of the current node. The adaptive-octree can be quickly constructed from the point cloud.

Octree-Graph Applications. Based on the Octree-Graph, we offer object retrieval and path planning functionalities which are critical for embodied agents. For object retrieval, two types of queries are supported, *i.e.,* Query $(target)$ and Query $(reference, relation, target)$. The former allows for directly locating an object by comparing the similarity between queries and stored semantics. The latter supports complex queries by sequentially locating the reference object, the edge that matches the described relation, and finally the target. For more complex queries, we leverage the reasoning capabilities of LLMs to decompose the task and flexibly call two types of functions to achieve the goal.

In path planning tasks, querying occupancy information is fundamental. The proposed Octree-Graph supports such queries, enabling us to easily implement path planning algorithms like classical A^* [9] and the recent [52].

4. Experiment

To validate the versatility and effectiveness of our method, we carry out extensive experiments, including semantic segmentation, instance segmentation, text-based object retrieval, and path planning. We compare our method with different SOTA methods in these tasks, and conduct comprehensive ablation studies to investigate several key components, demonstrating the effectiveness of our designs.

4.1. Implementation Details

We use CropFormer [32] as the 2D proposal generator following [48]. To extract visual features, we test two commonly used VLMs, *i.e.,* CLIP ViT-H [33] and OVSeg ViT-L [22]. We adopt TAP [30] for generating the mask caption. Additionally, we filtered out masks with pixels less than 25 and segments with points less than 50. The similarity threshold τ_d is empirically set to 0.7. The group split interval I, the under-segment filtering threshold τ_u, and the decay parameter θ_i are set to 200, 0.02, and 0.8 through hyper-parameter experiments. Considering the dimensions of indoor objects, we set the maximum depth $L_{\max}$ of the adaptive-octree to 4.

4.2. Dataset and Evaluation Metrics

Dataset. For zero-shot 3D semantic segmentation, we evaluate our method on common scenes following [7, 15, 44], *i.e.*, 8 scenes from Replica [39] dataset and 5 scenes from ScanNet [2]. For zero-shot 3D instance segmentation, we assess our method on the widely-used ScanNet200 [36] benchmark, including a validation set of 312 scenes with 200 categories. For text-based object retrieval, we test our method on Sr3D [1] dataset, and follow the experiment setting of BBQ [23] that subsampled 526 free-form queries from 8 scenes. For the path planning task, we employ the HM3DSem [46] dataset used in HOV-SG [44], where 8 scenes are selected for evaluation. Moreover, we also conduct real-world experiments to validate our effectiveness.

Evaluation Metrics. Following the mainstream evaluation metrics [44], we assess 3D semantic segmentation results via commonly used mean IoU (mIoU), frequency-weighted mean IoU (F-mIoU), and mean Accuracy (mAcc). For 3D instance segmentation, we report the standard Average Precision (AP) at IoU thresholds 25% and 50%, along with the mean of AP from 50% to 95% at 5% interval. For text-based object retrieval, we follow BBQ [23], using Acc@0.1 and Acc@0.25 as evaluation metrics where retrieval is treated as a true positive if the IoU between the predicted object's bounding box and the ground-truth bounding box surpasses 0.1 and 0.25, respectively. For the path planning task, we randomly select positions in the empty areas of a scene as the starting point and destination. When the endpoint of navigation is within a threshold s (*i.e.*, 1m, 0.5m, and 0.25m) from the destination, the path planning is considered successful.

Besides, to quantify spatial representation accuracy, we introduce the Effective Occupancy Ratio (EOR) as a metric. The occupancy range O_{pc} of a point cloud is determined by expanding this point cloud with a dilation $\triangle r = 0.005$, and the occupancy range of the octree is denoted as O_{oct}. Then, the EOR is calculated as $\text{EOR} = \frac{O_{\text{oct}} \cap O_{\text{pc}}}{O_{\text{oct}}}$. We denote the mean EOR for all objects in a scene as mEOR.

Method	CLIP Backbone	Replica			ScanNet		
		mIoU↑	F-mIoU↑	mAcc↑	mIoU↑	F-mIoU↑	mAcc↑
ConceptFusion [15]	OVSeg	0.10	0.21	0.16	0.08	0.11	0.15
	Vit-H-14	0.10	0.18	0.17	0.11	0.12	0.21
ConceptGraph [7]	OVSeg	0.13	0.27	0.21	0.15	0.18	0.23
	Vit-H-14	0.18	0.23	0.30	0.16	0.20	0.28
HOV-SG [44]	OVSeg	0.144	0.255	0.212	0.214	0.258	0.420
	Vit-H-14	0.231	0.386	0.304	0.222	0.303	0.431
Ours	OVSeg	**0.320**	**0.553**	**0.414**	**0.393**	**0.508**	**0.601**
	Vit-H-14	0.263	0.479	0.387	0.356	0.477	0.574

Table 1. Zero-shot 3D semantic segmentation results on Replica and ScanNet benchmark.

Method	AP↑	AP50↑	AP25↑
sup. mask + sup. semantic			
Mask3D [38]	26.9	36.2	41.4
sup. mask + z.s. semantic			
Open3DIS [29]	**23.7**	29.4	32.8
Open3DIS [29] (3D only)	18.6	23.1	27.3
OpenMask3D [40] (Mask3D)	15.4	19.9	23.1
Ours (Mask3D)	23.2	**30.3**	**33.3**
z.s. mask + z.s. semantic			
OVIR-3D [25]	9.3	18.7	25.0
SAM3D [51]	9.8	15.2	20.7
SAI3D [53]	12.7	18.8	24.1
Mask-Clustering [48]	12.0	23.3	30.1
Ours	**14.3**	**25.8**	**33.6**

Table 2. 3D instance segmentation results on ScanNet200. $sup.$ means supervised training, $z.s.$ denotes the zero-shot setting.

Method	Retrieval Algorithm	Acc↑ @0.1	Acc↑ @0.25
ConceptGraphs [7]	*Deductive* [23]	0.15	0.08
Open-Fusion [47]	CLIP [33]	0.13	0.02
BBQ [23]	CLIP [33]	0.10	0.06
BBQ [23]	*Deductive* [23]	0.23	0.18
Ours	Octree-Graph+LLM	**0.26**	**0.23**

Table 3. Text-based object retrieval results on the Sr3D dataset.

4.3. Quantitative Comparison

3D Semantic Segmentation. Tab. 1 reports the numerical results for zero-shot 3D semantic segmentation on Replica and ScanNet datasets. In this experiment, we compare the results generated by our CGSM and IFA with other works. It can be seen that our method significantly outperforms existing methods across all metrics on both datasets, demonstrating the effectiveness of the proposed CGSM and IFA. Compared to the existing SoTA 3D scene graph, HOV-SG [44], we achieve **+8.9%** mIoU and **+11.0%** mAcc on the Replica dataset. Similarly, we present a **+17.1%** mIoU and a **+17.0%** mAcc on ScanNet with the same settings.

Method	SR(s=1.0m)	SR(s=0.5m)	SR(s=0.25m)
HOV-SG [44]	55.25	46.75	32.16
Ours	**97.88**	**96.88**	**96.38**

Table 4. Path planning results on HM3DSem. SR denotes success rate (%). s is the threshold within which the distance between the navigation endpoint and the destination is considered successful.

3D Instance Segmentation. The quantitative results for 3D instance segmentation are shown in Tab. 2. We follow [48] to categorize all methods into 3 groups based on whether the proposal generation and semantic prediction are trained. Under the fully zero-shot setting, our method surpasses the previous most advanced method with gains of **2.3%**, **2.5%**, and **3.5%** in AP, AP25 and AP50, respectively. These results further demonstrate the effectiveness of the proposed CGSM and IFA. When using supervised 3D models for proposal generation, our method significantly outperforms OpenMask3D [40] and the Open3DIS [29] variant with only the 3D proposals, validating the superiority of our feature aggregation method IFA. Besides, our method achieves comparable results with the corresponding SOTA method Open3DIS [29], which specially designs a combination of 2D and 3D proposals.

Text-based Object Retrieval. Tab. 3 presents the comparison results of text-based object retrieval on Sr3D [1] dataset. Our method outperforms the SOTA method BBQ [23] by **3.0%** and **5.0%** in terms of Acc@0.1 and Acc@0.25, respectively. We attribute the performance gain to our accurate semantic object segmentation and the rich relations stored in the Octree-Graph.

Merging Strategy	mIoU↑	F-mIoU↑	mAcc↑
Frame-wise	0.323	0.439	0.519
Global-wise	0.286	0.414	0.476
Ours (CGSM) I=400	0.338	0.463	0.555
Ours (CGSM) I=200	**0.356**	**0.477**	**0.574**
Ours (CGSM) I=100	0.344	0.462	0.571

Table 5. Ablation study on the segment merging strategy and different temporal intervals for group partitioning of our CGSM.

under-segment filtering	threshold decay	mIoU↑	F-IoU↑	mAcc↑
✗	✗	0.337	0.460	0.547
✓	✗	0.346	0.471	0.557
✓	✓	**0.356**	**0.477**	**0.574**

Table 6. Ablations study on the strategies for segment merging.

Aggregation Strategy	mIoU↑	F-mIoU↑	mAcc↑
Average	0.338	0.453	0.563
Top-5	0.342	0.470	0.570
DBSCAN	0.345	0.459	0.566
Ours (IFA)	**0.356**	**0.477**	**0.574**

Table 7. Ablation study on various feature aggregation strategies.

	Replica		ScanNet	
Method	Storage↓	mEOR↑	Storage↓	mEOR↑
Point cloud	18.5MB		6.4MB	
Octree	17.6KB	0.0057	41.1KB	0.0041
Adaptive-Octree	29.8KB	**0.0108**	69.3KB	**0.0070**

Table 8. Ablation study on the efficiency of the adaptive-octree.

Path Planning. For each sense in the HM3DSem [46] dataset, we randomly select 100 pairs of starting points and destinations in navigable areas. HOV-SG [44] can be directly used for path planning, thus it is evaluated and compared with our method in this task. Tab. 4 shows the results, from which we can see that our method significantly surpasses HOV-SG, especially when the threshold s is small. This is because HOV-SG relies on Voronoi graph [41] for path planning, where the waypoints and paths are pre-calculated, making it improper for precise navigation. In contrast, our Octree-Graph supports navigation to any empty area, unless the destination is mistakenly occupied by the adaptive-octree.

Method	Structure	Storage↓	time↓
A*	Octree-Graph	268.41Kb	0.032s
A*	Point Cloud	71.16Mb	2.154s
Jump Point Search	Octree-Graph	268.41Kb	0.081s
Jump Point Search	Point Cloud	71.16Mb	2.153s

Table 9. Ablation study on path planning efficiency.

4.4. Ablation Studies

We analyze the impact of our key designs via zero-shot semantic segmentation experiments on ScanNet.

Effect of Group-Wise Split. We compare the proposed Chronological Group-wise Segment Merging (CGSM) with the vanilla frame-wise and global-wise merging strategies. As shown in Tab. 5, we achieve notable gains compared to both methods, with **+3.3%** mIoU over frame-wise merging and **+7.0%** mIoU over global-wise merging. We also analyze the impact of hyper-parameter I, and the results in Rows 3-5 show that our method exhibits robustness to I ranging from 100 to 400.

Analysis of Designs on Segment Merging. Tab. 6 presents the results with two key components for merging a single group. Row 0 serves as a fixed-threshold group-wise merging baseline with no extra design. Row 1 is equipped with our semantic-guided under-segment filtering, and achieves **+0.9**% mIoU and **+1.0**% mAcc. Row 2 further incorporates the threshold decay strategy, resulting in an additional **1.0**% mIoU and **1.7**% mAcc gains. These validate the effectiveness of the two strategies during each group-wise merging.

Effect of Instance Feature Aggregation. In Tab. 7, we compare the proposed Instance Feature Aggregation (IFA) method with several commonly used methods. For fairness, all methods use the same features as our IFA. Row 1 simply averages features across all views, yielding unsatisfactory results. Row 2 and Row 3 leverage Top-5 criterion [38, 40] and DBSCAN algorithm [44] to select the predominant feature, achieving gains of **+0.4**% and **+0.7**% mIoU, respectively. By contrast, our IFA achieves an improvement of **1.8**% mIoU over Row 1.

Adaptive-Octree Efficiency. Based on the results of instance generation in Replica and ScanNet, Tab. 8 provides a comparison of different spatial representations with respect to storage space and the accuracy of occupancy. For the same scene, octree and our adaptive-octree consume two orders of magnitude less storage compared to point clouds. Our adaptive-octree requires a bit more storage than the traditional octree due to its additional record of bounding boxes for each object. However, at the same depth, the adaptive-octree exhibits a much higher mEOR compared to the octree. This means that the space described by the adaptive-octree is more closely aligned with the target re-

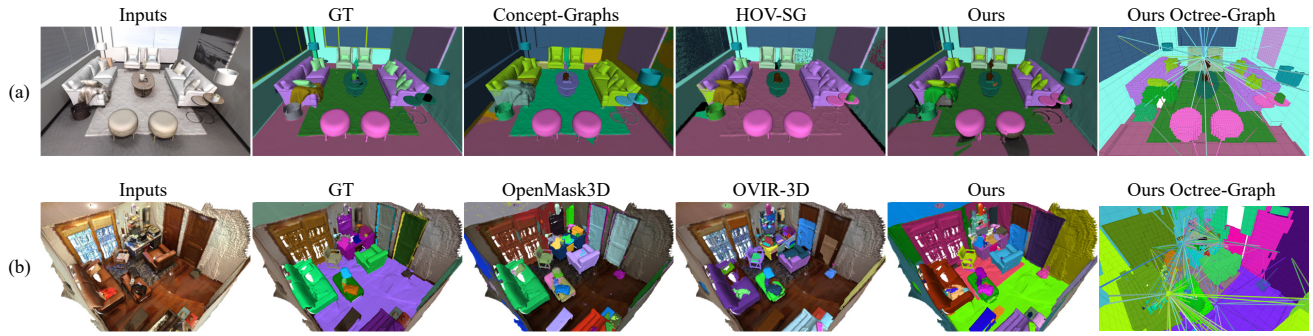

Figure 6. Visual comparisons. (a) Semantic segmentation results on Replica. (b) Instance segmentation results on ScanNet200.

gions. In summary, the adaptive-octree requires much less storage space than point clouds and provides more accurate occupancy information than a traditional octree.

Path Planning Efficiency based on Octree-Graph. To further verify the efficiency of our method, we conduct path planning experiments using A* [9] and Jump Point Search [52] algorithms on the HM3DSem [46] dataset, where the point cloud representation and our Octree-Graph are compared. It can be seen from Tab. 9 that Octree-Graph uses much less storage space and spends much less time. This is crucial for real-world deployment.

4.5. Qualitative Analysis

Visualization Results. Fig. 6 visualizes the results of semantic segmentation and instance segmentation, respectively. We can see that our method exhibits more accurate object semantics and fewer incorrect segments than comparison methods. Fig. 8 demonstrates the segment merging results of our CGSM and its baseline (*i.e.,* frame-wise sequential merging), where CGSM correctly resolves the over-segmented long table without introducing excessive merges between different objects.

Object Retrieval and Path Planning. We also conduct real-world experiments to further validate the effectiveness of our method. Fig. 7 presents the results where a real scene is set up and scanned by an Intel Realsense D435i camera. Then we reconstruct the colored point cloud and establish the Octree-Graph. Based on these, we deploy our method on a robotic dog and a drone with NVIDIA Orin NX as onboard computers. As shown in Fig. 7, robots can accurately find the target and successfully navigate to it relying on our Octree-Graph. The dynamic process of this experiment can be found in the supplementary video demo.

5. Conclusion

In this paper, we propose Octree-Graph, a novel scene representation for open-vocabulary 3D scene understanding. Specifically, an adaptive-octree structure is devised to characterize the occupancy of an object, which acts as the node

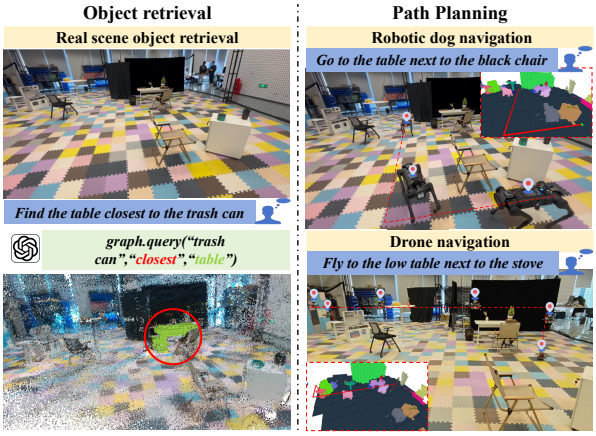

Figure 7. Visualization of the real-world experiment. The left column shows a real scene and the reconstructed colored point cloud with the retrieval target highlighted by our adaptive-octree. The right column presents path planning using a robotic dog and a drone based on our Octree-Graph.

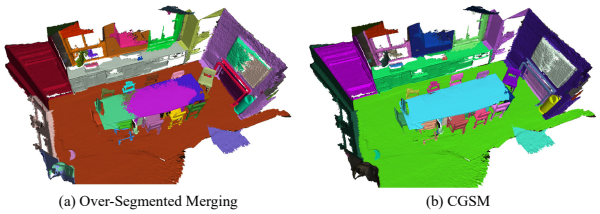

Figure 8. Segment merging comparison.

of the Octree-Graph. The edges describe rich relations among objects for spatial reasoning. For Octree-Graph construction, we also develop a training-free pipeline to conduct semantic object segmentation, where a Chronological Group-wise Segment Merging (CGSM) strategy is designed to alleviate inaccurate segment proposals, and an Instance Feature Aggregation (IFA) method is devised to get a semantic feature both representative and distinctive. Extensive evaluations on several tasks validate the versatility and effectiveness of our Octree-Graph.

Acknowledgments

This work is supported by the Shanghai AI Laboratory, the National Natural Science Foundation of China (62376222), and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 422–440, 2020. 5, 6

[2] 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. 5

[3] Yinan Deng, Jiahui Wang, Jingyu Zhao, Jianyu Dou, Yi Yang, and Yufeng Yue. Openobj: Open-vocabulary object-level neural radiance fields with fine-grained understanding. *IEEE Robotics Autom. Lett.*, 10(1):652–659, 2025. 2

[4] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Lowis3d: Language-driven open-world instance-level 3d scene understanding. *CoRR*, abs/2308.00353, 2023. 1

[5] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[6] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, pages 540–557, 2022. 2

[7] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.16650*, 2023. 1, 2, 3, 4, 5, 6

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. 2

[9] Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4 (2):100–107, 1968. 5, 8

[10] Shuting He, Henghui Ding, and Wei Jiang. Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19498–19507, 2023. 2

[11] Armin Hornung, Kai M Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. Octomap: An efficient probabilistic 3d mapping framework based on octrees. *Autonomous robots*, 34:189–206, 2013. 4

[12] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 10608–10615, 2023. 2

[13] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. *CoRR*, abs/2309.00616, 2023. 2

[14] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7010–7021, 2022. 2

[15] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems*, 2023. 1, 2, 5, 6

[16] Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3d semantic segmentation with foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21284–21294, 2024. 1

[17] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: language embedded radiance fields. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19672–19682, 2023. 1

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1, 2

[19] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J. Davison. vmap: Vectorised object mapping for neural field SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 952–961, 2023. 1

[20] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. 2

[21] Kejie Li, Daniel DeTone, Steven Chen, Minh Vo, Ian Reid, Hamid Rezatofighi, Chris Sweeney, Julian Straub, and Richard A. Newcombe. ODAM: object detection, association, and mapping using posed RGB video. In *2021*

IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, pages 5978–5988, 2021. 1

[22] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 7061–7070, 2023. 2, 5

[23] Sergey Linok, Tatiana Zemskova, Svetlana Ladanova, Roman Titkov, and Dmitry Yudin. Beyond bare queries: Open-vocabulary object retrieval with 3d scene graph. *arXiv preprint arXiv:2406.07113*, 2024. 5, 6

[24] 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. 2

[25] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *7th Annual Conference on Robot Learning*, 2023. 1, 2, 3, 4, 6

[26] Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. Octreeocc: Efficient and multi-granularity occupancy prediction using octree queries. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 2

[27] Dominic Maggio, Yun Chang, Nathan Hughes, Matthew Trang, J. Daniel Griffith, Carlyn Dougherty, Eric Cristofalo, Lukas Schmid, and Luca Carlone. Clio: Real-time task-driven open-set 3d scene graphs. *IEEE Robotics Autom. Lett.*, 9(10):8921–8928, 2024. 2

[28] Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part X*, pages 728–755, 2022. 2

[29] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 6

[30] Ting Pan, Lulu Tang, Xinlong Wang, and Shiguang Shan. Tokenize anything via prompting. *arXiv preprint arXiv:2312.09128*, 2023. 2, 5

[31] Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas A. Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 815–824, 2023. 1, 2

[32] Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4024–4033, 2023. 2, 3, 5

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, pages 8748–8763, 2021. 1, 2, 5, 6

[34] Damien Robert, Bruno Vallet, and Loïc Landrieu. Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5565–5574. 1

[35] Dávid Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, pages 125–141, 2022. 2

[36] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022. 5

[37] Erich Schubert, Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.*, 42(3):19:1–19:21, 2017. 3

[38] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. 2023. 1, 6, 7

[39] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[40] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems*, 2023. 2, 4, 6, 7

[41] Sebastian Thrun and Arno Bücken. Integrating grid-based and topological maps for mobile robot navigation. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, AAAI 96, IAAI 96, Portland, Oregon, USA, August 4-8, 1996, Volume 2*, pages 944–950, 1996. 7

[42] Khoi Nguyen Tuan Duc Ngo, Binh-Son Hua. Isbnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[43] Vibashan VS, Ning Yu, Chen Xing, Can Qin, Mingfei Gao, Juan Carlos Niebles, Vishal M. Patel, and Ran Xu. Mask-free OVIS: open-vocabulary instance segmentation without man-

ual mask annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23539–23549, 2023. 2

[44] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. *Robotics: Science and Systems*, 2024. 1, 2, 3, 4, 5, 6, 7

[45] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18113–18123, 2022. 2

[46] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Théophile Gervet, John M. Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-matterport 3d semantics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 4927–4936, 2023. 5, 7, 8

[47] Kashu Yamazaki, Taisei Hanyu, Khoa Vo, Thang Pham, Minh Tran, Gianfranco Doretto, Anh Nguyen, and Ngan Le. Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9411–9417. IEEE, 2024. 6

[48] Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28274–28284, 2024. 1, 2, 3, 5, 6

[49] Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[50] Xi Yang, Xu Gu, Xingyilang Yin, and Xinbo Gao. SA3DIP: segment any 3d instance with potential 3d priors. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 2

[51] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 2, 6

[52] Zhen Yang, Junli Li, Liwei Yang, and Hejiang Chen. A smooth jump point search algorithm for mobile robots path planning based on a two-dimensional grid model. *J. Robotics*, 2022:7682201:1–7682201:15, 2022. 5, 8

[53] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3292–3302, 2024. 2, 4, 6

[54] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5732–5741. IEEE, 2021. 2

[55] Jiazhao Zhang, Liu Dai, Fanpeng Meng, Qingnan Fan, Xuelin Chen, Kai Xu, and He Wang. 3d-aware object goal navigation via simultaneous exploration and identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6672–6682, 2023. 1