MaskClustering: View Consensus based Mask Graph Clustering for Open-Vocabulary 3D Instance Segmentation

Mi Yan[1,2] Jiazhao Zhang[1,2] Yan Zhu[1] He Wang[1,2,3,†]
[1]CFCS, School of CS, Peking University [2]Beijing Academy of Artificial Intelligence [3]Galbot

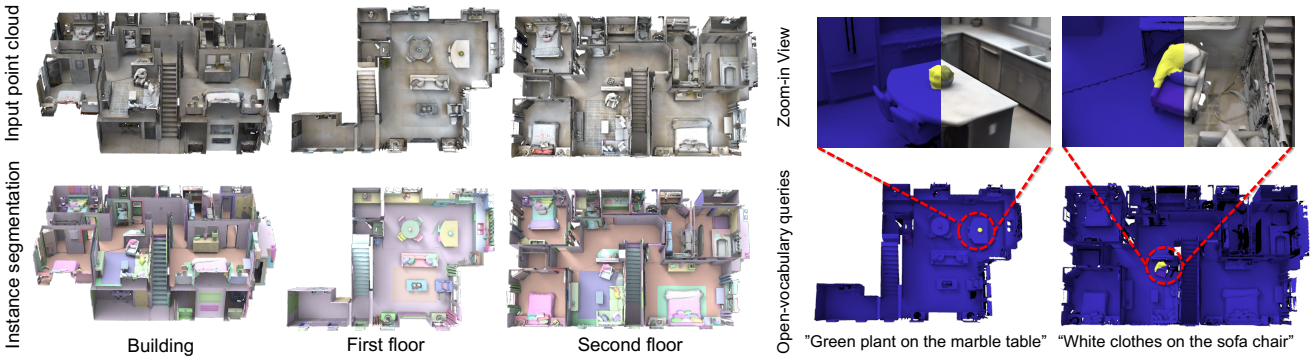

Figure 1. Our method tackles the challenges of open-vocabulary instance segmentation. It achieves detailed segmentation across objects of varying scales and can query these objects using open-vocabulary text.

Abstract

Open-vocabulary 3D instance segmentation is cutting-edge for its ability to segment 3D instances without predefined categories. However, progress in 3D lags behind its 2D counterpart due to limited annotated 3D data. To address this, recent works first generate 2D open-vocabulary masks through 2D models and then merge them into 3D instances based on metrics calculated between two neighboring frames. In contrast to these local metrics, we propose a novel metric, view consensus rate, to enhance the utilization of multi-view observations. The key insight is that two 2D masks should be deemed part of the same 3D instance if a significant number of other 2D masks from different views contain both these two masks. Using this metric as edge weight, we construct a global mask graph where each mask is a node. Through iterative clustering of masks showing high view consensus, we generate a series of clusters, each representing a distinct 3D instance. Notably, our model is training-free. Through extensive experiments on publicly available datasets, including ScanNet++, ScanNet200 and MatterPort3D, we demonstrate that our method achieves state-of-the-art performance in open-vocabulary 3D instance segmentation. Our project page is at https://pku-epic.github.io/MaskClustering.

1. Introduction

Open-vocabulary 3D instance segmentation tackles the problem of predicting 3D object instance masks and their corresponding categories from reconstructed 3D scenes, without relying on a predefined list of categories. This is an essential task for 3D scene understanding [4, 12, 35], robotics [8, 15, 50] and VR/AR applications [22, 45]. However, this task is more challenging than its established 2D counterpart, open-vocabulary 2D instance segmentation [11, 20, 39, 43, 44], primarily due to the lack of large-scale open-world 3D data. Consequently, most current methods [18, 28, 37] in this field divide this task into two stages: zero-shot 3D instance mask prediction, followed by open-vocabulary semantic queries. In this work, we primarily focuses on obtaining high-quality, zero-shot 3D instance masks.

Existing approaches for zero-shot 3D instance mask prediction primarily follow two paths. 3D-to-2D projection-based methods [18, 19] leverage existing 3D instance segmentation algorithms to generate 3D masks. However, this approach is fundamentally constrained by the quality of 3D reconstructions and the relatively modest capabilities of current 3D instance segmentation tools. As a result, these methods often struggle to accurately segment small

†: He Wang is the corresponding author.

objects, leading to a significant loss of detail in complex scenes. In contrast, 2D-to-3D region grow-based methods [28, 46] leverage 2D segmentation models to process frames sequentially and update a list of 3D instances simultaneously. They merge new 2D masks with existing 3D instances based on geometric overlap and semantic similarity for each frame. However, we find that such online processing lacks global optimality across all frames, often resulting in incorrect merging.

To address these limitations, we propose a novel approach that improves global consistency via multi-view verification, inspired by bundle adjustment [38]. Unlike prior methods that rely on local metrics calculated between adjacent frames to decide whether a mask pair should be merged, our method introduces a new global metric, the view consensus rate, which measures the proportion of frames supporting their merging. Here, a frame t supports merging only if another 2D mask within frame t contains this mask pair. In this way, the same-instance relationship of two view-consensus masks are indeed supported by multi-view observation.

Utilizing the same-instance relationship, we build a global mask graph wherein each node is a mask, with edges added between high view consensus mask pairs. Following this, mask pairs exhibiting high view consensus are prioritized for merging into a mask cluster, and the view consensus between this mask cluster and other mask clusters will be updated. This iterative clustering and updating process yields a final list of clusters, each containing multiple masks and denoting a 3D instance. For each 3D instance, its point cloud and semantic feature are the aggregated partial point clouds and open-vocabulary features derived from individual 2D masks, respectively.

Our method, validated on ScanNet++ [47], Matterport3D [1], and ScanNet200 [35] benchmarks, achieves state-of-the-art results in zero-shot mask prediction and open-vocabulary instance understanding, surpassing existing methods, especially in segmenting fine-grained objects.

Our contributions can be concluded as follows:

- A novel graph clustering based methodology to merge 2D masks for 3D open-vocabulary instance segmentation.
- A novel view consensus metric for evaluating the relationship between 2D masks, effectively leveraging global information from input image sequences.
- A SOTA open-vocabulary 3D instance segmentation method, which demonstrates superior performance on many publicly available datasets.

2. Related Works

Closed-set 3D instance segmentation. Since the emergence of 3D scene datasets [4, 11], the computer vision community has witnessed a large literature of 3D segmentation methods [3, 9, 13, 14, 16, 26, 34, 36, 41, 42]. These methods tackle this problem either in online [17, 27, 29, 49, 51] or offline [34, 36, 41, 42] manner, representing the scene as points cloud, voxels, and more recently neural field [40, 52]. Though significant progress has been made, these methods are limited to a closed-set category list which is pre-defined in certain dataset, suffering poor performance in open-vocabulary settings as tail classes that have few or no training examples. In contrast, our method aims to tackle open-vocabulary 3D instance segmentation that segment objects of arbitrary category.

Open-vocabulary 2D instance segmentation. The recent advances in large visual foundation models [2, 7, 24, 25, 32, 33] have enabled a remarkable level of robustness of 2D understanding tasks. Typical tasks include zero-shot 2D segmentation [2, 24, 32], open-vocabulary 2D image understanding [7, 25, 33], and open-vocabulary 2D object detection [21, 23, 53]. Recently, many works [11, 20, 39, 43, 44] focus on the open-vocabulary 2D segmentation task, which requires predicting the open-vocabulary feature at the pixel level. These methods encode 2D images and align open-vocabulary pixel features with them. However, due to the lack of large-scale 3D annotated data, end-to-end open-vocabulary 3D instance segmentation is in slow progress. In this work, we tackle the open-vocabulary 3D instance segmentation by leveraging the prior from large 2D vision-language models.

Open-vocabulary 3D instance segmentation. There are two types of methods: (1) 3D-to-2D projection methods and (2) 2D-to-3D region grow-based methods. (1) 3D-to-2D projection methods [19, 31, 37] directly conduct 3D instance segmentation [19, 36] on 3D indoor scene input. They project the 3D instance objects to 2D frames, and extract open-vocabulary features for final aggregation. However, these types of methods are limited to well-reconstructed scene and detailed objects are easily missed if the geometry details are poor. (2) 2D-to-3D region grow-based methods [8, 28] propose to online fuse 2D observation to 3D instance segmentation. By back-projecting the 2D mask to 3D point cloud, these methods leverage clustering algorithm [5] or geometry overlapping to find corresponding 3D instances. The open-vocabulary feature is also aggregated during the back-projection. However, these types of methods consider the associations between historical constructed 3D instances with live frame, lacking a global understanding of all observed frames.

Concurrently, SAI3D[48] and Open3DIS[30] propose merging 3D superpoints[6] guided by predictions from SAM[24], showing robust performance in open-vocabulary 3D instance segmentation. However, we diverge from their approach by avoiding reliance on 3D superpoints, which face challenges in distinguishing geometrically-homogeneous objects like posters on walls or rows of similar medicine boxes.

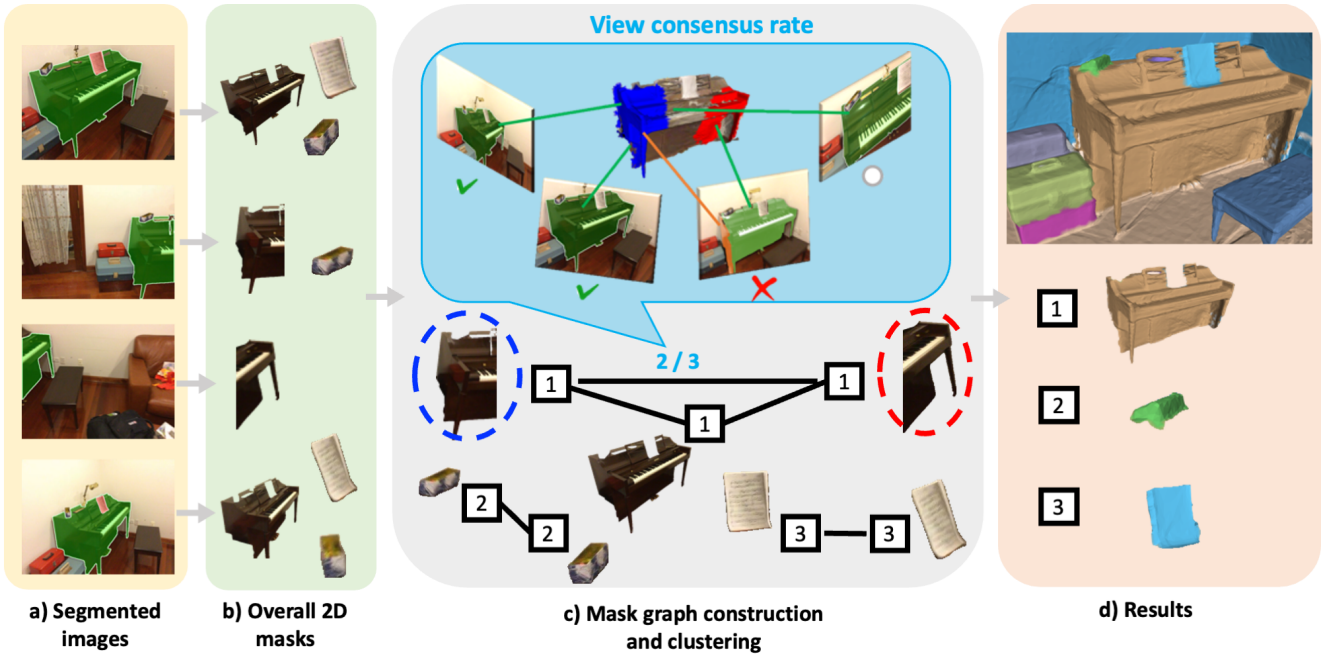

Figure 2. Overview pipeline of our method: a) We take segmented image sequences as input and b) extract all 2D masks from the input. c) To merge them, we build a global graph with each node as a mask. We use the view consensus rate, which is defined as the proportion of frames supporting the merging, to add edges between nodes. Each frame supports the merging only if there is a mask in this frame containing both nodes. d) Each mask cluster is merged into a 3D instance. For clarity, we only visualize three objects in the figure.

3. Method

3.1. Problem Formulation and Method Overview

Given a set of posed color images $\{I_1^c, I_2^c, \ldots, I_T^c\}$, their corresponding depths $\{I_1^d, I_2^d, \ldots, I_T^d\}$, and the reconstructed point cloud P of a scene, our algorithm outputs a list of 3D instances along with their open-vocabulary semantics fused from 2D mask proposals.

We initially employ an off-the-shelf, class-agnostic mask predictor to process each color image I_t^c and derive the 2D masks $\{m_{t,i} \mid i = 1, 2, \ldots, n_t\}$ where n_t denotes the number of masks in frame t. We assume the mask predictor to generate entity-level panoptic segmentation masks, indicating that each mask approximates one object with nearly all pixels assigned to a single mask. This assumption aligns with capabilities of advanced segmentation tools like CropFormer[32].

The overview pipeline of our method is illustrated in Fig. 2. To fuse these 2D masks from different frames into 3D instances, we propose to construct a mask graph $G = (V, E)$. Each node in V corresponds to a mask $m_{t,i}$, and an edge in E indicates that two masks are part of the same instance and should be merged. To assess edge connectivity, we propose to leverage consensus cues from multi-view observations and therefore introduce view consensus rate as a criterion (Sec. 3.2).

Once the mask graph is established, we initiate an iterative process to cluster masks and update edges, with a priority on merging mask pairs displaying solid view consensus (Sec. 3.3). The result of this iterative process is a list of clusters, each denoting a 3D instance and containing multiple masks. Within such a cluster, we aggregate the corresponding partial point clouds from the individual masks to form the ultimate 3D instance. Building on these correspondences between 2D masks and 3D instances, we perform feature fusion for a more comprehensive representation, which aids in open-vocabulary semantic prediction (Sec. 3.4).

3.2. Mask Graph Construction

In this subsesction, we introduce view consensus rate, which serves as the criterion to determine edge connectivity between two masks (Sec.3.2.1). We then propose an efficient method for calculating this rate (Sec.3.2.2) and leverage this rate to filter out under-segmented masks (Sec.3.2.3).

Notations and Definitions Given the reconstructed point cloud P and frame index t, for a mask $m_{t,i}$, we can obtain the mask point cloud $P_{t,i}$ by projecting onto P the backprojected point cloud of $m_{t,i}$ from I_t^d. Then we define the frame point cloud P_t as the union of all $P_{t,i}$s for $i = 1, 2, ..., n_t$, yielding $P_{t,i} \subset P_t \subset P$. We define a point p to be visible at frame t if $p \in P_t$. We then define a mask $m_{t',i}$ to be visible at frame t if at least $\tau_{vis} = 0.3$ of its total

points from $P_{t',i}$ are visible and denote the visible part as $P^t_{t',i}$. We denote the set of frames where $m_{t',i}$ is visible as $F(m_{t',i})$. Finally, we define the approximate containment relationship of one point clouds P_i by another point cloud P_j as $P_i \sqsubset P_j$, if at least $\tau_{contain} = 0.8$ of the total points in P_i lie within P_j.

3.2.1 View Consensus Rate

The cornerstone of our method lies in determining if two masks belong to the same instance by utilizing 2D predictions across all frames. In this context, we propose to leverage view consensus cues, as detailed below.

To assess the relationship between two masks, specifically $m_{t',i}$ and $m_{t'',j}$, where t' and t'' may be the same or different frames, we utilize the masks $\{m_{t,k}\}$ from relevant views. The goal is to check if there is substantial consensus among the views supporting that these two masks represent the same 3D instance.

To be more specific, we first find all the frames O in which both of the two masks are visible, serving as the observers to the two masks, *i.e.*, $O(m_{t',i}, m_{t'',j}) = F(m_{t',i}) \cap F(m_{t'',j})$. And we denote the number of observers in O as $n(m_{t',i}, m_{t'',j}) = |O(m_{t',i}, m_{t'',j})|$, where $|\cdot|$ represents the cardinality of the set.

We then check whether an observer frame $t \in O$ supports the merging of these two masks. For an observer frame $t \in O$, if there exists a mask $m_{t,k}$ whose corresponding point cloud $P_{t,k}$ approximately contains both the point clouds $P^t_{t',i}$ of $m_{t',i}$ and $P^t_{t'',j}$ of $m_{t'',j}$, i.e., $P^t_{t',i} \sqsubset P_{t,k}$ and $P^t_{t'',j} \sqsubset P_{t,k}$, then this observer supports that the two masks are components of the same instance. The total number of supporters would be $n_{supporter}(m_{t',i}, m_{t'',j}) = |\{t \in O(m_{t',i}, m_{t'',j}) \mid \exists k, s.t.\ P^t_{t',i}, P^t_{t'',j} \sqsubset P_{t,k}\}|$. The proportion of supporters among all observers is subsequently defined as the **view consensus rate** c, as illustrated below:

$$c(m_{t',i}, m_{t'',j}) = \frac{n_{supporter}(m_{t',i}, m_{t'',j})}{n(m_{t',i}, m_{t'',j})} \quad (1)$$

An illustration of thie view consensus rate can be found in Fig. 3.

Employing the consensus rate as a criterion, we connect edges between mask pairs whose view consensus rates exceeding a predefined threshold $\tau_{rate} = 0.9$. This procedure yields the set of edges E as follows:

$$E = \{\{m_{t',i'}, m_{t'',i''}\} \mid c(m_{t',i'}, m_{t'',i''}) \geq \tau_{rate}\} \quad (2)$$

Leveraging predictions across the entire sequence of images, our criterion shows enhanced robustness against oversegmentation errors compared to approaches that solely depend on local geometric overlap. Illustrated in Fig. 3, the two masks exhibit low geometric overlap despite belonging to the same armchair. However, our approach identify a high consensus rate for them. This is attributed to the outstanding overall performance of modern mask predictors, which consistently segment this armchair comprehensively in most frames, encompassing both parts and thus yielding a high view consensus rate.

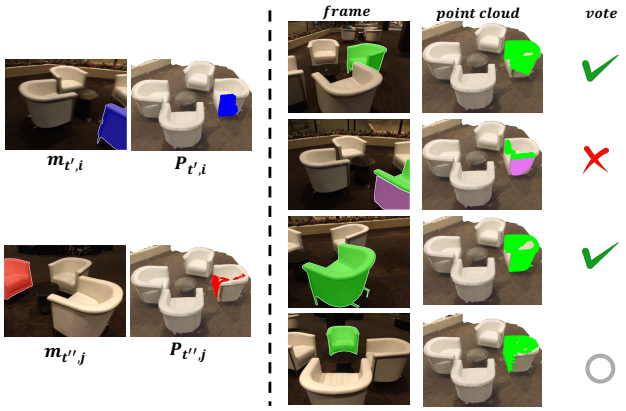

Figure 3. **View consensus rate.** Masks $m_{t',i}$ and $m_{t'',j}$ (side and frontal view of an armchair) are both visible in three frames, with two supporting them belonging to the same instance, resulting in a 2/3 consensus rate. Each mask is accompanied by its respective mask point cloud, displayed on the right. All point clouds are rendered under a consistent camera pose for clarity.

3.2.2 Efficient Computation of View Consensus Rate

Naively computing view consensus rates for all mask pairs can be untractable with a time complexity of $\mathcal{O}(N^2T)$, where N represents the total number of masks, i.e., $N = \sum_t n_t$. To speed up, we initially calculate and store the intermediate result to eliminate redundant computations.

Specifically, for each mask $m_{t',i}$, we first find $F(m_{t',i})$ and then identify all the *masks that approximately contain it*, denoted as $M(m_{t',i}) = \{m_{t,k} \mid t \in F(m_{t',i}) \text{ and } P^t_{t',i} \sqsubset P_{t,k}\}$. With these intermediate results, the computation of equation 1 can be simplified as,

$$c(m_{t',i}, m_{t'',j}) = \frac{|M(m_{t',i}) \cap M(m_{t'',j})|}{|F(m_{t',i}) \cap F(m_{t'',j})|} \quad (3)$$

In this way, all the operations in this expression have been simplified to simple set intersection operations involving only a few dozen elements, leading to a significant reduction in computational complexity.

We now introduce the efficient computation of $M(m_{t',i})$. Initially, we examine the mask ID distribution of $P^t_{t',i}$ at frame t. If this distribution is concentrated, with more than $\tau_{contain} = 0.8$ of elements equalling k, it indicates that $P^t_{t',i}$ primarily constitutes a part of the k-th instance at frame t. By definition, $P^t_{t',i} \sqsubset P_{t,k}$. The mask ID distribution is denoted as $d(m_{t',i}, t)$, and we elaborate

on its efficient calculation through a space-time trade-off in the supplementary material.

3.2.3 Under-Segment Mask Filtering

We can also identify whether a mask is under-segmented based on the mask ID distribution $d(m_{t',i}, t)$. If $d(m_{t',i}, t)$ exhibits a very diverse distribution, it signifies that $P_{t',i}$ comprises multiple instances at frame t, making it highly likely that $m_{t',i}$ is an under-segmented mask. Assuming most 2D mask predictor outputs are correct, we ignore the alternative explanation that $m_{t',i}$ is accurate but the mask predictor over-segments this object consistently in other views.

Therefore, under-segmentation is marked by frequent distinction of $P_{t',i}$ into parts. We track the frequency of such occurrences (number of frames with diverse distributions in $d(m_{t',i}, t)$ / $|F(m_{t',i})|$). If this frequency exceeds $\tau_{filter} = 0.2$, we classify the mask as under-segmented and filter it out. Specifically, we remove it from the mask graph. Additionally, to prevent this mask from erroneously inflating the consensus rate between two masks belonging to different instances, we also eliminate it from all $M(m_{t'',j})$ and remove t' from $F(m_{t'',j})$.

3.3. Iterative Graph Clustering

Building upon the mask graph, we introduce an iterative graph clustering technique to merge masks and update the graph structure alternately. In the last iteration, each cluster denotes an instance.

When determining which masks to merge, we consider two strategies: 1) merging each maximal clique (where a clique is a subset of the graph with an edge between every pair of nodes); 2) merging each connected component (where a connected component is a subset of the graph with a path between every pair of nodes). The first approach, though precise, tends to be overly stringent, often leading to insufficient merging and excessive over-segmentation. The second approach, more permissive, relies on the correctness of every pair-wise identified same-instance relationship, which can be less reliable when the number of observers n—the denominator of c—is low.

To balance these strategies, we modify the second approach to prioritize merging masks with a high number of observers first, postponing less reliable connections to later iterations.

As illustrated in Fig.4, in each iteration k, we set an observer threshold n_k and edges with $n < n_k$ are disconnected. We then identify connected components in the graph and merge them into new nodes. For a newly formed mask m_{new} from a set of masks $\{m_{t_1,i_1}, m_{t_2,i_2}, \ldots, m_{t_s,i_s}\}$, its point cloud P_{new} is the union of $\{P_{t_1,i_1}, P_{t_2,i_2}, \ldots, P_{t_s,i_s}\}$.

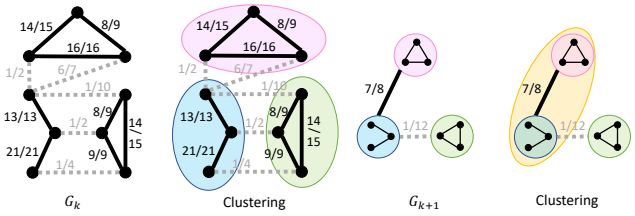

Figure 4. Illustration of iterative clustering. Node pairs with more observers are prioritized clustered (G_k). Then, view consensus of grouped masks is updated for the next clustering with more confident view consensus measurements. The text on the edge means $n_{support}/n$.

Subsequent to these node merging operations, updating edges requires recalculating the view consensus rate for the new mask in relation to others. Referring to equation 3, we calculate $F(m_{new})$ and $M(m_{new})$. While these two sets can be computed using the same technique as introduced in Sec.3.2.2, we propose a method to accelerate this calculation while achieving comparable results through a straightforward approximation. Specifically, we approximate $F(m_{new})$ as $F(m_{t_1,i_1}) \cup F(m_{t_2,i_2}) \ldots \cup F(m_{t_s,i_s})$ and $M(m_{new})$ as $M(m_{t_1,i_1}) \cup M(m_{t_2,i_2}) \ldots \cup M(m_{t_s,i_s})$. This approximation is justified since masks merged due to high consensus rates often share containment by the same mask in frames where they both appear. The quantitative impact of this approximation is presented in Table 4.

After each iteration k, a new graph G_{k+1} is formed. The observer threshold n_k is adjusted downwards over several iterations to avoid neglecting smaller objects visible in fewer frames. We adopt a decreasing n_k schedule, ranging from the top 5%, 10%, to 95% of observer counts across all mask pairs.

3.4. Open-Vocabulary Feature Aggregation

After multiple iterations of clustering, we have obtained a conclusive list where each entry represents a 3D instance proposal. Simultaneously, we maintain a corresponding list of masks associated with each instance. This 2D-3D relationship allows us to directly select representative masks and fuse their semantic features to create an open-vocabulary feature for this instance. Following OpenMask3D[37], we first pick the top-5 masks that best cover the instance. Subsequently, we crop the original RGB image at multiple scales around each mask and input these image crops into CLIP[33] to extract open-vocabulary features. The final instance feature is derived from the average pooling result of these features.

3.5. Implementation Details

In order to obtain object-level masks, we use CropFormer [32] as our 2D mask predictor. For open-vocabulary feature extraction, we use CLIP[33] ViT-H. To get mask point cloud

Table 1. **Zero-shot 3D instance segmentation results on ScanNet++ and MatterPort3D.** We report both semantic and class-agnostic performance. Our method outperform all baselines on all metrics significantly.

Model	ScanNet++						MatterPort3D					
	Semantic			Class-agnostic			Semantic			Class-agnostic		
	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}
Mask3D	3.6	5.1	6.7	22.8	33.3	45.7	2.5	4.5	6.7	4.4	9.8	20.6
OpenMask3D	2.0	2.7	3.4	22.8	33.3	45.7	4.6	8.5	13.0	4.4	9.8	20.6
OVIR-3D	3.6	5.7	7.3	19.4	34.1	46.5	6.3	16.4	24.4	5.9	13.9	24.6
Ours	**7.8**	**11.9**	**13.2**	**24.6**	**40.3**	**51.5**	**9.2**	**19.7**	**26.5**	**8.3**	**18.9**	**33.4**

$P_{t,i}$, we first back-project each mask to get the raw point cloud and then ball query the reconstructed point cloud with a radius equal to 3cm. We adopt the post-processing approach from OVIR-3D[28] to refine the output 3D instances by using DBSCAN algorithm to separate disconnected point clusters into distinct instances.

4. Experiments

In this section, we extensively evaluate our proposed method by comparing it with previous state-of-the-art methods on publicly available 3D instance segmentation benchmarks. The experimental setup is detailed in Section 4.1, and the statistics are comprehensively analyzed in Section 4.2. Following that, we showcase the remarkable visual outcomes of our approach across a diverse range of complex scenes in Section 4.4. The validation of all the components of our method is presented in Section 4.3.

4.1. Experimental setup

Dataset ScanNet++[47] is a recently released high-quality benchmark that comprises 1554 classes with fine-grained annotation, making it an optimal choice for assessing open-vocabulary 3D instance segmentation. We also assess our method on two widely-used benchmarks: ScanNet200[4, 35], which focuses on room-level evaluations, and MatterPort3D[1], designed for building-level evaluations with sparser viewpoints. We utilize the validation sets of ScanNet++ and ScanNet, along with the testing set of MatterPort3D.

Baselines We select the recent SOTA methods on both supervised closed-set 3D instance segmentation and open-vocabulary 3D instance segmentation. Mask3D [36] stands out as a state-of-the-art method which requires supervised training on ScanNet200. OpenMask3D [37] leverages supervised mask proposals from Mask3D and employs CLIP for open-vocabulary semantics aggregation. Different from our setting, both of them rely on supervised mask. OVIR-3D [28] utilize both zero-shot masks and semantics, merging zero-shot 2D masks with large geometric and semantic overlap and using K-Means to choose the most representative features from the per-frame semantic feature.

Metrics We report the standard Average Precision (AP) at 25% and 50% IoU and the mean of AP from 50% to 95% at 5% intervals. In addition to the conventional semantic instance segmentation setting, we also test in a class-agnostic setting, disregarding semantic labels and solely assessing mask quality. This setting offers a precise assessment of the zero-shot mask prediction capability.

Table 2. **3D instance segmentation results on ScanNet200.** Mask3D and OpenMask3D both require supervised (sup.) training on ScanNet200. In fully zero shot (z.s.) setting, our method surpass OVIR-3D by a large margin on all metrics.

Model	Class-agnostic			Semantic		
	AP	AP_{50}	AP_{25}	AP	AP_{50}	AP_{25}
sup. mask + sup. semantic						
Mask3D	39.7	53.6	62.5	26.9	36.2	41.4
sup. mask + z.s. semantic						
OpenMask3D	39.7	53.6	62.5	15.1	19.6	22.6
z.s. mask + z.s. semantic						
OVIR-3D	14.4	27.5	38.8	9.3	18.7	25.0
Ours	**19.7**	**36.4**	**51.4**	**12.0**	**23.3**	**30.1**

4.2. Quantitative Comparison.

ScanNet++ and MatterPort3D. We directly test all methods on ScanNet++ and MatterPort3D in a zero-shot manner. As shown in Table 1, our method outperforms all baselines by a large margin. In comparison to OVIR-3D, the most akin work to ours, we achieve +4.2% and +5.2% AP on ScanNet++ semantic and class-agnostic setting, respectively. Similarly, we demonstrate +2.9% and +2.4% AP on MatterPort3D in the same settings, validating our globally optimal association design.

OpenMask3D shares Mask3D's mask predictor, rendering their performance identical in the class-agnostic setting. We observe that this mask predictor, trained on ScanNet200, has limited generalizability. While it shows impressive results within the confines of ScanNet200, its performance suffers significantly when evaluated on new benchmarks, as demonstrated in Table 1 and Fig. 5. Additionally, it exhibits sensitivity to point distribution patterns. For instance, in the class-agnostic setting of ScanNet++, its AP is a mere 13.6% when using the raw point cloud as input. Interestingly, a simple preprocessing step such as uniform sampling significantly boosts performance to 22.7%.

ScanNet200. As the mask predictor and semantic head of Mask3D are trained specifically on ScanNet200, we classify methods according to their train-test settings. In comparison to the fully zero-shot method, OVIR-3D, our approach exhibits a significant performance advantage, surpassing it by +5.3% in average precision (AP), +8.9% in AP_{50}, and +12.6% in AP_{25} in the class-agnostic setting. This further underscores the effectiveness of our proposed globally optimal merging mechanism. Moreover, our method even outperforms OpenMask3D, which relies on a supervised mask predictor, by a substantial margin of +3.7% in AP_{50} and +7.5% in AP_{25}.

4.3. Ablation Studies

In Table 3, we analyze key components of our method on ScanNet200—under-segment mask filtering and iterative clustering. Starting with a baseline using view consensus rate, we merge masks within connected components. This simple approach matches OVIR-3D baseline performance. Upon adding under-segment mask filtering and iterative clustering, performance steadily rises from 10.0% AP to 11.7% AP, reaching peak performance when both modules are combined.

Table 3. **Ablation study on under-segment mask filtering and iterative clustering on ScanNet200.**

under. filtering	iter. clustering	AP	AP_{50}	AP_{25}
✘	✘	10.0	19.1	24.2
✔	✘	11.0	21.2	27.5
✘	✔	11.7	22.3	29.2
✔	✔	**12.0**	**23.3**	**30.1**

We compare various clustering algorithms, including clustering cliques or connected components as discussed in Section 3.3, and also clustering a relaxation of cliques using the Highly Connected Sub-graphs (HCS) algorithm [10]. We also show the impact of the approximation introduced in Section 3.3. As shown in Table 4, our proposed iterative clustering method outperforms all other trials. Comprehensive statistics are available in the supplementary material.

Table 4. **Ablation study on clustering methods.**

Clustering Algorithm	AP	AP_{50}	AP_{25}
Connected component	11.0	21.2	27.5
Clique	11.3	22.0	29.4
Quasi-Clique (HCS)	11.9	22.9	29.7
Ours w/o approximation	11.8	23.1	**30.4**
Ours	**12.0**	**23.3**	30.1

We conducted additional evaluations to assess the robustness of our algorithm to variations in hyperparameters. For the mask visibility threshold τ_{vis} ranging from 0.6 to 0.8, the under-segment mask filtering threshold τ_{filter} ranging from 0.2 to 0.4, the consensus rate threshold τ_{rate} ranging from 0.8 to 1 and the approximate containment threshold $\tau_{contain}$ ranging from 0.7 to 0.9, our method consistently demonstrates satisfying performance.

Table 5. **Ablation study on Hyperparameters on ScanNet200.**

	AP	AP_{50}	AP_{25}
$\tau_{vis}(0.6-0.8)$	11.9 ± 0.06	23.2 ± 0.09	30.1 ± 0.07
$\tau_{filter}(0.2-0.4)$	11.9 ± 0.05	23.3 ± 0.19	30.0 ± 0.18
$\tau_{rate}(0.8-1)$	11.8 ± 0.20	22.7 ± 0.52	28.9 ± 0.83
$\tau_{contain}(0.7-0.9)$	11.9 ± 0.10	23.3 ± 0.22	30.3 ± 0.20

4.4. Qualitative Results.

In Fig. 6, we present the similarity heatmaps for a wide range of open-vocabulary queries, showcasing the remarkable capabilities of our open-vocabulary segmentation system. Additionally, in Fig. 5, we offer a visual comparison of our algorithm against all baseline methods. Our method shows excellent ability to segment small objects, *e.g.*, items on the counter in ScanNet a), boxes on the shelf in ScanNet b). These small objects are simply labeled as part of its containers in the ground truth, which cause the AP at higher IoU threshold of our method drops severely.

Compared to OVIR-3D, our method has two main advantages: i) OVIR-3D can't merges masks that have low geometric overlap but correspond to a same object well. For example, in ScanNet (b), items on the coffee table split the table point cloud into two pieces, making OVIR-3D fail to merge these two parts together. So do the sofa chair in ScanNet b) and the right rug in the MatterPort3D example. In the contrary, our method merges these objects well based on view consensus as explained in Section 3.2.1. ii) The strict filtering process in OVIR-3D falsely filter out many objects, *e.g.*, counter and pictures in ScanNet a) while our method only conservatively filter out under-segment masks.

4.5. Limitations

While our approach demonstrates remarkable performance, it is important to acknowledge two notable limitations. Firstly, this work assumes near-perfect 2D segmentation and 2D-3D correspondence, which may not always be the case in certain applications. Presently, we only generate object-level masks, whereas real-world applications may necessitate multi-level masks spanning from parts and objects to clusters.

5. Conclusion

In this work, we propose a view consensus based mask graph clustering algorithm for open-vocabulary 3D instance segmentation. Specifically, our method constructs a global mask graph and leverages the view consensus to cluster

Figure 5. Comparison of 3D zero-shot segmentation performance. We compare our methods with OpenMask3D [37] and OVIR-3D [28] on ScanNet [4] and Matterport3D [1].

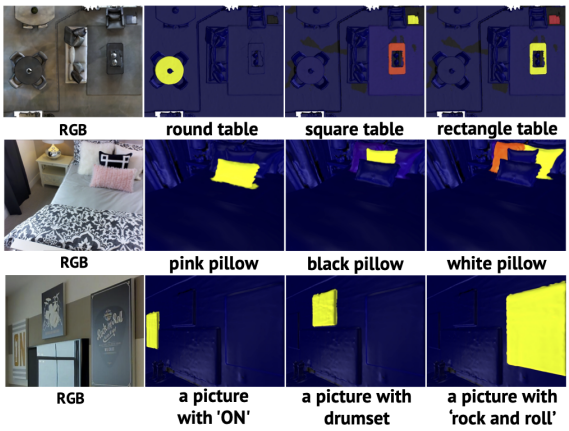

Figure 6. **Open-vocabulary queries of different shapes, colors and contents.**

the masks belonging to the same 3D instances. Besides, the mask clustering guided the clustering of the open-vocabulary features for text queries. The results demonstrate that our method achieves SOTA performance on zero-shot mask prediction and open-vocabulary understating. In the future, we would like to investigate the application of the proposed method on robotic tasks, such as open-vocabulary object navigation.

6. Acknowledgements

This work was supported in part by National Key R&D Program of China 2022ZD0160801.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2, 6, 8

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[3] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 2

[4] 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. 1, 2, 6, 8

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 2

[6] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 2

[7] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, 2021. 2

[8] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Ramalingam Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *ArXiv*, abs/2309.16650, 2023. 1, 2

[9] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020. 2

[10] Erez Hartuv and Ron Shamir. A clustering algorithm based on graph connectivity. *Information processing letters*, 76(4-6):175–181, 2000. 7

[11] Shuting He, Henghui Ding, and Wei Jiang. Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19498–19507, 2023. 1, 2

[12] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4421–4430, 2019. 1

[13] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14373–14382, 2021. 2

[14] Zeyu Hu, Xuyang Bai, Jiaxiang Shang, Runze Zhang, Jiayu Dong, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15488–15498, 2021. 2

[15] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023. 1

[16] Jingwei Huang, Haotian Zhang, Li Yi, Thomas Funkhouser, Matthias Nießner, and Leonidas J Guibas. 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*, pages 4440–4449, 2019. 2

[17] Shi-Sheng Huang, Ze-Yu Ma, Tai-Jiang Mu, Hongbo Fu, and Shi-Min Hu. Supervoxel convolution for online 3d semantic segmentation. *ACM Transactions on Graphics (TOG)*, 40(3): 1–15, 2021. 2

[18] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. *ArXiv*, abs/2309.00616, 2023. 1

[19] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. *arXiv preprint arXiv:2309.00616*, 2023. 1, 2

[20] Dat T. Huynh, Jason Kuen, Zhe nan Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7010–7021, 2021. 1, 2

[21] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In *International Conference on Machine Learning*, pages 15946–15969. PMLR, 2023. 2

[22] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 1

[23] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11144–11154, 2023. 2

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[25] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic seg-

mentation. In *International Conference on Learning Representations*, 2022. 2

[26] Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, and Dan Zhang. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11809–11818, 2022. 2

[27] Leyao Liu, Tian Zheng, Yun-Jou Lin, Kai Ni, and Lu Fang. Ins-conv: Incremental sparse convolution for online 3d segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18975–18984, 2022. 2

[28] 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 *Conference on Robot Learning*, pages 1610–1620. PMLR, 2023. 1, 2, 6, 8

[29] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4205–4212. IEEE, 2019. 2

[30] Phuc DA Nguyen, Tuan Duc Ngo, Chuang Gan, Evangelos Kalogerakis, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. *arXiv preprint arXiv:2312.10671*, 2023. 2

[31] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4047–4056, 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, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5

[34] Damien Robert, Bruno Vallet, and Loic Landrieu. Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5575–5584, 2022. 2

[35] 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. 1, 2, 6

[36] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 2, 6

[37] 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 (NeurIPS)*, 2023. 1, 2, 5, 6, 8

[38] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pages 298–372. Springer, 2000. 2

[39] VS Vibashan, 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 manual mask annotations. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23539–23549, 2023. 1, 2

[40] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021. 2

[41] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, Junyeong Kim, and Chang D Yoo. Softgroup++: Scalable 3d instance segmentation with octree pyramid grouping. *arXiv preprint arXiv:2209.08263*, 2022. 2

[42] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 2

[43] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. *ArXiv*, abs/2310.15308, 2023. 1, 2

[44] Jianzong Wu, Xiangtai Li, Henghui Ding, Xia Li, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21938–21948, 2023. 1, 2

[45] 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. *arXiv preprint arXiv:2310.03923*, 2023. 1

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

[47] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 2, 6

[48] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any in-

stance in 3d scenes. *arXiv preprint arXiv:2312.11557*, 2023. 2

[49] Jiazhao Zhang, Chenyang Zhu, Lintao Zheng, and Kai Xu. Fusion-aware point convolution for online semantic 3d scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4534–4543, 2020. 2

[50] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6672–6682, 2023. 1

[51] Lintao Zheng, Chenyang Zhu, Jiazhao Zhang, Hang Zhao, Hui Huang, Matthias Niessner, and Kai Xu. Active scene understanding via online semantic reconstruction. In *Computer Graphics Forum*, pages 103–114. Wiley Online Library, 2019. 2

[52] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 2

[53] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 2