SAM-Zero3D: Extending Segment Anything to Zero Shot 3D Scene Segmentation via Iterative Global–Local Interaction

Dejun Zhang, *Member, IEEE,* Shifeng Xu, Yanzi Bai, Yiqi Wu, *Member, IEEE,* and Jun Liu, *Senior Member, IEEE*

***Abstract*—Lifting multi-view 2D masks generated by the Segment Anything Model (SAM) into 3D space offers a promising direction for zero-shot 3D scene segmentation, but view-dependent occlusions and limited fields of view often cause incomplete observations and cross-view inconsistencies, resulting in fragmented semantics and geometric misalignment. To address this, we propose SAM-Zero3D, which extends SAM to the 3D domain through a structured fusion pipeline with two complementary branches. The global anchor point-guided branch projects 3D anchors into multi-view masks to construct a cross-view affinity graph, identifies consistent mask groups via connected component analysis, and assigns 3D masks via majority voting and nearest-neighbor propagation. The local geometry-driven branch partitions the point cloud into fine-grained regions, estimates region-level semantic similarity from aggregated mask distributions, and progressively merges similar regions through a multi-stage merging strategy. An iterative global–local interaction further refines both branches by aligning global semantic priors with local geometric cues. Extensive experiments on ShapeNetPart, ScanNetV2, and ScanNet200 show that SAM-Zero3D significantly outperforms existing zero-shot baselines, achieving accurate and structure-aware segmentation without any 3D training or supervision.**

***Index Terms*—point cloud, zero-shot 3D segmentation, segment anything, multi-view.**

I. Introduction

Point cloud semantic segmentation, a core task in 3D scene understanding, plays a vital role in applications such as scene-level semantic parsing [1], point cloud compression [2], and autonomous driving [3]. This task aims to assign semantic labels to each point in an unordered 3D point cloud, thereby enabling detailed parsing of complex spatial structures. While deep learning has achieved remarkable progress in this field, current approaches heavily rely on large-scale annotated datasets [4]. However, acquiring accurate 3D annotations is notoriously labor-intensive and demands significant domain expertise, which severely hinders scalability and deployment in real-world scenarios [5].

To alleviate the annotation burden, recent efforts have explored few-shot and zero-shot 3D segmentation by leveraging cross-modal semantic embedding [6], language-guided supervision [7], or multi-view 2D detection [8]. Despite promising results, these methods face three key limitations: (1) a strong dependence on large-scale pretrained language models, which constrains generalization to unseen categories and open-vocabulary scenarios; (2) the reliance on task-specific and pretrained point cloud encoders, which limits adaptability in genuine zero-shot conditions; and (3) insufficient cross-view alignment and limited 3D structural reasoning, often leading to performance degradation in occluded or geometrically complex scenes.

Segment Anything Model (SAM) [9] has emerged as a powerful foundation model for zero-shot 2D image segmentation. It generates high-quality segmentation masks from simple prompts without relying on language inputs or additional training, demonstrating strong generalization to unseen categories and domains. This capability motivates an alternative paradigm that leverages SAM for training-free 3D scene segmentation by lifting multi-view 2D masks into 3D space [10], as illustrated in Fig. 1. This paradigm has gained increasing attention due to its simplicity, efficiency, and potential for general-purpose 3D perception. However, point cloud sparsity and view-dependent occlusions often lead to semantic inconsistency and poor alignment when directly fusing multi-view 2D masks, severely limiting segmentation accuracy.

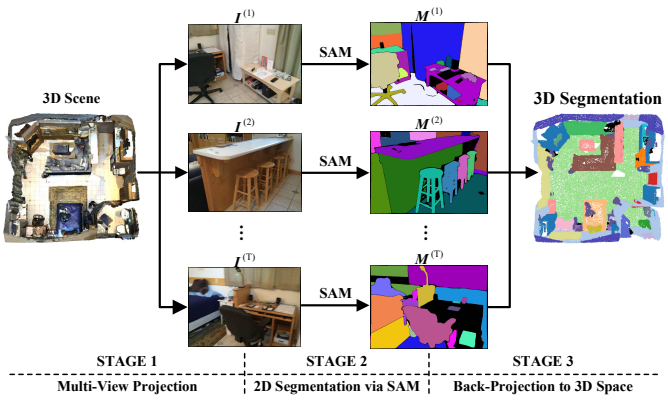

Fig. 1. An alternative paradigm for zero-shot 3D segmentation via multi-view masks from SAM. Given a 3D scene, SAM performs 2D segmentation on rendered views, and the resulting masks are back-projected into 3D space to enable training-free segmentation.

To address these challenges, we propose **SAM-Zero3D**, a training-free framework for zero-shot 3D point cloud segmentation by transferring multi-view 2D masks generated by the Segment Anything Model (SAM) into the 3D domain. Instead of relying on trained 3D encoders or language supervision, SAM-Zero3D introduces a structured fusion pipeline that

Dejun Zhang, Shifeng Xu, Yanzi Bai, and Yiqi Wu are with China University of Geosciences, Wuhan 430078, China (e-mail: zhangdejun@cug.edu.cn; xushifeng777@cug.edu.cn; baiyanzi123@cug.edu.cn; wuyq@cug.edu.cn).

Jun Liu is with Lancaster University, Lancaster LA1 4WA, United Kingdom of Great Britain and Northern Ireland (e-mail: j.liu81@lancaster.ac.uk). (*Corresponding author: Jun Liu*)

This work is supported by Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2023-B12).

Manuscript received April 19, 2021; revised August 16, 2021.

enforces both semantic consistency and geometric alignment across views. Specifically, we develop two complementary branches: a **global anchor point-guided branch**, which samples high-visibility 3D anchor points, projects them into multiple views, and constructs an affinity graph to align 2D masks across views; and a **local geometry-driven branch**, which partitions the point cloud into fine-grained geometric regions, projects them into multiple views, estimates semantic similarity from aggregated 2D mask distributions, and iteratively merges similar regions via a threshold-controlled strategy. To integrate global structure with local detail, we further introduce an **iterative global–local interaction** module that jointly refines both branches in a mutually reinforcing manner. Through this design, SAM-Zero3D achieves accurate and spatially coherent segmentation without any 3D training or supervision.

SAM-Zero3D directly addresses the challenges of lifting multi-view 2D masks into 3D space. Occlusions and incomplete observations are mitigated by sampling high-visibility 3D anchor points and leveraging their multi-view projections as stable correspondence cues. Cross-view inconsistencies are alleviated through an anchor-induced affinity graph over 2D mask patches, enabling robust cross-view association. Fragmented semantics are handled by a local geometry-driven branch that progressively merges regions based on the similarity between aggregated multi-view mask distributions. Geometric misalignment and boundary ambiguities are further reduced via an iterative global–local interaction, in which global instance assignments guide local merging and refined local regions update the global instance regions.

We validate SAM-Zero3D on three representative benchmarks—ShapeNetPart [11], ScanNetV2 [12], and ScanNet200 [13]—covering synthetic geometry, real-world indoor scenes, and long-tail open-vocabulary categories. The results show that SAM-Zero3D consistently outperforms existing zero-shot baselines, demonstrating strong generalization, robustness to occlusion, and scalability for training-free 3D segmentation.

Our main contributions are summarized as follows:

- SAM-Zero3D, a training-free framework, is proposed to transfer multi-view 2D masks from the Segment Anything Model (SAM) into 3D space, enabling semantic understanding without training, annotations, or 3D proposals.
- A global anchor point-guided branch is introduced to sample 3D anchor points, project them into multiple views, and construct an affinity graph over 2D masks for robust cross-view mask alignment and consistent 3D mask assignment.
- A local geometry-driven branch is developed to partition the point cloud into fine-grained geometric regions, aggregate multi-view 2D masks for estimating region-level semantic similarity, and merge similar regions through a multi-stage merging strategy.
- An iterative global–local interaction mechanism is proposed to integrate global semantic priors and local geometric details, where the former reweights region-level similarity and the latter updates global 3D masks via region-based overlap analysis.

The remainder of this paper is organized as follows. Section II reviews related work. Section III details the proposed SAM-Zero3D framework. Section IV introduces the datasets and implementation details, followed by quantitative and qualitative results, comparisons with state-of-the-art methods, and ablation studies. Finally, Section V concludes the paper.

II. Related Work

We review related work in three areas relevant to our approach: (1) geometry-aware methods for supervised point cloud segmentation, (2) zero-shot and open-vocabulary segmentation methods for unseen categories, and (3) SAM-based training-free 3D segmentation frameworks.

A. Point Cloud Semantic Segmentation

Point cloud semantic segmentation aims to assign a semantic label to each point in a 3D space and serves as a fundamental task in 3D scene understanding. Early works typically transformed point clouds into structured formats through voxelization or 2D projections. Examples include voxel-based methods like VoxNet [14] and projection-based models such as SqueezeSeg [15]. However, these approaches often suffer from spatial resolution loss and geometric distortion.

PointNet [16] pioneered methods for directly learning point-wise features from unordered 3D point sets, removing the need for voxelization or mesh conversion. PointNet++ [17] extended this by hierarchically aggregating neighborhood features to capture local geometric structures. KPConv [18] introduced deformable convolutional kernels for improved local shape modeling. PointTransformer [19] employed attention mechanisms to enhance structural and semantic representation. More recently, SGIFormer [4] addressed instance query initialization and global–local feature fusion by incorporating semantic-guided queries and geometry-enhanced interleaving transformer blocks.

In recent years, few-shot point cloud segmentation has gained attention, focusing on fast generalization to new categories with limited labeled support. For example, attMPTI [20] proposes a prototype-query attention propagation mechanism for label transfer; PAP [21] introduces prototype alignment to address intra-class viewpoint variations; QGE [22] leverages background prototypes for context-aware refinement. More recent works such as CoNet [23] and SQFI [24] focus on bridging the feature gap between support and query sets through co-occurrence mining and feature interaction, respectively. However, these methods still rely on explicit supervision from the support set and are not applicable to open-world, fully zero-shot scenarios.

B. Zero-Shot and Open-Vocabulary 3D Segmentation

To bypass the need for annotated 3D data, researchers have explored zero-shot 3D recognition [25] and segmentation [6], [26] by transferring knowledge from vision-language models. PointCLIP [6] first adapted CLIP [27] to 3D by rendering multi-view images and aligning them with text embeddings for open-vocabulary segmentation. ULIP [26] extended this

paradigm by jointly optimizing point cloud, image, and text encoders to improve cross-modal generalization.

Subsequent methods such as CG3D [28] introduced learnable visual prompt tuning to alleviate spatial misalignment between 2D and 3D features. CLIP2Point [29] and Openscene [30] leveraged pixel-level feature distillation and multi-view retraining to improve semantic transfer. PointCLIPv2 [31] further integrated GPT-generated prompts and dense rendering to enhance 2D–3D consistency and segmentation granularity under open-vocabulary conditions.

Additionally, GeoZe [32] incorporated geometric priors like surface normals and curvature into the feature fusion process, improving zero-shot generalization on unstructured point clouds. Nonetheless, two major limitations remain: (1) several methods still require fine-tuning on point cloud encoders (e.g., ULIP [26], CG3D [28], GeoZe [32]), limiting deployment flexibility; (2) distillation-based approaches from 2D to 3D (e.g., CLIP2Point [29], OpenScene [30]) are vulnerable to semantic degradation and loss of structural details under occlusions or missing views.

C. Segment Anything and 3D Extension

The Segment Anything Model (SAM) [9], developed by Meta AI, is a vision foundation model capable of zero-shot 2D segmentation from arbitrary prompts. Its promptable design and strong generalization ability have enabled broad adoption across diverse image segmentation tasks. However, as a generic 2D model, SAM still exhibits limitations in complex semantic reasoning, fine-grained mask quality, and consistency across temporal sequences.

Recent works have extended SAM along these directions in 2D images and videos. LISA [33] enhances SAM with large language models to support reasoning-based segmentation from implicit and complex textual queries. HQ-SAM [34] improves mask fidelity by refining SAM's output for objects with intricate structures while preserving its zero-shot and promptable nature. VRS-HQ [35] extends SAM to video reasoning segmentation by modeling spatiotemporal consistency across frames. While these methods significantly advance SAM in reasoning, mask quality, and temporal modeling, they primarily focus on 2D or video domains and do not address the problem of multi-view 2D-to-3D semantic transfer.

To adapt SAM for 3D scene understanding, recent works explore lifting its 2D masks from multi-view images into 3D space. OpenMask3D [36] extends SAM to open-vocabulary 3D instance segmentation by generating class-agnostic masks and computing CLIP-based features via multi-view refinement, enabling instance retrieval with free-form textual queries. SAM3D [10] transfers SAM-generated masks from posed RGB images to the 3D point cloud, followed by iterative bidirectional merging to construct full-scene instance masks without any training. SAMPro3D [37] further improves cross-view consistency by selecting and aligning prompt locations in 3D, consolidating multiple 2D segments into coherent 3D instances. Collectively, these efforts demonstrate the potential of leveraging 2D foundation models such as SAM for 3D segmentation with minimal or no 3D supervision.

Different from prior SAM3D-style approaches that lift SAM masks to 3D via frame-wise projection and bottom-up merging, our method adopts a scene-centric structured fusion paradigm. By introducing anchor-based cross-view 2D mask alignment, region-level similarity estimation, and iterative global–local interaction, the proposed method explicitly enforces global consistency and geometric coherence in 3D space, rather than relying on local frame adjacency. This structured fusion design enables more stable and interpretable 3D segmentation, particularly in cluttered indoor scenes.

III. METHOD

This section first revisits the Segment Anything Model (SAM)(Sec. III-A), which produces 2D segmentation masks from multi-view images. We then present the overall pipeline of our framework, SAM-Zero3D (Sec. III-B), which transfers these masks into 3D point cloud segmentations in a training-free, zero-shot manner. Sec. III-C introduces the global anchor point-guided branch for multi-view mask alignment. Sec. III-D describes the local geometry-driven branch for region-level semantic merging. Finally, Sec. III-E outlines the iterative global–local interaction module that integrates global semantics with local geometric details for joint refinement.

A. A Revisit of Segment Anything

The Segment Anything Model (SAM) [9] is a recently proposed foundation model for general-purpose image segmentation. It adopts a promptable Transformer architecture comprising an image encoder (typically a Vision Transformer [38]), a prompt encoder, and a mask decoder. Given user-defined prompts such as points, boxes, or coarse masks, SAM generates high-quality 2D segmentation masks without fine-tuning and demonstrates strong zero-shot generalization across categories and domains.

Benefiting from pretraining on large-scale natural image datasets, SAM is capable of recognizing and segmenting novel objects in unseen scenes without additional supervision. This makes it particularly attractive for 3D perception tasks where semantic annotations are costly or unavailable.

Formally, given a 2D image $\boldsymbol{I}^{(t)}$ from the t-th view, SAM predicts the corresponding 2D segmentation mask $\boldsymbol{M}^{(t)}$ as:

$$\boldsymbol{M}^{(t)} = \text{SAM}(\boldsymbol{I}^{(t)}), \tag{1}$$

where $\text{SAM}(\cdot)$ denotes the segmentation model. Applying SAM to all T image views yields a set of 2D masks: $\mathcal{M} = \left\{ \boldsymbol{M}^{(t)} \in \mathbb{R}^{H \times W} \right\}_{t=1}^{T}$.

However, as a purely 2D model, SAM lacks the capability for multi-view reasoning and 3D geometric understanding. Directly applying it to point cloud segmentation is suboptimal, as it fails to ensure spatial consistency and cross-view alignment. To address these limitations, we propose a structured fusion pipeline that transfers rich semantic cues from 2D masks into 3D space.

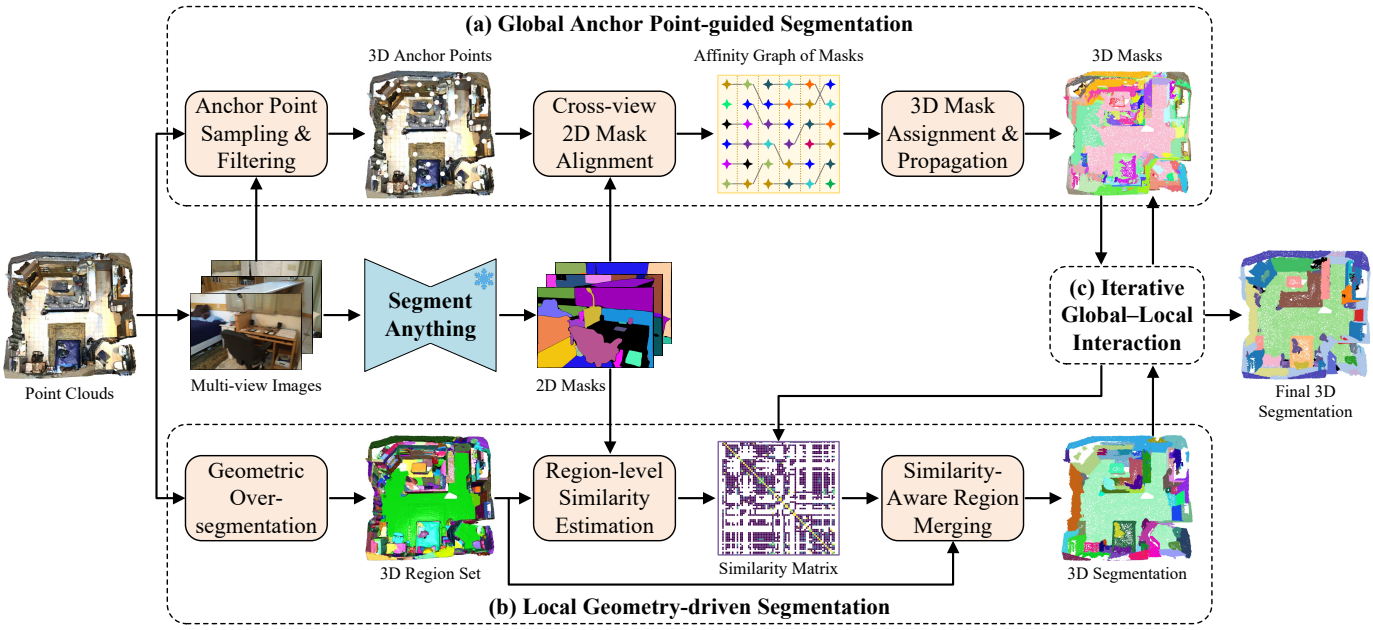

Fig. 2. Overview of the SAM-Zero3D framework. Multi-view images are rendered from the input point cloud and segmented by SAM. The 2D masks are fused via (a) a global anchor point-guided branch and (b) a local geometry-driven branch. Their outputs are refined through (c) an iterative global–local interaction to produce the final segmentation.

B. Overview of the SAM-Zero3D Framework

SAM-Zero3D performs zero-shot 3D scene segmentation by transferring multi-view 2D masks into 3D space, as illustrated in Fig. 2. Given a 3D point cloud and its corresponding multi-view RGB images, we first use SAM to automatically generate class-agnostic 2D segmentation masks for each view without any manual prompts. The point cloud, images, and masks are then processed in parallel by two complementary branches in our framework.

In the *global branch*, we sample a set of 3D anchor points via farthest point sampling and filter them based on multi-view visibility. The projected anchors guide cross-view mask association through overlap analysis, forming an affinity graph over 2D mask patches. A graph-based propagation strategy then assigns consistent 3D masks across the point cloud.

In the *local branch*, the point cloud is partitioned into regions using the region growing segmentation algorithm [39]. To estimate semantic similarity between regions, we project each region into multiple views and aggregate the corresponding 2D mask distributions. A multi-stage, threshold-controlled merging strategy then combines similar regions into fine-grained 3D segments.

Finally, the *iterative global–local interaction* module bridges the two branches. Global 3D masks are used to reweight local region similarities, while refined local segments update the global 3D masks. This iterative feedback improves both semantic consistency and boundary accuracy in the final 3D segmentation.

C. Global Anchor Point-guided Segmentation

To align multi-view 2D masks in 3D space, we construct a global anchor point-guided segmentation branch. It comprises three key steps: sampling high-visibility anchor points, associating 2D mask patches through anchor overlaps, and assigning 3D masks via an affinity graph.

1) Anchor Point Sampling and Filtering: Given a 3D scene $\mathcal{P} = \{\boldsymbol{p}_i \in \mathbb{R}^3\}_{i=1}^{M}$, where M denotes the number of points in the input point cloud, and its corresponding multi-view image sequence $\mathcal{I} = \{\boldsymbol{I}^{(t)} \in \mathbb{R}^{H\times W\times 3}\}_{t=1}^{T}$, where T denotes the number of rendered views, we first sample a set of representative 3D anchor points $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$, where N denotes the number of sampled anchor points, from the point cloud $\mathcal{P}$ using farthest point sampling, ensuring global coverage across the scene.

To facilitate cross-view 2D masks alignment, each 3D anchor point is projected into every view. For datasets with known camera intrinsics $\boldsymbol{K}^{(t)}$ and extrinsics $\boldsymbol{E}^{(t)}$, the projection of $\boldsymbol{x}_i$ into view t is computed as:

$$\boldsymbol{y}_i^{(t)} = \boldsymbol{K}^{(t)} \cdot \boldsymbol{E}^{(t)} \cdot \begin{bmatrix} \boldsymbol{x}_i \\ 1 \end{bmatrix}. \tag{2}$$

If camera parameters are unavailable, we apply the sparse visual projection module from PointCLIP-v2 [31] to obtain:

$$\boldsymbol{y}_i^{(t)} = \mathrm{SVP}^{(t)}(\boldsymbol{x}_i), \tag{3}$$

where $\mathrm{SVP}^{(t)}(\cdot)$ denotes a learned projection from 3D coordinates to 2D image space.

Figure 3 illustrates the anchor point generation process. Since each view only captures part of the scene due to occlusions and limited fields of view, anchor point visibility may vary across views. We define the visibility of $\boldsymbol{x}_i$ in view t as $\mathcal{V}(\boldsymbol{y}_i^{(t)}) \in \{0, 1\}$, where 1 indicates that the anchor point is visible. The average visibility of anchor point $\boldsymbol{x}_i$ across all views is:

$$V_i = \frac{1}{T}\sum_{t=1}^{T} \mathcal{V}\left(\boldsymbol{y}_i^{(t)}\right). \tag{4}$$

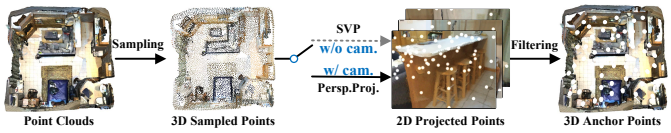

Fig. 3. Anchor point generation process. Sampled 3D points are projected into multi-view images using either perspective projection (with camera parameters) or a learned sparse visual projection module (without camera parameters), and visible points are retained for cross-view 2D mask alignment.

We retain only those anchor points with visibility score $V_i \geq \tau_v$ for subsequent steps to ensure robustness.

2) Cross-View 2D Mask Alignment: To estimate the affinity between 2D mask patches across different views, visible 3D anchor points are used as geometric correspondence cues. For any two patches $M_i^{(t)}$ and $M_{i'}^{(t')}$, the anchor-based affinity is defined as:

$$\mathcal{A}\left(M_i^{(t)}, M_{i'}^{(t')}\right) = \frac{\left|\mathcal{S}\left(M_i^{(t)}\right) \cap \mathcal{S}\left(M_{i'}^{(t')}\right)\right|}{\min\left(\left|\mathcal{S}\left(M_i^{(t)}\right)\right|, \left|\mathcal{S}\left(M_{i'}^{(t')}\right)\right|\right)}, \quad (5)$$

where $\mathcal{S}(\cdot)$ denotes the set of visible 3D anchor points whose projections fall inside the corresponding 2D mask patch.

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed, where each vertex represents a 2D mask patch, and an edge is added if the affinity between two patches exceeds a threshold τ_a:

$$\mathcal{E}\left(M_i^{(t)}, M_{i'}^{(t')}\right) = \begin{cases} 1, & \text{if } \mathcal{A}\left(M_i^{(t)}, M_{i'}^{(t')}\right) \geq \tau_a, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Depth-first search is applied to identify the connected components $\{C_k\}_{k=1}^{K}$ of the graph $\mathcal{G}$. Each connected component C_k corresponds to a cross-view consistent 3D instance hypothesis induced by multi-view mask agreement and is assigned a unique instance index k. Accordingly, each 2D mask patch $M_i^{(t)}$ is associated with an instance index:

$$\pi_i^{(t)} \leftarrow k \quad \text{if} \quad M_i^{(t)} \in C_k. \quad (7)$$

This process establishes cross-view consistent instance identities for all 2D mask patches, which are subsequently used for point-wise 3D mask assignment.

3) 3D Mask Assignment & Propagation: To assign a 3D mask to each point in the point cloud, we back-project each 3D point $\boldsymbol{p}_i$ into all views where it is visible. Let $\mathcal{V}_i$ denote the set of views in which the 3D point $\boldsymbol{p}_i$ has a valid projection. For each view $t \in \mathcal{V}_i$, if the projection of $\boldsymbol{p}_i$ falls inside a 2D mask patch $M_j^{(t)}$, the corresponding instance index $\pi_j^{(t)}$ is retrieved. The final 3D mask of $\boldsymbol{p}_i$ is determined by a voting scheme over all visible views:

$$M_i^{3D} \leftarrow \arg\max_k \sum_{t \in \mathcal{V}_i} \mathbb{I}\left(\pi_j^{(t)} = k\right), \quad (8)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

After this step, some points may still remain unassigned due to invisibility in all views. For these points, we apply nearest-neighbor propagation:

$$M_i^{3D} \leftarrow M_{\arg\min_j \|\boldsymbol{p}_i - \boldsymbol{p}_j\|_2}^{3D}, \quad (9)$$

where $\boldsymbol{p}_j$ denotes the nearest point with an assigned 3D mask. This post-processing step completes the 3D mask assignment by covering the remaining points without valid projections.

D. Local Geometry-driven Segmentation

To capture fine-grained 3D structures, we introduce a local geometry-driven segmentation branch that first partitions the 3D scene into regions, then estimates their semantic similarity by aggregating multi-view 2D masks. A multi-stage merging strategy is subsequently applied to progressively group similar regions into coherent instances.

1) Local Geometric Over-segmentation: We first partition the 3D point cloud $\mathcal{P}$ into a set of local regions $\mathcal{R} = \{R_i\}_{i=1}^{D}$ using region growing segmentation [39], which clusters points based on normal similarity and spatial continuity. This over-segmentation produces geometrically consistent patches well-suited for region-level semantic analysis. To support efficient neighborhood queries, we build a KD-Tree [40] over region centroids and construct a spatial adjacency graph.

2) Region-level Semantic Similarity Estimation: To estimate the semantic similarity between regions, we utilize segmentation information provided by SAM. Specifically, each 3D region R_i is projected into the image plane of view t, yielding a 2D area $A_i^{(t)}$. We then extract the corresponding 2D masks, and obtain a mask vector $\boldsymbol{v}_i^{(t)}$ by counting the frequency of each 2D mask within $A_i^{(t)}$. This vector characterizes the semantic composition of region R_i in the given view.

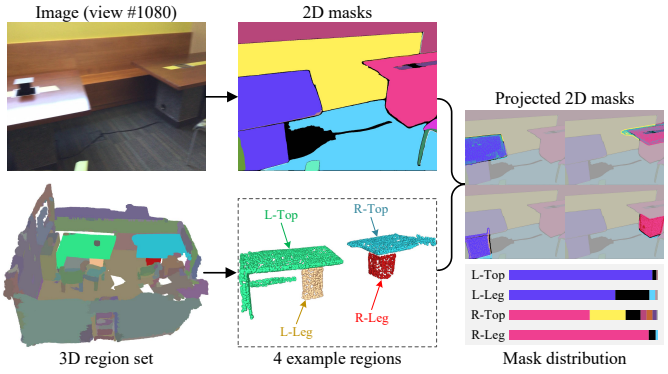

Fig. 4. Example of region-level semantic similarity estimation. Four representative 3D regions, namely left tabletop (L-Top), left table leg (L-Leg), right tabletop (R-Top), and right table leg (R-Leg), are projected onto a single image view for illustration.

As illustrated in Fig. 4, semantic similarity between regions can be analyzed by examining the distributions of their associated 2D masks. In this example, four representative regions corresponding to the left tabletop, left table leg, right tabletop, and right table leg are visualized. For each region, the distribution of associated 2D masks is presented as a stacked bar chart, which enables an intuitive comparison of semantic composition across regions. Regions belonging to the same object category exhibit similar distribution patterns, whereas semantically distinct regions show clearly different profiles. This observation motivates grouping semantically similar regions while maintaining separation between unrelated regions during the region merging process.

To quantify the semantic similarity between two regions R_i and R_j in view t, we compute the cosine similarity between their mask vectors:

$$S_{i,j}^{(t)} = \frac{\boldsymbol{v}_i^{(t)} \cdot \boldsymbol{v}_j^{(t)}}{\|\boldsymbol{v}_i^{(t)}\| \cdot \|\boldsymbol{v}_j^{(t)}\|}. \tag{10}$$

Since not all regions are visible in every view due to occlusions or limited fields of view, we aggregate similarity values across views using an average over valid observations:

$$\bar{S}_{i,j} = \frac{\sum_{t=1}^{T} \mathbb{I}(S_{i,j}^{(t)} > 0) \cdot S_{i,j}^{(t)}}{\sum_{t=1}^{T} \mathbb{I}(S_{i,j}^{(t)} > 0) + \epsilon}, \tag{11}$$

where $\mathbb{I}(\cdot)$ is the indicator function and ϵ is a small constant to avoid division by zero. The resulting similarity matrix $\bar{\boldsymbol{S}} \in \mathbb{R}^{D \times D}$ captures pairwise semantic similarity and supports the subsequent region merging process.

3) Similarity-Aware Region Merging: As illustrated in Fig. 2(b), the similarity-aware region merging (SRM) module takes two inputs: the 3D region set $\mathcal{R} = \{R_i\}_{i=1}^{D}$ obtained from geometric over-segmentation, and the region-level semantic similarity matrix $\bar{\boldsymbol{S}} \in \mathbb{R}^{D \times D}$ computed by the region similarity estimation module.

The similarity matrix $\bar{\boldsymbol{S}}$ is computed once from the initial region set and reused throughout the merging process. Based on this matrix, the SRM module performs a multi-stage merging procedure internally by applying a sequence of decreasing similarity thresholds to iteratively merge semantically consistent regions. Specifically, the merging process proceeds over K stages using a descending threshold schedule $\tau_1 > \tau_2 > \cdots > \tau_K$.

At each stage k, region pairs (R_i, R_j) are merged if their similarity exceeds the current threshold, i.e., $\bar{S}_{i,j} \geq \tau_k$. This scheme ensures that high-confidence, fine-grained regions are merged first, while more relaxed merging in later stages promotes the consolidation of semantically related but fragmented regions.

By progressively lowering the merging threshold, our strategy avoids premature over-merging and adapts to varying region scales and structures, ultimately producing accurate and structurally consistent 3D segmentation.

E. Iterative Global–Local Interaction

The global branch (Sec. III-C) ensures semantic consistency by aligning multi-view 2D masks via anchor-based fusion, while the local branch (Sec. III-D) improves boundary accuracy through geometry-aware region merging. To leverage their complementary strengths—global structure perception and local detail modeling—we introduce an iterative global–local interaction mechanism that enables mutual enhancement between the two branches.

As summarized in Algorithm 1, the global–local interaction is performed for U iterations. In each iteration, a global-to-local (G-to-L) step is first applied to update region-level similarity and the local segmentation, followed by a local-to-global (L-to-G) step to refine the global 3D masks.

1) Global-to-Local: To incorporate global mask-level information into region-level similarity estimation, we leverage the 3D masks produced by the global branch to guide the local branch.

Specifically, the global branch assigns each point in the point cloud to a 3D mask, resulting in a set of disjoint 3D masks. For each geometric region R_i in the over-segmented region set $\mathcal{R} = \{R_i\}_{i=1}^{D}$ (Sec. III-D1), the composition of 3D masks within the region is quantified by computing the proportion of points assigned to each mask:

$$\eta_i^{(k)} = \frac{N_i^{(k)}}{\sum_k N_i^{(k)}}, \tag{12}$$

where $N_i^{(k)}$ denotes the number of points in region R_i assigned to the k-th 3D mask.

For any region pair (R_i, R_j), their consistency with respect to 3D masks is assessed based on the product of their mask-wise point proportions. To suppress spurious region–mask associations caused by projection noise, partial visibility, or boundary inaccuracies, only the dominant contribution is retained:

$$w_{i,j} = \max_k \left(\eta_i^{(k)} \cdot \eta_j^{(k)} \right). \tag{13}$$

The weight $w_{i,j}$ is used to refine the region-level similarity $\bar{S}_{i,j}$ computed in Sec. III-D2:

$$\hat{S}_{i,j} = (1 + w_{i,j}) \cdot \bar{S}_{i,j}. \tag{14}$$

The new similarity matrix $\hat{\boldsymbol{S}}$, together with the fixed over-segmented region set $\mathcal{R}$, is fed into the similarity-aware region merging module (Sec. III-D3), denoted as $\mathrm{SRM}(\cdot)$, to update the local geometric regions.

2) Local-to-Global: To improve the quality of 3D masks, especially near object boundaries, we leverage refined local geometry to update the 3D masks. Specifically, we propose a 3D mask correction mechanism based on region overlap.

Let $\mathcal{G} = \{G_i\}$ denote the set of regions induced by the global branch, where each region is obtained by grouping points that share the same 3D mask assignment, and let $\mathcal{L} = \{L_j\}$ denote the set of geometric regions produced by the local branch. For each local region L_j, we compute its overlap with a global region G_i as the ratio $|G_i \cap L_j|/|G_i|$. If the overlap exceeds a threshold τ_o, L_j is considered to correspond to the same 3D instance as G_i. Accordingly, the global region G_i is refined by incorporating the points contained in L_j, resulting in an updated global segmentation.

The updated global segmentation is then used as the global input for the next iteration of the global-to-local update. After U iterations, the final global segmentation $\mathcal{G}$ is taken as the output.

The global-to-local and local-to-global steps establish a bidirectional interaction between global 3D masks and local geometric details, leading to more stable region merging and more accurate refinement of global 3D masks, particularly in boundary regions and geometrically complex areas.

Algorithm 1 Iterative Global–Local Refinement

Require: 3D region set $\mathcal{R}$, global segmentation $\mathcal{G}$, local segmentation $\mathcal{L}$, similarity matrix $\bar{\boldsymbol{S}}$
Ensure: Final 3D segmentation $\mathcal{G}$

for $u = 1$ to U **do**
  for each region $R_i \in \mathcal{R}$ **do**
    for $k = 1$ to K **do**
      $\eta_i^{(k)} \leftarrow N_i^{(k)} / \sum_k N_i^{(k)}$
    end for
  end for
  for each region pair (R_i, R_j) **do**
    $w_{i,j} \leftarrow \max_k(\eta_i^{(k)} \cdot \eta_j^{(k)})$
    $\hat{S}_{i,j} \leftarrow (1 + w_{i,j}) \cdot \bar{S}_{i,j}$
  end for
  $\mathcal{L} \leftarrow \mathrm{SRM}(\mathcal{R}, \hat{\boldsymbol{S}})$
  for each local region $L_j \in \mathcal{L}$ **do**
    for each global region $G_i \in \mathcal{G}$ **do**
      $\gamma \leftarrow |G_i \cap L_j| / |G_i|$
      if $\gamma \geq \tau_o$ **then**
        $G_i \leftarrow G_i \cup L_j$
      end if
    end for
  end for
end for
return $\mathcal{G}$

G-to-L

L-to-G

IV. Experiments

We evaluate the effectiveness and generalization ability of SAM-Zero3D from multiple perspectives. Sec. IV-A introduces the three benchmark datasets: ShapeNetPart, ScanNetV2, and ScanNet200. Sec. IV-B outlines the experimental settings and implementation details. Sec. IV-C presents quantitative results and comparisons with state-of-the-art methods. Finally, Sec. IV-D provides ablation studies to analyze the contribution of each component and validate the overall framework design.

A. Datasets

To evaluate the generalization and robustness of SAM-Zero3D, we conduct experiments on three representative datasets: ShapeNetPart (CAD-based), ScanNetV2 (real indoor), and ScanNet200 (open-vocabulary). These datasets cover diverse segmentation scenarios, ranging from clean synthetic shapes to cluttered real-world scenes and from limited to long-tailed category distributions.

ShapeNetPart [11] is a standard benchmark for part-level segmentation, containing 16 object categories and 50 part labels over 16,881 CAD-based point clouds, each annotated with 2,000-3,000 points including XYZ coordinates and normals. With its clean geometry and consistent structure, it is well-suited for evaluating fine-grained segmentation. We generate multi-view projections using a 3D perception module and evaluate SAM-Zero3D under a no-camera-parameter setting, focusing on view alignment and geometric reasoning in synthetic scenes.

ScanNetV2 [12] is a large-scale real-world RGB-D dataset for indoor semantic understanding, comprising 1,513 scan sequences and 707 scenes with over 2.5 million RGB-D frames. Each scene includes camera poses and 3D annotations for 20 semantic categories. With rich clutter, complex layouts, and diverse semantics, it serves as a strong benchmark for testing robustness in noisy and occluded environments. We use ScanNetV2 to assess SAM-Zero3D's ability to fuse multi-view masks and preserve semantic consistency in realistic scenarios.

ScanNet200 [13] extends ScanNet by introducing 200 semantic classes ranging from furniture and decor to small everyday objects, significantly increasing granularity and label openness. It features severe class imbalance and long-tail distributions, along with many zero- and few-shot categories. This dataset challenges the model's generalization, open-vocabulary segmentation, and cross-domain adaptation. We evaluate SAM-Zero3D on ScanNet200 to validate its performance in zero-shot segmentation and its robustness under fine-grained, large-vocabulary conditions.

B. Implementation Details

All experiments are conducted on a desktop workstation equipped with an Intel Core i5-13600K CPU (5.1 GHz), 32 GB RAM, and an NVIDIA RTX 4090 GPU with 24 GB VRAM. The framework is implemented in PyTorch, and the code is publicly available on the project website.[1] Key experimental settings are summarized in Table I.

TABLE I
Experimental settings for zero-shot segmentation with SAM-Zero3D across different datasets.

Configuration	ShapeNetPart	ScanNetV2	ScanNet200
Input points	2048	All	All
Multi-view #	10	1/100	1/100
Image resolution	$224\times224\times3$	$1296\times968\times3$	$1296 \times 968 \times 3$
3D anchors #	64	1/10	1/10
Threshold τ_a	0.4	0.4	0.4
Merge steps	5	5	5
Merge threshold	[0.9:0.1:0.5]	[0.9:0.1:0.5]	[0.9:0.1:0.5]
Threshold τ_o	0.3	0.3	0.3
Metric	mIoU	AP	AP, long-tail AP

C. Experimental Results

1) Results on ShapeNetPart: Since SAM-Zero3D performs instance-level segmentation without predicting explicit semantic labels, we follow the language-guided evaluation protocol of PointCLIP-V2 [31] to assess its zero-shot capability on ShapeNetPart [11]. Specifically, we first apply SAM-Zero3D to segment the point cloud into instance-level regions. These regions are then projected into multiple 2D views and classified using CLIP by comparing image features with part-level text prompts. This process enables semantic labeling without requiring any annotations or fine-tuning.

Quantitative results are shown in Table II. SAM-Zero3D achieves a mean IoU of 59.9%, outperforming PointCLIP-V2 (48.4%) and PointCLIP (31.0%) by 11.5 and 28.9 percentage

[1] https://github.com/djzgroup/SAM-Zero3D

TABLE II
ZERO-SHOT PART SEGMENTATION RESULTS ON THE SHAPENETPART DATASET. WE REPORT PER-CATEGORY MIOU (%) AND THE OVERALL MIOU. OPENSCENE AND PARTDISTILL REPORT RESULTS ON ONLY 10 CATEGORIES, WHILE OTHER METHODS ARE EVALUATED ON ALL 12 CATEGORIES.

	mIoU	Airp.	Bag	Cap	Chair	Earp.	Guit.	Knife	Lap.	Mug	Rock.	Skate	Table
# Shapes	2874	341	14	11	704	14	159	80	83	38	12	31	848
PointCLIP [6]	31.0	22.0	44.8	13.4	18.7	28.3	22.7	24.8	22.9	48.6	22.7	42.7	45.5
PointCLIP-V2 [31]	48.4	35.7	53.3	53.1	51.9	48.1	59.1	66.7	61.8	45.5	46.7	45.8	49.8
OpenScene [30]	52.9	34.4	**63.8**	56.1	**59.8**	**62.6**	69.3	70.1	65.4	51.0	-	-	60.4
PartDistill [41]	53.8	**37.5**	62.6	55.5	56.4	55.6	71.7	76.9	67.4	**53.5**	-	-	**62.9**
SAM-Zero3D	**59.9**	33.2	63.0	**64.7**	53.2	51.0	**85.7**	**80.7**	**73.7**	47.9	**55.0**	**50.2**	60.4

points, respectively. The performance gains are especially significant in categories with clear part structure, such as Guitar (85.7%), Knife (80.7%), and Laptop (73.7%). These improvements demonstrate that the high-quality 3D segmentation provided by SAM-Zero3D substantially enhances downstream zero-shot semantic classification.

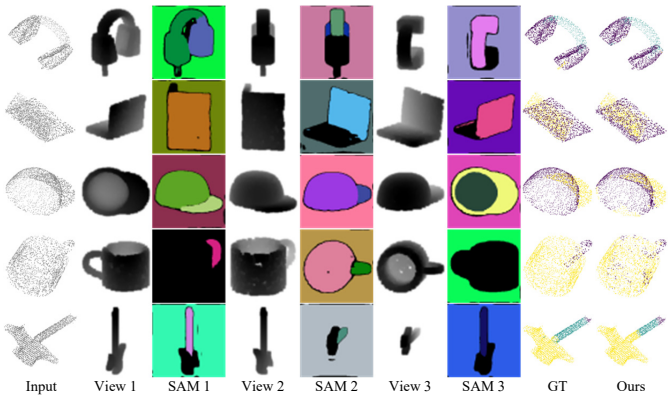

Fig. 5. Qualitative results of zero-shot 3D segmentation on ShapeNetPart. Each column shows the input point cloud, three view projections with corresponding 2D masks, ground-truth, and the SAM-Zero3D prediction. For simplicity, only three views are visualized, while our method uses 10 views per scene during inference.

Figure 5 further visualizes representative examples to qualitatively validate our approach. We make two key observations: (1) Although 2D masks are often fragmented and inconsistent across views, SAM-Zero3D effectively integrates them into unified 3D segments by leveraging anchor-based fusion and geometry-aware refinement; (2) Compared to ground-truth, our results exhibit sharper boundaries and more semantically coherent regions, especially for small parts and intricate structures.

These results confirm that SAM-Zero3D, by combining global anchor-guided alignment and local geometric reasoning, can perform high-quality zero-shot 3D segmentation even in the absence of supervision and camera parameters.

2) Results on ScanNetV2: We evaluate SAM-Zero3D for open-vocabulary 3D instance segmentation under a zero-shot setting using the ScanNetV2 [12] dataset. Since SAM-Zero3D does not output semantic labels, we incorporate OVSeg [42] to assign categories based on user-defined textual prompts. Given a prompt (e.g., "sofa"), OVSeg produces 2D masks for the input RGB images. These masks are back-projected into 3D space and matched with SAM-Zero3D's instance predictions via Intersection over Union (IoU). A match is considered valid if the IoU exceeds 0.5, allowing us to infer category labels without any 3D supervision.

TABLE III
ZERO-SHOT OPEN-VOCABULARY 3D INSTANCE SEGMENTATION ON SCANNETV2 DATASET.

Method	Open-vocab	Train Set	AP	AP@50	AP@25
Mask3D [43]	×	S3DIS	31.1	44.9	58.0
Mask3D [43]	×	ScanNetV2	65.7	83.1	91.0
SPFormer [44]	×	ScanNetV2	56.3	73.9	82.9
SGIFormer [4]	×	ScanNetV2	58.9	78.4	86.2
UnScene3D [45]	√	–	15.9	32.2	58.5
SAM3D [10]	√	–	20.2	34.0	53.3
SAI3D [46]	√	–	30.8	50.5	70.6
SAMPro3D [37]	√	–	33.8	**56.2**	75.3
SAM-Zero3D	√	–	**35.7**	55.0	**76.5**

As shown in Table III, SAM-Zero3D achieves a mean AP of 35.7%, outperforming the current best zero-shot method SAMPro3D (33.8%) by 1.9%. Other baselines such as UnScene3D, SAM3D, and SAI3D perform notably worse. Despite being a training-free framework, SAM-Zero3D also approaches the performance of fully supervised methods like Mask3D, demonstrating its strong generalization and robustness in complex indoor scenes. This performance stems from the effective integration of global anchor guidance and local geometric cues, which enhances structure recovery in occluded or cluttered environments.

Figure 6 visualizes representative results. Compared to the fragmented and noisy 2D masks, SAM-Zero3D produces consistent and structurally complete 3D segmentations, validating its effectiveness in open-vocabulary indoor settings with limited supervision.

3) Results on ScanNet200: To further assess the open-vocabulary semantic instance segmentation capability of SAM-Zero3D under more challenging conditions, we conduct zero-shot experiments on the ScanNet200 dataset. Compared to ScanNetV2, ScanNet200 significantly increases the difficulty by introducing 200 fine-grained categories with highly imbalanced, long-tailed distributions. This makes it a more representative benchmark for open-world 3D scene understanding.

Since SAM-Zero3D does not produce semantic labels, we follow OpenMask3D [36] to assign open-vocabulary categories to the predicted 3D segments. Specifically, OpenMask3D's semantic labeling module matches the input cat-

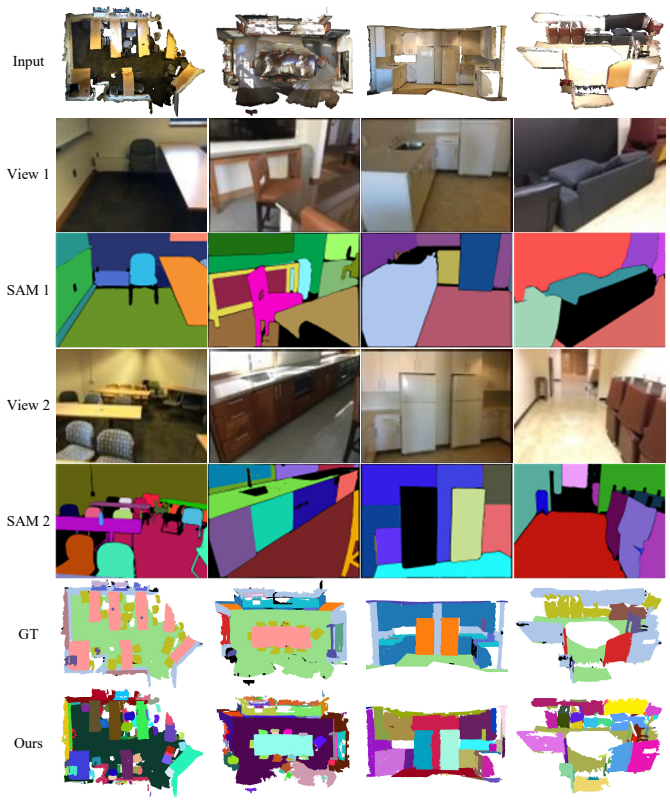

Fig. 6. Qualitative results of zero-shot 3D segmentation on ScanNetV2. Each row shows the input point cloud, two RGB views with 2D masks, ground-truth, and our SAM-Zero3D results. For clarity, only two views are visualized, while our method actually utilizes 1% of the available multi-view images during inference.

egory text with multi-view projections of the predicted segments, enabling fully annotation-free semantic segmentation.

TABLE IV
SEMANTIC SEGMENTATION RESULTS ON THE SCANNET200 DATASET. "O.V.", "3D P.", "HED.", "COM." AND "TAL." RESPECTIVELY INDICATE OPEN-VOCABULARY, 3D PROPOSAL, HEAD, COMMON, AND TAIL.

Method	O.V.	3D P.	AP	AP@50	AP@25	Hed.	Com.	Tal.
ISBNet [47]	×	–	24.5	32.7	37.6	38.6	20.5	12.5
SPFormer [44]	×	–	25.2	33.8	39.6	-	-	-
Mask3D [43]	×	–	26.9	36.2	41.4	39.8	21.7	17.9
SGIFormer [4]	×	–	28.9	38.6	43.6	-	-	-
OpenIns3D [48]	√	√	8.8	10.3	14.4	16.0	6.5	4.2
OpenScene [30]	√	√	11.7	15.2	17.8	13.4	11.6	9.9
OVIR-3D [49]	√	√	13.0	24.9	32.0	14.4	12.7	11.7
OpenMask3D [36]	√	√	15.4	19.9	23.1	17.1	14.1	14.9
OpenScene [30]	√	×	2.8	7.8	18.6	2.7	3.1	2.6
SAMPro3D [37]	√	×	2.7	13.3	20.0	-	-	-
SAM3D [10]	√	×	6.1	14.2	21.3	7.0	6.2	4.6
SAM-Zero3D	√	×	**12.3**	**18.5**	**23.2**	**13.2**	**10.5**	**15.7**

As shown in Table IV, SAM-Zero3D achieves 12.3% AP, 18.5% AP@50, and 23.2% AP@25 without using any 3D annotations or training data. It outperforms existing open-vocabulary baselines such as SAM3D and OpenScene by a clear margin. More importantly, on rare (tail) categories, SAM-Zero3D attains 15.7% AP—surpassing all prior open-vocabulary methods, including proposal-based approaches like OpenMask3D (14.9%) and OVIR-3D (11.7%).

These results validate the effectiveness of SAM-Zero3D in capturing fine-grained semantics and generalizing to infrequent classes, even in cluttered indoor scenes. Although it lags behind fully supervised methods like ISBNet and Mask3D in overall accuracy, SAM-Zero3D demonstrates strong potential for open-world 3D segmentation in realistic zero-shot settings.

D. Ablation Study

To comprehensively evaluate the effectiveness and design rationality of each component in SAM-Zero3D, we conduct a series of ablation experiments on the ScanNetV2 dataset. All experiments are performed under consistent hardware and input configurations to ensure fair comparison.

1) Component Ablation of SAM-Zero3D: To evaluate the effectiveness of each module in SAM-Zero3D, we conduct ablation studies on the ScanNetV2 dataset under consistent hardware and input settings. SAM-Zero3D comprises a global branch, a local branch, and an iterative global–local interaction module. The global branch further includes 3D anchor point sampling and cross-view 2D mask alignment. As shown in Table V, "GB", "LB", "G–L", "Anc." and "Align." respectively indicate global branch, local branch, global–local interaction, 3D anchor points, and cross-view 2D mask alignment.

TABLE V
ABLATION RESULTS OF SAM-ZERO3D COMPONENTS ON THE SCANNETV2 DATASET.

GB		LB	G–L	AP	AP@50	AP@25
Anc.	Align.					
×	√	×	×	20.3	33.4	40.6
√	√	×	×	22.5	40.2	51.1
×	×	√	×	27.8	43.5	56.3
×	√	√	√	32.6	50.7	73.8
√	√	√	√	**35.7**	**55.0**	**76.5**

Using only the cross-view 2D mask alignment module from the global branch yields limited performance, reflecting the challenge of relying purely on multi-view masks without geometric anchoring. Introducing 3D anchor points notably improves segmentation quality, confirming their role in enhancing cross-view consistency and reducing projection ambiguity. When only the local geometry-driven branch is retained, performance surpasses both global-only settings, demonstrating the advantage of geometric cues in capturing fine structures beyond what 2D semantics can offer.

Further combining the cross-view 2D mask alignment module and the local branch with the interaction mechanism yields a clear improvement. Even without 3D anchor points, the global–local interaction enables more consistent semantics and sharper boundaries by bridging global and local segmentation. The complete SAM-Zero3D—equipped with all components—achieves the best performance, indicating that integrating global priors with local geometry in an iterative manner is essential for robust segmentation under the zero-shot setting.

2) Sensitivity to the Number of Global–Local Interaction Iterations: To evaluate the impact of iterative interaction between the global and local branches, we conduct a controlled study on the ScanNetV2 dataset by varying the number of interaction rounds while keeping all other settings fixed. Specifically, we set the number of iterations from 0 to 7 to identify an optimal balance between performance and computational cost.

TABLE VI
QUANTITATIVE ANALYSIS OF THE NUMBER OF GLOBAL–LOCAL INTERACTION ITERATIONS ON THE SCANNETV2 DATASET.

Iteration Rounds	AP	AP@50	AP@25
0	22.5	40.2	51.1
1	29.7	45.3	60.3
2	32.2	48.2	65.7
3	33.4	50.8	70.3
4	34.6	52.9	73.4
5	**35.7**	**55.0**	76.5
6	35.5	53.7	**76.8**
7	35.2	53.1	76.6

The results are summarized in Table VI. When the number of iterations is set to 0, SAM-Zero3D reduces to the global branch alone, yielding suboptimal results. As the number of iterations increases, the model consistently improves across all metrics. The best performance is observed at 5 iterations, reaching 35.7% AP, 55.0% AP@50, and 76.5% AP@25. Although using 6 iterations slightly improves AP@25, the overall AP drops marginally, indicating diminishing returns.

3) Sensitivity to Anchor Point Number: The effect of anchor point density is evaluated by varying the sampling ratio, as shown in Table VII. As the sampling ratio increases from 1/50 to 1/10, AP, AP@50, and AP@25 consistently improve, indicating that denser anchor points facilitate more reliable cross-view instance alignment. When the sampling ratio is further increased, performance gains gradually saturate and exhibit minor fluctuations. This suggests that once sufficient geometric coverage is achieved, increasing the number of anchor points yields diminishing returns. Overall, SAM-Zero3D achieves a favorable balance between performance and efficiency at moderate anchor point densities.

TABLE VII
SENSITIVITY OF SAM-ZERO3D TO THE ANCHOR POINT SAMPLING RATIO ON SCANNETV2.

Sampling Ratio	AP	AP@50	AP@25
1/50	29.0	42.6	64.8
1/30	31.4	49.3	70.5
1/20	33.6	52.1	73.9
1/15	34.3	54.4	75.7
1/10	**35.7**	55.0	76.5
1/8	35.5	**55.3**	76.6
1/6	35.2	54.8	76.7
1/5	34.9	54.7	**77.0**

4) Sensitivity to the Number of Views: Table VIII reports the performance under different numbers of views. Increasing the number of views leads to notable improvements in all metrics, demonstrating that richer multi-view observations enhance cross-view 2D mask alignment. The best performance is achieved at a moderate view sampling ratio (around 1/100). When the number of views continues to increase, performance degrades, likely due to redundant or inconsistent 2D masks introduced by excessive views. These results indicate that an appropriate number of views is crucial for balancing global coverage and cross-view consistency.

TABLE VIII
SENSITIVITY OF SAM-ZERO3D TO THE NUMBER OF VIEWS FOR MULTI-VIEW MASK FUSION ON SCANNETV2.

Sampling Ratio	AP	AP@50	AP@25
1/10	24.6	40.9	66.4
1/25	33.6	49.9	74.8
1/50	34.1	52.1	75.4
1/100	**35.7**	**55.0**	**76.5**
1/125	30.4	45.4	74.1
1/150	28.7	41.3	73.5
1/200	18.6	33.6	66.3
1/250	13.5	31.0	65.4

5) Sensitivity to the Merging Thresholds: The influence of merging thresholds in the similarity-aware region merging module is summarized in Table IX. Single-step merging exhibits high sensitivity to threshold selection: overly high thresholds result in fragmented regions, while overly low thresholds cause excessive merging. In contrast, the proposed multi-stage merging strategy with decreasing thresholds progressively enhances semantic consistency while preserving geometric details. Performance improves steadily as the number of merging stages increases, and the best results are achieved with five stages using thresholds [0.9:0.1:0.5]. This demonstrates that progressive region merging provides a more stable and effective mechanism for integrating local geometry and global semantics.

TABLE IX
SENSITIVITY TO MERGING THRESHOLD STRATEGIES IN THE SIMILARITY-AWARE REGION MERGING MODULE ON SCANNETV2.

Strategy	Thresholds	AP	AP@50	AP@25
Single-stage	[0.9]	17.7	36.7	65.8
	[0.8]	25.9	37.9	68.6
	[0.7]	25.6	46.7	74.8
	[0.6]	25.1	36.6	71.0
	[0.5]	24.7	31.8	65.0
Multi-stage	[0.9:0.1:0.8]	25.5	35.6	66.7
	[0.9:0.1:0.7]	34.3	52.3	70.8
	[0.9:0.1:0.6]	34.8	53.3	75.0
	[0.9:0.1:0.5]	**35.7**	**55.0**	**76.5**

6) Runtime and GPU Memory Analysis: We analyze the runtime efficiency and GPU memory usage of SAM-Zero3D on the ScanNetV2 dataset. Table X reports a module-wise breakdown of runtime and GPU memory consumption. The overall runtime is dominated by the multi-view 2D inference of SAM, while the global branch, local branch, and iterative global–local interaction introduce only limited additional overhead. Regarding GPU memory usage, the peak consumption is mainly incurred by the SAM inference stage. The global branch requires moderate GPU memory for affinity

computation, whereas the local geometry-driven branch and the global–local interaction are executed on the CPU and therefore introduce negligible GPU memory overhead. This hybrid computation strategy enables stable execution under limited GPU memory without out-of-memory issues.

TABLE X
RUNTIME BREAKDOWN AND PEAK GPU MEMORY USAGE OF SAM-ZERO3D ON SCANNETV2.

Stage	Time (s/scene)	GPU Memory (MB)
SAM	45	8,192
Global branch	15	4,620
Local branch	1	0
Global–local interaction	3	0
SAM-Zero3D(Total)	64	8,192 (max)

E. Intermediate Result Visualization and Analysis

To better illustrate how each component contributes to the final segmentation, we provide qualitative visualizations of key intermediate results along the pipeline.

1) Effect of Cross-view Mask Alignment: To illustrate the effect of cross-view 2D mask alignment, Fig. 7 presents a comparison of 2D mask patches before and after alignment across viewpoints. In the second row, SAM-generated masks use view-dependent colors, causing 2D mask patches corresponding to the same 3D instance to appear with different colors across views. After applying the proposed anchor-based affinity graph and connected component analysis, mask patches associated with the same 3D instance are consistently grouped across views, as shown in the third row. The resulting aligned mask groups provide reliable cross-view consistency for subsequent 3D mask assignment.

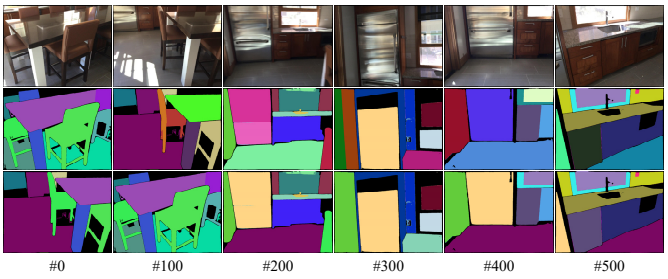

Fig. 7. Cross-view grouping of 2D masks under different viewpoints. The first row shows input RGB images, the second row shows class-agnostic 2D masks from SAM, and the third row shows cross-view aligned mask groups.

2) Effect of Similarity-Aware Region Merging: To demonstrate the effectiveness of the similarity-aware region merging strategy, we visualize intermediate results from multiple merging stages on several scenes. As shown in Fig. 8, the merging process is performed in five stages with progressively relaxed thresholds. In early stages, only highly similar regions are merged, yielding fine-grained but semantically consistent structures. As the merging proceeds, semantically related yet spatially separated regions are gradually integrated into more complete object instances. These results show that the proposed multi-stage strategy mitigates region fragmentation while avoiding premature over-merging.

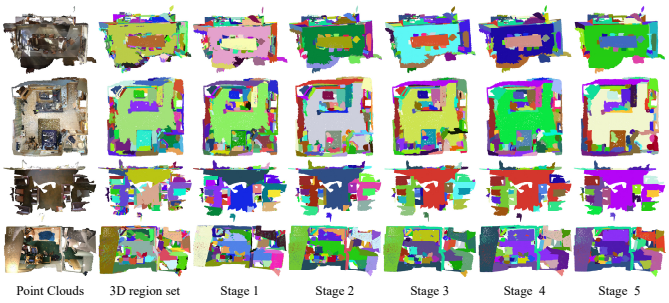

Fig. 8. Intermediate results of multi-stage region merging. Starting from the initial 3D region set, regions are progressively merged over five steps, showing the gradual aggregation from fragmented local regions to coherent object-level segments.

3) Effect of Iterative Global–Local Interaction: We further visualize the intermediate results of the iterative global–local interaction process. As shown in Fig. 9, the interaction is performed for five iterations. In each iteration, the global segmentation is used to refine the region-level similarity matrix, followed by similarity-aware region merging to update local regions, which are then used to refine the global segmentation. As iterations proceed, the similarity matrix exhibits clearer block structures, local regions become more semantically coherent, and global segmentation boundaries are progressively refined. These visualizations provide intuitive evidence of the mutual refinement between global semantics and local geometry.

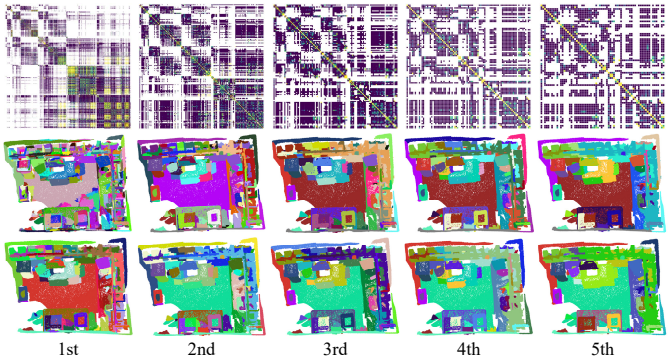

Fig. 9. Evolution of segmentation results during iterative refinement. Each row shows the region-level similarity matrix, global segmentation, and local segmentation at different iterations.

V. CONCLUSION

In this paper, we introduced SAM-Zero3D, a training-free framework for zero-shot 3D scene segmentation that transfers 2D masks from the SAM into 3D point clouds through structured multi-view fusion. Unlike existing approaches that rely on supervised 3D encoders or language-driven models, SAM-Zero3D eliminates the need for annotations, training, or 3D proposals by combining two complementary branches: a global anchor point-guided branch for cross-view mask alignment, and a local geometry-driven branch for fine-grained region merging. An iterative global–local interaction further integrates global semantics with local geometric structure to

enhance segmentation accuracy and consistency. Extensive experiments on ShapeNetPart, ScanNetV2, and ScanNet200 demonstrate that SAM-Zero3D consistently outperforms existing zero-shot baselines, particularly in open-vocabulary and long-tailed settings, achieving accurate and structure-aware 3D segmentation without any 3D supervision.

Despite these promising results, SAM-Zero3D is primarily designed for offline, training-free, and open-vocabulary 3D scene segmentation. Its performance depends on several factors inherent to this setting, including multi-view availability, 2D mask quality, and the choice of hyperparameters. In large-scale outdoor scenes, point clouds are typically sparser and effective views are more limited, which poses additional challenges for multi-view mask lifting and cross-view semantic consistency. As a result, the current framework is not directly optimized for such scenarios without further adaptation. Future work will explore lightweight learnable components to improve robustness under sparse-view and sparse-point conditions, while preserving the low-annotation-cost advantage of the proposed approach.

REFERENCES

[1] F. Song, G. Li, X. Yang, W. Gao, and S. Liu, "Block-adaptive point cloud attribute coding with region-aware optimized transform," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4294–4308, 2023.

[2] Y. Shao, X. Yang, W. Gao, S. Liu, and G. Li, "3d point cloud attribute compression using diffusion-based texture-aware intra prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 9633–9646, 2024.

[3] L. Zhan, Y. Du, J. Jiang, Y. Wei, T. Zhou, and Z. Yang, "Bgc-net: Bilateral graph convolutional network for weakly-supervised semantic segmentation of large-scale point clouds," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2025.

[4] L. Yao, Y. Wang, M. Liu, and L.-P. Chau, "Sgiformer: Semantic-guided and geometric-enhanced interleaving transformer for 3d instance segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 3, pp. 2276–2288, 2025.

[5] Y. Wu, X. Chen, X. Huang, K. Song, and D. Zhang, "Unsupervised distribution-aware keypoints generation from 3d point clouds," *Neural Netw.*, vol. 173, p. 106158, 2024.

[6] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "Pointclip: Point cloud understanding by clip," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 8542–8552.

[7] J. Yang, R. Ding, W. Deng, Z. Wang, and Q. Xiaojuan, "Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 19 823–19 832.

[8] A. Abdelreheem, I. Skorokhodov, M. Ovsjanikov, and P. Wonka, "Satr: Zero-shot semantic segmentation of 3d shapes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 15 120–15 133.

[9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 3992–4003.

[10] D. Zhang, D. Liang, H. Yang, Z. Zou, X. Ye, Z. Liu, and X. Bai, "Sam3d: zero-shot 3d object detection via the segment anything model," *Sci. China Inf. Sci.*, vol. 67, no. 4, p. 149101, 2024.

[11] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2463–2471.

[12] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2432–2443.

[13] D. Rozenberszki, O. Litany, and A. Dai, "Language-grounded indoor 3d semantic segmentation in the wild," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2022, pp. 125–141.

[14] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2015, pp. 922–928.

[15] B. Wu, A. Wan, X. Yue, and K. Keutzer, "Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 1887–1893.

[16] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 77–85.

[17] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[18] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6410–6419.

[19] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 16 239–16 248.

[20] N. Zhao, T.-S. Chua, and G. H. Lee, "Few-shot 3d point cloud semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 8869–8878.

[21] S. He, X. Jiang, W. Jiang, and H. Ding, "Prototype adaption and projection for few- and zero-shot 3d point cloud semantic segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 3199–3211, 2023.

[22] Z. Ning, Z. Tian, G. Lu, and W. Pei, "Boosting few-shot 3d point cloud segmentation via query-guided enhancement," in *Proc. ACM Int. Conf. Multimedia (MM)*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1895–1904.

[23] G. Zhu, Y. Zhou, R. Yao, and H. Zhu, "Information gap narrowing for point cloud few-shot segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4421–4433, 2024.

[24] C. Zheng, L. Liu, Y. Meng, X. Peng, and M. Wang, "Few-shot point cloud semantic segmentation via support-query feature interaction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 11, pp. 10 753–10 763, 2024.

[25] D. Song, X. Fu, N. Liu, W.-Z. Nie, W.-H. Li, L.-J. Wang, Y. Yang, and A.-A. Liu, "Mv-clip: Multi-view clip for zero-shot 3d shape recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 9, pp. 8767–8779, 2025.

[26] L. Xue, M. Gao, C. Xing, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. Savarese, "Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 1179–1189.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PmLR, 2021, pp. 8748–8763.

[28] D. Hegde, J. M. Jose Valanarasu, and V. M. Patel, "Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2023, pp. 2020–2030.

[29] T. Huang, B. Dong, Y. Yang, X. Huang, R. W. Lau, W. Ouyang, and W. Zuo, "Clip2point: Transfer clip to point cloud classification with image-depth pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 22 100–22 110.

[30] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, "Openscene: 3d scene understanding with open vocabularies," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 815–824.

[31] X. Zhu, R. Zhang, B. He, Z. Guo, Z. Zeng, Z. Qin, S. Zhang, and P. Gao, "Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 2639–2650.

[32] G. Mei, L. Riz, Y. Wang, and F. Poiesi, "Geometrically-driven aggregation for zero-shot 3d point cloud understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 27 896–27 905.

[33] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 9579–9589.

[34] L. Ke, M. Ye, M. Danelljan, Y. liu, Y.-W. Tai, C.-K. Tang, and F. Yu, "Segment anything in high quality," in *Adv. Neural Inf. Process. Syst.*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 29 914–29 934.

[35] S. Gong, Y. Zhuge, L. Zhang, Z. Yang, P. Zhang, and H. Lu, "The devil is in temporal token: High quality video reasoning segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 29 183–29 192.
[36] A. Takmaz, E. Fedele, R. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "Openmask3d: Open-vocabulary 3d instance segmentation," in *Adv. Neural Inf. Process. Syst.*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 68 367–68 390.
[37] M. Xu, X. Yin, L. Qiu, Y. Liu, X. Tong, and X. Han, "Sampro3d: Locating sam prompts in 3d for zero-shot instance segmentation," in *Proc. Int. Conf. 3D Vis. (3DV)*, 2025, pp. 1222–1232.
[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.
[39] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 1–4.
[40] K. Zhou, Q. Hou, R. Wang, and B. Guo, "Real-time kd-tree construction on graphics hardware," *ACM Trans. Graph.*, vol. 27, no. 5, Dec. 2008.
[41] A. Umam, C.-K. Yang, M.-H. Chen, J.-H. Chuang, and Y.-Y. Lin, "Partdistill: 3d shape part segmentation by vision-language model distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 3470–3479.
[42] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7061–7070.
[43] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3d: Mask transformer for 3d semantic instance segmentation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2023, pp. 8216–8223.
[44] J. Sun, C. Qing, J. Tan, and X. Xu, "Superpoint transformer for 3d scene instance segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, 2023, pp. 2393–2401.
[45] D. Rozenberszki, O. Litany, and A. Dai, "Unscene3d: Unsupervised 3d instance segmentation for indoor scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 19 957–19 967.
[46] Y. Yin, Y. Liu, Y. Xiao, D. Cohen-Or, J. Huang, and B. Chen, "Sai3d: Segment any instance in 3d scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 3292–3302.
[47] T. D. Ngo, B.-S. Hua, and K. Nguyen, "Isbnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 13 550–13 559.
[48] Z. Huang, X. Wu, X. Chen, H. Zhao, L. Zhu, and J. Lasenby, "Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 169–185.
[49] S. Lu, H. Chang, E. P. Jing, A. Boularias, and K. Bekris, "Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data," in *Proc. Conf. Robot Learn. (CoRL)*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., vol. 229. PMLR, 06–09 Nov 2023, pp. 1610–1620.

Dejun Zhang (Member, IEEE) received the Ph.D. degree from the Department of Computer Science, Wuhan University, China, in 2015. He is currently an associate professor in the College of Computer Science, China University of Geosciences, China. Since 2017, he has been serving as a senior member of the China Society for Industrial and Applied Mathematics (CSIAM) and a committee member of the Geometric Design & Computing division of CSIAM. Since 2020, he has been serving as a senior member of the China Computer Federation (CCF). Additionally, he is a member of the Institute of Electrical and Electronics Engineers (IEEE). Since 2021, he has been serving as a committee member of the Technical Committee on CAD and Graphics of CCF. His research areas include computer graphics, 3D vision, image processing, and video processing.

Shifeng Xu received the B.S. degree from Sichuan Agricultural University, China, in 2024. He is currently pursuing the M.S. degree in the College of Computer Science, China University of Geosciences, China. He is a member of the China Computer Federation (CCF) and the China Society for Industrial and Applied Mathematics. His research interests include point cloud analysis, such as point cloud segmentation and point cloud based retrieval for place recognition.

Yanzi Bai received the B.S. degree from China University of Geosciences, China, in 2022, and the M.S. degree in computer science from the same university in 2025. She is currently working at Li Auto Inc. She is a member of the China Computer Federation (CCF) and the China Society for Industrial and Applied Mathematics. Her research interests include 3D point cloud perception, object detection, and artificial intelligence.

Yiqi Wu (Member, IEEE) received the B.S. degree from Huazhong University of Science and Technology in 2007 and the M.S. degree from China University of Geosciences in 2011. He received the Ph.D. degree from the Department of Computer Science, Wuhan University, China, in 2017. Since 2021, he has been serving as a member of the China Society for Industrial and Applied Mathematics (CSIAM) and a committee member of the Geometric Design & Computing division of CSIAM. Since 2022, he has been serving as a member of the China Computer Federation (CCF). Additionally, he is a member of the Institute of Electrical and Electronics Engineers (IEEE). He is currently an associate professor in the College of Computer Science, China University of Geosciences, China. His research interests include computer graphics, computer vision, and computer-aided design.

Jun Liu (Senior Member, IEEE) is a Professor and Chair in Digital Health at School of Computing and Communications in Lancaster University. He got the PhD degree from Nanyang Technological University in 2019. He obtained the best paper awards from PREMIA in 2016 and 2019, the Best Thesis Award from EEE at NTU in 2020, the IEEE VSPC Rising Star Honorable Mention Award in 2024, and is listed in the top 2% scientists for both career-long and single-year categories by Stanford University. He is an Associate Editor of IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Industrial Informatics, IEEE Transactions on Biometrics, Behavior and Identity Science, ACM Computing Surveys, and Pattern Recognition. He has served as an Area Chair of CVPR, ECCV, ICML, NeurIPS, ICLR and MM. His research interests include computer vision, machine learning and digital health.