IEEE Transactions on Robotics (T-RO) paper, presented at ICRA 2026, Vienna, Austria. Cite as T-RO paper.

RAZER: Robust Accelerated Zero-Shot 3-D Open-Vocabulary Panoptic Reconstruction With Spatio-Temporal Aggregation

Naman Patel, Prashanth Krishnamurthy, *Member, IEEE*, and Farshad Khorrami, *Fellow, IEEE*

***Abstract*—Mapping and understanding complex 3-D environments is fundamental to how autonomous systems perceive and interact with the physical world, requiring both precise geometric reconstruction and rich semantic comprehension. While existing 3-D semantic mapping systems excel at reconstructing and identifying predefined object instances, they lack the flexibility to efficiently build semantic maps with open-vocabulary during online operation. Although recent vision-language models (VLMs) have enabled open-vocabulary object recognition in 2-D images, they haven't yet bridged the gap to 3-D spatial understanding. The critical challenge lies in developing a training-free unified system that can simultaneously construct accurate 3-D maps while maintaining semantic consistency and supporting natural language interactions in real time. In this article, we develop a zero-shot framework that seamlessly integrates GPU-accelerated geometric reconstruction with open-vocabulary VLMs through online instance-level semantic embedding fusion, guided by hierarchical object association with spatial indexing. Our training-free system achieves superior performance through incremental processing and unified geometric-semantic updates, while robustly handling 2-D segmentation inconsistencies. The proposed general-purpose 3-D scene understanding framework can be used for various tasks including zero-shot 3-D instance retrieval, segmentation, and object detection to reason about previously unseen objects and interpret natural language queries.**

***Index Terms*—RGB-D Perception, recognition, simultaneous localization and mapping (SLAM), semantic scene understanding.**

I. Introduction

THE ability to create semantically meaningful 3-D maps of dynamic environments is crucial for applications ranging from robotic navigation and manipulation to augmented reality and scene understanding. While significant advances have been made in geometric 3-D reconstruction and 2-D semantic understanding separately, combining these capabilities into a real-time system that can handle arbitrary objects and support natural language interactions remains a fundamental challenge in computer vision and robotics. This integration is essential for enabling intelligent systems to not only perceive the geometric structure of their environment, but also understand and reason about the objects and their relationships within it.

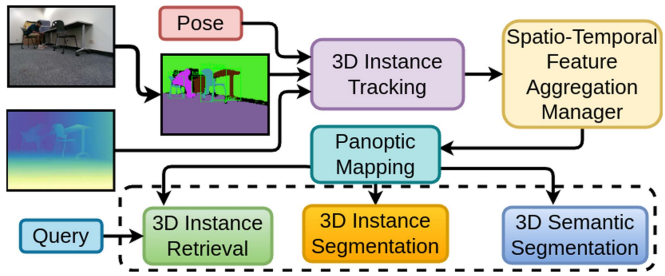

Fig. 1. *Pipeline overview of our proposed 3-D scene understanding framework.* Our system processes posed RGB-D inputs through open-vocabulary segmentation for robust 3-D instance tracking. Spatio-temporal feature aggregation fuses and prunes tracks while updating a panoptic map that enables online text-based 3-D instance retrieval and segmentation tasks.

Traditional 3-D mapping systems have primarily focused on geometric accuracy, employing techniques like simultaneous localization and mapping (SLAM) and dense reconstruction to create precise spatial representations. However, these approaches typically lack semantic understanding, limiting their utility in applications requiring object-level reasoning. 3-D semantic-instance mapping addresses this limitation by simultaneously detecting and segmenting objects with their semantic and instance labels to generate a comprehensive 3-D semantic map of the surrounding environment. Existing approaches for 3-D scene understanding broadly fall into two categories: *3D-to-3D* and *2D-to-3D* methods. The *3D-to-3D* approach operates directly on dense 3-D point clouds or volumetric representations, typically acquired through depth sensors or multiview reconstruction, to perform object or concept level segmentation of the scene in 3-D space. While these approaches benefit from having complete 3-D geometric information available, they often struggle with computational efficiency and real-time performance due to the dense nature of point cloud processing as well as a scarcity of large-scale training data. In contrast, *2D-to-3D* methods analyze a set of 2-D images and project their predictions onto the 3-D

Received 17 May 2025; revised 16 November 2025; accepted 30 November 2025. Date of publication 12 January 2026; date of current version 13 February 2026. This work was supported in part by the Army Research Office under Grant W911NF-21-1-0155 and in part by the New York University Abu Dhabi (NYUAD) Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award, under Grant CG010. This article was recommended for publication by Associate Editor S. Song and Editor J. Civera upon evaluation of the reviewers' comments. *(Corresponding author: Farshad Khorrami.)*

The authors are with the Control/Robotics Research Laboratory (CRRL), Department of Electrical and Computer Engineering, NYU Tandon School of Engineering, Brooklyn, NY 11201 USA (e-mail: naman.patel@nyu.edu; prashanth.krishnamurthy@nyu.edu; khorrami@nyu.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TRO.2026.3651674, provided by the authors.

Digital Object Identifier 10.1109/TRO.2026.3651674

map, performing reconstruction and semantic mapping synchronously. These approaches can leverage state-of-the-art 2-D semantic-instance segmentation algorithms, partially addressing the limitations of *3D-to-3D* methods. Recent works have attempted to bridge this gap by integrating semantic information into 3-D reconstructions. However, these approaches face several critical challenges: they often require extensive training data, operate offline, or struggle with maintaining consistency when processing streaming data. In addition, they frequently fail to handle the inherent inconsistencies in 2-D segmentation outputs, leading to fragmented or incorrect 3-D semantic maps. Furthermore, these approaches typically operate with predefined categories, limiting their applicability in open-world scenarios, where systems must recognize and reason about previously unseen objects.

While recent breakthroughs in vision-language models (VLMs) have enabled remarkable open-vocabulary recognition capabilities in 2-D images, these models lack the ability to reason about 3-D structure and spatial relationships, creating a crucial gap in current advances. The fundamental challenges in creating a unified 3-D semantic mapping system span multiple critical aspects of computer vision and robotics. A primary challenge lies in maintaining temporal consistency of object instances across frames without access to complete mapping history, as real-world applications often require processing streaming data with limited memory resources. This is compounded by the need to handle inconsistent labels and masks from 2-D segmentation models in real-time, where prediction errors and uncertainties must be robustly managed. In addition, enabling natural language interactions with the 3-D environment without requiring task-specific training presents significant difficulties in bridging the gap between language understanding and spatial reasoning. These challenges are further complicated by the requirement to integrate geometric and semantic information in a computationally efficient manner suitable for real-time applications, where processing constraints demand careful optimization of both memory and computational resources.

In this article, we present a novel zero-shot framework that addresses these challenges by seamlessly integrating geometric reconstruction with open-vocabulary VLMs. Our key insight is that by maintaining a unified semantic embedding space and employing efficient spatial indexing strategies, we can achieve robust real-time performance while handling the uncertainties inherent in 2-D segmentation outputs. This approach enables natural language interaction with the 3-D environment while maintaining both geometric and semantic consistency. Our framework addresses these challenges through several key innovations that enable real-time, zero-shot 3-D semantic mapping, with pipeline overview provided in Fig. 1. This capability is particularly relevant for applications such as robot navigation in unknown environments, where robots need to efficiently construct semantic maps of their environments in real-time as well as robustly accommodate sensor/segmentation noise and adapt to changing environments. The core of our approach lies in maintaining temporal consistency without requiring global optimization, achieved through an instance-level semantic embedding fusion combined with efficient spatial indexing and association strategies for fast 3-D tracking. Our key technical contributions can be summarized as follows.

1) A modular zero-shot 3-D semantic mapping framework leveraging pretrained VLMs to perform various open-vocabulary 3-D scene understanding tasks without training or fine-tuning.
2) An online, instance-level geometric and semantic fusion algorithm for Red Green Blue - Depth (RGB-D) streams enabling real-time mapping without global optimization.
3) A robust object association strategy combining R-tree spatial indexing with minimum-cost bipartite matching for fast 3-D tracking to effectively handle inconsistent masks and labels from 2-D segmentation.
4) A unified geometric-semantic update mechanism ensuring temporal consistency via instance-level tracking and supporting natural language interactions with the 3-D environment.

We demonstrate the effectiveness of our approach through extensive experiments on multiple open-vocabulary 3-D benchmarks, showing superior performance in tasks, such as instance retrieval, semantic segmentation, and instance segmentation. We demonstrate that our framework successfully bridges the gap between geometric reconstruction and semantic understanding, enabling robust 3-D scene understanding with natural language interaction capabilities in unconstrained environments. The ability to handle arbitrary objects without requiring task-specific training, combined with real-time processing of streaming data while maintaining both geometric and semantic consistency, makes our framework particularly suitable for real-world applications in environments that could dynamically change over time. In particular, considering robot operation in scenarios where environment geometries could have changed between successive visits to a region, our framework intrinsically addresses such environment changes through its robust object association, consistency-based pruning, and multihypothesis embedding banks. The rest of this article is organized as follows: Section II covers related work, Section III presents the problem and Section IV describes our framework, Section V shows experimental results. Finally, Section VI concludes this article.

II. Related Work

Traditional 3-D scene understanding methods remain constrained by closed-set categories predefined during training [1], [2], [3], [4], [5], limiting their adaptability to new object classes. Recent research leverages VLMs to attain *open-vocabulary* recognition in 3-D, expanding beyond fixed labels. For instance, OpenScene projects 3-D points into contrastive language-image pretraining (CLIP) embedding space, enabling text-driven queries with zero-shot capability [6], while Open3DIS and OpenMask3D aggregate multiview 2-D instance masks into coherent 3-D segments and associate them with language embeddings [7], [8]. These approaches achieve instance-level segmentation of novel classes but often rely on offline or batch processing, making them less suitable for real-time use. Open-vocabulary *3-D instance retrieval* has also been explored, notably by methods that fuse 2-D text-aligned proposals with

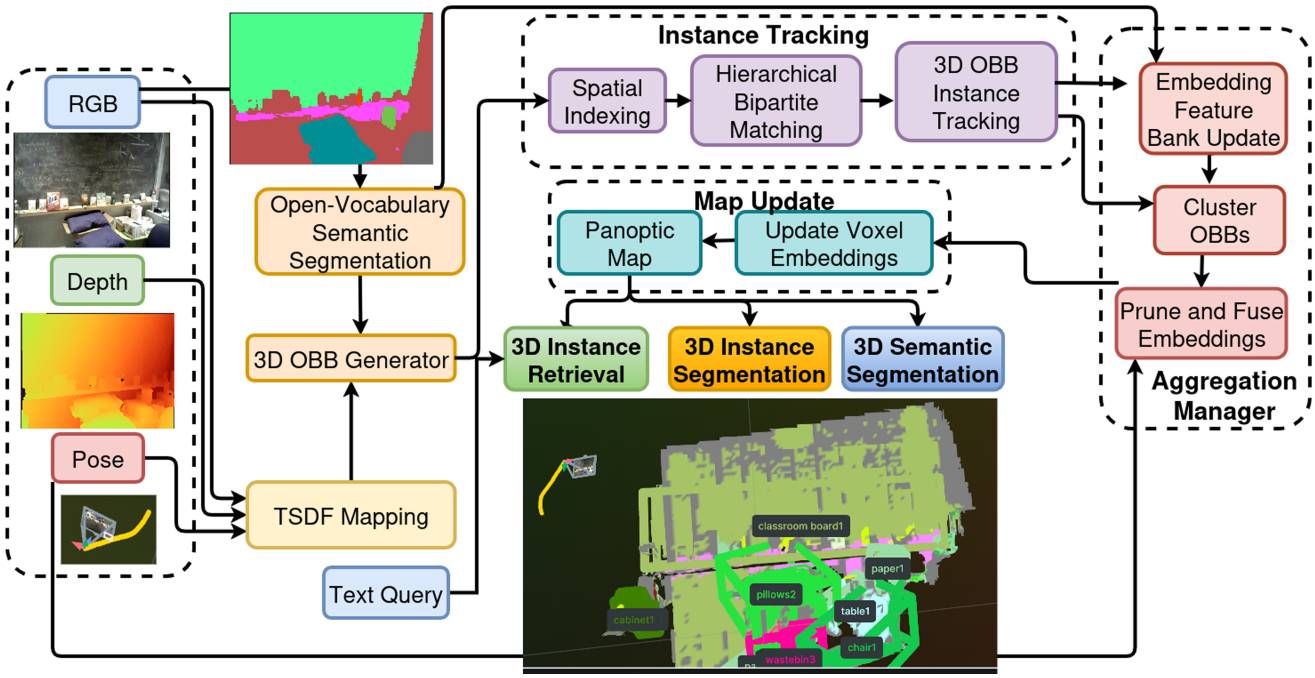

Fig. 2. *System-level architecture of our RAZER framework*. It processes RGB, depth, and pose inputs through three modules: (1) *Instance Tracking* to enable efficient feature updates, (2) *Aggregation Manager* to aggregate and fuse/prune instances and their corresponding coarse features, and (3) *Map Update* to update features at voxel level and their corresponding labels, thus generating a panoptic map that enables 3-D scene understanding.

incremental 3-D mapping [9], yet many do not offer online performance for robotics.

Incremental mapping integrates 2-D semantic cues (e.g., segmentation masks) into an evolving 3-D representation as a robot explores [10], [11]. Early systems like SemanticFusion [12] demonstrated real-time 3-D semantic mapping by fusing per-frame labels, while Voxblox++ [10] and Panoptic-Fusion [11] extended this to instance-level and panoptic segmentation. Although effective for known classes, these pipelines cannot accommodate unseen categories. Recent work has introduced post-hoc optimization or multiview consistency [13], [14] to mitigate frame-level label noise, but still within a closed-set taxonomy. Our approach builds on the strengths of real-time fusion but adopts an open-vocabulary 2-D model for segmentation, enabling online handling of novel objects without retraining. VLMs such as CLIP [15] and bootstrapping language-image pre-training (BLIP) [16] have accelerated progress in zero-shot classification and retrieval for 2-D images. Their extension to 3-D data includes methods like PointCLIP [17] and unified representation of language, images, and point clouds (ULIP) [18], which unify point cloud and text embeddings. These methods can recognize categories not present in any 3-D training set, but often operate offline and are computationally heavy. Meanwhile, approaches like ConceptGraphs [19] use multiview images and 2-D VLM predictions to label 3-D points or clusters, yielding open-vocabulary scene representations. However, they typically rely on batch (offline) processing or expensive clustering, limiting real-time deployment. The recent INS-CONV [20] combines the advantages of 2D–3D and 3D–3D procedures. While it builds the 3-D model of the environment incrementally from RGB-D frames, it performs segmentation directly in the 3-D space. This, however, comes at the cost of requiring 3-D ground truth annotations for training.

Recent works have significantly improved the performance of open-vocabulary 3-D segmentation using point clouds from LiDAR or depth data from RGB-D cameras [21], [22], [23], [24], [25], [26], [27], [28], [29]. In addition, there have also been approaches fusing embeddings from open world VLMs like CLIP with neural radiance fields, Gaussian splats or implicit neural network based representations [30], [31], [32], [33] for generating a consistent segmentation of the scene. The semantic information aids in both understanding the scene as well as improving the accuracy of the map and odometry for semantic SLAM [12], [34], [35], [36], [37], [38], [39], [40], [41], [42]. End-to-end 3-D methods that predict 3-D semantic information directly from sensor modalities like camera and depth, or from intermediate representations like reconstructed point clouds or voxel-based representations, can capture complex 3-D relationships. The recently proposed MaskClustering framework [43] is an offline, batch processing method that builds a global graph of all 2-D masks across a complete sequence and using multiview "view consensus" to cluster them for better segmentation through global optimization. Recent efforts also combine SLAM-based mapping with language models to produce richer 3-D scene graphs for tasks like grounded query and high-level planning [9], [19], [44], [45]. By coupling object-level mappings with language-based descriptions or relational knowledge, these systems enable queries such as *"retrieve the red chair in the corner"* and facilitate complex reasoning. However, computational overhead and reliance on offline graph construction remain common hurdles. Semantic mapping techniques, vital

IEEE Transactions on Robotics (T-RO) paper, presented at ICRA 2026, Vienna, Austria. Cite as T-RO paper.

for scene graph generation, extend beyond 2-D segmentation to encompass complex 3-D structures. Reconstruction of 3-D structures with object recognition and grouping remains a fundamental challenge in computer vision, especially in the context of 3-D scene understanding. These methods help a large language model (LLM) understand the scene through reasoning [46], [47] for building grounded 3-D LLMs [48], [49], [50], [51], [52], [53], [54].

The open vocabulary allows for greater flexibility in real-world applications and can help an LLM to understand the scene with multiple objects, which multimodal LLMs hallucinate or fail to perceive [55]. However, these methods do not scale well in larger scenes, making them computationally intensive and memory-consuming, rendering them impractical for real-time segmentation in robotic applications. In parallel, LLMs like GPT [56] have shown potential for reasoning over text-based representations of a scene [57], but bridging real-time 3-D reconstruction with LLM-driven semantics remains under-explored.

Our work aims to unify these threads by performing online, incremental 3-D semantic instance mapping *without* closed-set constraints, using a pretrained VLM for open-vocabulary segmentation. In contrast to methods that require domain-specific 3-D supervision or post-hoc processing, we achieve zero-shot instance detection and tracking on the fly. Also, in contrast to offline methods such as MaskClustering, our framework is an online, incremental processing method that performs per-frame, per-segment clustering of 2-D masks to generate 3-D instances. By maintaining object-level embeddings in a continuously updated 3-D map, we further enable tasks such as real-time retrieval of novel objects, making it practical for robotic semantic mapping for navigation.

III. Problem Formulation

We present an open-vocabulary 3-D scene understanding system that processes RGB-D streams with known camera poses to build semantically rich 3-D maps. Unlike traditional systems constrained to a fixed taxonomy of object categories, our approach enables unrestricted object recognition and tracking through continuous semantic embedding spaces. At each timestep t, our system processes an RGB image $I_t^{\text{rgb}} \in \mathbb{R}^{H\times W\times 3}$, a depth image $I_t^{\text{depth}} \in \mathbb{R}^{H\times W}$, and a camera pose $P_t \in \text{SE}(3)$, where H and W denote the image height and width, respectively.

Unlike traditional closed-set methods that rely on a predefined label space, our zero-shot approach harnesses a *pretrained* vision-language segmentation model capable of recognizing a wide range of object categories. Specifically, rather than fine-tuning or distilling from this model, we directly leverage its zero-shot capability to produce semantic embeddings for any concept of interest. A key innovation lies in maintaining a unified representation that encodes both geometric and semantic properties in a continuous embedding space. Rather than performing discrete classification into predefined categories, we leverage VLMs to embed objects in a high-dimensional semantic space $\mathcal{S} \in \mathbb{R}^d$, where semantically similar concepts are naturally clustered together, enabling new categories to be recognized at runtime without retraining. Our system therefore supports the following.

1) *Volumetric reconstruction* of the scene using a truncated signed distance function (TSDF) $\Phi : \mathbb{R}^3 \to \mathbb{R}$ with truncation distance τ.
2) *Object detection and 6-DoF tracking* with oriented bounding boxes (OBBs) $\mathcal{B} = \{b_i\}_{i=1}^N$, where $b_i = (c, R, s) \in \mathbb{R}^3 \times SO(3) \times \mathbb{R}^3$ denotes the center, orientation, and scale, respectively.
3) *Semantic embeddings* $\mathcal{E} = \{e_i\}_{i=1}^N$, where $e_i \in \mathbb{R}^d$ accumulates open-vocabulary features for object i using a *pretrained* VLM.
4) *Probabilistic object pruning* through confidence scores $\alpha_i \in [0, 1]$, updated based on temporal consistency and semantic stability.
5) *Online fusion of new observations* to incrementally refine both geometry and semantics, ensuring robust instance identity across multiple viewpoints.

As our system operates in a zero-shot manner, at each timestep, the system extracts semantic features from the incoming RGB-D frame using a frozen, pretrained 2-D model, and fuses them with previously observed data in 3-D space. This approach allows the model to generalize to novel object categories so long as they are recognizable by the pretrained 2-D backbone.

A fundamental challenge in this work is maintaining robust object tracking and coherent semantic labeling under such unconstrained open-vocabulary settings. When the system first detects an object, the initial embedding e_i^0 may reflect incomplete or occluded views. As more viewpoints become available, the system refines e_i^t through a temporal fusion function $e_i^t = f\left(e_i^{t-1}, z_i^t\right)$, where z_i^t denotes newly extracted semantic features at time t. This process ensures that object instance identity is preserved via geometric consistency and occlusion-aware updates, even when only partial observations are available in any single frame.

IV. Approach

A. Volumetric Scene Reconstruction

We discretize the environment into a volumetric grid and maintain a TSDF volume $V : \mathbb{R}^3 \to \mathbb{R}^6$. For each voxel at position $\mathbf{p} \in \mathbb{R}^3$, the stored tuple includes

$$V(\mathbf{p}) = \left\{ d, w, \mathbf{c}, l, h \right\} \tag{1}$$

where $d \in [-\tau, \tau]$ is the truncated signed distance to the nearest surface, $w \in [0, 1]$ is the accumulated confidence of d, $\mathbf{c} \in \mathbb{R}^3$ is an RGB color, $l \in \mathbb{N}$ is an instance label index, and h is a histogram tracking how frequently each instance label has been observed in that voxel.

At each incoming frame $(I_t^{\text{rgb}}, I_t^{\text{depth}})$, we fuse new depth measurements into the TSDF volume via standard weighted averaging

$$d_{\text{new}} = \omega_{\text{old}}\, d_{\text{old}} + \omega_{\text{new}}\, d_{\text{obs}}$$

where d_{obs} is the signed distance derived from the depth map, and $\omega_{\text{old}}, \omega_{\text{new}}$ are confidence weights. The voxel color $\mathbf{c}$ and instance histogram h are also updated if the voxel lies within

the truncated distance bound of a newly observed surface. This ensures that regions recognized as belonging to a particular object accumulate a consistent label history.

Through this process, our system continuously refines a dense 3-D representation that captures both geometry and instance labels. Unlike conventional methods that only update a fixed label, our open-vocabulary setting allows for dynamic incorporation of semantic cues from the pretrained model, even if new object classes appear over time.

B. 3-D Object Detection

Our detection and tracking system integrates both geometric and semantic cues by first processing each RGB frame with an open-vocabulary vision-language segmentation model $\mathcal{F}$, described further below. This model generates instance masks M_t for *any* objects it can visually separate, without being restricted to a finite set of categories

$$M_t = \mathcal{F}(I_t^{\text{rgb}}). \tag{2}$$

Since this is a zero-shot approach, we do not perform any additional training or distillation; instead, we rely on the model's ability to segment objects of interest based on its pretrained knowledge from vision-language data. In cluttered scenes, the model can produce a large number of masks, potentially including false positives. Hence, a confidence-based threshold or heuristic filters can be applied to remove implausible masks.

a) Vision-Language Segmentation Model: takes RGB frames as input and produces object masks at ~ 30 Hz on a DGX Spark. It is optimized for memory and compute via vectorized operations and TensorRT with FP16 quantization, and recognizes 560 object classes and 1306 text categories, while allowing new classes to be added on the fly. We use FC-CLIP [58] with a CLIP [15] ConvNeXt [59] backbone, which maintains translation equivariance and supports multiscale feature fusion for high-resolution robotic scene understanding.

The model $\mathcal{F}$ comprises a class-agnostic mask generator, an in-vocabulary classifier, and an out-of-vocabulary classifier, all operating on multiscale feature maps from a frozen ConvNeXt-L CLIP backbone. Following Mask2Former [60], the mask generator applies a pixel decoder with multiscale deformable attention, and a mask decoder that takes 250 learnable object queries and produces 100 candidate masks. Decoupling the number of queries and masks allows flexible object counts while generating masks only for valid predictions. For zero-shot classification, each class is represented by 768-dimensional text embeddings obtained using CLIP-style prompts (e.g., "a photo of a <label>", "an image of a <label>"), as illustrated in Fig. 3. For each class, its name and close synonyms which are the text categories (e.g., cabinet, cupboard, bureau, China cabinet) are inserted into these templates, tokenized [61], and encoded with the CLIP text encoder. Each synonym yields a distinct embedding, but all map back to a single class (e.g., "cabinet").

The out-of-vocabulary classifier pools features from the frozen CLIP backbone using mask pooling to preserve open-vocabulary behavior, while the in-vocabulary classifier pools features from the pixel decoder to improve performance on

Category Text Prompt Templates

- <label>.
- a photo of a <label>.
- a dark photo of a <label>.
- This is a photo of a <label>
- There is a <label> in the scene
- There is the <label> in the scene
- a photo of a <label> in the scene
- a photo of a small <label>.
- a photo of a medium <label>.
- a photo of a large <label>.
- This is a photo of a small <label>.
- This is a photo of a medium <label>.
- This is a photo of a large <label>.
- There is a small <label> in the scene.
- There is a medium <label> in the scene.
- There is a large <label> in the scene.
- a photo of a hard to see <label>.
- a bright photo of a <label>.
- a close-up photo of a <label>.
- a black and white photo of <label>.
- a bright photo of the <label>.
- a blurry photo of the <label>.
- a photo of the <label>.
- a close-up photo of the <label>.
- a low resolution photo of a <label>.
- a blurry photo of a <label>.

Fig. 3. Prompt templates used to generate text embeddings for each category. Each "<label>" is replaced with the corresponding category name.

training classes. Both compute cosine similarity between pooled features and all synonym embeddings, followed by a softmax with temperature 0.07; the highest-scoring synonym determines the predicted class. The two classifiers are combined via geometric ensembling, improving performance on both seen and unseen classes. Candidate masks are discarded if classified as background, have low confidence, or lack a dominant class. Large structural elements (ceilings, floors, walls) are segmented but excluded from object mapping. The label space includes COCO Panoptic classes [62], [63], [64] and additional indoor categories with synonyms, all mapped to consistent general names, enabling robust zero-shot segmentation and stable semantics for object mapping.

b) 3D Point Extraction: For each mask m_i in M_t, we extract the corresponding 3-D points from the depth map I_t^{depth}. Specifically, for every pixel (x, y) that lies within mask m_i, we back-project it into 3-D space

$$\mathbf{P}_i = \{\pi^{-1}(x, y, I_t^{\text{depth}}(x, y)) \mid (x, y) \in m_i\} \tag{3}$$

where π^{-1} is the camera intrinsics-based inverse projection function. Accurate calibration is assumed for correct 3-D point placement.

c) 3-D Clustering for Object Separation: Although a single instance mask in 2-D often corresponds to a single object, occlusions and overlapping masks can lead to mismatches between 2-D and 3-D object boundaries. To address this, we apply DBSCAN clustering in 3-D on the points $\mathbf{P}_i$ to distinguish distinct object clusters based on spatial density. DBSCAN parameters (e.g., ε, the neighborhood radius, and the minimum point count) must be tuned to balance merging and splitting errors.

d) OBB Fitting: For each resulting 3-D cluster, we compute an OBB using Principal Component Analysis (PCA). Given $\mathbf{P}_i$ for a cluster, we first compute the centroid $\mathbf{c}$

$$\mathbf{c} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{P}_j \tag{4}$$

and the covariance matrix

$$\mathbf{C} = \frac{1}{N-1}\sum_{j=1}^{N}(\mathbf{P}_j - \mathbf{c})(\mathbf{P}_j - \mathbf{c})^T. \tag{5}$$

Eigen-decomposition $\mathbf{C} = \mathbf{R}\mathbf{\Lambda}\mathbf{R}^T$ yields the principal axes $\mathbf{R} = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$. We ensure a right-handed coordinate system by reorienting $\mathbf{v}_3$ if necessary via $\mathbf{v}_3 \leftarrow \mathbf{v}_1 \times \mathbf{v}_2$. We compute the OBB extents by projecting the points onto each principal axis $\mathbf{v}_i$

$$e_i = \max_j \big(\mathbf{v}_i^T \mathbf{P}_j\big) - \min_j \big(\mathbf{v}_i^T \mathbf{P}_j\big). \tag{6}$$

Each object detection can thus be represented by the tuple $(\mathbf{c}, \mathbf{R}, \mathbf{s})$, describing the center, principal axes, and box dimensions, respectively.

C. R-Tree Based Hierarchical Association for Tracking

Having obtained new OBB detections, we must associate them with tracked objects from previous frames. This must be efficient, as real-time systems can track dozens of objects simultaneously.

a) R-Tree Organization: We store bounding boxes of tracked objects in an R-tree, which enables spatial queries in expected $\mathcal{O}(\log n)$ time. Each tracked object's OBB $b_i = (c_i, R_i, s_i)$ is converted to an axis-aligned bounding box (AABB)

$$\mathbf{b}_{\min}, \mathbf{b}_{\max} = \text{AABB\_Enclose}\big(c_i, R_i, s_i\big) \tag{7}$$

and this $(\mathbf{b}_{\min}, \mathbf{b}_{\max})$ is stored in a leaf node of the R-tree. Internal nodes recursively store the minimal AABBs enclosing their children, creating a spatial hierarchy to prune searches for overlapping or nearby objects.

b) Association via R-Tree Query: For a new detection $b_j^{(\text{new})}$, we generate its AABB and query the R-tree to retrieve only those tracked objects whose AABBs intersect or lie within a small distance. This reduces the candidate set from the entire pool of tracked objects to a manageable subset.

c) Candidate Matching and Hungarian Algorithm: Within this candidate subset, we resolve final matches using a bipartite matching approach

$$M_{ij} = w_v\, V_{ij} + w_s\, S_{ij}, \quad \min_{\mathbf{X}} \sum_{i,j} M_{ij}\, X_{ij} \tag{8}$$

where

1) $V_{ij} = 1 - \text{IoU}(\text{OBB}_i, \text{OBB}_j^{(\text{new})})$ is the geometric dissimilarity based on 3-D IoU (intersection over union).
2) $S_{ij} = \|\mathbf{e}_i - \mathbf{e}_j^{(\text{new})}\|_2$ represents a semantic distance measure. If the system has an embedding vector for each object, S_{ij} could combine both shape and open-vocabulary features. See Section IV-E for more details.
3) $X_{ij} \in \{0, 1\}$ are the elements of assignment matrix $\mathbf{X}$ that impose a one-to-one matching constraint.

We solve for the assignment matrix $\mathbf{X}$ to match candidate detections with tracked object OBBs using the Hungarian (Kuhn–Munkres) algorithm [65] in $\mathcal{O}(m^3)$ time, where m is the number of candidates. Due to the R-tree query, m is usually small, making real-time matching computationally feasible. Unmatched new detections spawn new tracks, while previously tracked objects that remain unmatched for k consecutive frames are pruned unless their semantic confidence is high (to handle long occlusions). This pruning applies only within the current camera frustum; global object persistence is maintained through the semantic map, which accounts for long-term occlusions, viewpoint changes, and spatial displacement as the robot navigates across rooms.

D. Incremental OBB Updates

For each matched object, we refine its OBB incrementally by assimilating newly observed 3-D points. This is particularly important when objects are only partially visible or change orientation over time.

a) Incremental covariance calculation: Let an object o_i at frame $t-1$ have a scatter matrix $\mathbf{S}_{t-1}$, centroid $\mathbf{c}_{t-1}$, and N_{t-1} points accumulated thus far. A new detection in frame t contributes N_{new} points $\mathbf{P}_{\text{new}}$, with centroid $\mathbf{c}_{\text{new}}$ and scatter matrix $\mathbf{S}_{\text{new}}$. We update

$$N_t = N_{t-1} + N_{\text{new}} \tag{9}$$

$$\mathbf{c}_t = \frac{N_{t-1}\,\mathbf{c}_{t-1} + N_{\text{new}}\,\mathbf{c}_{\text{new}}}{N_t} \tag{10}$$

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{S}_{\text{new}} + \frac{N_{t-1}\,N_{\text{new}}}{N_t}\big(\mathbf{c}_{t-1} - \mathbf{c}_{\text{new}}\big)\big(\mathbf{c}_{t-1} - \mathbf{c}_{\text{new}}\big)^T. \tag{11}$$

The new covariance matrix is $\mathbf{C}_t = \mathbf{S}_t \,/\, (N_t - 1)$. By performing eigen decomposition on $\mathbf{C}_t$, we derive updated principal axes and extents, which can expand or contract the bounding box based on the newly visible parts of the object.

b) R-Tree Synchronization: Finally, once OBB_i is updated to $(\mathbf{c}_t, \mathbf{R}_t, \mathbf{e}_t)$, we convert it back to an AABB via

$$\text{AABB\_Enclose}\big(\mathbf{c}_t, \mathbf{R}_t, \mathbf{e}_t\big)$$

and update o_i's entry in the R-tree accordingly. This ensures that future queries accurately reflect the object's most current spatial extent. Since no additional supervision or distillation is used, this approach handles unseen object types in a zero-shot manner and remains flexible across changing scene conditions.

E. Open-Vocabulary Semantic Embedding Management

At the core of our system is the management of open-vocabulary semantic embeddings for tracked objects. Each object o_i maintains a semantic state consisting of an embedding bank $\mathcal{E}_i = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ containing up to three concept embeddings, along with corresponding confidence scores $\mathcal{C}_i = \{c_1, c_2, c_3\}$ where $c_j \in [0, 1]$. This multiembedding approach enables maintaining uncertain hypotheses about an object's identity, crucial for handling ambiguous cases and novel objects.

When an object is first detected at time t, we initialize its semantic representation by extracting vision-language features $\mathbf{f}_t \in \mathbb{R}^{H \times W \times d}$ from the backbone network. Given the instance mask m_t, we compute the initial embedding through mask-guided feature aggregation

$$\mathbf{e}_t = \frac{1}{|m_t|} \sum_{(x,y) \in m_t} \mathbf{f}_t(x, y) \tag{12}$$

where $|m_t|$ denotes the number of valid pixels in the mask. This normalized pooling operation produces a fixed-dimensional embedding $\mathbf{e}_t \in \mathbb{R}^d$ that captures the object's semantic properties while being invariant to mask size.

As we accumulate observations, the embedding bank is updated according to semantic similarity and confidence scores. For a new observation embedding $\mathbf{e}_{\text{new}}$ with confidence c_{new}, we first compute similarities to existing embeddings

$$s_j = \cos(\mathbf{e}_{\text{new}}, \mathbf{e}_j), \quad j \in \{1, 2, 3\}. \tag{13}$$

If $\max_j s_j > \sigma_{\text{sim}}$ for similarity threshold σ_{sim}, we update the most similar existing embedding

$$\mathbf{e}_j = \frac{c_j \mathbf{e}_j + c_{\text{new}} \mathbf{e}_{\text{new}}}{c_j + c_{\text{new}}}, \quad c_j = c_j + c_{\text{new}}. \tag{14}$$

Otherwise, if $|\mathcal{E}_i| < 3$, we add $\mathbf{e}_{\text{new}}$ as a new hypothesis. This mechanism allows maintaining multiple semantic interpretations while consolidating consistent observations.

F. Semantic Map Management

The system maintains semantic consistency at both the voxel and object level. Each voxel v maintains a histogram h_v over observed instance labels and a maximum likelihood label l_v. For computational efficiency, we update these statistics only for voxels within object OBBs.

For tracked objects, we employ a support-based pruning mechanism. Given an object o_i with bounding box volume $|B_i|$, we compute its voxel support ratio

$$r_i = \frac{|\{v \in V | l_v = i\}|}{|B_i|}. \tag{15}$$

Objects with consistently low support ($r_i < \tau_{\text{supp}}$ for k consecutive frames) are candidates for pruning. However, for objects with high semantic confidence ($\max_j c_j > \tau_{\text{conf}}$), we maintain tracking even with temporarily low support to handle partial occlusions.

G. System Integration

The complete system operates as a tightly coupled pipeline that maintains both geometric and semantic consistency. Each incoming RGB-D frame $(I_t^{\text{rgb}}, I_t^{\text{depth}})$ first updates the volumetric reconstruction V using weighted averaging of signed distances to maintain an accurate geometric foundation. The VLM $\mathcal{F}$ then processes I_t^{rgb} to produce instance masks M_t, which are lifted to 3-D using the corresponding depth information from I_t^{depth}. The resulting 3-D point clouds undergo DBSCAN clustering and PCA-based OBB computation as described earlier.

Object tracking leverages the R-tree spatial index for efficient candidate selection, with final associations determined through Hungarian matching of the cost matrix $\mathbf{M}$. For successfully matched objects, the system performs a series of synchronized updates. The geometric state is refined through incremental covariance computation, maintaining accurate OBB estimates without storing historical point clouds. Simultaneously, the semantic state is updated by integrating new observations into the embedding bank $\mathcal{E}_i$ based on observation quality and similarity metrics. The system also updates voxel label histograms within the refined OBB boundaries to maintain spatial semantic consistency.

The map maintenance phase evaluates object persistence using the support ratio r_i and semantic confidence scores $\mathcal{C}_i$. This integrated approach enables robust open-vocabulary mapping by leveraging complementary strengths: geometric consistency guides object tracking and segmentation, while semantic embeddings resolve ambiguities and maintain object identity through significant viewpoint changes. The multihypothesis embedding bank is particularly crucial for handling uncertainty during partial observations while allowing refinement as more evidence becomes available. The tight coupling between geometric and semantic components enables the system to handle challenging scenarios such as object occlusions, novel object categories, and viewpoint variations while maintaining consistent semantic map.

The unified representation of geometry and open-vocabulary semantics enables a range of higher-level applications including *online 3-D instance segmentation*, *3-D instance retrieval*, and *3-D visual grounding* without requiring additional 3-D-domain training.

H. Online 3-D Instance Segmentation

While our approach already maintains *instance-level* object bounding boxes and per-voxel label histograms, it can directly provide a *3-D instance segmentation* of the scene as follows:

1) Each tracked object o_i has an identifier ID_i and a bounding box $b_i = (c_i, R_i, s_i)$. During the volumetric fusion step, all voxels within b_i are labeled with ID_i in their histograms h_v.
2) Whenever multiple objects overlap in 3-D, we maintain up to three hypothesis, allowing multiple semantic interpretations by fusing them over time as explained in Section IV-E, which are subsequently pruned based on voxel support ratio explained in Section IV-F.
3) Hence, at any point in time, *each voxel* in the TSDF volume carries the instance label $l \in \mathbb{N}$ (from ID_i). By aggregating all voxels labeled with the same ID_i, we obtain a complete 3-D instance mask for object o_i.

As this process is performed incrementally for each new RGB-D frame, it yields an *online* 3-D instance segmentation: after receiving t frames, the system can query the TSDF volume to retrieve the current segmentation. This is particularly helpful for applications like robotic manipulation, where a robot needs to know the volumetric extent of each object in real time. Notably, if a novel category appears (e.g., an object not in any fixed taxonomy), the pretrained model $\mathcal{F}$ can still segment it in 2-D, and our pipeline will produce a corresponding 3-D instance in the map.

I. 3-D Instance Retrieval

The open-vocabulary embeddings maintained for each object enable flexible 3-D instance retrieval through both text and visual queries. Each tracked object o_i stores multiple semantic embeddings $\mathbf{e}_{i,j}$ generated by a pretrained VLM $\mathcal{F}$ (see

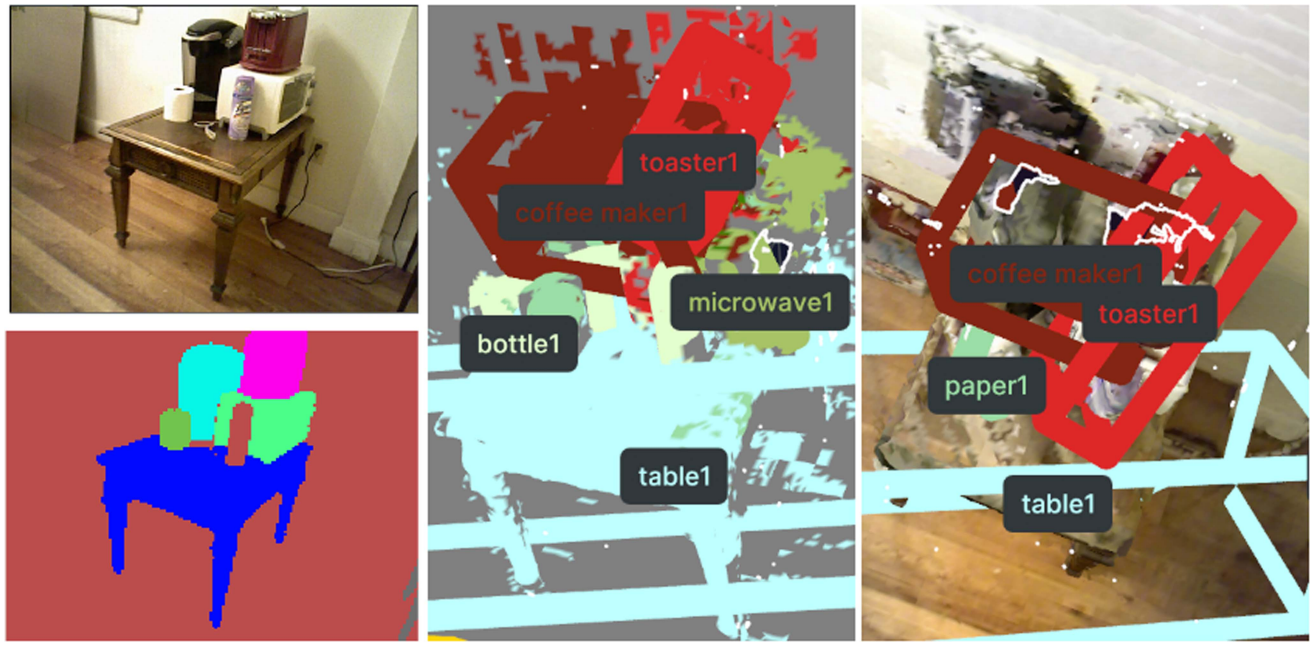

Fig. 4. Qualitative example of our system's output. *Left:* input RGB frame and corresponding 2-D open-vocabulary segmentation. *Center:* 3-D panoptic map with instance-labeled voxels reconstructed from the RGB–D stream. *Right:* 3-D OBBs aligned with the reconstructed geometry, demonstrating accurate 3-D localization of tabletop objects.

Section IV-E). For text-based retrieval, the query text and its associated prompt are processed through $\mathcal{F}$'s language encoder to produce a query embedding $\mathbf{e}_{\text{query}} \in \mathbb{R}^d$. The similarity between the query and each tracked object o_i is computed as the maximum cosine similarity across the object's embeddings. Objects can then be either ranked by similarity score or filtered using a threshold σ_{sim}, with objects exceeding this threshold considered matches. Matched objects can be visualized in 3-D using their bounding boxes or instance segmentation masks, facilitating physical interaction by users or robotic systems. The zero-shot nature of these embeddings enables retrieval using arbitrary natural language descriptions without requiring additional training.

V. Experiments

We demonstrate the modularity and effectiveness of our proposed framework by evaluating it across multiple benchmarks, including 3-D instance segmentation, instance retrieval, and semantic segmentation. We use five well-established indoor datasets for these tasks: SceneNN [66], ScanNet [67], ScanNetv2 [68], ScanNet200 [69], and Replica [70]. Next, we describe each dataset, outline evaluation metrics, and present comprehensive results comparing the performance of our method against recent approaches.

A. Qualitative Results for Representative Example

Fig. 4 illustrates how the proposed system in Fig. 2 transforms RGB–D observations into a 3-D semantic map with object-centric geometry. The left column shows the input RGB frame and the corresponding 2-D open-vocabulary segmentation produced by $\mathcal{F}$. The center view visualizes the reconstructed 3-D panoptic map with instance-labeled voxels, while the right view overlays the recovered 3-D OBBs on the map. Together, these views highlight the system's ability to consistently segment and localize tabletop objects (e.g., table, coffee maker, toaster, bottle, paper) in 3-D from a single egocentric observation stream.

Our system incrementally builds a large-scale 3-D semantic map as the agent explores the environment, as illustrated in Fig. 5. The TSDF map is updated online from the RGB–D stream and fused with open-vocabulary semantics, yielding a dense voxel map where each cell is assigned a semantic label. The three top–down views correspond to different spatial extents and stages of exploration, showing how the yellow trajectory progressively covers the scene while maintaining consistent labels for structural elements (e.g., floors, walls) and room-scale objects.

B. Mapping-Based 3-D Instance Segmentation

3-D instance segmentation is a key task for scene understanding, requiring systems to identify individual object instances in 3-D space. This task differs from semantic segmentation by distinguishing between multiple objects of the same class, which is particularly challenging in complex indoor environments. We aim to demonstrate the superior performance of our framework in the context of 3-D instance segmentation compared to existing volumetric mapping techniques, including TSDF-based methods, graph-based super-point strategies, and geometric-semantic fusion approaches. As illustrated in our evaluation, our approach outperforms these conventional methods significantly in terms of accuracy and efficiency.

SceneNN is an RGB-D dataset comprising over 100 reconstructed indoor scenes captured as RGB-D videos. Each scene

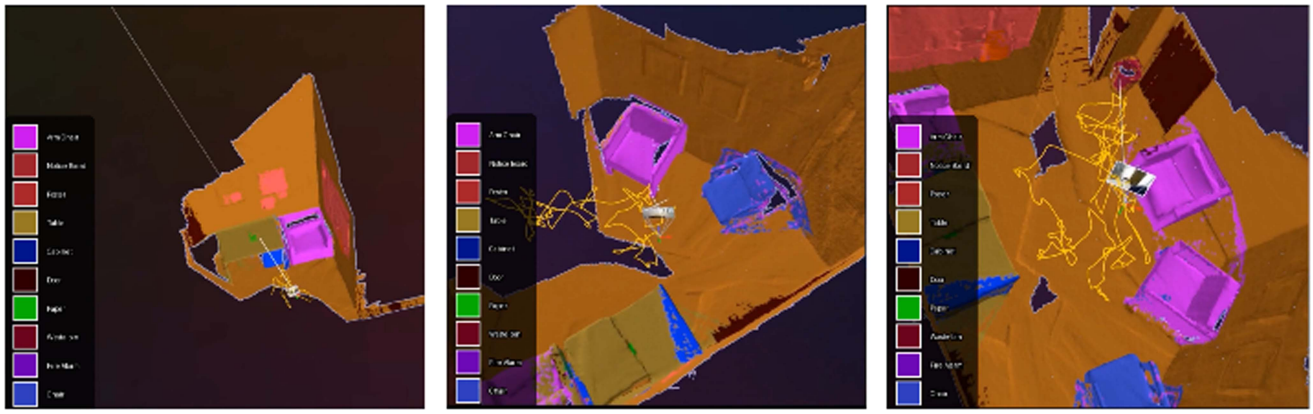

Fig. 5. Qualitative visualization of large-scale 3-D semantic mapping. From left to right, top–down views show the progressively completed TSDF map as the agent traverses the environment (yellow trajectory). Voxels are colored by open-vocabulary semantic labels using the legend on the left of each view, illustrating that our system maintains consistent semantics for floors, walls, furniture, and objects across the entire scene while the map is built online.

TABLE I
RESULTS FOR 3-D INSTANCE SEGMENTATION (MAP@50) ON THE SCENENN DATASET WITH GROUND TRUTH (TOP) AND ORB-SLAM3 (BOTTOM) POSES

Method / Seq.	11	16	30	61	78	86	96	206	223	255	**Avg.**
Ground truth trajectory											
Voxblox++	75	48.2	62.4	66.7	55.8	20	34.6	79.6	43.8	75	56.11
Han et al.	65.8	50	66.6	43.3	**100**	56.9	22.8	**92.1**	46.7	33	57.72
Wang et al.	62.2	43	60.7	36.3	49.3	45.8	32.7	46.6	56.4	47.9	47.90
Li et al.	78.6	25	58.6	46.6	69.8	47.2	26.7	78.0	48.5	75	55.40
Mascaro et al.	**100**	**75**	72.5	50	50	50	51.3	74.1	45.8	**100**	66.87
INS-CONV	**100**	62	83.4	**69.8**	93.7	60	57.6	56.7	**78.6**	**100**	76.18
VolumePanoptic	**100**	73.3	91.7	62.4	87.5	61.7	66.7	83.3	60	**100**	78.66
Ours	**100**	74.1	**91.8**	63.4	88.4	**62.3**	**66.8**	83.5	61.2	**100**	**79.15**
With ORB-SLAM3 Trajectory											
Voxblox++ [10]	61.5	38.9	50	58.4	44.3	16.4	27.6	48.7	40.7	33.6	42.01
Han et al. [72]	53.4	43.2	50	37.6	75.6	48.2	13.4	57.8	44.1	24.7	44.8
INS-CONV [20]	75	46.7	56.4	57.1	**83.7**	22.4	**48.1**	28.1	50	28.1	53.27
VolumePanoptic [75]	75	56.7	72.3	62.4	68.9	55.6	33.2	40.8	**65.2**	63.3	58.82
Ours	**75.3**	**57.6**	**73.2**	**63.4**	69.2	**59.2**	41.7	**58.8**	61.2	**63.4**	**62.3**

The bold and underlined values indicate the best and second-best results, respectively.

is provided as a textured triangle mesh with pervertex semantic and instance annotations. The dataset includes detailed object instance labels, camera trajectories, bounding boxes, and raw RGB-D frames, making it particularly useful for evaluating 3-D instance segmentation, semantic segmentation, and instance retrieval tasks. We conduct experiments and compare the proposed method with multiple state-of-the-art frameworks [10], [20], [71], [72], [73], [74], [75] on the SceneNN [66] dataset following the same setting proposed in [75]. We use mean average precision (mAP) metric to compare accuracy, computed by thresholding the IoU at thresholds 0.5. As per standard practice, we run methods using GT camera poses. In addition, we run all approaches on poses estimated by ORB-SLAM3 [76] to demonstrate their effectiveness in real-world settings. For SLAM-based experiments, we limit our evaluations to [10], [20], [72], [75], similar to the evaluation in [75].

As demonstrated in Table I, our method achieves superior performance on the SceneNN dataset. With ground truth trajectories, our approach achieves 79.15% mAP@50, outperforming the previous state-of-the-art VolumePanoptic (78.66%) framework. The performance gain is consistent across individual sequences, with our method achieving perfect scores on sequences 11 and 255. Using ORB-SLAM3 estimated trajectories, our method maintains robust performance (62.3% mAP@50), significantly surpassing VolumePanoptic (58.82%) and other methods. This improvement demonstrates our framework's effectiveness in real-world scenarios with imperfect pose estimation. However, it is to be noted that pose inaccuracies when using SLAM-generated trajectories does result in reduced performance, especially in the cluttered scenes in this dataset, since pose inaccuracies affect the accuracy and robustness of matching over successive frames and the spatio-temporal consistency in map generation.

C. 3-D Open Vocabulary Instance Segmentation

We evaluate the performance of our framework on the 3-D instance segmentation task using the ScanNet200 dataset,

TABLE II
3-D INSTANCE SEGMENTATION RESULTS ON THE SCANNET200 VALIDATION SET ON HEAD, COMMON, AND TAIL CLASSES

Method	mAP	mAP50	mAP25	Head	Common	Tail	Time/scene (s)
SAM3D	6.1	14.2	21.3	7.0	6.2	4.6	482.60
OVIR-3D	13.0	24.9	32.3	14.4	12.7	11.7	466.80
Open3DIS	23.7	29.4	32.8	**27.8**	21.2	**21.8**	360.12
OpenScene (2D Fusion)	11.7	15.2	17.8	13.4	11.6	9.9	46.45
OpenScene (Ensemble)	5.3	6.7	8.1	11.0	3.2	1.1	46.78
OpenMask3D	15.4	19.9	23.1	17.1	14.1	14.9	553.87
RAZER	**24.7**	**31.7**	**36.2**	**27.8**	**24.3**	21.6	**24.32**

The bold and underlined values indicate the best and second-best results, respectively.

TABLE III
3-D SEMANTIC SEGMENTATION RESULTS ON REPLICA AND SCANNET

Method	CLIP Backbone	Replica			ScanNet		
		mIoU	f-mIoU	f-mAcc	mIoU	f-mIoU	f-mAcc
ConceptFusion	OVSeg	0.10	0.21	0.16	0.08	0.11	0.15
	ViT-H-14	0.10	0.18	0.17	0.11	0.12	0.21
ConceptGraph	OVSeg	0.13	0.27	0.21	0.15	0.18	0.23
	ViT-H-14	0.18	0.23	0.30	0.16	0.20	0.28
HOV-SG	OVSeg	0.144	0.255	0.212	0.214	0.258	0.420
	ViT-H-14	0.231	0.386	0.304	0.222	0.303	0.431
Ours	OVSeg	**0.320**	**0.553**	**0.414**	**0.393**	**0.508**	**0.601**

The bold and underlined values indicate the best and second-best results, respectively.

employing IoU and average precision (AP) metrics. IoU measures the overlap between predicted and ground-truth instances. AP summarizes performance across multiple IoU thresholds (at 25% and 50%), integrating precision and recall into a single metric. These metrics provide insights into the model's ability to accurately segment individual object instances in complex 3-D scenes. Finally, we also report the average time required to compute the scene's representations, measuring clock wall time on a GPU RTX-4090, and for our method, we report in seconds the average time spent to process a scene. The ScanNet200 dataset encompasses 200 diverse semantic classes, categorized based on their frequency into head (66 most frequent classes), common (68 moderately frequent classes), and tail (66 least frequent classes), covering a wide range of indoor object categories and facilitating a thorough evaluation of segmentation performance across realistic scenarios.

As demonstrated in Table II, our proposed method, RAZER, achieves state-of-the-art results with 24.7% mAP, 31.7% mAP50, and 36.2% mAP25, surpassing existing methods on the majority of metrics. Specifically, RAZER shows strong performance in head and common categories (27.8% and 24.3% respectively), while remaining competitive in tail categories (21.6%). In addition, our method is the most computationally efficient, processing each scene in only 24.32 s, which is over an order of magnitude faster than previous approaches such as OpenMask3D (553.87 s) and SAM3D (482.60 s).

D. 3-D Open Vocabulary Segmentation

We demonstrate the efficacy of our framework for 3-D open-vocabulary segmentation task by demonstrating its performance on ScanNet and Replica datasets. For semantic segmentation tasks, we use mean intersection over union (mIoU) and pixel-wise accuracy. mIoU calculates the average overlap between predicted and ground-truth segmentation across all classes, effectively balancing performance evaluation for both frequent and rare classes. Pixel-wise accuracy measures the overall fraction of correctly predicted pixels, providing a straightforward but less class-sensitive performance measure. These metrics collectively capture the detailed performance characteristics of semantic segmentation models. The quantitative performance is evaluated by labeling the vertices of ground-truth meshes, and computing 3-D mIoU and mean accuracy (mAcc) versus ground-truth labels. We also report the metrics weighted by the frequency of the labels in the ground-truth (f-mIoU and f-mAcc).

ScanNet presents particular challenges for semantic segmentation due to its diverse indoor environments, varying lighting conditions, and complex spatial arrangements. The dataset's rich variety of object categories with different sizes, shapes, and textures makes it an ideal testbed for evaluating the generalization capabilities of open-vocabulary approaches. Furthermore, the presence of partial occlusions and varying object densities across scenes tests a model's ability to resolve contextual relationships. Table III demonstrates our method's exceptional performance on ScanNet, achieving 0.393 mIoU, 0.508 f-mIoU, and 0.601 f-mAcc with the OVSeg backbone. These results represent substantial improvements over previous approaches – our method nearly doubles the mIoU score of HOV-SG with ViT-H-14 (0.222 mIoU). This performance gap illustrates the effectiveness of our framework's semantic feature propagation mechanism, which better preserves fine-grained details and handles class boundaries more precisely.

Replica is a dataset of highly realistic indoor scenes designed primarily for simulation and embodied perception tasks. It includes 18 densely reconstructed environments provided as high-resolution textured meshes with semantic and instance annotations. Replica facilitates realistic simulation and evaluation of semantic segmentation and instance segmentation algorithms, serving as a critical dataset for evaluating models aimed at real-world applicability through simulation-to-real transfer. We evaluate performance using mIoU and mAcc, which are standard metrics for open-vocabulary 3-D semantic segmentation. We report f-mIoU and f-mAcc metrics that exclude background classes, similar to prior work. As shown in Table III, our method achieves 0.320 mIoU, 0.553 f-mIoU, and 0.414 f-mAcc with the OVSeg backbone, substantially outperforming previous state-of-the-art methods. Compared to HOV-SG with ViT-H-14 (0.231 mIoU, 0.386 f-mIoU, and 0.304 f-mAcc), our approach demonstrates improvements across all metrics.

E. Instance Retrieval

ScanNet [67] is an RGB-D video dataset consisting of approximately 1500 room scans reconstructed into textured meshes with detailed semantic and instance-level annotations. ScanNetv2 [68], an updated version with refined annotations, comprises 1,513 scenes commonly split into training, validation, and test sets. ScanNet200 [69] extends the dataset by providing annotations for 200 detailed semantic classes, significantly increasing annotation granularity. These datasets support benchmarking for 3-D semantic segmentation, instance segmentation, and instance

TABLE IV
PERFORMANCE ON SCANNETV2 3-D INSTANCE RETRIEVAL TASK IN TERMS OF TOP-1 ACCURACY (%) OF INSTANCE CLASSIFICATION

Method	Avg.	Bed	Cab	Chair	Sofa	Tabl	Door	Wind	Bksf	Pic	Cntr	Desk	Curt	Fridg	Bath	Showr	Toil	Sink
PointCLIP	6.3	0	0	0	0	0.7	0	0	**91.8**	0	0	0	15	0	0	0	0	0
PointCLIP V2	11.0	0	0	23.8	0	0	0	7.8	0	**90.7**	0	0	0	0	64.4	0	0	0
CLIP2Point	24.9	20.8	0	85.1	43.3	26.5	69.9	0	20.9	1.7	31.7	27	0	1.6	46.5	0	22.4	25.6
PointCLIP w/ TP.	26.1	0	55.7	72.8	5.1	1.7	0	**77.2**	0	0	51.7	0	40.3	**85.3**	4.9	0	0	34.9
CLIP2Point w/ TP.	35.2	11.8	45.1	27.6	10.5	61.5	2.7	1.9	0.3	33.6	29.9	4.7	11.5	72.2	**92.4**	86.1	3	34
CLIP^2	38.5	32.6	**67.2**	69.3	42.3	18.3	19.1	4	62.6	1.4	12.7	52.8	40.1	9.1	59.7	4	17.1	45.5
Uni3D	45.8	58.5	3.7	78.8	**83.7**	54.9	31.3	39.4	70.1	35.1	1.9	27.3	**94.2**	13.8	38.7	10.7	88.1	47.6
OpenIns3D	60.8	85.2	27.4	**87.6**	77.3	46.9	54.8	64.2	71.4	9.9	80.8	82.7	71.6	61.4	38.7	0	87.9	85.7
Ours	**61.2**	**85.3**	36.9	86.3	74.2	**76.5**	**72.3**	74.6	73.2	35.6	**81.2**	**83.4**	74.8	71.4	58.2	**87.2**	**92.3**	**86.8**

The bold and underlined values indicate the best and second-best results, respectively.

TABLE V
RUNTIME (IN MS) ON THE 10 SEQUENCES OF THE SCENENN DATASET

	Volume Panoptic		**RAZER**	**RAZER**
	Quadro RTX 5000		Quadro RTX 5000	DGX Spark
2D Inst. seg.	216.0	2D Inst. Seg.	82.3	32.9
Super-point seg.	70.3	3D OBB Det.	1.7	1.7
Graph update	127.2	3D		
Semantic reg.	324.0 (once per map)	OBB	18.4	25.9
Instance ref.	9.4 (once per map)	Tracking		
Emb. update	–	Emb. Update	0.8	1.2
Average Total	413.5 (per frame) + 333.4		103.2	61.7

retrieval tasks due to their detailed labeling in a variety of indoor environments.

We evaluate our framework's 3-D instance retrieval performance on ScanNetv2. Following the setting similar to [77], the "other furniture" class in ScanNetv2 is excluded and evaluated in terms of the Top-1 accuracy of instance classification. The instance classification is directly performed on the feature embeddings corresponding to the OBBs generated from our framework. Table IV demonstrates our method's state-of-the-art performance in 3-D instance retrieval, achieving 61.2% average Top-1 accuracy across all classes. Our approach outperforms OpenIns3D (60.8%) and significantly surpasses other methods like Uni3D (45.8%) and CLIP^2 (38.5%). Notably, our method excels in several challenging categories, achieving the highest accuracy for beds (85.3%), tables (76.5%), doors (72.3%), windows (74.6%), bookshelves (73.2%), counters (81.2%), desks (83.4%), curtains (74.8%), bathtubs (58.2%), showers (87.2%), toilets (92.3%), and sinks (86.8%). This consistent performance across diverse object categories demonstrates the robustness of our feature representation. The framework's integrated 3-D OBB detection and tracking mechanism plays a crucial role in this superior performance, enabling more precise object localization and maintaining temporal consistency of instance identities throughout the scene reconstruction process.

F. Runtime Analysis

We conduct a detailed runtime analysis of our framework to evaluate its computational efficiency compared to the prior state-of-the-art semantic mapping framework on SceneNN, VolumePanoptic [75]. Table V presents a component-wise breakdown of processing times for both approaches on the NVIDIA Quadro RTX 5000 GPU. Separately, we also tested our framework on the NVIDIA DGX Spark as also summarized in Table V.

For VolumePanoptic [75], the computational pipeline consists of multiple stages: 2-D instance segmentation (216.0 ms) which handles the initial detection and segmentation of objects in RGB images, super-point segmentation (70.3 ms) for grouping 3-D points into coherent surface patches, graph update (127.2 ms) to maintain the hierarchical scene representation, semantic regularization (324.0 ms, performed once per map) for refining semantic labels across super-points, and instance refinement (9.4 ms, performed once per map) to resolve instance ambiguities. In contrast, our method (RAZER) achieves significantly faster processing across all components. For 2-D instance segmentation, we reduce computation time to 82.3 ms through our optimized architecture. Our 3-D OBB detection component operates at just 1.7 ms per frame, while our efficient 3-D OBB tracking requires only 18.4 ms. The embedding update module runs at a remarkable 0.8 ms, demonstrating the lightweight nature of our feature propagation mechanism.

Our approach replaces the computation-heavy super-point segmentation with a more efficient OBB detection algorithm operating at just 1.7 ms per frame. This results in a 41$\times$ speedup for this component by directly estimating geometric primitives from point clouds rather than performing dense point-wise grouping. The 3-D OBB tracking component (18.4 ms) represents a fundamental departure from VolumePanoptic's graph-based approach. Where VolumePanoptic requires maintaining and updating a complex graph structure (127.2 ms) with nodes representing super-points and edges encoding spatial-semantic relationships, our OBB tracking employs a more direct geometric approach. By representing objects as OBBs, we perform efficient spatial association and motion estimation without the overhead of graph operations. This simplification not only reduces computation time by approximately 85%, but also improves robustness by eliminating the cascading errors that can occur in graph-based representations when initial segmentations are noisy. Our approach further benefits from the inherent geometric constraints of rigid objects, allowing for more consistent tracking over time without relying on potentially unstable point-wise feature correspondences.

Our embedding update module runs at a remarkable 0.8 ms, demonstrating the lightweight nature of our feature propagation mechanism. Rather than propagating features through a complex graph network requiring multiple message-passing iterations, we directly update our compact OBB-based

TABLE VI
AVERAGE RUNTIME ON THE REPLICA SCENES DATASET (ON QUADRO RTX 5000)

Method	Time/scene
HOV-SG	11h 12m
OpenNeRF	19m 3s
OVO-mapping	8m 17s
Ours	**3m 48s**

representation with new observations, maintaining semantic consistency through efficient feature averaging and outlier rejection.

Overall, our framework achieves an average total runtime of 103.2 ms per frame on the Quadro RTX 5000, representing a 4× speedup compared to VolumePanoptic's 413.5 ms (plus an additional 333.4 ms for one-time map processing). This substantial efficiency improvement makes our approach more suitable for real-time applications while maintaining superior performance as demonstrated in Table I, making it particularly valuable for 3-D scene understanding tasks in robotics applications. The combination of improved accuracy and significantly reduced computational requirements enables deployment on platforms with limited resources, opening possibilities for autonomous navigation, manipulation, and human-robot interaction in complex environments that require detailed semantic understanding.

Table VI further highlights our method's efficiency on the Replica dataset. Our approach completes scene processing in just 3 min and 48 s, substantially outperforming competing methods such as OVO-mapping (8 m 17 s), OpenNeRF (19 m 3 s), and especially HOV-SG (11 h 12 m). This dramatic reduction in processing time demonstrates the exceptional computational efficiency of our framework, making it practical for large-scale deployment in real-world scenarios. These computational performance improvements come from the modularity of our framework that enables it to update and compute scene and object feature embeddings swiftly for 3-D open-vocabulary semantic segmentation.

G. Qualitative Ablations

Our aggregation framework maintains multiple semantic hypotheses for each object and postpones hard decisions until sufficient evidence is available. Instead of committing to a single label, we keep several candidate identities in the embedding feature bank (e.g., *chair* vs. *armchair*) and update their scores as new views are integrated. When one hypothesis becomes consistently better supported, the system smoothly switches the active label while preserving the same underlying 3-D geometry and track, avoiding brittle label flips and enabling more reliable long-horizon reasoning.

While Fig. 6 illustrates how our aggregation manager keeps multiple competing semantic hypotheses for each object and switches labels only when one becomes dominant, a single-hypothesis baseline behaves quite differently. As shown in Fig. 7, committing to a single label at each update causes the object to oscillate between visually similar categories (e.g., *chair* and *armchair*), leading to temporal label flickering and inconsistent semantics in the 3-D map.

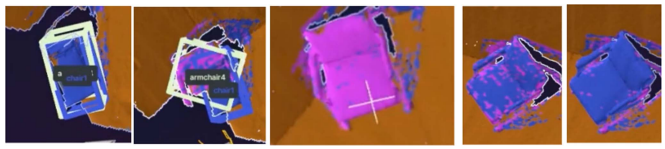

Fig. 6. Qualitative example of our multihypothesis semantic tracking. From left to right, the same 3-D object is represented by overlapping OBBs with different candidate labels (e.g., *chair* and *armchair*). Our method maintains these alternative hypotheses as voxel features and, as more observations are fused, can switch the active label from one hypothesis to another without breaking the underlying 3-D track.

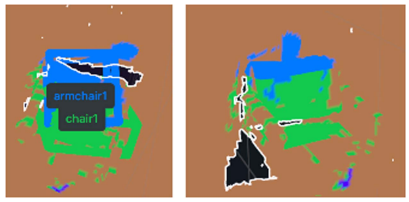

Fig. 7. Single-hypothesis semantic aggregation baseline. The object is assigned a single label at each update, causing it to alternate between categories such as *chair* and *armchair* across views. Without maintaining multiple hypotheses, the system cannot reconcile these conflicting predictions, resulting in unstable and inconsistent semantics in the reconstructed 3-D map.

Single-metric association like spatial/proximity only or feature based only strategies exhibit systematic failure modes that our multimetric fusion explicitly addresses. In an office scene with four identical chairs around a conference table, where adjacent chairs are very close to each other, a spatial-only approach incorrectly merges nearby chairs whose centers fall below the distance threshold, failing to maintain them as distinct instances. A feature-only approach can separate the chairs when viewed from the front (different 3-D positions induce distinct appearance embeddings), but catastrophically fragments the *same* chair when the camera rotates 180°: the front and back views look so different that they are treated as separate objects, producing spurious identity switches. A geometric-only approach based on 3-D OBB IoU similarly fails for two monitors of identical size placed side-by-side, where high overlap leads to an erroneous merge into a single instance. Our multimetric fusion resolves all three failure modes: spatial proximity and geometric consistency maintain track continuity for each physical object, while feature similarity disambiguates side-by-side chairs and monitors in appearance space. When the camera rotates 180°, spatial and geometric cues preserve the identity of each chair despite degraded feature similarity, correctly associating the back-view observations with the same chair seen from the front.

VI. CONCLUSION

We present a novel zero-shot framework for real-time 3-D semantic mapping that bridges geometric reconstruction and semantic understanding through a unified embedding space.

By combining efficient spatial indexing with instance-level semantic fusion, we demonstrate superior performance in handling streaming data without requiring global optimization. The framework processes inconsistent 2-D segmentation outputs while maintaining both geometric and semantic coherence in real-time, representing a significant advancement in open-vocabulary 3-D scene understanding. Experimental results validate our approach's effectiveness across multiple benchmarks while maintaining real-time performance suitable for robotics applications. This work opens new avenues for research in embodied AI systems that integrate geometric, semantic, and linguistic understanding, enabling more sophisticated human–robot interactions in unconstrained environments.

Limitations: While our framework demonstrates robust performance across various benchmarks, there are a few limitations to be addressed in future research. The system assumes relatively static environments and may have limited efficacy in fast-changing scenes with dynamic objects, as the temporal consistency mechanisms rely on stable object persistence across frames. Furthermore, deformable objects or items that undergo state changes (e.g., doors opening/closing, robot arms, articulated furniture) pose challenges for the rigid bounding box representation and incremental covariance updates. Also, inconsistent segmentation over successive frames can particularly impact smaller objects that are intermittently labeled resulting in them sometimes being removed during thresholding for consistency of the constructed model. In environments with sparse visual features or semantically ambiguous regions (e.g., textureless walls, uniform surfaces), the VLM may fail to generate meaningful segmentations, limiting the system's ability to build comprehensive semantic maps. In addition, while our approach includes mechanisms to handle inconsistent segmentations through multihypothesis embedding banks and support-based pruning, severe or sustained segmentation inconsistencies over multiple frames can impact accuracy and mapping stability. We plan to address these in future work.

REFERENCES

[1] X. Wang, S. Liu, X. Shen, C. Shen, and J. Jia, "Associatively segmenting instances and semantics in point clouds," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 4096–4105.

[2] C. Elich, F. Engelmann, T. Kontogianni, and B. Leibe, "3D bird's-eye-view instance segmentation," in *Proc. German Conf. Pattern Recognit.*, G. A. Fink, S. Frintrop, and X. Jiang, Eds. Dortmund, Germany, Sep. 2019, pp. 48–61.

[3] J. Lahoud, B. Ghanem, M. R. Oswald, and M. Pollefeys, "3D instance segmentation via multi-task metric learning," in *Proc. Int. Conf. Comput. Vis.*, Seoul, South Korea, Oct. 2019, pp. 9255–9265.

[4] J. Hou, A. Dai, and M. Nießner, "3D-SIS: 3D semantic instance segmentation of RGB-D scans," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 4421–4430.

[5] B. Yang et al., "Learning object bounding boxes for 3D instance segmentation on point clouds," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Vancouver, BC, Canada, Dec. 2019, pp. 6737–6746.

[6] J. Wu et al., "Towards open vocabulary learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 5092–5113, Jul. 2024.

[7] C. Zhu and L. Chen, "A survey on open-vocabulary detection and segmentation: Past, present, and future," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 8954–8975, Dec. 2024.

[8] M. Xu et al., "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Oct. 2022, pp. 736–753.

[9] D. Maggio et al., "Clio: Real-time task-driven open-set 3D scene graphs," *IEEE Robot. Automat. Lett.*, vol. 9, no. 10, pp. 8921–8928, Oct. 2024.

[10] M. Grinvald et al., "Volumetric instance-aware semantic mapping and 3D object discovery," *IEEE Robot. Automat. Lett.*, vol. 4, no. 3, pp. 3037–3044, Jul. 2019.

[11] G. Narita, T. Seno, T. Ishikawa, and Y. Kaji, "Panopticfusion: Online volumetric semantic mapping at the level of stuff and things," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Macau, SAR, China, Nov. 2019, pp. 4205–4212.

[12] M. Rünz, M. Buffier, and L. Agapito, "Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects," in *Proc. Int. Symp. Mixed Augmented Reality*, D. Gabbard, J. L. Chu, J. Grubert, and H. Regenbrecht, Eds. Munich, Germany, Oct. 2018, pp. 10–20.

[13] L. Han, T. Zheng, L. Xu, and L. Fang, "Occuseg: Occupancy-aware 3D instance segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 2937–2946.

[14] S. Chen, J. Fang, Q. Zhang, W. Liu, and X. Wang, "Hierarchical aggregation for 3D instance segmentation," in *Proc. Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 15447–15456.

[15] A. Radford, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, Vienna, Austria, Jul. 2021, pp. 8748–8763.

[16] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, Honolulu, HI, USA, Jul. 2023, pp. 19730–19742.

[17] R. Zhang et al., "PointCLIP: Point cloud understanding by CLIP," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 2022, pp. 8542–8552.

[18] L. Xue et al., "ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 1179–1189.

[19] Q. Gu et al., "Conceptgraphs: Open-vocabulary 3D scene graphs for perception and planning," in *Proc. Int. Conf. Robot. Automat.*, Yokohama, Japan, May 2024, pp. 5021–5028.

[20] L. Liu, T. Zheng, Y. Lin, K. Ni, and L. Fang, "Ins-conv: Incremental sparse convolution for online 3D segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 2022, pp. 18953–18962.

[21] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "PLA: Language-driven open-vocabulary 3D scene understanding," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 7010–7019.

[22] Z. Jin, M. Hayat, Y. Yang, Y. Guo, and Y. Lei, "Context-aware alignment and mutual masking for 3D-language pre-training," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 10984–10994.

[23] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li, "3D-VisTA: Pre-trained transformer for 3D vision and text alignment," in *Proc. Int. Conf. Comput. Vis.*, Paris, France, Oct. 2023, pp. 2911–2921.

[24] S. Chen et al., "Vote2Cap-DETR : Decoupling localization and describing for end-to-end 3D dense captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 11, pp. 7331–7347, Nov. 2024.

[25] D. Z. Chen, A. Gholami, M. Nießner, and A. X. Chang, "Scan2Cap: Context-aware dense captioning in RGB-D scans," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, Jun. 2021, pp. 3193–3203.

[26] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3D object localization in RGB-D scans using natural language," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 202–221.

[27] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3D: Mask transformer for 3D semantic instance segmentation," in *Proc. Int. Conf. Robot. Automat.*, London, U.K., May 2023, pp. 8216–8223.

[28] Y. Yue et al., "AGILE3D: Attention guided interactive multi-object 3D segmentation," in *Proc. Int. Conf. Learn. Representations*, Vienna, Austria, May 2024.

[29] J. Zhou, J. Wang, B. Ma, Y.-S. Liu, T. Huang, and X. Wang, "Uni3D: Exploring unified 3D representation at scale," in *Proc. Int. Conf. Learn. Representations*, Vienna, Austria, May 2024.

[30] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "LERF: Language embedded radiance fields," in *Proc. Int. Conf. Comput. Vis.*, Paris, France, Oct. 2023, pp. 19672–19682.

[31] H. Zhang, F. Li, and N. Ahuja, "Open-NeRF: Towards open vocabulary NeRF decomposition," in *Proc. Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, Jan. 2024, pp. 3444–3453.

[32] Y. Wang, H. Chen, and G. H. Lee, "GOV-NeSF: Generalizable open-vocabulary neural semantic fields," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2024, pp. 20443–20453.

[33] T. Nguyen, A. Bourki, M. Macudzinski, A. Brunel, and M. Bennamoun, "Semantically-aware neural radiance fields for visual scene understanding: A comprehensive review," 2024, *arXiv:2402.11141*.

[34] A. Rosinol et al., "Kimera: From SLAM to spatial perception with 3D dynamic scene graphs," *Int. J. Robot. Res.*, vol. 40, no. 12-14, pp. 1510–1546, 2021.

[35] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. J. Davison, and S. Leutenegger, "Mid-fusion: Octree-based object-level multi-instance dynamic SLAM," in *Proc. Int. Conf. Robot. Automat.*, Montreal, QC, Canada, May 2019, pp. 5231–5237.

[36] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented SLAM," *IEEE Robot. Automat. Lett.*, vol. 4, no. 1, pp. 1–8, Jan. 2019.

[37] N. Patel, P. Krishnamurthy, and F. Khorrami, "Semantic segmentation guided SLAM using vision and LiDAR," in *Proc. Int. Symp. Robot.*, Munich, German, Jun. 2018, pp. 1–7.

[38] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison, "SLAM : Simultaneous localisation and mapping at the level of objects," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 1352–1359.

[39] X. Kong, S. Liu, M. Taher, and A. J. Davison, "Vmap: Vectorised object mapping for neural field SLAM," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 952–961.

[40] N. Patel, F. Khorrami, P. Krishnamurthy, and A. Tzes, "Tightly coupled semantic RGB-D inertial odometry for accurate long-term localization and mapping," in *Proc. Int. Conf. Adv. Robot.*, Belo Horizonte, Brazil, Dec. 2019, pp. 523–528.

[41] X. Han, H. Liu, Y. Ding, and L. Yang, "RO-MAP: Real-time multi-object mapping with neural radiance fields," *IEEE Robot. Automat. Lett.*, vol. 8, no. 9, pp. 5950–5957, Sep. 2023.

[42] J. McCormac, R. Clark, M. Bloesch, A. J. Davison, and S. Leutenegger, "Fusion : Volumetric object-level SLAM," in *Proc. Int. Conf. 3D Vis.*, Verona, Italy, Sep. 2018, pp. 32–41.

[43] M. Yan, J. Zhang, Y. Zhu, and H. Wang, "Maskclustering: View consensus based mask graph clustering for open-vocabulary 3D instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2024, pp. 28274–28284.

[44] S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "SceneGraphFusion: Incremental 3D scene graph prediction from RGB-D sequences," *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 7515–7525.

[45] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3D scene graphs for language-grounded robot navigation," in *Proc. Robotics: Sci. Syst.*, Delft, Netherlands, Jul. 2024.

[46] J. Wang and L. Ke, "LLM-Seg: Bridging image segmentation and large language model reasoning," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2024, pp. 1765–1774.

[47] H. Shi, S. D. Dao, and J. Cai, "Llmformer: Large language model for open-vocabulary semantic segmentation," *Int. J. Comput. Vis.*, vol. 133, pp. 742–759, Aug. 2025.

[48] X. Ma et al., "SQA3D: Situated question answering in 3D scenes," in *Proc. Int. Conf. Learn. Representations*, Kigali, Rwanda, May 2023.

[49] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "ScanQA: 3D question answering for spatial scene understanding," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 2022, pp. 19107–19117.

[50] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. J. Guibas, "ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 422–440.

[51] Y. Chen et al., "Grounded 3D-LLM with referent tokens," 2024, *arXiv:2405.10370*.

[52] Y. Hong et al., "3D-LLM: Injecting the 3D world into large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, New Orleans, LA, USA, Dec. 2023, pp. 20482–20494.

[53] S. Chen et al., "LL3DA: Visual interactive instruction tuning for omni-3D understanding reasoning and planning," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2024, pp. 26428–26438.

[54] J. Yang et al., "LLM-Grounder: Open-vocabulary 3D visual grounding with large language model as an agent," in *Proc. Int. Conf. Robot. Automat.*, Yokohama, Japan, May 2024, pp. 7694–7701.

[55] X. Chen et al., "Multi-object hallucination in vision-language models," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2024, pp. 44393–44418.

[56] J. Achiam et al., "GPT-4 Technical Report," 2023, *arXiv:2303.08774*.

[57] S. Koch, N. Vaskevicius, M. Colosi, P. Hermosilla, and T. Ropinski, "Open3DSG: Open-vocabulary 3D scene graphs from point clouds with queryable objects and open-set relationships," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2024, pp. 14183–14193.

[58] Q. Yu, J. He, X. Deng, X. Shen, and L. Chen, "Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP," in *Proc. Adv. Neural Inf. Process. Syst.*, New Orleans, LA, USA, Dec. 2023, pp. 32215–32234.

[59] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020 s," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 2022, pp. 11966–11976.

[60] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 2022, pp. 1280–1289.

[61] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*,, Aug. 2016, pp. 1715–1725.

[62] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, vol. 8693, Zurich, Switzerland, Sep. 2014, pp. 740–755.

[63] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1209–1218.

[64] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 9404–9413.

[65] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. logistics Quart.*, vol. 2, no. 1-2, pp. 83–97, 1955.

[66] B. Hua, Q. Pham, D. T. Nguyen, M. Tran, L. Yu, and S. Yeung, "Scenenn: A scene meshes dataset with annotations," in *Proc. Int. Conf. 3D Vis.*, Stanford, CA, USA, Oct. 2016, pp. 92–101.

[67] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 2432–2443.

[68] C. Yeshwanth, Y. Liu, M. Nießner, and A. Dai, "ScanNet++: A high-fidelity dataset of 3D indoor scenes," in *Proc. Int. Conf. Comput. Vis.*, Paris, France, Oct. 2023, pp. 12–22.

[69] D. Rozenberszki, O. Litany, and A. Dai, "Language-grounded indoor 3D semantic segmentation in the wild," in *Proc. Eur. Conf. Comput. Vis.*, G. J. Avidan, M. Brostow, G. M. Cissé, S. Farinella, and T. Hassner, Eds. Tel Aviv, Israel, Oct. 2022, pp. 125–141.

[70] J. Straub et al., "The replica dataset: A digital replica of indoor spaces," 2019, *arXiv:1906.05797*.

[71] R. Mascaro, L. Teixeira, and M. Chli, "Volumetric instance-level semantic mapping via multi-view 2D-to-3D label diffusion," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 3531–3538, Apr. 2022.

[72] M. Han et al., "Reconstructing interactive 3D scenes by panoptic mapping and CAD model alignments," in *Proc. Int. Conf. Robot. Automat.*, Xi'an, China, May 2021, pp. 12199–12206.

[73] L. Wang et al., "Multi-view fusion-based 3D object detection for robot indoor scene perception," *Sensors*, vol. 19, no. 19, 2019, Art. no. 4092.

[74] W. Li, J. Gu, B. Chen, and J. Han, "Incremental instance-oriented 3D semantic mapping via RGB-D cameras for unknown indoor scene," *Discrete Dyn. Nature Soc.*, vol. 2020, pp. 1–10, 2020.

[75] Y. Miao, I. Armeni, M. Pollefeys, and D. Barath, "Volumetric semantically consistent 3D panoptic mapping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* 2024, pp. 12924–12931.

[76] C. Campos, R. Elvira, J. J. G. Rodrıguez, J. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[77] 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.*, 2024, pp. 169–185.

Naman Patel (Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from NYU Tandon School of Engineering, Brooklyn, NY, USA, in 2016 and 2021, respectively.

His research interests include robot perception, machine learning, computer vision, and security. He has worked on a breadth of different problems related to perception-based autonomy for different types of robots with varying sensor modalities during his Ph.D. His current research focuses on finding shortcomings in perception algorithms for robotics systems, such as autonomous ground and aerial vehicles, and developing robust and secure algorithms to overcome them.

Prashanth Krishnamurthy (Member, IEEE) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Chennai, India, in 1999, and the M.S. and Ph.D. degrees in electrical engineering from New York University Tandon School of Engineering (formerly Polytechnic University), Brooklyn, NY, USA, in 2002 and 2006, respectively.

He is currently a Senior Research Scientist and Adjunct Faculty with the Department of Electrical and Computer Engineering, NYU Tandon School of Engineering. He has coauthored more than 175 journal and conference papers. He has also coauthored the book *Modeling and Adaptive Nonlinear Control of Electric Motors* (Springer Verlag, 2003). His research interests include autonomous vehicles and robotic systems, multiagent systems, nonlinear control, resilient control, sensor data fusion, machine learning, real-time embedded systems, cyber-physical systems and cyber-security, and decentralized and large-scale systems.

Farshad Khorrami (Fellow, IEEE) received dual bachelor's degrees in mathematics and electrical engineering and the master's degree in mathematics and the Ph.D. degree in electrical engineering from The Ohio State University, Columbus, OH, USA, in 1982, 1984, 1984, and 1988, respectively.

He is currently a Professor of Electrical and Computer Engineering Department, NYU, Brooklyn, NY, USA, where he joined as an Assistant Professor in 1988. He has developed and directed the Control/Robotics Research Laboratory, Polytechnic University (Now NYU), and the Codirector of the Center in AI and Robotics (CAIR), NYU Abu Dhabi. He has also commercialized UAVs as well as development of auto-pilots for various autonomous vehicles. He has authored or coauthored more than 400 refereed journal and conference papers in these areas and a book *Modeling and Adaptive Nonlinear Control of Electric Motors* (Springer, 2003). He also has 15 U.S. patents on novel smart micro-positioners, control systems, cyber security, and wireless sensors and actuators. His research interests include adaptive and nonlinear controls, robotics and automation, autonomous vehicles, cyber security for CPS, embedded systems security, machine learning, and large-scale systems and decentralized control.

Dr. Khorrami's research has been supported by the DOE, ARO, NSF, ONR, DARPA, ARL, AFRL, and several corporations. He was Conference Organizing Committee Member of several international conferences.