2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

Open-Fusion: Real-time Open-Vocabulary 3D Mapping and Queryable Scene Representation

Kashu Yamazaki[1], Taisei Hanyu[1], Khoa Vo[1], Thang Pham[1], Minh Tran[1],
Gianfranco Doretto[2], Anh Nguyen[3], Ngan Le[1]

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10610193

***Abstract*— Precise 3D environmental mapping with semantics is essential in robotics. Existing methods often rely on predefined concepts during training or are time-intensive when generating semantic maps. This paper presents Open-Fusion, an approach for real-time open-vocabulary 3D mapping and queryable scene representation using RGB-D data. Open-Fusion harnesses the power of a pretrained vision-language foundation model (VLFM) for open-set semantic comprehension and employs the Truncated Signed Distance Function (TSDF) for swift 3D scene reconstruction. By leveraging the VLFM, we extract region-based embeddings and their associated confidence maps. These are then integrated with the 3D knowledge from TSDF using an enhanced Hungarian-based feature-matching mechanism. In particular, Open-Fusion delivers outstanding annotation-free 3D segmentation for open vocabulary query without the need for additional 3D training. Benchmark tests on the ScanNet dataset against leading zero-shot methods highlight Open-Fusion's superiority. Furthermore, it seamlessly combines the strengths of region-based VLFM and TSDF, facilitating real-time 3D scene comprehension that includes object concepts and open-world semantics. We encourage the readers to view the demos on our project page: `https://uark-aicv.github.io/OpenFusion`**

I. INTRODUCTION

Real-time 3D scene understanding, crucial in computer vision, involves discerning object semantics, locations, and geometric attributes from RGB-D data in unstructured environments [1]. Despite its diverse applications in virtual reality, robotics, and augmented reality, traditional training methods face significant challenges [2]. These include the need for extensive human annotations, limited closed-set semantic information, and the demand for real-time performance in applications like robotics and augmented reality.

In recent years, the convergence of language and robotics has garnered significant attention, driven by the promise it holds in enabling robots to interpret and act upon straightforward natural language commands. This benefit from the emergence of large-scale vision-language foundation models (VLFMs) such as CLIP [3], ALIGN [4], BLIP [5], GLIP [6], RegionCLIP [7], etc. Those models are learned in unsupervised manner using massive image-text pairs from the internet and have showcased remarkable capabilities in zero-shot learning and open-vocab reasoning. However, integrating VLFMs into robotics requires addressing scalability and real-time processing concerns. Scalability is essential to avoid exponential data growth in large environments, while real-time capability is vital for instant decision-making. Achieving these goals necessitates efficient data extraction and integration without undue delays.

Despite the impressive qualities exhibited by these VLFMs, there remains a significant untapped potential for their integration into robotic applications, particularly in the context of 3D mapping and understanding. The primary bottleneck in leveraging VLFMs for robotics stems from the fact that most foundation models consume images and produce only a single vector encoding of the entire image within an embedding space. This approach falls short of meeting the stringent demands of robotic perception systems, which require precise reasoning at point-level or object-level granularity across a diverse spectrum of concepts. This is crucial for tasks involving interaction with the external 3D environment, such as navigation and manipulation. Moreover, it is essential to acknowledge that applying VLFMs at the point-level can be computationally intensive and time-consuming, rendering it unsuitable for meeting the real-time demands of real-world applications. Therefore, to fully harness the potential of VLFMs in robotics, there is a pressing need to develop more efficient and effective techniques that enable these models to operate in real-time while delivering the required level of precision for tasks in complex 3D environments.

In response to the aforementioned challenges, we present Open-Fusion, a queryable semantic representation rooted in VLFMs. Open-Fusion facilitates real-time 3D scene reconstruction, incorporating semantics, through the use of the Truncated Signed Distance Function (TSDF). Our work demonstrates that Open-Fusion excels in the efficient zero-shot reconstruction and understanding of 3D scenes, offering queryable scene representations for enhanced understanding and interaction. To summarize, we make the following contributions: **1) Real-time 3D Scene Reconstruction**: We extend TSDF to achieve effective real-time 3D scene reconstruction. **2) Semantic-aware Region-based Feature Matching**: We extend Hungarian matching to seamlessly match features from the VLFM into the 3D scene representation, enabling incremental semantic reconstruction. **3) Embedding Dictionary for Efficiency**: To reduce memory consumption during scene reconstruction and facilitate open-vocab scene queries, we implement an embedding dictionary. **4) Open-Fusion**: As a result, we propose Open-Fusion, a real-time 3D map reconstruction and scene representation with open-vocab query capabilities. This framework promises to advance the field of real-time 3D scene understanding for robotics.

[1]AICV Lab, Department of EECS, University of Arkansas, USA. kyamazak@uark.edu
2 Department of CSCE, West Virginia University, USA.
3 Department of CS, University of Liverpool, UK.

TABLE I

HIGH-LEVEL COMPARISON BETWEEN OUR OPEN-FUSION AND EXISTING SOTA QUERYABLE SCENE REPRESENTATIONS. P DENOTES THE NUMBER OF POINTS IN A MAP, M IS THE NUMBER OF OBJECTS IN THE SCENE.

Map	Method	Representation	Foundation Model	Feature Level	Real-time [1]	Scene-specific	Sem-Query [2]
2D	CoW [8]	point	CLIP [3] + GradCAM [9]	point	-	✗	O(P)
	NLMap [10]	point	ViLD [11] + CLIP [3]	bbox	-	✗	O(P)
	VLMap [12]	point	LSeg [13]	point	✗	✗	O(P)
3D	CLIP-Fields [14]	NeRF	Detic [15] + CLIP [3]	bbox	✗	✓	-
	LERF [16]	NeRF	CLIP [3]	image patch	✗	✓	-
	SemAbs[17]	occupancy	CLIP [3] + GradCAM [9]	point	✗	✗	O(P)
	ConceptFusion [18]	point	SAM [19] + CLIP [3]	bbox	✗	✗	O(P)
	Open-Fusion	TSDF	SEEM [20]	region	✓	✗	O(M)

II. RELATED WORKS

A. Vision-Language Foundation Models (VLFMs).

VLFMs have brought about a revolution in the field of perception by enabling open-set inference using natural language. These models, renowned for their robust generalization capabilities, owe their success to the extensive datasets and model parameters that drive them. VLFMs can be broadly categorized into three groups based on the level of resolution in vision-language alignment as follows: (i) Image-Level Aligned Models (ILAMs), (ii) Pixel-Level Aligned Models (PLAMs), and (iii) Region-Level Aligned Models (RLAMs). Specifically, ILAMs (UniCL [21], CLIP [3], ALIGN [4], BLIP [5], BLIPv2 [22], etc.) generate a single vector representation for the entire image that can correlate with text embeddings. PLAMs (LSeg[13], MaskCLIP [23], etc.), on the other hand, produce vector representation for each pixel of an image. Similarlly, RLAMs (GLIP [6], GLIPv2 [24], RegionCLIP [7], ODISE [25], SEEM [20], HIPIE [26], SemanticSAM [27], etc.) offer the representation for each region within an image.

Image-level representations offer the advantage of computational efficiency but are limited by their provision of coarser semantic insights. This limitation becomes pronounced in contexts demanding finer semantic information, necessitating the integration of auxiliary modules such as Grad-CAM [9] to imbue spatial knowleadge. However, this integration incurs a substantial increase in computational overhead, rendering such approaches impractical for applications where real-time processing is desired. In contrast, pixel-level and region-level Aligned Models are inherently endowed with spatial awareness, positioning them as more apt solutions for tasks requiring granular semantic insights. Recognizing the necessity for both open-set semantics and computational expediency, we have opted to leverage SEEM, a region-level VLFM with masks. SEEM strikes a balance between the demand for nuanced semantic understanding and the imperative of time-efficient processing, all while maintaining scalability.

B. Queryable scene representation.

To offer a detailed survey of the current landscape in semantic mapping methodologies, we examine both two-dimensional (2D) and three-dimensional (3D) approaches.

[1]Real-time: the real-time requirement for 3D scene reconstruction.

[2]Sem-Query: the time for open-vocab semantic query.

2D Mapping: CoW [8] and NLMap [10] are notable examples, harnessing the open-set features derived from CLIP to construct 2D map for exploration. CoW employs Grad-CAM [9] to extract spatial knowledge from CLIP whereas NLMap integrates ViLD [11] to crop objects before applying CLIP. VLMaps [12] stands out by utilizing pixel-aligned features from LSeg [13] to enable the creation of bird's-eye view 2D maps, specifically designed for efficient landmark querying.

NeRF-based 3D Mapping: CLIP-Fields [14] trains a NeRF-inspired implicit representation network that maps spatial coordinates (x, y, z) to vectors enriched with semantic information through MLPs. Remarkably, this approach is scene-specific, with direct supervision from semantic vectors obtained from CLIP or other models like Sentence BERT [28]. LERF [16], while also drawing inspiration from CLIP, focuses primarily on object localization. It trains a neural field through knowledge distillation from multi-scale CLIP features and DINO. However, it is worth noting that LERF may struggle with capturing precise object boundaries due to its primary emphasis on object localization.

Non-NeRF 3D Mapping: SemAbs [17] proposed to incorporate semantics from CLIP with GradCAM and the 3D completion module to produce semantic-aware occupancy. While it showcases promising results in 3D scene understanding, it cannot run in real-time. ConceptFusion [18] introduces a unique paradigm by employing off-the-shelf foundation models to construct 3D maps with open-set features. While this approach exhibits great potential for open-vocab 3D scene understanding, their integration of semantics to every points in space make the method resource intensive.

While NeRF-based methods excel in achieving photorealistic scene reconstruction, they require retraining for each new scene and are constrained in the volume they can render. As a result, they tend to be customized for specific scenes, which limits their applicability to real-world scenarios. Conversely, non-NeRF-based methods have the potential to capture more generalizable representations as they might not require retraining for each new scene. However, previous works focus on the offline generation of the queryable map mainly due to the time-consuming computational requirements posed by their point-based approach. This drawback makes them less suitable for real-time robotics applications. To address this challenge, we introduce Open-Fusion, which is a non-NeRF methods for real-time processing, resulting in an open-vocab 3D scene representation suitable for robotics.

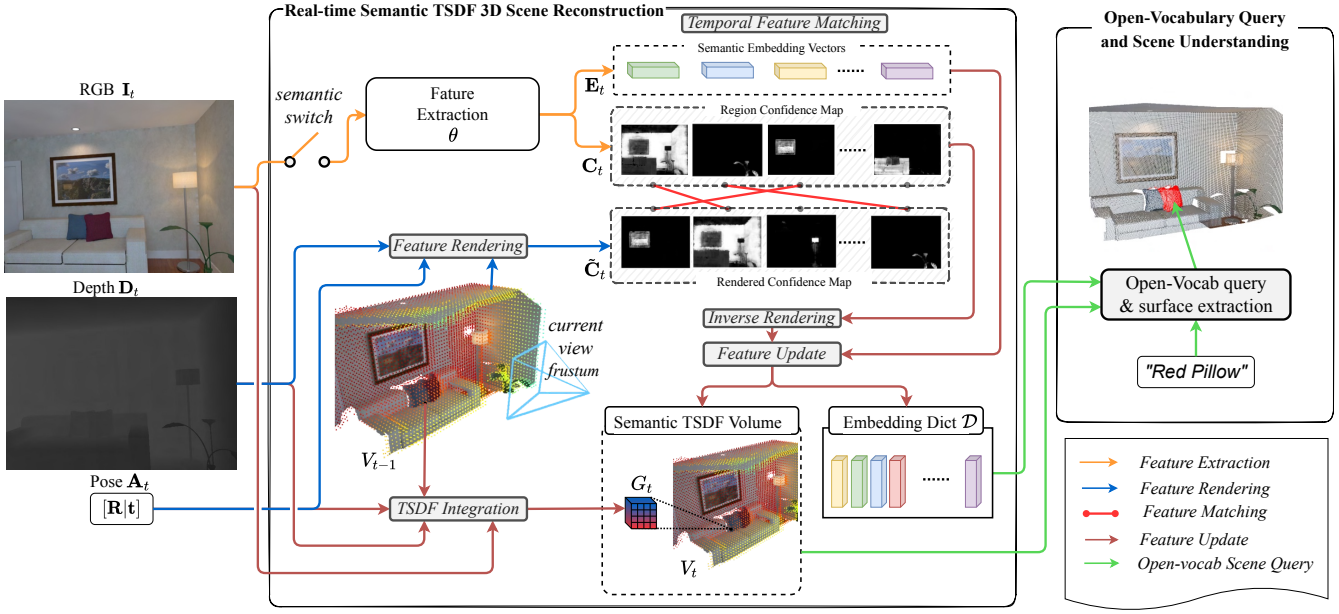

Fig. 1. The overall pipeline of Open-Fusion, which contains two modules. *Real-time Semantic TSDF 3D Scene Reconstruction Module*: This module takes in a stream of RGB-D images ($\mathbf{I}_t$, $\mathbf{D}_t$) and the corresponding camera pose ($\mathbf{A}_t$). It incrementally reconstructs the 3D scene, representing it as a semantic TSDF volume V_t at time t. *Open-Vocabulary Query and Scene Understanding Module*: In the second module, Open-Fusion accepts open-vocab queries as inputs and provides corresponding scene segmentations in response, which can serve as an eye for language base robot commanding.

III. METHODOLOGY

A. Problem Setup

Consider a sequence of T RGB-D observations obtained from an environment, which can be represented as $\{(\mathbf{I}_t, \mathbf{D}_t, \mathbf{A}_t)\}_{t=0}^{T}$. Here, $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$ represents an RGB frame, $\mathbf{D}_t \in \mathbb{R}^{H\times W}$ indicates a depth frame, and $\mathbf{A}_t = [\mathbf{R}_t|\mathbf{t}_t] \in \mathbb{R}^{3\times 4}$ denotes the associated camera pose with rotation $\mathbf{R} \in \mathbb{R}^{3\times 3}$ and translation $\mathbf{t} \in \mathbb{R}^{3\times 1}$. Additionally, we have the camera's intrinsic parameters represented as $\mathbf{K} \in \mathbb{R}^{3\times 3}$. Our primary objective is to construct a language-queryable 3D map denoted as $\mathcal{M}$ in real-time. In this context, we define a queryable map as a 2D/3D representation of the environment that incorporates both physical and semantic features. These features can be extracted using a query vector $\mathbf{q} \in \mathbb{R}^d$. Notably, various entities such as images, text, coordinates, etc., can be transformed into the query vector by encoding them into a shared embedding space using an encoding function.

Our proposed Open-Fusion, as depicted in Fig. 1, comprises two main modules: 1) *Real-time Semantic TSDF 3D Scene Reconstruction*: this module consists of two sub modules i) Feature Extraction: this module aims to extract region-based feature including confidence map and embedding map ii) Real-time Semantic 3D Scene Reconstruction: this module facilitates the integration of an incoming frame at time t into the current semantic STDF volume V_{t-1} while updating the embedding dictionary ($\mathcal{D}_t$). Consequently, it generates a 3D scene representation V_T and an updated embedding dictionary ($\mathcal{D}_T$) after the integration of T frames. The second module consists of three components of Feature Rendering by TSDF, Region-based Semantic Feature Matching, and Feature Update. 2) *Open-Vocab Query and Scene Understanding*: this module is designed to localize and segment objects in the scene based on user queries and open-vocab semantics.

B. Region-based Feature Extractor

Given the RGB frame of the current view $\mathbf{I}_t$ at time t, employ the SEEM model [20], denoted as θ, for encoding. Unlike the widely adopted CLIP model, SEEM produces *region-level* aligned feature. This aims to eliminate the need for the class agnostic mask proposal generator in two-stage setup [18] or attention-explainability model to localize the relevant regions like [8], [17]. Considering the real-time constraints, avoiding the inclusion of such expensive models in a sequence of function calls is of utmost importance.

For each $\mathbf{I}_t$, the model θ generates region confidence maps $\mathbf{C}_t \in \mathbb{R}^{|Q|\times H/4\times W/4}$ at a quarter of the input resolution. Additionally, it produces corresponding semantic embedding vectors, denoted as $\mathbf{E}_t \in \mathbb{R}^{|Q|\times d}$, tailored for the predefined number of object queries $|Q|$, where d is feature dimension. The feature extraction at time t can be formulated as $\mathbf{C}_t, \mathbf{E}_t = \theta(\mathbf{I}_t)$. In practice, the region-based feature extraction process is specifically for semantic-related tasks and may pose a bottleneck due to SEEM's time consumption at 4.5 FPS. If a task doesn't require semantics, this process can be skipped. Additionally, given the substantial overlap between two consecutive frames, it's feasible to omit some frames. To enhance the flexibility and efficiency of our OpenFusion, we have implemented a semantic switch, as depicted in Fig. 1.

C. Real-time 3D Scene Reconstruction with Semantics

Every time-frame, we incorporate the incoming observation $(\mathbf{I}_t, \mathbf{D}_t, \mathbf{A}_t)$ into an implicit surface using the Truncated

Signed Distance Function (TSDF). Specifically, we integrate $(\mathbf{I}_t, \mathbf{D}_t, \mathbf{A}_t)$ into the TSDF volume V_{t-1} to create the TSDF volume at time t, denoted as V_t. It is important to emphasize that the TSDF volume V_t comprises a set of M volumetric blocks, represented as $V_t = \{G_i\}_{i=1}^{M}$. The TSDF is an extension of the Signed Distance Function (SDF) ϕ, which is a function that provides the shortest distance to any surface for every 3D point. The sign indicates whether the point is located in front of or behind the surface. In the context of scene reconstruction, the points of interest typically reside on the boundary $\delta\Omega$. For a distance function d and a point $\mathbf{p} \in \mathbb{R}^3$, the SDF $\phi : \mathbb{R}^3 \to \mathbb{R}$ defines the signed distance to the surface as follows:

$$\phi(\mathbf{p}) = \begin{cases} -d(\mathbf{p}, \delta\Omega) & \text{if } \mathbf{p} \in \Omega \\ d(\mathbf{p}, \delta\Omega) & \text{if } \mathbf{p} \in \Omega^c \end{cases} \tag{1}$$

This means that points located inside the surface have negative values, while the surface itself lies precisely at the zero crossing point between positive and negative values. The TSDF truncates all values above with a specified threshold τ, with τ chosen as four times the voxel size.

As the reconstruction of the 3D scene essentially represents a local 2D surface—a 2D manifold embedded in 3D space—we can efficiently embed the 3D scene using globally sparse but locally dense voxel blocks. These voxel blocks exhibit a distinctive characteristic where they are globally present only near the surface of interest (while other parts remain void). Within each block, we maintain a dense voxel grid typically sized at $r \times r \times r$. Following the approach in [29], we construct *semantic TSDF volume* as a set of globally sparse volumetric blocks $V_t = \{G_i\}_{i=1}^{M}$, each containing a locally dense voxel grid $G_i = \{p_j\}_{j=1}^{r^3}$ and the information in $p_j = \{(RGB_j, w_j, \hat{\phi}_j, k_j, c_j)\}$ includes color RGB, weight w for TSDF updates, TSDF value $\hat{\phi}$, embedding key k, and confidence score c.

Notably, unlike previous approaches like [18] that store the semantic embedding for each point, we opt to store only the keys for embedding and the associated confidence scores for each pixel. The actual embedding information is maintained separately within the dedicated embedding dictionary $\mathcal{D} : k \to \mathbf{E}$. Given our utilization of region-based embedding for the scene, it's important to highlight that the number of embeddings required for the entire scene is significantly reduced compared to point-based counterparts. In addition to the surface and color data, we also incorporate semantics into the TSDF volume. However, to optimize computation and memory usage in subsequent modules, we limit the storage of semantics to points near the surface. These points are strategically sampled based on the TSDF values, resulting in a more efficient representation.

As a result, to integrate $(\mathbf{I}_t, \mathbf{D}_t, \mathbf{A}_t)$ at time t into semantic TSDF volume V_{t-1} at time $t-1$, consisting of M volumetric blocks $\{G_i\}_{i=1}^{M}$, we perform the following steps:

1) *Feature Rendering*: This initial step involves generating a rendered confidence map $\tilde{\mathbf{C}}_t$ and retrieving corresponding embedding $\tilde{\mathbf{E}}_t$ from the existing TSDF volume V_{t-1}.
2) *Region-based Matching*: In this step, we establish the correspondence between the confidence map $\mathbf{C}_t$ and the rendered confidence map $\tilde{\mathbf{C}}_t$ for the update.
3) *Feature Update*: This step focuses on updating the TSDF volume V_{t-1} at time t and concurrently updating the embedding dictionary $\mathcal{D}_t$ based on the matching.

Feature Rendering by TSDF: We render confidence map $\tilde{\mathbf{C}}_t$ with its corresponding embeding map $\tilde{\mathbf{E}}_t$ from the TSDF volume with the current camera pose $\mathbf{R}_t|\mathbf{t}_t$ and depth image $\mathbf{D}_t$ at time t. Given the semantic TSDF volume V_{t-1} accumulated from time 0 to $t-1$ and the current observation $(\mathbf{I}_t, \mathbf{D}_t, \mathbf{A}_t)$, our integration process involves several key steps: (i) Conversion of depth image $\mathbf{D}_t$: Initially, we convert the depth value $\mathbf{D}_t^{i,j}$ obtained from the 2D depth image at the location of pixel coordinates i, j within the image, into a 3D coordinates (x, y, z) using Eq.2.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{R}_t^{-1} \left(\mathbf{D}_t^{i,j} \mathbf{K}^{-1} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix} \right) - \mathbf{t}_t \ , \tag{2}$$

where $\mathbf{R}_t$ and $\mathbf{t}_t$ are the rotation and translation component of camera pose $\mathbf{A}_t$, and $\mathbf{K}$ represents intrinsic parameters. (ii) Identifying relevant blocks: Next, we identify the set of volumetric blocks $\mathcal{G}_{active}$ that contain points unprojected from the current depth image. We determine these active blocks within the current viewing frustum by examining whether the 3D coordinates (x, y, z) fall within the boundaries of these blocks or not. (iii) Projection of semantic information: Subsequently, we project the voxels G_j within the active blocks that possess semantic keys and confidence scores onto the image plane, as defined by Eq.3.

$$\begin{bmatrix} u \\ v \\ \hat{d} \end{bmatrix} = \hat{\mathbf{K}} \left(\mathbf{R}_t \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \mathbf{t}_t \right) \ , \tag{3}$$

where $\hat{\mathbf{K}}$ represents the rescaled intrinsic parameters obtained by scaling $\mathbf{K}$ to the θ's output reslution and the coordinate of the valid voxel (x, y, z) are mapped to the pixel location $(u/\hat{d}, v/\hat{d})$ subjected to $\Big(\hat{d} > 0\Big) \wedge \Big(0 \leq u/\hat{d} < W/4\Big) \wedge (0 \leq v/d' < H/4)$.

This projection is a crucial step in incorporating semantic information into the current frame's representation. Building upon the rendering operation described above, we generate confidence maps $\tilde{\mathbf{C}}_t \in \mathbb{R}^{m \times H/4 \times W/4}$ within the current field of view (FoV).

Region-based Temporal Feature Matching: This step aims to find fusion candidates by matching pairs between the confidence map $\mathbf{C}_t$ and the rendered confidence map $\tilde{\mathbf{C}}_t$, which casts the knowledge of objects accumulated until $t-1$ in the semantic TSDF volume from the current FoV. We formulate this feature matching as a 2D rectangular assignment problem, with the goal of identifying the assignment $\mathcal{S}^*$ that maximizes the soft-IoU [30] between $\mathbf{C}_t$ and $\tilde{\mathbf{C}}_t$.

$$\mathcal{S}^* = \arg\max_{\mathcal{S}} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathcal{L}_{match} \langle \mathbf{C}_t, \tilde{\mathbf{C}}_t \rangle_{i,j} \sigma_{i,j} \ , \tag{4}$$

Here, n represents the number of semantic regions in the current frame, and m is the number of rendered regions within the current FoV. $\mathcal{L}_{match}$ calculates the soft-IoU of $\mathbf{C}_t$ and $\tilde{\mathbf{C}}_t$. The matrix $\mathcal{S}$ represents a set of all $\sigma_{i,j}$ values, subject to the constraints $\sum_{j=1}^{m} \sigma_{i,j} = 1; \forall i$, $\sum_{i=1}^{n} \sigma_{i,j} \leq 1; \forall j$, and $\sigma_{i,j} \in \{0, 1\}$. If $\sigma_{i,j} = 1$, it signifies that the prediction in row i is assigned to the rendered embedding in column j. To solve this problem, we employ a modified Jonker-Volgenant algorithm [31] (extension version of Hungarian). We discard the match if the soft-IoU score is below 0.10. This operation helps us to avoid fusing poor quality masks of the same object due to occlusion or blur.

Feature Update and Inverse Rendering: In this step, information, i.e., $(\mathbf{I}_t, \mathbf{D}_t, \mathbf{C}_t, \mathbf{E}_t)$, we obtain from the current time frame is integrated into the semantic TSDF volume V_{t-1} to create V_t. First, each voxel p_j within the active volumetric blocks $\mathcal{G}_{active}$ undergoes the standard TSDF integration process [29], where the stored color RGB_j and TSDF values $\hat{\phi}_j$ are updated using weighted average. Using Eq. 3 with the actual camera intrinsic $\mathbf{K}$, we can obtain (u, v, d) and the update is summarized as:

$$RGB_j \leftarrow \frac{w_j \cdot RGB_j + \mathbf{I}_t^{u/d, v/d}}{w_j + 1} \tag{5}$$

$$\hat{\phi}_j \leftarrow \frac{w_j \cdot \hat{\phi}_j + \Psi(\mathbf{D}_t^{u/d, v/d} - d, \tau)}{w_j + 1} \tag{6}$$

$$w_j \leftarrow w_j + 1 \tag{7}$$

where $\Psi(\cdot)$ is the truncation operation that is applied to SDF to obtain TSDF (Section III-C).

The dictionary $\mathcal{D}$ and the confidence score c_j and the associated key k_j will be updated according to the matching $\mathcal{S}^*$. If the new region is matched to the existing one, only the confidence map is updated with weighted average while unmatched candidates also update the dictionary as a new region. The confidence maps $\mathbf{C}_t$ are inversely rendered by applying Eq.3.

D. Querying Semantics from the 3D Map

At any time t, we can extract the corresponding point cloud or mesh from the semantic TSDF volume V_t by querying it with a vector $\mathbf{q}$. Our querying method involves a similarity calculation between the query and the semantic embeddings stored in the dictionary $\mathcal{D}_t$. This approach is significantly faster and more memory-efficient than previous methods that store embeddings for individual points. Specifically, we calculate the cosine similarity $\cos\langle \mathbf{E}, \mathbf{q} \rangle$ between the semantic embeddings $\mathbf{E} \in \mathbb{R}^{R \times d}$ in the dictionary $\mathcal{D}_t$ and the query vector $\mathbf{q} \in \mathbb{R}^d$, which is obtained using a modality-specific encoder trained in a shared embedding space with the semantic vectors $\mathbf{E}$, and select the most relevant region as the object proposal. After the query, Marching Cubes is applied to extract surfaces or point clouds from the semantic TSDF volumes to indicate the queried region. For a resource constraint environment, one can simply use the semantic TSDF voxel coordinates as the approximation of the region.

TABLE II

QUANTITATIVE COMPARISON OF **OPEN-SET SEMANTIC SEGMENTATION** AND **3D SCENE REPRESENTATION TIME** BETWEEN OPEN-FUSION AND EXISTING METHODS ON THE SCANNET DATASET.

	Method	Time (FPS)↑		Accuracy↑	
		3D-Rec.[1]	Sem-3D-Rec[2]	mAcc	f-mIoU
Priv.[3]	LSeg	-	-	0.70	0.63
	OpenSeg	-	-	0.63	0.62
	CLIPSeg (rd64-uni)	-	-	0.41	0.34
	CLIPSeg (rd16-uni)	-	-	0.41	0.36
ZS.[4]	MaskCLIP	-	-	0.24	0.28
	ConceptFusion	1.5	0.15	**0.63**	0.58
	Open-Fusion	**50**	**4.5**	0.62	**0.59**

IV. EXPERIMENTS

In this section, we conduct a comprehensive evaluation of Open-Fusion's performance through both quantitative and qualitative assessments on the ScanNet [32] and Replica [33] datasets, specifically focusing on open-set semantic segmentation tasks. In this work, we will focus our quantitative results and comparisons exclusively on the ScanNet dataset and we will provide qualitative results and comparisons for the Replica dataset. Furthermore, we showcase the real-world applicability of Open-Fusion by seamlessly integrating it into the Kobuki platform, enabling real-time 3D scene representation.

A. Quantitative Benchmarks

Our quantitative experimental benchmarks are conducted on the ScanNet dataset, a comprehensive RGB-D video dataset with annotations for 3D camera poses, surface reconstructions, and instance-level semantic segmentations. Following the methodology of ConceptFusion [18], we select room-scale indoor scenes for evaluation of our research. For each selected scene, we utilize the semantic categories provided in the scene annotations as text-prompted queries for performing open-set segmentation tasks.

Consistent with the evaluation methodology introduced in ConceptFusion [18], we assess both performance and time efficiency of Open-Fusion in the context of open-set semantic 3D scene understanding. Our evaluation encompasses a dual focus: performance and time efficiency. To assess accuracy, we employ the mean accuracy (mAcc) and frequency mean Intersection over Union (f-mIoU) metrics. In addition, we measure time consumption for 3D scene representation in frames per second (FPS). The measurements were done on single RTX 3900. Table II offers a comprehensive comparative analysis between Open-Fusion against existing SOTA methods in terms of mAcc, f-mIoU, and FPS. Thanks to the utilization of region-based embedding and TSDF, Open-Fusion achieves nearly real-time performance at 4.5 FPS, which is 30 times faster than the runner-up ConceptFusion.

[1] 3D-Rec.: 3D scene reconstruction only.
[2] Sem-3D-Rec: 3D scene reconstruction with semantics.
[3] Priv.: finetuned VLFMs specifically for semantic segmentation.
[4] ZS: zero-shot approaches.

Fig. 2. Qualitative comparison of 3D object query results on Replica dataset. While ConceptFusion failed to pinpoint the object location, Open-Fusion can estimate more precise location from language queries.

While excelling in time efficiency, Open-Fusion maintains competitive performance levels with the existing SOTA ConceptFusion in terms of both mAcc and f-mIoU metrics. This experiment underscores the efficiency and effectiveness of Open-Fusion in the realm of open-set semantic 3D scene understanding. Open-Fusion represents a significant advancement, establishing itself as the new SOTA in terms of both performance and efficiency.

B. Qualitative Results

We conducted a qualitative evaluation of Open-Fusion on the Replica dataset [33], as illustrated in Fig. 2. In this experiment, we demonstrated the semantic segmentation performance with queries involving various object sizes, from small objects like vases and lamps to larger ones like sofas and cabinets. Our Open-Fusion not only achieves significantly faster processing times (30x faster) but also delivers more accurate queryable semantic segmentation results.

C. Real-World Experiment

In this section, we present a real-world demonstration of real-time queryable scene reconstruction. Our experiment was conducted using the Kobuki platform, equipped with an RGB-D camera setup. Specifically, we utilized the *Azure Kinect Camera* to capture RGB-D images at a downsampled resolution of 360×630, along with the *Intel T265 Camera* for capturing corresponding camera poses. To ensure accurate alignment between camera poses and image streams, we synchronized them based on timestamps and filtered out any images without a matching pose recorded within a 10 ms

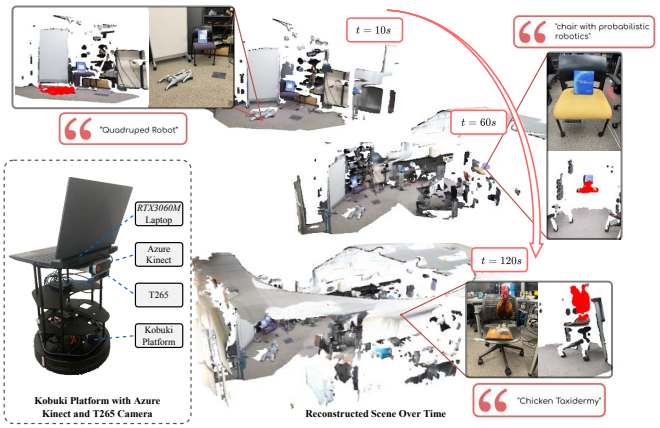

Fig. 3. The Kobuki platform is equipped with an Azure Kinect Camera and an Intel T265 Camera to demonstrate real-time mapping in a real-world environment. This system enables interaction with the world through natural language queries. The system is able to highlight the novel objects like the "quadruped robot" or "chicken taxidermy".

timeframe. Fig. 3 provides a visual representation of our real-world experimental setup using the Kobuki platform.

As it is difficult to obtain the ground truth semantic mask for real environment, we visually compare the suggested region by the model with the known environment setup. Fig. 3 displays the 3D map reconstruction generated at 50 FPS and semantics updated at 4.5 FPS running on two threads by the Kobuki platform. In this demonstration, we emphasize long-tailed reasoning like "quadruped robot" or "chicken taxidermy".

V. CONCLUSION & DISCUSSION

In this paper, we have introduced Open-Fusion, an efficient approach for real-time open-vocabulary 3D mapping and queryable scene representation from RGB-D data. Open-Fusion leverages the VLFM to extract region-based embeddings and employs TSDF, along with an extended version of Hungarian matching, for 3D semantic representation. We conducted both qualitative and quantitative benchmarks to assess our performance. In a qualitative evaluation, we compared Open-Fusion with ConceptFusion using the Replica dataset, demonstrating superior object segmentation results and real-time efficiency. In a quantitative assessment, we compared Open-Fusion with SOTA methods using the ScanNet dataset, achieving competitive results in terms of mean accuracy (mAcc) and surface mean Intersection over Union (f-mIoU) while Open-Fusion is 30x faster than ConceptFusion. Additionally, we conducted a real-world experiment with the Kobuki platform, highlighting Open-Fusion's capability in practical applications.

It is worth noting that our dependency on SEEM could limit the audio queries or multiple language (e.g., Spanish and French) queries presented in ConceptFusion.

ACKNOWLEDGMENT

This material is based upon work supported by the NSF 2223793 EFRI BRAID, NSF OIA-1946391 RII Track-1, NSF 2236302, and UoA ERISF.

REFERENCES

[1] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

[2] M. Naseer, S. Khan, and F. Porikli, "Indoor scene understanding in 2.5/3d for autonomous agents: A survey," *IEEE access*, vol. 7, pp. 1859–1887, 2018.

[3] 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 *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[4] C. Jia, Y. Yang, Y. Xia, *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICLR*, pp. 4904–4916, PMLR, 2021.

[5] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, pp. 12888–12900, PMLR, 2022.

[6] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

[7] Y. Zhong, J. Yang, *et al.*, "Regionclip: Region-based language-image pretraining," in *CVPR*, pp. 16793–16803, 2022.

[8] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23171–23181, 2023.

[9] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization. arxiv 2016," *arXiv preprint arXiv:1610.02391*, 2022.

[10] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler, "Open-vocabulary queryable scene representations for real world planning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11509–11522, IEEE, 2023.

[11] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[12] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10608–10615, IEEE, 2023.

[13] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven Semantic Segmentation," *arXiv e-prints*, p. arXiv:2201.03546, Jan. 2022.

[14] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam, "Clip-fields: Weakly supervised semantic fields for robotic memory," *arXiv preprint arXiv:2210.05663*, 2022.

[15] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *European Conference on Computer Vision*, pp. 350–368, Springer, 2022.

[16] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," *arXiv preprint arXiv:2303.09553*, 2023.

[17] H. Ha and S. Song, "Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models," in *6th Annual Conference on Robot Learning*, 2022.

[18] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, *et al.*, "Conceptfusion: Open-set multimodal 3d mapping," *arXiv preprint arXiv:2302.07241*, 2023.

[19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[20] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," *arXiv preprint arXiv:2304.06718*, 2023.

[21] J. Yang, C. Li, *et al.*, "Unified contrastive learning in image-text-label space," in *CVPR*, pp. 19163–19173, 2022.

[22] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[23] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *European Conference on Computer Vision*, pp. 696–712, Springer, 2022.

[24] H. Zhang, P. Zhang, *et al.*, "Glipv2: Unifying localization and vision-language understanding," *NIPS*, 2022.

[25] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2955–2966, 2023.

[26] X. Wang, S. Li, K. Kallidromitis, Y. Kato, K. Kozuka, and T. Darrell, "Hierarchical open-vocabulary universal image segmentation," *arXiv preprint arXiv:2307.00764*, 2023.

[27] F. Li, H. Zhang, P. Sun, X. Zou, S. Liu, J. Yang, C. Li, L. Zhang, and J. Gao, "Semantic-sam: Segment and recognize anything at any granularity," *arXiv preprint arXiv:2307.04767*, 2023.

[28] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[29] W. Dong, J. Park, Y. Yang, and M. Kaess, "Gpu accelerated robust scene reconstruction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7863–7870, IEEE, 2019.

[30] Y. Huang, Z. Tang, D. Chen, K. Su, and C. Chen, "Batching soft iou for training semantic segmentation networks," *IEEE Signal Processing Letters*, vol. 27, pp. 66–70, 2019.

[31] D. F. Crouse, "On implementing 2d rectangular assignment algorithms," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 4, pp. 1679–1696, 2016.

[32] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

[33] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.