

SGS-3D: High-Fidelity 3D Instance Segmentation via Reliable Semantic Mask Splitting and Growing

Chaolei Wang¹, Yang Luo¹, Jing Du^{2*}, Siyu Chen^{3*}, Yiping Chen^{1†}, Ting Han^{1†}

¹Sun Yat-sen University, Zhuhai, China

²University of Waterloo, Waterloo, Canada

³Jimei University, Xiamen, China

{wangchlei5, hant23}@mail2.sysu.edu.cn, chenyp79@mail.sysu.edu.cn

Abstract

Accurate 3D instance segmentation is crucial for high-quality scene understanding in the 3D vision domain. However, 3D instance segmentation based on 2D-to-3D lifting approaches struggle to produce precise instance-level segmentation, due to accumulated errors introduced during the lifting process from ambiguous semantic guidance and insufficient depth constraints. To tackle these challenges, we propose Splitting and Growing reliable Semantic mask for high-fidelity 3D instance segmentation (SGS-3D), a novel "split-then-grow" framework that first purifies and splits ambiguous lifted masks using geometric primitives, and then grows them into complete instances within the scene. Unlike existing approaches that directly rely on raw lifted masks and sacrifice segmentation accuracy, SGS-3D serves as a training-free refinement method that jointly fuses semantic and geometric information, enabling effective cooperation between the two levels of representation. Specifically, for semantic guidance, we introduce a mask filtering strategy that leverages the co-occurrence of 3D geometry primitives to identify and remove ambiguous masks, thereby ensuring more reliable semantic consistency with the 3D object instances. For the geometric refinement, we construct fine-grained object instances by exploiting both spatial continuity and high-level features, particularly in the case of semantic ambiguity between distinct objects. Experimental results on ScanNet200, ScanNet++, and KITTI-360 demonstrate that SGS-3D substantially improves segmentation accuracy and robustness against inaccurate masks from pre-trained models, yielding high-fidelity object instances while maintaining strong generalization across diverse indoor and outdoor environments.

Code — <https://github.com/wangchaolei7/SGS-3D>

Introduction

Arbitrary instances segmentation within 3D scenes constitutes a crucial task in various domains, including autonomous driving, virtual reality, and multi-modality scene understanding. While recent methods (Felzenszwalb 2004; Chen et al. 2021; Vu et al. 2022; Liang et al. 2021; Engelmann et al. 2020; Lu et al. 2023a; Schult et al. 2023; Yu

et al. 2023; Han et al. 2024; Luo et al. 2025) yield impressive segmentation results when trained on specifically annotated datasets (Dai et al. 2017; Armeni et al. 2016; Robert, Raguet, and Landrieu 2023; Yeshwanth et al. 2023; Liao, Xie, and Geiger 2022), their generalization to open-world environments remains limited. In contrast to the inherent difficulties in obtaining extensive 3D annotations, foundation models trained on massive 2D image datasets have shown remarkable performance and strong zero-shot generalization abilities. This observation motivates a promising direction: *utilizing 2D pre-trained foundation models for 3D scene perception and interaction*.

Currently, two predominant paradigms exist for executing the 3D instance segmentation: feature-based methods and mask-based methods. Feature-based methods (Peng et al. 2023; Takmaz et al. 2023; Ding et al. 2023; Hegde, Valanarasu, and Patel 2023; Yang et al. 2024; Lee, Zhao, and Lee 2024; Huang et al. 2024b) primarily focus on learning robust feature representations from 3D point clouds or by lifting features from 2D modalities (Oquab et al. 2023). Instance segmentation is subsequently achieved by classifying, grouping, or decoding these learned features into instance-level masks. However, these methods commonly suffer from inefficient training and error propagation, which arising from ambiguous feature representations stemming from 2D-to-3D lifting or the inherent high-dimensional feature space.

To mitigate these concerns, mask-based methods leverage the semantic masks extracted from input registered images using Segment Anything Model (SAM) (Kirillov et al. 2023) to derive instance-level assignments for each 3D geometric primitive in the 3D scene, consolidating these 3D primitives iteratively based on their projected masks (Yang et al. 2023; Lu et al. 2023b; Xu et al. 2023; Nguyen et al. 2024; Huang et al. 2024a). Despite this advancement, the robust propagation of semantic information from these projected masks is often overlooked, with a primary focus on optimizing strategies for merging 3D geometric primitives while neglecting fine-grained geometric cues. This oversight leads to continuous error accumulation during scene segmentation, ultimately yielding imprecise results, particularly for adjacent but separate instances, as illustrated in Figure 1. Although some existing methods incorporate depth information from sensors to enhance back-projection accuracy (Guo

*These authors contributed equally.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

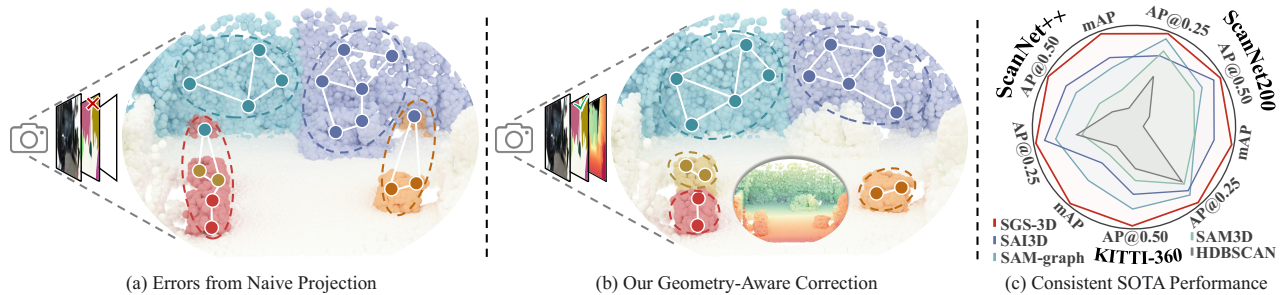


Figure 1: SGS-3D: High-Fidelity segmentation by overcoming ambiguous 2D-to-3D lifting. Previous methods suffer from flawed instance grouping (a), caused by ambiguous 2D semantics and inadequately unhandled occlusions during projection. SGS-3D overcomes this by establishing reliable 3D semantics via occlusion-aware mapping and a novel ”split-then-grow” refinement (b). This dual-refinement strategy achieves state-of-the-art accuracy across diverse scenes, especially in challenging, depth-less outdoor environments (c).

et al. 2024; Yin et al. 2024; Zhao et al. 2025), this reliance is problematic for two key reasons: depth sensors inherently struggle on textureless and highly reflective surfaces, and moreover, they are often entirely unavailable in wild scenes, severely compromising segmentation performance.

To confront these issues, we propose SGS-3D, a training-free refinement and segmentation framework that jointly fuse semantic and geometric information to accurately register 2D semantic instance masks to 3D geometric primitives. Inspired by prior work, we introduce 3D geometric over-segmentation to decompose objects within the scene into primitives, which serve as the basic units for instance segmentation. Moreover, we reveal the significance of precise depth constraints, which serve as a powerful tool for propagating reliable and fine-grained semantic cues to 3D point clouds. Specifically, during the mask refinement phase, we first effectively establish a point-to-pixel mapping, without relying on true depth maps and without sacrificing precision. Leveraging this mapping, the co-occurrence of 3D geometric primitives effectively prevents the accumulation of ambiguous semantic errors. In the geometry refinement phase, we split 3D semantic masks within the density space to further refine semantic guidance and maintain semantic quality. After geometric optimization, our method consolidates the refined 3D semantic masks with geometric primitives by performing feature similarity matching, which enhances the completeness of object instances while reducing redundancy. To summarize our contributions in a few words:

- We propose SGS-3D, a novel training-free framework that achieves high-fidelity 3D instance segmentation through a principled ”split-then-grow” strategy, jointly leveraging semantic and geometric cues to overcome error accumulation in existing 2D-to-3D lifting approaches.
- We introduce an occlusion-aware point-image mapping that ensures accurate mask-to-point correspondence without ground-truth depth, and develop a visibility-based co-occurrence filtering that effectively removes ambiguous 2D masks while achieving $4\times$ computational speedup and robust cross-scene generalization.

- We design a semantic-guided aggregation pipeline featuring density-based spatial splitting to generate pure semantic-geometric seeds, followed by feature-guided growing that intelligently aggregates fragmented instances into complete objects, particularly excelling in challenging interwoven scenarios.

Methodology

Method Overview and Preliminaries

The proposed SGS-3D comprises three main components: (1) Point-Image Mapping, (2) 2D Mask Proposal Module, and (3) Semantic-Guided Aggregation, as illustrated in Figure 2. Our primary objective is to generate a set of high-quality, class-agnostic 3D instance proposals by leveraging semantic and spatial cues from multiple 2D views. Our method takes as input a 3D point cloud $P = \{p_i \in \mathbb{R}^6\}_{i=1}^N$ (XYZRGB) and a set of T registered images $\{I_t\}_{t=1}^T$ with their corresponding camera parameters, including intrinsics \mathbf{K}_t and pose \mathbf{T}_t .

We first over-segment the point cloud P into a set of U superpoints, $\mathcal{S} = \{s_u\}_{u=1}^U$, which serve as our fundamental processing primitives. These spatially coherent and geometrically compact groups of points improve computational efficiency and help maintain structural consistency. To adapt to diverse 3D scenarios, we employ method in (Dai et al. 2017) for indoor mesh inputs and (Robert, Raguét, and Landrieu 2023) for outdoor unstructured point clouds. Furthermore, for each superpoint s_u , we assume access to a pre-computed feature vector $\mathbf{f}_u \in \mathbb{R}^D$, derived from a pre-trained segmentation model (Schult et al. 2023), which captures its high-dimensional feature.

Point-Image Mapping

We start by establishing a robust and efficient mapping between the points in P and the images in I . A point-image pair (p, I_t) is considered compatible if the point p is visible in the image I_t . That is, p lies within the camera’s viewing frustum and is not occluded by any other point that is closer to the camera. For each compatible pair, we define

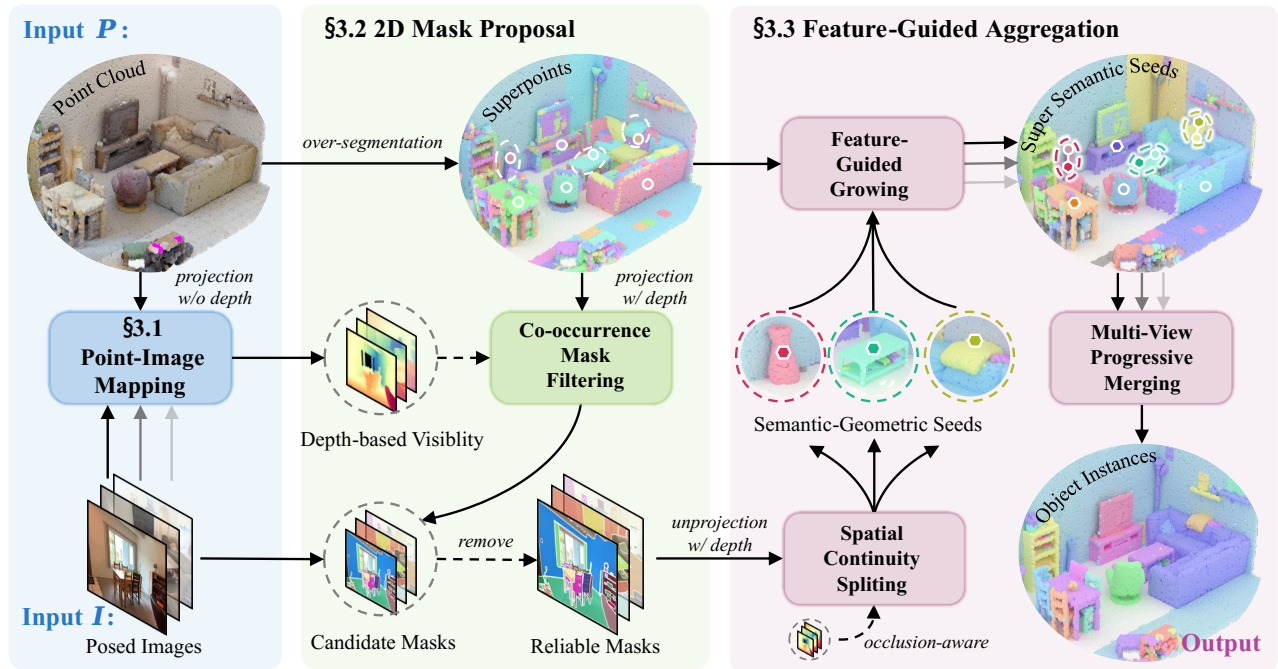


Figure 2: Overview of the training-free SGS-3D pipeline. Our method begins by computing robust, occlusion-aware point-image mappings without requiring ground-truth depth (§ 3.1). In the 2D Mask Proposal stage (§ 3.2), these mappings guide a Co-occurrence Mask Filtering process to prune ambiguous candidate masks, yielding a set of reliable 2D masks. These are then lifted to 3D and fed into the Feature-Guided Aggregation stage (§ 3.3), which first uses Spatial Continuity Splitting to generate pure semantic-geometric seeds. Subsequently, Feature-Guided Growing expands these seeds into complete instances, which are finally consolidated via Multi-View Progressive Merging to produce the final, high-fidelity object instances.

its re-projection $\text{pix}(p, I_t)$ as the 2D pixel coordinate in image I_t where the point p is projected. However, computing this mapping is non-trivial, especially in multi-view scenarios where occlusions are common.

Depth Construction In contrast to existing methods that rely on ground-truth depth maps from dedicated sensors, we propose an efficient mapping strategy to compute visibility directly from the point cloud and camera parameters using Z-buffering (Ravi et al. 2020). For a given image I_t with its associated camera parameters $(\mathbf{K}_t, \mathbf{T}_t)$, the entire point cloud P is projected into the camera’s image plane. The transformation π maps a homogeneous world point p_i to pixel coordinates (u_i, v_i) with depth z_i , according to $z_i[u_i, v_i, 1]^T = \mathbf{K}_t \mathbf{T}_t^{-1} \mathbf{p}_i$. To resolve occlusions, a depth buffer $\mathbf{D}_t \in \mathbb{R}^{H \times W}$ is initialized with infinite values. The final value at each pixel (u, v) is populated via a highly parallel rasterization process on the GPU, which efficiently solves for the minimum depth over the set of all points projecting to that location:

$$\mathbf{D}_t(u, v) = \min\{z_i \mid p_i \in P, \pi(p_i, \mathbf{K}_t, \mathbf{T}_t) = (u, v)\}. \quad (1)$$

Viewing Visibility Verification Subsequent to the generation of the metric depth buffer \mathbf{D}_t , we finalize the visibility status for each point. The per-point visibility for a view t is encapsulated in a binary tensor $\mathcal{V}_t \in \{0, 1\}^N$, whose elements are computed via a composite frustum and occlusion

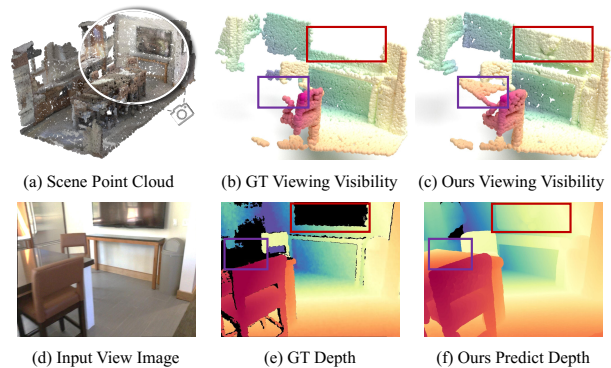


Figure 3: Our method constructs valid visibility mapping for each point. While depth sensors struggle on textureless and highly reflective surfaces, our approach remains effective.

check:

$$\mathcal{V}_t(i) = \mathbf{1}(0 \leq u_i < W \wedge 0 \leq v_i < H) \cdot \mathbf{1}(|z_i - \mathbf{D}_t(u_i, v_i)| \leq \tau_{vis}), \quad (2)$$

where (u_i, v_i) and z_i are the projected pixel coordinates and true computed camera-space depth of point p_i , respectively, $\mathbf{1}(\cdot)$ is the indicator function, and τ_{vis} is the depth congruence tolerance. Simultaneously, we maintain a definitive point-to-pixel correspondence map $\mathcal{M}_{pix,t} \in \mathbb{Z}^{N \times 3}$, which

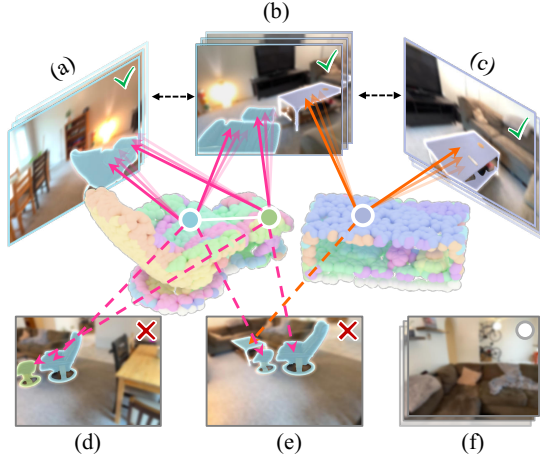


Figure 4: Co-occurrence mask filtering strategy. Co-occurrence scores between superpoints are constructed from accurate 2D mask sets (a-c). Over-segmented (d) and under-segmented (e) masks exhibiting low scores are then removed from the candidate masks list, where image sets with $\mathcal{P}_{vis,m}^j = 0$ (f) are excluded from the calculation.

stores the valid mapping $[u_i, v_i, 1]^T$ for each point where $\mathcal{V}_t(i) = 1$. This verification yields the precise, per-view mapping essential for subsequent sections, as illustrated in Figure 3.

2D Mask Proposal

Mask Generation Mask-based 3D instance segmentation involves generating masks for the target objects based on input images. For each image I_t in a scene sequence $I = \{I_t\}_{t=1}^T$, the objective is to produce a corresponding set of binary masks, $\mathcal{M}_{2D,t} = \{M_{t,k}\}_{k=1}^{K_t}$. With the emergence of foundational models like SAM (Kirillov et al. 2023), mask prediction has become highly precise. However, SAM lacks targeted guidance and often segments objects that are not of interest. Therefore, we apply Grounding-DINO (Liu et al. 2024) to extract box prompts from the images with text prompts. The high-confidence box prompts among them are then used to generate the initial mask predictions.

Co-occurrence Mask Filtering Through introducing text prompts via Grounding-DINO, the number of semantically ambiguous masks should progressively decline. However, experiments reveal that although the quantity of major ambiguous masks reduces, some minor ambiguous ones still remain.

In fact, the binary masks $\mathcal{M}_{2D,t}$ predicted by the trained 2D vision foundation model exhibit discreteness, which often leads to inaccuracies and inconsistencies in object boundary area predictions across different views. This limitation may increase the instability of the semantic guidance of 2D masks. Unlike Maskclustering (Yan et al. 2024), which purely relies on image pixel consistency between views, we leverage viewing visibility of superpoints to exclusively enhance robustness against inaccurate masks.

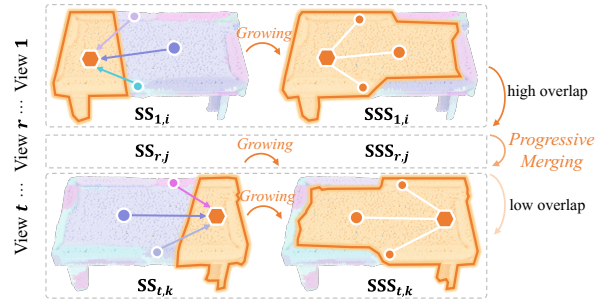


Figure 5: Semantic-guided aggregation. Within a single view, a Semantic-geometric Seed (SS) (orange) is expanded into a Super Semantic Seed (SSS) by merging with neighboring superpoints (colorful). Subsequently, these proposals from other views are progressively merged to form the final object instance.

Specifically, for superpoints belonging to the same instance, we accumulate their mutual visibility across multiple masks and views in Figure 4. We leverage our pre-computed point-to-pixel mappings $\mathcal{M}_{pix,t} \in \mathbb{Z}^{N \times 3}$ to establish the association between a superpoint and a 2D mask. A point p_i is considered to fall within a mask M_m in view j if it is physically visible ($\mathcal{V}_j(i) = 1$) and its projected pixel coordinates (u_i, v_i) lie inside the mask area. Let the set of all superpoints in the scene be $\mathcal{S} = \{s_u\}_{u=1}^U$. We define the association between a mask M_m and the superpoints it makes visible in view j , denoted as the set $\mathcal{P}_{vis,m}^j$:

$$\mathcal{P}_{vis,m}^j = \{s_u \in \mathcal{S} \mid \frac{|\{p_i \in s_u \mid \mathcal{V}_j(i)=1 \wedge M_m(\pi_j(p_i))=1\}|}{|s_u|} > 0.5\}. \quad (3)$$

Based on this, we define the visibility co-occurrence score c_m for each mask, which measures its average consistency with all other masks in the candidate set:

$$c_m = \frac{1}{(K_{2D} - 1)T} \sum_{n=1, n \neq m}^{K_{2D}} \sum_{j=1}^T \frac{|\mathcal{P}_{vis,m}^j \cap \mathcal{P}_{vis,n}^j|}{\sqrt{|\mathcal{P}_{vis,m}^j| \cdot |\mathcal{P}_{vis,n}^j|}}, \quad (4)$$

where $\mathcal{P}_{vis,m}^j$ represents the set of visible superpoints for mask M_m in view j , T is the total number of views, and K_{2D} is the total number of masks in the candidate set. This principled score allows us to rank masks by their cross-view consistency. To prevent error propagation, masks exhibiting low co-occurrence scores, which indicate inconsistent visibility patterns, are identified as outliers and pruned from the candidate set.

Semantic-Guided Aggregation

Existing methods often compromise segmentation quality by either directly unprojecting 2D masks onto the point cloud (Yang et al. 2023; Yan et al. 2024) or by solely using them as a semantic prior to coarsely guide the iterative merging of superpoints (Yin et al. 2024; Nguyen et al. 2024). In such rough methods, the interplay between semantic and geometric information is not deeply exploited and leads to

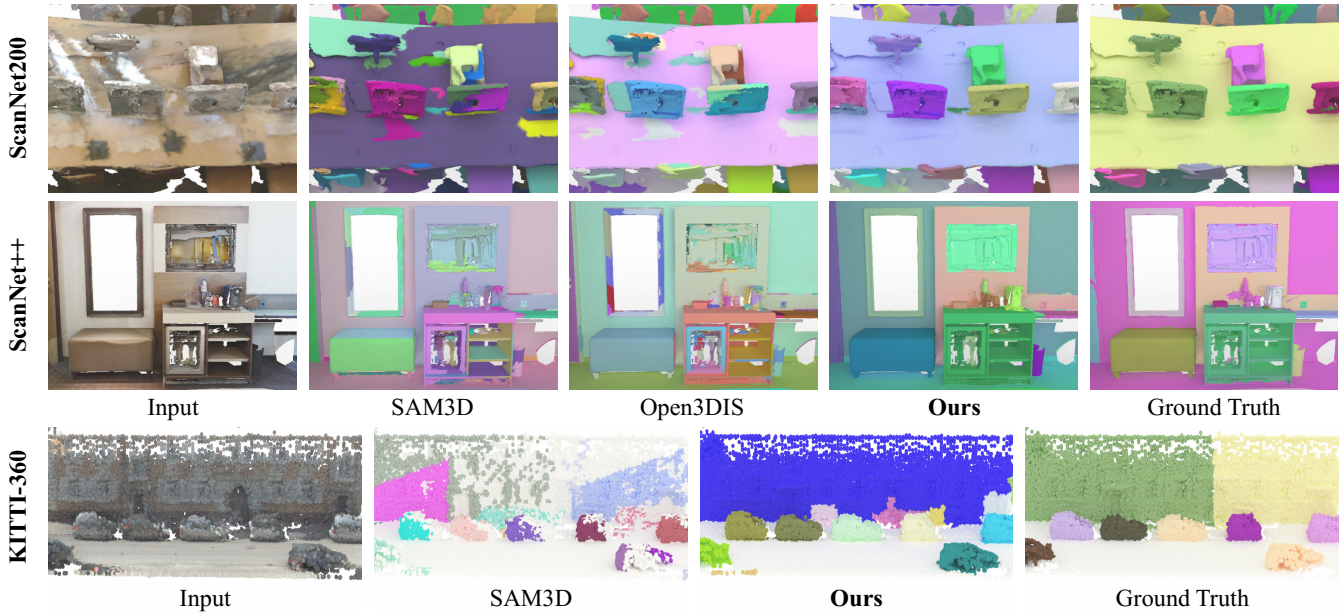


Figure 6: Qualitative comparison on indoor and outdoor datasets. Our method obtains more accurate segmentation results and significantly reduces over-segmentation or fragmented instances compared to previous method.

over-segmentation of object instances. To solve this problem, our unique insight is that leveraging reliable 3D semantics can enhance 3D segmentation and improve the robustness of instance aggregation.

Spatial Continuity Splitting Leveraging the occlusion-aware viewing visibility, we lift the filtered 2D masks from each view t onto point cloud. These point sets, aggregated by their shared semantic identity, are termed 3D semantic masks, $\mathcal{M}_{3D,t} = \{M_{t,k}\}_{k=1}^{K_t}$. While the 2D semantic consistency derived from vision models has proven effective in prior work, the intrinsic geometric coherence of the 3D point cloud itself is often overlooked. We experimentally find that objects with similar appearances are erroneously grouped into a single instance, despite being spatially distinct in the point cloud (see supplementary, Figure 10). As a solution, we apply HDBSCAN (Campello, Moulavi, and Sander 2013) on each of the 3D semantic masks ($M_{t,k} \in \mathcal{M}_{3D,t}$) which splits them into new clusters based on spatial contiguity. We refer to these dense clusters as semantic-geometric seeds, forming a new collection of fine-grained masks $\mathcal{M}_{seed,t} = \{M'_{t,j}\}_{j=1}^{K'_t}$. This density-based splitting utilizes geometric properties to refine the semantics of the segmentation results, ensuring a fine-grained distinction between instances and mitigating the propagation of ambiguous semantics. The effectiveness of this geometric refinement in resolving semantic ambiguity is significant, as our ablation study shows it single-handedly surpasses existing state-of-the-art methods (see Table 2).

Single-View Feature-Guided Growing While the preceding density-based splitting excels at purifying seeds, producing $\mathcal{M}_{seed,t}$, the challenge is ensembling these over-

segmented superpoints into complete instances. To address this, we introduce a feature-guided growing process. We leverage rich pre-trained features (Schult et al. 2023) to bridge the spatial gaps between these semantic-geometric seeds and their corresponding object parts.

Rather than applying rigid thresholds, our growing process is guided by a unified affinity score that naturally balances semantic coherence with spatial adjacency, as illustrated in Figure 5. For a given seed $M'_{t,j} \in \mathcal{M}_{seed,t}$ and a candidate neighbor superpoint s_u , we define this score as:

$$\text{Affinity}(M'_{t,j}, s_u) = \text{sim}(\bar{\mathbf{f}}_{t,j}, \mathbf{f}_u) \cdot \text{Overlap}(M'_{t,j}, s_u), \quad (5)$$

where $\text{sim}(\bar{\mathbf{f}}_{t,j}, \mathbf{f}_u)$ is the cosine similarity between the seed’s mean feature and the candidate’s mean feature, and $\text{Overlap}(\cdot, \cdot)$ measures their spatial overlap, calculated as the IoU of their 3D point sets. This multiplicative form ensures that a high affinity is only achieved when both semantic similarity and spatial overlap are strong.

At each expansion step, we merge the unassigned neighbor with the highest affinity score into the seed. This iterative process continues until the affinity scores of all remaining neighbors become negligible. By optimizing for the highest affinity, object instances can correctly associate disparate parts and expand their boundaries for the final aggregation stage.

Multi-View Progressive Merging The single-view growing process generates high-fidelity instance proposals, which are essentially the grown semantic-geometric seeds from $\mathcal{M}_{seed,t}$. However, each proposal is inherently limited to the information available from its originating viewpoint. To reconcile these partial views into complete objects, we employ a progressive, multi-view merging strategy. Inspired

Methods	Proc. & Year	VFM	ScanNet200			ScanNet++			KITTI-360		
			mAP	AP ₅₀	AP ₂₅	mAP	AP ₅₀	AP ₂₅	mAP	AP ₅₀	AP ₂₅
<i>training-dependent</i>											
UnScene3D	CVPR24	DINO	15.9	32.2	58.5	-	-	-	-	-	-
Segment3D	ECCV24	SAM	-	-	-	12.0	22.7	37.8	-	-	-
SAM-graph	ECCV24	SAM	22.1	41.7	<u>62.8</u>	15.3	27.2	44.3	<u>23.8</u>	<u>37.2</u>	<u>49.1</u>
<i>training-free</i>											
HDBSCAN	ICDMW17	-	2.9	8.2	33.1	4.3	10.6	32.3	9.3	18.9	39.6
Felzenszwalb	IJCV04	-	4.8	9.8	27.5	4.1	9.2	25.3	-	-	-
SAM3D	ICCVW23	SAM	20.9	34.8	51.4	9.3	16.6	29.5	13.0	24.2	41.1
Open3DIS	CVPR24	GD-SAM	29.7	45.2	56.8	18.5	35.5	44.3	-	-	-
SAI3D	CVPR24	SAM	28.2	<u>47.2</u>	48.5	17.1	31.1	49.5	16.5	30.2	45.6
SAM2Object	CVPR25	SAM2	-	-	-	20.2	34.1	48.7	-	-	-
Ours⁻	-	GD-SAM	<u>30.5</u>	<u>47.2</u>	62.2	<u>22.5</u>	<u>36.6</u>	<u>53.0</u>	32.9 _{+16.4}	43.7 _{+13.5}	53.4 _{+7.8}
Ours⁺	-	GD-SAM	34.3 _{+4.6}	51.3 _{+4.1}	64.6 _{+7.8}	23.7 _{+3.5}	39.6 _{+4.1}	54.3 _{+4.8}	N/A	N/A	N/A

Table 1: Quantitative results compared with conventional clustering methods and diverse 2D-to-3D lifting methods on ScanNet200, ScanNet++ and KITTI-360. "VFM" denotes the vision foundation models. Ours⁻ and Ours⁺ represent our method without and with ground-truth (GT) depth, respectively. N/A indicates results are not applicable as KITTI-360 lacks the required GT depth. All 2D-to-3D lifting baselines use GT depth in indoor datasets. Best results are in **bold** (%), and the second-best are underlined.

No.	CMF	SCP	FGG	mAP	AP ₅₀	AP ₂₅
1				25.9	41.9	57.3
2	✓			27.8	44.0	59.0
3	✓	✓		30.1	46.8	61.4
4	✓		✓	29.7	45.2	58.9
5	✓	✓	✓	34.3	51.3	64.6

Table 2: Ablation of the effectiveness of our proposed components. CMF indicates the co-occurrence mask filtering; SCP indicates the spatial continuity Splitting; FGG indicates the feature-guided growing.

by (Yin et al. 2024), we initially merge proposals from different views only if their 3D spatial overlap is very high, ensuring only unambiguous matches are made. Subsequent iterations systematically relax this overlap requirement, allowing fragmented parts of the same large instance to be fused. This hierarchical process robustly assembles the final, complete 3D object instances for the entire scene, effectively consolidating information from multiple views.

Experiments

Experimental Setup

We evaluate our method on three prevalent benchmarks: ScanNet200 (Dai et al. 2017), ScanNet++ (Yeshwanth et al. 2023) for indoor scenes, and KITTI-360 (Liao, Xie, and Geiger 2022) for outdoor environments. Following established protocols (Rozenberszki, Litany, and Dai 2024; Schult et al. 2023; Takmaz et al. 2023), we compute AP scores at 50% (AP₅₀) and 25% (AP₂₅) intersection-over-union (IoU) thresholds, as well as the mean AP (mAP) averaged over 50% to 95% IoU thresholds in 5% increments.

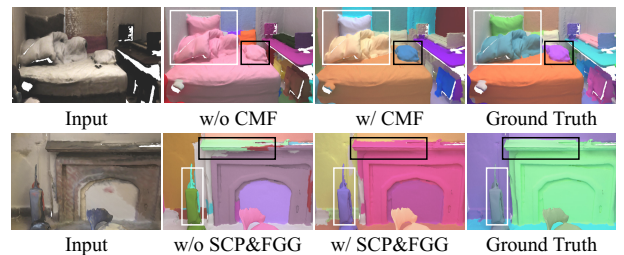


Figure 7: Ablation visualization results. Top: 2D mask proposal section; Bottom: semantic-guided aggregation section.

Analysis Experiments

Quantitative Results As shown in Table 1, SGS-3D achieves SOTA performance among training-free methods. On the challenging, depth-less KITTI-360 dataset, Ours⁻ establishes a remarkable +16.4% mAP lead over the next best competitor, SAI3D, validating our method’s robustness in outdoor scenes. On indoor benchmarks, Ours⁻ also consistently surpasses prior SOTA methods on ScanNet200 and ScanNet++. Furthermore, with GT depth, Ours⁺ sets a new performance ceiling (e.g., 34.3% mAP on ScanNet200), outperforming all training-free and even several training-dependent approaches, showcasing our framework’s strong potential.

Qualitative Results Figure 6 shows our method produces visually superior results with more complete and geometrically precise instances. This stems from our framework’s ability to mitigate semantic ambiguity through mask filtering and spatial splitting, yielding cleaner and more coherent segmentations.

Methods	Proc. & Year	3D Mask	mAP	AP ₅₀	AP ₂₅	mAP _{head}	mAP _{common}	mAP _{tail}
OpenMask3D	NeurIPS23	Supervised	15.4	19.9	23.1	17.1	14.1	14.9
OpenIns3D	ECCV24	Supervised	15.9	20.6	23.3	19.2	14.2	14.2
OVIR-3D	CoRL23	Zero-shot	9.3	18.7	<u>25.0</u>	-	-	-
SAM3D	ICCVW23	Zero-shot	9.8	15.2	<u>20.7</u>	9.2	8.3	12.3
SAI3D	CVPR24	Zero-shot	12.7	18.8	24.1	<u>12.1</u>	<u>10.4</u>	<u>16.2</u>
SAM2Object	CVPR25	Zero-shot	<u>13.3</u>	<u>19.0</u>	23.8	-	-	-
Ours	-	Zero-shot	21.1 _{+7.8}	29.4 _{+10.4}	35.0 _{+10.0}	21.6 _{+9.5}	20.0 _{+9.6}	22.0 _{+5.8}

Table 3: Semantic results on ScanNet200. Head, common, and tail categories represent object classes with high, medium, and low frequencies in the dataset, respectively. OpenMask3D (Takmaz et al. 2023) and OpenIns3D (Huang et al. 2024b) requires supervised training on ScanNet200. In fully supervised and zero-shot setting, our method surpass previous SOTA methods.

Method	Img. Prop.	mAP	AP ₅₀	AP ₇₅	Time	Inst. Pred.
Open3DIS	10%	29.7	45.2	56.8	9.18s	102
Ours	10%	34.3	51.3	64.6	9.51s	62
	5%	33.7	51.1	64.5	4.94s	66
	2.5%	31.8	48.8	63.2	2.42s	68

Table 4: Influence of input image proportion on segmentation efficiency and performance. Inst. Pred. indicates predicted instance counts per scene.

VFM	mAP	AP ₅₀	AP ₂₅	c_m	mAP	AP ₅₀	AP ₂₅
Cropformer	<u>33.6</u>	<u>50.7</u>	63.1	0.4	33.8	50.2	63.1
SAM	32.3	47.8	<u>63.8</u>	0.3	34.0	50.9	63.2
YoloW-SAM	32.4	49.5	<u>62.3</u>	0.2	34.3	51.6	64.6
GD-SAM	34.3	51.3	64.6	0.1	32.0	48.3	61.8

Table 5: Ablation of different 2D vision models.

Table 6: Ablation of co-occurrence score c_m .

Ablation Studies Table 2 details the contribution of each component. CMF provides an initial boost. Critically, adding SCP without growing achieves 30.1% mAP, already surpassing SOTA methods like SAM-graph and proving the power of our semantic purification. Finally, integrating FGG elevates performance to our full model’s 34.3% mAP, confirming a powerful synergy where our purification stages create high-quality seeds for FGG to complete (Figure 7).

Efficiency and Robustness Table 4 highlights SGS-3D’s exceptional efficiency. Our method achieves superior accuracy to Open3DIS while using only 2.5% of the images (vs. 10%) and being nearly four times faster. It also significantly reduces over-segmentation by 40%.

Our pipeline demonstrates robustness across diverse vision foundation models (VFMs, Table 5) (Qi et al. 2022; Cheng et al. 2024) and is insensitive to the co-occurrence threshold c_m within a reasonable range (Table 6). As shown in Table 7, using a protocol inspired by (Naseer et al. 2021), our method is also resilient to occlusion, maintaining strong performance even when up to 50% of foreground masks are removed.

Percentage (%)	0	5	10	30	50	60	70	90
mAP	34.3	34.1	34.1	33.4	29.6	24.2	14.8	0.1

Table 7: Robustness to simulated occlusion.



Figure 8: Open-vocabulary 3D object search. Given a text prompt, our method can accurately locate the corresponding object instances (pink) in a 3D scene.

Open-Set Scene Understanding Application

Our high-quality, class-agnostic proposals form a strong basis for open-set understanding. By labeling them with the off-the-shelf vision-language models (Radford et al. 2021; Zhai et al. 2023), we extend our method to open-vocabulary 3D segmentation. As shown in Table 3, our geometrically precise instance masks reduce semantic ambiguity, enabling SGS-3D to significantly outperform previous zero-shot methods and support applications like text-based object search within complex 3D environments, as illustrated in Figure 8.

Conclusion

We have presented SGS-3D, a training-free framework for high-fidelity 3D instance segmentation. Our ”split-then-grow” strategy first purifies noisy proposals into high-quality 3D seeds using geometric and spatial cues, then grows them into complete instances via a feature-guided process. Experiments show SGS-3D substantially outperforms SOTA methods, particularly on challenging depthless dataset. Our work provides a robust bridge between 2D semantics and 3D geometry, advancing class-agnostic segmentation and open-set understanding. Future directions include extending our framework to dynamic scenes and optimizing its multi-stage design for real-time performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 42371343, and in part by the Basic and Applied Basic Research Foundation of Guangdong Province under Grant 2024A1515010986.

References

- Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1534–1543.
- Campello, R. J.; Moulavi, D.; and Sander, J. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, 160–172. Springer.
- Chen, S.; Fang, J.; Zhang, Q.; Liu, W.; and Wang, X. 2021. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15467–15476.
- Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16901–16911.
- Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.
- Ding, R.; Yang, J.; Xue, C.; Zhang, W.; Bai, S.; and Qi, X. 2023. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7010–7019.
- Engelmann, F.; Bokeloh, M.; Fathi, A.; Leibe, B.; and Nießner, M. 2020. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9031–9040.
- Felzenszwalb, e. a. 2004. Efficient graph-based image segmentation. *International journal of computer vision*, 59: 167–181.
- Guo, H.; Zhu, H.; Peng, S.; Wang, Y.; Shen, Y.; Hu, R.; and Zhou, X. 2024. Sam-guided graph cut for 3d instance segmentation. In *European Conference on Computer Vision*, 234–251. Springer.
- Han, T.; Chen, Y.; Ma, J.; Liu, X.; Zhang, W.; Zhang, X.; and Wang, H. 2024. Point cloud semantic segmentation with adaptive spatial structure graph transformer. *International Journal of Applied Earth Observation and Geoinformation*, 133: 104105.
- Hegde, D.; Valanarasu, J. M. J.; and Patel, V. 2023. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2028–2038.
- Huang, R.; Peng, S.; Takmaz, A.; Tombari, F.; Pollefeys, M.; Song, S.; Huang, G.; and Engelmann, F. 2024a. Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels. In *European Conference on Computer Vision*, 278–295. Springer.
- Huang, Z.; Wu, X.; Chen, X.; Zhao, H.; Zhu, L.; and Lasenby, J. 2024b. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *European Conference on Computer Vision*, 169–185. Springer.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.
- Lee, S.; Zhao, Y.; and Lee, G. H. 2024. Segment any 3d object with language. *arXiv preprint arXiv:2404.02157*.
- Liang, Z.; Li, Z.; Xu, S.; Tan, M.; and Jia, K. 2021. Instance segmentation in 3d scenes using semantic superpoint tree networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2783–2792.
- Liao, Y.; Xie, J.; and Geiger, A. 2022. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3292–3310.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55. Springer.
- Lu, J.; Deng, J.; Wang, C.; He, J.; and Zhang, T. 2023a. Query refinement transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18516–18526.
- Lu, S.; Chang, H.; Jing, E. P.; Boularias, A.; and Bekris, K. 2023b. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *Conference on Robot Learning*, 1610–1620. PMLR.
- Luo, Y.; Han, T.; Liu, Y.; Su, J.; Chen, Y.; Li, J.; Wu, Y.; and Cai, G. 2025. CSFNet: Cross-modal Semantic Focus Network for Semantic Segmentation of Large-Scale Point Clouds. *IEEE Transactions on Geoscience and Remote Sensing*.
- Naseer, M. M.; Ranasinghe, K.; Khan, S. H.; Hayat, M.; Shahbaz Khan, F.; and Yang, M.-H. 2021. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34: 23296–23308.
- Nguyen, P.; Ngo, T. D.; Kalogerakis, E.; Gan, C.; Tran, A.; Pham, C.; and Nguyen, K. 2024. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4018–4028.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of*

- the *IEEE/CVF conference on computer vision and pattern recognition*, 815–824.
- Qi, L.; Kuen, J.; Guo, W.; Shen, T.; Gu, J.; Jia, J.; Lin, Z.; and Yang, M.-H. 2022. High-quality entity segmentation. *arXiv preprint arXiv:2211.05776*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Ravi, N.; Reizenstein, J.; Novotny, D.; Gordon, T.; Lo, W.-Y.; Johnson, J.; and Gkioxari, G. 2020. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*.
- Robert, D.; Raguét, H.; and Landrieu, L. 2023. Efficient 3d semantic segmentation with superpoint transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17195–17204.
- Rozenberszki, D.; Litany, O.; and Dai, A. 2024. Unscene3d: Unsupervised 3d instance segmentation for indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19957–19967.
- Schult, J.; Engelmann, F.; Hermans, A.; Litany, O.; Tang, S.; and Leibe, B. 2023. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 8216–8223. IEEE.
- Takmaz, A.; Fedele, E.; Sumner, R. W.; Pollefeys, M.; Tombari, F.; and Engelmann, F. 2023. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*.
- Vu, T.; Kim, K.; Luu, T. M.; Nguyen, T.; and Yoo, C. D. 2022. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2708–2717.
- Xu, M.; Yin, X.; Qiu, L.; Liu, Y.; Tong, X.; and Han, X. 2023. Sampro3d: Locating sam prompts in 3d for zero-shot scene segmentation. *arXiv preprint arXiv:2311.17707*.
- Yan, M.; Zhang, J.; Zhu, Y.; and Wang, H. 2024. 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*, 28274–28284.
- Yang, J.; Ding, R.; Deng, W.; Wang, Z.; and Qi, X. 2024. 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*, 19823–19832.
- Yang, Y.; Wu, X.; He, T.; Zhao, H.; and Liu, X. 2023. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*.
- Yeshwanth, C.; Liu, Y.-C.; Nießner, M.; and Dai, A. 2023. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12–22.
- Yin, Y.; Liu, Y.; Xiao, Y.; Cohen-Or, D.; Huang, J.; and Chen, B. 2024. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3292–3302.
- Yu, Z.; Chen, M.; Zhang, Z.; You, S.; Rao, R.; Agarwal, S.; and Ren, F. 2023. Transupr: A transformer-based plug-and-play uncertain point refiner for lidar point cloud semantic segmentation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5864–5869. IEEE.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.
- Zhao, J.; Zhuo, J.; Chen, J.; and Ma, H. 2025. SAM2Object: Consolidating View Consistency via SAM2 for Zero-Shot 3D Instance Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19325–19334.