2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MTE-SLAM: Multi-Tier Feature Fusion for Efficient Neural Semantic SLAM

Danqi Lu, Changxin Huang, Zhuangzhuang Chen, Zhiliang Lin, Dachong Li, Yanbin Chang, Jianqiang Li*

***Abstract*— Neural implicit representations have demonstrated excellent performance in Simultaneous Localization and Mapping (SLAM) by virtue of their ability to jointly model geometry, color and camera poses. Recent studies have attempted to integrate scene semantic information into implicit representation frameworks, significantly improving the ability of environmental understanding. Nevertheless, most existing methods rely on direct semantic coloring or rough fusing other modalities, resulting in underutilized semantic clues. This further causes problems such as blurred small objects, loss of fine structures and unclear regional boundaries. Additionally, redundant features introduced in the process reduce system efficiency. To address these challenges, we propose MTE-SLAM, an accurate and efficient end-to-end neural RGB-D semantic SLAM framework that synergizes Multi-Tier Feature Fusion (MTFF) and Feature Redundancy Suppressor (FRS). MTFF progressively fuses semantic features at global and local scales. The global context enhancement module captures scene-level semantic correlations, while the local continuity enhancement module refines neighborhood consistency, generating detailed and coherent semantic maps. FRS adaptively filters redundant features based on their importance and temporal variation, reducing parameters and computation while preserving representational power to accelerate training and inference. Comprehensive evaluations on Replica and ScanNet demonstrate that MTE-SLAM achieves centimeter-level reconstruction, state-of-the-art tracking and semantic accuracy, and runs up to four times faster than existing semantic SLAM systems.**

I. INTRODUCTION

Simultaneous Localization and Mapping (SLAM) enables autonomous agents to localize themselves while constructing maps of unknown environments, and is widely used in robotics [15], [20], augmented reality and autonomous driving [12], [26]–[28]. Traditional SLAM methods, however, rely heavily on hand-crafted features and are sensitive to illumination changes, dynamic objects and complex scene geometries [2], [24], [32]. Neural Radiance Fields (NeRF) [16] represent 3D scenes using neural implicit functions, enabling photorealistic view synthesis from sparse inputs.Their continuous spatial encoding supports high-fidelity surface reconstruction, making them well-suited for detailed, dense mapping [14]. Integrating NeRF into SLAM has led to neural implicit SLAM, where geometry and appearance are jointly optimized via neural networks. Recent research increasingly incorporates semantic understanding into implicit representations to improve environmental comprehension and localization robustness [5]. Studies show that combining pre-trained semantic predictions with implicit function-based mapping, which is augmented by multi-view fusion, can jointly optimize geometry, appearance and semantics in 3D, resulting in dense structural reconstruction, compact representation and semantic consistency [4]. Semantic features are high-dimensional, strongly context-dependent and structurally complex. When semantic information is introduced into neural implicit SLAM systems, the need to process such high-dimensional and complex features often leads to a significant drop in overall runtime efficiency. In addition, most existing methods adopt relatively simple semantic processing strategies. For example, vMap [8] only embeds semantic information into object-level neural field representations but does not construct a semantic map; in NIDS-SLAM [5], semantic information is treated as RGB data and used for simple coloring; SNI-SLAM [31] introduces a joint mapping

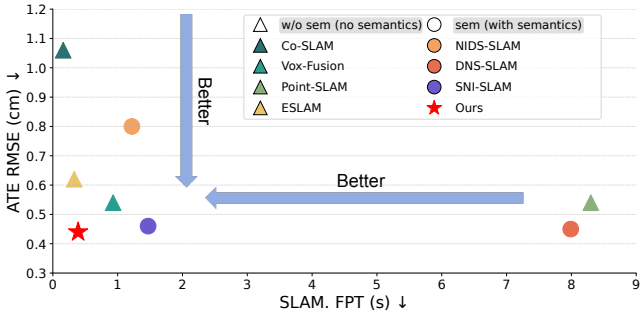

Fig. 1: MTE-SLAM delivers the best trajectory accuracy and offers the fastest runtime among semantic neural implicit SLAM methods, approaching the speed of leading non-semantic systems and achieving a balance between localisation precision and real-time performance. In addition, MTE-SLAM produces cleaner and more consistent object-level semantic boundaries, exhibiting noticeably fewer semantic-segmentation errors than competing semantic baselines.

This work is supported in part by the National Natural Science Funds for Distinguished Young Scholar (No. 62325307), in part by the National Natural Science Foundation Major Scientific Research Instrument Development Project (No. 62527809), in part by the Shenzhen Key Industry R&D Program Project (No. ZDCY20250901102300001), in part by the National Natural Science Foundation of China (Nos. 62373257, 62473264, 62203134, 62403325), in part by the Natural Science Foundation of Guangdong Province (Nos. 2023B1515120038, 2026A1515011532), in part by Shenzhen Science and Technology Innovation Commission (No. KJZD20230923113801004), in part by the Key-Area Research and Development Program of Guangdong Province (No. 2025B0909020002), and in part by the Open Project of State Key Laboratory of Synthetical Automation for Process Industries (No. SAPI-2025-KFKT-11). This work is supported by the Intelligent Computing Center of Shenzhen University.

All authors are with the School of Artificial Intelligence and National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518061, China.

*Corresponding author: Jianqiang Li(lijq@szu.edu.cn)

mechanism of semantics and geometry, but its fusion strategy places multi-modal features side by side at a single scale, without explicitly considering the distribution differences of semantic features in the global context and local boundary continuity. As a result, although the generated maps contain some semantic labels, the overall system becomes slower, and the semantic details in the maps remain coarse and blurred. There is a pressing need for more fine-grained and efficient fusion strategies that can leverage semantic information without introducing prohibitive computational overhead.

To address the aforementioned challenges, we propose a multi-tier semantic feature fusion mechanism that fully leverages semantic information across modalities and scales. Specifically, our approach integrates semantic features at both global and local levels: global semantic context is first used to enhance representations by capturing scene-level cues, while local refinement enforces spatial continuity through neighborhood consistency. This global-local fusion strategy ensures effective utilization of semantic information at both macro and micro levels, thereby significantly enriching the coherence and expressiveness of the resulting semantic maps. In parallel, we introduce a feature redundancy suppressor designed to perform dynamic compression and refinement of semantic features. This module adaptively identifies and discards less informative components based on feature importance and temporal variation, thus reducing redundancy. As a result, it curbs the growth of model complexity and mitigates the computational overhead associated with iterative training, effectively preventing efficiency degradation due to unbounded complexity growth.

Overall, our main contributions include:

- We propose a multi-tier semantic feature fusion mechanism that enhances semantic representations by leveraging scene-level global context and enforcing continuity constraints within local spatiotemporal neighborhoods, thereby mitigating missing edge details and blurred semantic boundaries to improve reconstruction quality.
- We utilize a feature redundancy suppressor that dynamically prunes redundant information based on feature importance and temporal variation, compressing feature representation while preserving semantic expressiveness to improve training efficiency and real-time performance.
- Our method is validated on Replica [22] and ScanNet [3], outperforming existing semantic SLAM approaches in tracking, mapping, segmentation and real-time performance.

II. Related Work

a) Neural Implicit SLAM: Traditional feature-based SLAM [1], [17]–[19] depends on sparse handcrafted features and degrades in textureless or dynamic environments [10]. Recent advances introduce neural implicit representations for dense, end-to-end scene modeling. DROID-SLAM [25] combines learned motion and depth estimation with dense bundle adjustment. iMAP [24] employs an MLP for unified, scene-specific representation in real-time tracking and mapping. NICE-SLAM [32] improves scalability and detail through hierarchical encoding and geometric priors. Vox-Fusion [30] adopts voxel partitioning with octree optimization for efficient reconstruction. Point-SLAM [21] anchors features on point clouds to support adaptive resolution. Co-SLAM [29] leverages multi-resolution hash grids and volumetric encoding for faster convergence and higher fidelity. ESLAM [7] incorporates neural radiance fields and multi-scale feature planes for efficient dense mapping. QQ-SLAM [6] addresses the optimization inefficiency in neural SLAM by introducing quantized feature queries, facilitating faster convergence during per-frame optimization. However, it focuses only on compressing visual features without semantic information. In contrast, we achieve faster convergence during per-frame optimization, but we propose a semantic-aware compression strategy specifically for semantic information, enabling efficient and consistent semantic optimization at each frame.

b) Semantic Neural Implicit SLAM: Despite progress in dense neural implicit SLAM, most methods emphasize geometric and visual reconstruction, lacking explicit semantic understanding. This limits their utility in downstream tasks like semantic navigation and object-level manipulation. Recent works integrate semantics into neural SLAM frameworks. vMap [8] presents an object-centric system modeling each object instance with a lightweight MLP, supporting online, watertight object-level mapping. NIDS-SLAM [5] combines classical SLAM pipelines with neural implicit networks to jointly learn geometry, appearance and semantics at scale. DNS-SLAM [11] enhances reconstruction and semantic coherence by incorporating 2D semantic priors and multi-view constraints. SNI-SLAM [31] introduces a hierarchical semantic representation with cross-modal feature collaboration, achieving robust mapping and tracking under occlusion or noise. However, it treats feature fusion as a single-stage process, lacking attention to fine-grained semantic consistency and boundary precision. To address this, we propose a multi-tier semantic fusion strategy: global fusion captures scene-level context for structural coherence, while local fusion refines neighborhood continuity to improve boundary sharpness and occlusion robustness.

III. Method

This section explains how to effectively leverage semantic information in neural implicit semantic SLAM to improve overall performance. First, a multi-tier feature fusion mechanism enhances semantic representation by integrating global context and local spatiotemporal constraints, reducing boundary ambiguity and detail loss. Second, the feature redundancy suppressor and its working process are described. The overall pipeline of our method is shown in Fig. 2.

A. Multi-Tier Feature Fusion

In multimodal perception, geometric, semantic and appearance features complement each other to enable accurate object understanding. Geometric features convey spatial

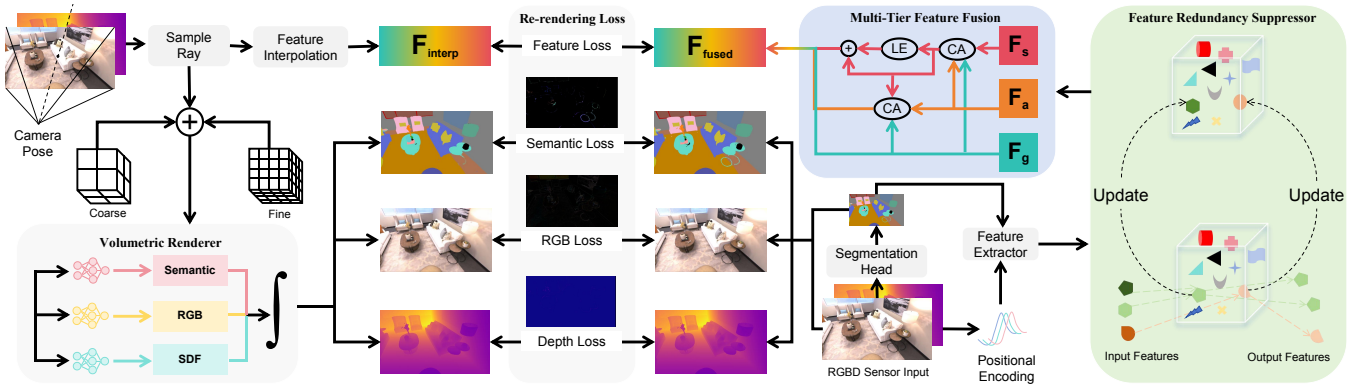

Fig. 2: Overview of our MTE-SLAM framework. Our framework takes an RGB-D stream as input, samples 3D rays, and extracts multi-modal features via a feature extractor and segmentation head. Semantic features are projected into an updatable space and refined by the feature redundancy suppressor to remove redundant information. Appearance, geometry and semantic features are then fused through a multi-tier module for enhanced representation. This fused output, along with RGB-D images and segmentation results, provides supervision. Interpolated features are decoded to generate RGB, depth and semantic predictions, which combined with supervision signals, define the loss function. The entire system is jointly optimized to estimate camera poses and update a dense 3D map.

structure, aiding localization and mapping. Semantic features provide category-level recognition, essential for task planning and decision-making. Appearance features offer rich textures and visual details that support semantic reasoning and enhance recognition. Fine textures help distinguish similar-looking regions, improving recognition accuracy and geometric reconstruction, while reducing redundant mapping and registration errors. To harness the correlations and constraints among these features, we propose a hierarchical multimodal fusion mechanism that integrates structural cues from geometry, relational context from semantics and detailed textures from appearance.

a) Global Fusion: To fully exploit the complementarity among geometric, semantic and appearance features, we design a two-stage cross-modal attention fusion mechanism. In the first stage, geometric features guide the enhancement of semantic features. The structural information from geometry helps highlight components of the semantic representation that are closely aligned with the scene's physical layout, thereby improving the accuracy and reliability of the semantic encoding. In the second stage, semantic features are leveraged to refine appearance features. With the globally consistent context provided by semantics, the network can guide the appearance representation to focus on visual details that are coherent with semantic categories. This hierarchical fusion pipeline enables progressive information exchange from geometry to semantics and from semantics to appearance through cross-modal attention.

We extract features from input RGB-D frames $I = \{R_i, D_i\}_{i=1}^{n}$, representing geometric, semantic and appearance features as matrices $\mathbf{G} \in \mathbb{R}^{N\times C_g}$, $\mathbf{S} \in \mathbb{R}^{N\times C_s}$ and $\mathbf{A} \in \mathbb{R}^{N\times C_a}$, where N is the number of feature elements and C_g, C_s and C_a are the feature dimensions of each modality. We project each modality to a shared embedding space of dimension d:

$$\mathbf{Q}_1 = W_Q^{(1)}\mathbf{G}, \quad \mathbf{K}_1 = W_K^{(1)}\mathbf{S}, \quad \mathbf{V}_1 = W_V^{(1)}\mathbf{A}, \tag{1}$$

where $W_Q^{(1)} \in \mathbb{R}^{d\times C_g}$, $W_K^{(1)} \in \mathbb{R}^{d\times C_s}$ and $W_V^{(1)} \in \mathbb{R}^{d\times C_a}$ as learnable projections. The fusion attention matrix is computed via $\mathbf{A}_{(1)} = \text{softmax}(\mathbf{Q}_1\mathbf{K}_1^\top/\sqrt{d})$, yielding geometry-guided semantic features:

$$\mathbf{T}_s = \mathbf{A}_{(1)}\mathbf{V}_1 \in \mathbb{R}^{N\times d}. \tag{2}$$

We adopt a cross-modal attention fusion mechanism: geometric features serve as queries ($\mathbf{Q}_1$), semantic features as keys ($\mathbf{K}_1$) and appearance features as values ($\mathbf{V}_1$). The spatial structure in geometric features guides attention to consistent visual regions, suppressing noisy or irrelevant semantics and yielding more reliable semantic representation $\mathbf{T}_s$. In the second stage, semantic features $\mathbf{T}_s$ guide the refinement of appearance features. Specifically, $\mathbf{T}_s$ serves as queries and appearance features $\mathbf{A}$ as both keys and values:

$$\mathbf{Q}_2 = W_Q^{(2)}\mathbf{T}_s, \quad \mathbf{K}_2 = W_K^{(2)}\mathbf{A}, \quad \mathbf{V}_2 = W_V^{(2)}\mathbf{A}, \tag{3}$$

where $W_Q^{(2)}, W_K^{(2)}, W_V^{(2)}$ are independently learned projections. Attention weights are computed as $\mathbf{A}_{(2)} = \text{softmax}(\mathbf{Q}_2\mathbf{K}_2^\top/\sqrt{d})$, and the enhanced appearance features are:

$$\mathbf{T}_a = \mathbf{A}_{(2)}\mathbf{V}_2 \in \mathbb{R}^{N\times d}. \tag{4}$$

This two-stage design enables progressive, modality-aware fusion: the first stage improves semantic features under geometric constraints, while the second enriches appearance features using refined semantics without disrupting the established geometric-semantic structure.

b) Local Fusion: Although global semantic and appearance features are enhanced through cross-modal attention fusion, the local structural details of semantic features may still remain ambiguous. Semantic features often suffer

from oversmoothing near object boundaries. Then, we use a lightweight local modeling mechanism into the fusion module to further enhance the spatial resolution. By aggregating features within a local neighborhood to strengthen the semantic response around object boundaries and fine structures:

$$\mathbf{T}'_s = \mathrm{Conv}_2\left(\sigma\left(\mathrm{Conv}_1\left(\mathbf{T}_s\right)\right)\right), \tag{5}$$

where Conv_1 and Conv_2 denote depthwise or lightweight convolutions, and $\sigma(\cdot)$ is a nonlinear activation function. To enhance local semantic features with clearer region boundaries and finer structural details, we employ a depthwise-pointwise convolutional module defined as follows:

$$\mathbf{S}' = \sigma\left(\mathrm{Conv}^{\mathrm{dw}}_{3\times3}(\mathbf{T}_s)\right), \quad \hat{\mathbf{S}} = \sigma\left(\mathrm{Conv}^{\mathrm{pw}}_{3\times3}(\mathbf{S}')\right). \tag{6}$$

After processing through this module, we obtain the locally enhanced semantic feature $\hat{\mathbf{S}}$, which exhibits clearer region boundaries and finer structural details. However, simply stacking the enhanced semantic, appearance, and geometric features may introduce errors due to inconsistencies across modalities, so we need a more adaptive feature integration strategy that can handle spatially varying modality reliability. Then, we design a dynamic gating fusion mechanism that adaptively integrates multi-modal features:

$$[\Gamma_g(i), \Gamma_s(i), \Gamma_a(i)] = \sigma\left(W_g * [\mathbf{G}(i); \hat{\mathbf{S}}(i); \mathbf{T}_a(i)]\right), \tag{7}$$

where $i = 1, \ldots, N$, W_g denotes the convolution kernel weights used in the gating function, "$*$" represents the convolution operation, ";" indicates channel-wise feature concatenation, and σ is the sigmoid activation. The gating coefficients $\Gamma_g(i)$, $\Gamma_s(i)$ and $\Gamma_a(i)$ correspond to the weights for geometric, semantic and appearance modalities at spatial location i, and are constrained within the range $[0, 1]$. To ensure that the total contribution of different modalities is appropriate, we normalize the output gating coefficients and ensure that $\Gamma_g(i) + \Gamma_s(i) + \Gamma_a(i) = 1$. This mechanism adaptively adjusts fusion weights across spatial locations. In areas with reliable semantics but weak visual cues, higher weights are assigned to the semantic modality to enhance category-level understanding. In texture-rich yet semantically ambiguous regions, appearance features are emphasized for fine detail. At structural boundaries, geometric features dominate to preserve spatial accuracy. Finally, the fused feature representation at each spatial location is computed as:

$$\mathbf{F}_{fused}(i) = \Gamma_g(i)\,\mathbf{G}(i) + \Gamma_s(i)\,\hat{\mathbf{S}}(i) + \Gamma_a(i)\,\mathbf{T}_a(i), \tag{8}$$

where $\mathbf{F}_{fused}(i)$ denotes the fused implicit feature representation vector at spatial location i. After the two stage cross-modal attention and local gated fusion, we obtain a unified multi-modal feature representation.

B. Feature Redundancy Suppressor

Most SLAM systems construct the semantic feature space statically, using a fixed set of semantic prototypes predefined at initialization for feature alignment, matching and map annotation [9]. As frames accumulate, redundant or similar semantic features often emerge, leading to high-dimensional redundancy that degrades computational efficiency, real-time performance and map accuracy. Therefore we propose a feature redundancy suppressor that adaptively adjusts the semantic feature space during operation. This module dynamically compresses redundancy, restructures features, and enhances their discriminability and generalization by analyzing the temporal and spatial dynamics of the input data.

a) Feature Space Initialization: At the early stage of training neural implicit SLAM systems, semantic features extracted by the network are often unstable, leading to spatiotemporal drift in semantic representations. To provide a reliable structural foundation, we introduce an explicit feature space for initialization. The key idea is to predefine a set of semantic prototype vectors, each corresponding to a semantic category, serving as stable references from the start of training. We initialize this space by sampling representative feature vectors from the initial training data. Let the scene contain C semantic categories; the initial feature space is defined as $\mathcal{A}^{(0)} = \{\mathbf{u}_c^{(0)}\}_{c=1}^{C}$, where $\mathbf{u}_c^{(0)}$ denotes the initial prototype for category c. To ensure diversity and coverage, we adopt a balanced sampling strategy that selects representative features from each category, ensuring every class is associated with at least one prototype.

b) Semantic Feature Compression Mapping: As training progresses, new semantic features are continuously extracted from sensor data or the neural network. Directly incorporating these features into the map may result in jitter or drift in the semantic locations due to random noise or fluctuations in network weights. To ensure semantic consistency and mitigate drift in the feature space, we introduce a mechanism to robustly anchor each new feature to the existing semantic structure. Whenever a new semantic feature vector $\mathbf{f}_{j_*}^{\mathrm{new}}$ is generated, we search for the most similar prototype vector in the current explicit feature space $\mathcal{A}$ to serve as its projection anchor. Specifically, we adopt a nearest-neighbor matching strategy: based on a distance metric, we compute the similarity between $\mathbf{f}_{j_*}^{\mathrm{new}}$ and all existing anchors $\mathbf{u}_j$:

$$j^* = \arg\min_{j\in\{1,\ldots,|\mathcal{A}|\}} \|\mathbf{f}_{j_*}^{\mathrm{new}} - \mathbf{u}_j\|_2. \tag{9}$$

The anchor point $\mathbf{u}_{j^*}$ with the smallest distance is selected as the match, effectively performing vector quantization in feature space, where anchors serve as centroids and features are assigned to their nearest prototypes. This nearest-neighbor mapping anchors new semantic information to stable references, preventing independent drift. The matched anchor $\mathbf{u}_{j^*}$ participates in both training and inference. During training, it provides a supervisory signal that guides the network's output toward the semantic prototype, stabilizing feature learning. At inference, it acts as a semantic representative, enabling efficient prediction. For instance, when a known object is reobserved, its features likely match a stored anchor, instantly yielding the correct semantic label.

We define a compression loss that penalizes the discrepancy between the input feature vector and its corresponding

updated output. Specifically, the loss is defined as the squared ℓ_2 norm between the original feature vector $\mathbf{u}_j$ and the reconstructed feature $\mathbf{f}_{j_*}^{\text{new}}$ after compression:

$$\mathcal{L}_{\text{C}} = \left\| \mathbf{f}_{j_*}^{\text{new}} - \mathbf{u}_j \right\|_2^2, \tag{10}$$

which encourages feature stability across the optimization process and helps retain semantic consistency. We define the overall loss as a weighted sum of the following components $\mathcal{L} = \lambda_C \mathcal{L}_C + \mathcal{L}_{sni}$, where $\mathcal{L}_{sni}$ is consistent with those used in SNI-SLAM [31], which provides comprehensive supervision signals across geometry, semantics, features and appearance. We assign λ_C according to the optimization targets at different stages of training.

c) Adaptive Feature Space Refinement: As training progresses and the network converges, the distribution of semantic features stabilizes, making the initial anchor vectors less representative. If kept static, these anchors may fail to reflect the fine-grained semantics learned later in training. To address this, we propose an adaptive refinement mechanism that updates anchor vectors throughout training, ensuring alignment with the evolving semantic feature space. This leads to a more representative and compact set of prototypes. The core idea is to incrementally update each anchor after successful feature matching using a weighted rule. For an input feature $\mathbf{f}_{j_*}^{\text{new}}$ matched to anchor $\mathbf{u}_{j^*}$, the anchor is updated as:

$$\mathbf{u}_{j_*}^{\text{new}} \leftarrow \mathbf{u}_{j_*}^{\text{old}} + \beta \left(\mathbf{f}_{j_*}^{\text{new}} - \mathbf{u}_{j_*}^{\text{old}} \right), \tag{11}$$

where β is a small learning rate controlling the update magnitude, and which performs an exponential moving average update of the anchor using the newly observed feature $\mathbf{f}_{j_*}^{\text{new}}$. As more new features are continuously mapped to the same anchor, the anchor vector gradually converges to the "centroid" of these features. If we additionally let each anchor track the number of features it has matched, denoted as N_j, and set the update weight as $\beta = 1/(N_j + 1)$, the update becomes an incremental mean:

$$\mathbf{u}_{j^*}^{(t+1)} = \frac{N_{j^*}}{N_{j^*} + 1} \cdot \mathbf{u}_{j^*}^{(t)} + \frac{1}{N_{j^*} + 1} \cdot \mathbf{f}_{j_*}^{\text{new}}, \tag{12}$$

which precisely computes the running average of all matched samples. In this way, each anchor progressively approximates the true feature mean of its associated semantic category, gradually refining the anchor set in feature space during training through this optimization process. The initial set $\mathcal{A}^{(0)}$ evolves into $\mathcal{A}^{(t)}$, and eventually converges to $\mathcal{A}^*$, a sparse yet representative set of semantic prototypes. This compact set effectively captures the semantic patterns in the scene using a limited number of anchor vectors. This design greatly reduces storage and computational cost while maintaining semantic expressiveness. Since the number of anchors is much smaller than the total number of features, loss computation and matching operate only on anchor vectors, improving training efficiency. At inference, semantic recognition is performed via nearest-neighbor search over the anchor set, avoiding costly per-pixel network inference.

Method	0000	0059	0106	0207
NICE-SLAM	8.64	12.25	8.09	5.59
Co-SLAM	7.13	11.14	9.36	7.14
QQ-SLAM	6.99	9.47	8.22	8.49
ESLAM	7.30	8.50	7.50	5.70
SNI-SLAM	**6.90**	7.38	7.19	4.70
Ours	8.50	**7.10**	**6.84**	**4.53**

TABLE I: ATE RMSE (cm)↓ in tracking on ScanNet.

Method	room0	room1	room2	office0	office1	office2	office3	office4
NICE-SLAM	1.86	2.37	2.26	1.50	1.01	1.85	5.67	3.35
Co-SLAM	0.72	0.85	1.02	0.69	0.56	2.12	1.62	0.87
QQ-SLAM	0.58	1.16	0.87	0.52	0.48	1.74	1.12	0.73
ESLAM	0.71	0.70	0.52	0.57	0.55	0.58	0.72	0.63
SNI-SLAM	**0.50**	0.55	0.45	**0.33**	0.41	**0.32**	0.62	0.47
Ours	0.55	**0.44**	**0.43**	0.36	**0.39**	0.34	**0.59**	**0.45**

TABLE II: ATE RMSE (cm)↓ in tracking on Replica.

IV. Experiments

A. Experimental Setup

a) Datasets: We evaluate our method on synthetic and real-world datasets with semantic annotations. Reconstruction quality is quantitatively evaluated on 8 synthetic scenes from Replica [22]. Camera pose accuracy is quantitatively evaluated on both Replica [22] and ScanNet [3]. The ground-truth camera poses and semantic maps of Replica [22] are provided by simulation. The ground-truth camera poses of ScanNet [3] are obtained using BundleFusion.

b) Evaluation Metrics: We adopt the culling strategy and evaluation protocol from Co-SLAM [29], using the reconstruction accuracy metrics Depth L1 (cm), Accuracy (cm), Completion (cm) and Completion Ratio (%) with a threshold of 5 cm to assess the performance of our SLAM system. To evaluate the camera pose, we use the average absolute trajectory error (ATE RMSE (cm) [23]). Semantic segmentation is evaluated with respect to mIoU (%) [13]metric.

c) Baseline Methods: For SLAM performance, we compare our method with state-of-the-art NeRF-based dense visual SLAM methods NICE-SLAM [32], Co-SLAM [29], Vox-Fusion [30], Point-SLAM [21], ESLAM [7], QQS-LAM [6], vMap [8], NIDS-SLAM [5], DNS-SLAM [11] and SNI-SLAM [31].

d) Implementation Details: All experiments are conducted using a NVIDIA RTX 4090 GPU. All reported results are averaged 5 runs to ensure reliability. We adopt a Truncated Signed Distance Function (TSDF) representation and perform volume fusion incrementally with camera poses optimized via photometric tracking. For Replica, the TSDF mesh resolution is 0.01 m, while for ScanNet it is 0.02 m for efficiency. Tracking is initialized with 2000 sampled pixels and optimized for 8 iterations on Replica and 30 iterations on ScanNet, with learning rates $\text{lr}_T = 0.002$, $\text{lr}_R = 0.001$ for Replica and $\text{lr}_T = 0.0005$, $\text{lr}_R = 0.0025$ for ScanNet.

B. Tracking Evaluation

Quantitative tracking results are shown in Tab. I and Tab. II. In most scenes, our method achieves lower errors in ATE RMSE, significantly improving trajectory accuracy.

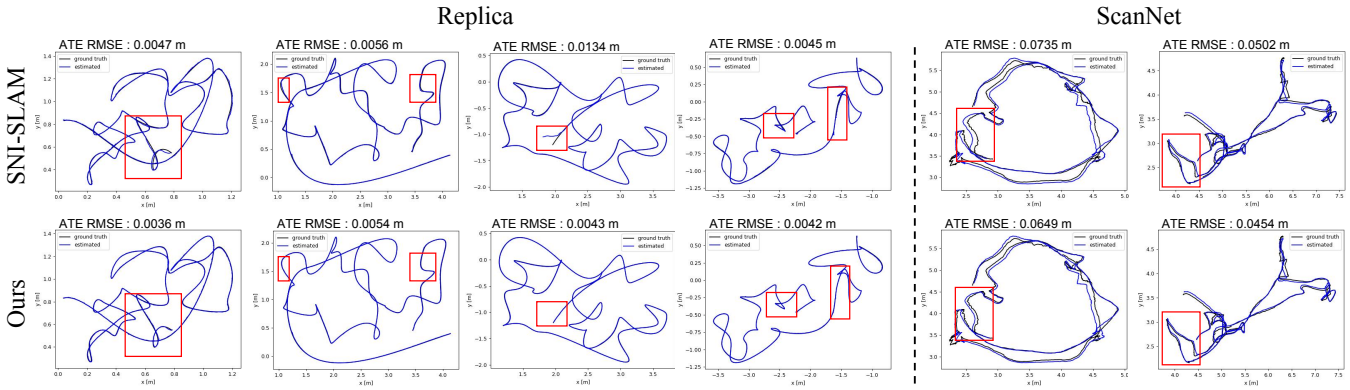

Fig. 3: Qualitative Tracking on Replica and ScanNet.

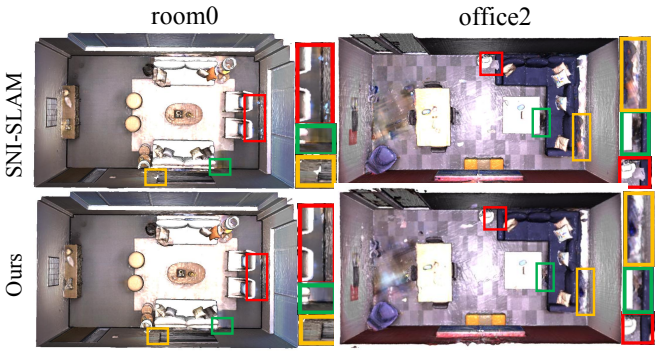

Fig. 4: Qualitative RGB Reconstruction on Replica.

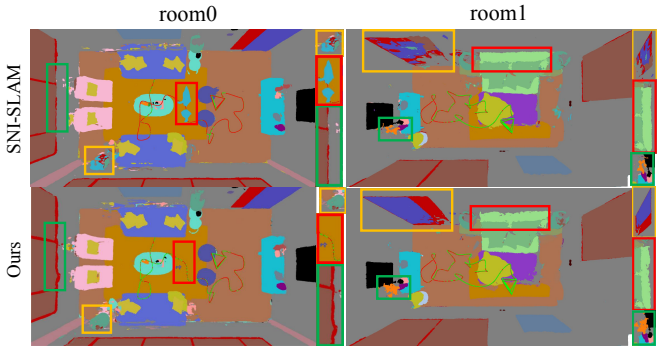

Fig. 5: Qualitative Semantic Reconstruction on Replica.

Method	Depth L1 (cm) ↓	Acc. (cm) ↓	Comp. (cm) ↓	Comp. Ratio(%) ↑
NICE-SLAM	3.53	2.85	3.00	89.33
Co-SLAM	1.62	2.10	2.08	93.44
QQ-SLAM	1.42	2.43	1.93	94.38
ESLAM	1.18	**0.97**	1.05	98.60
SNI-SLAM	0.77	1.94	1.69	96.62
Ours	**0.16**	1.06	**0.99**	**99.23**

TABLE III: Quantitative Reconstruction Comparison on Replica.

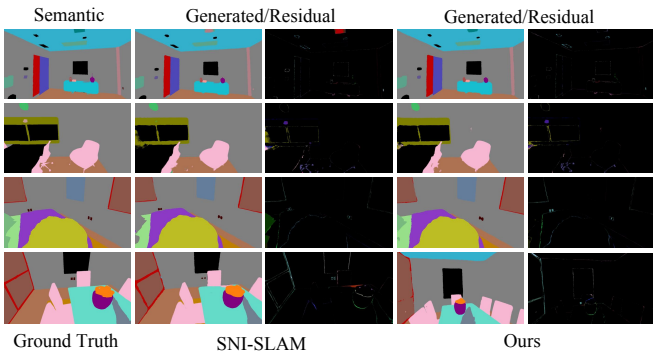

Fig. 6: Qualitative Semantic Segmentation on Replica.

Method	room0	room1	room2	office0	office1	office2	office3	office4
NIDS-SLAM	82.45	84.08	76.99	85.94	-	-	-	-
DNS-SLAM	88.32	84.90	81.20	84.66	-	-	-	-
SNI-SLAM	**88.42**	87.43	86.16	**87.63**	78.63	**86.49**	74.01	80.22
Ours	86.75	**88.85**	**86.31**	86.40	**90.21**	84.77	**81.40**	**80.46**

TABLE IV: Semantic Segmentation mIoU(%)↑ Comparison on Replica.

As shown in the qualitative comparison in Fig. 3, the camera trajectories generated by our method are smoother and exhibit less jitter compared to SNI-SLAM [31]. Stable and continuous tracking is maintained even under large-scale rapid motion or complex structural variations.

C. Reconstruction Evaluation

Quantitative and qualitative reconstruction results on the Replica [22] dataset are shown in Tab. III and Fig. 4. Our method achieves the lowest Depth L1 error, lowest completion error and highest completion ratio, indicating superior reconstruction fidelity and surface coverage. While ESLAM [7] slightly surpasses us in accuracy, our method offers a better trade-off between accuracy and completeness. Qualitative results further show clear improvements in local continuity and geometric completeness. As shown in the Fig. 4, in room0 scene, the reconstructed edges between walls and adjacent objects are smoother and more continuous. In office2 scene, the sofa exhibits finer geometric details and more coherent surface structure.

D. Semantics Evaluation

To evaluate the semantic segmentation performance of our method, Tab. IV reports the mIoU (%) results on the Replica [22] dataset across multiple indoor scenes. Our method achieves the best or highly competitive performance

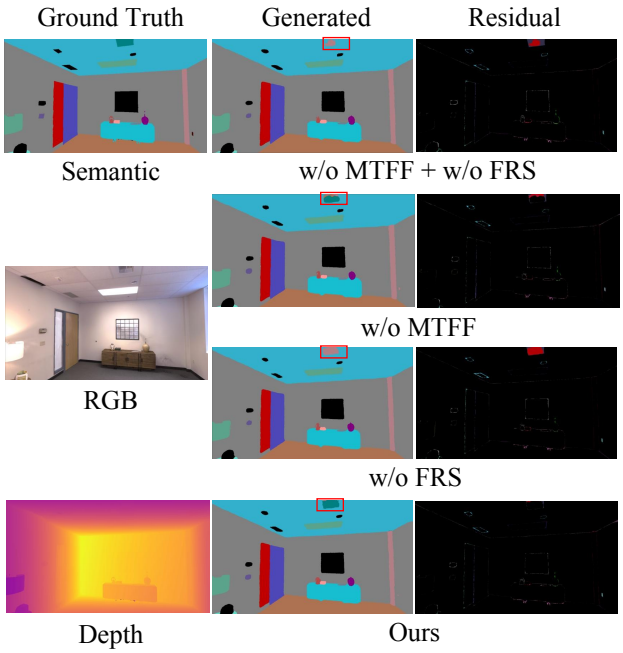

Fig. 7: Qualitative Ablation Study on Replica.

	Method	Track. FPT(s) ↓	Map. FPT(s) ↓	SLAM. FPT(s) ↓
w/o sem	NICE-SLAM	0.07	5.00	5.05
	Vox-Fusion	0.47	0.46	0.93
	Point-SLAM	0.85	7.15	8.23
	Co-SLAM	0.14	**0.10**	**0.16**
	ESLAM	**0.06**	0.28	0.33
sem	vMAP	0.15	1.72	1.66
	DNS-SLAM	0.36	7.58	7.99
	SNI-SLAM	0.06	1.40	1.47
	Ours	0.12	0.32	0.39

TABLE V: Runtime Comparison on Replica.

in most scenes. In the office1 scene, it attains an mIoU of 90.21%, surpassing the strongest baseline by more than 11.5%. In room1 and office3, the scores of 88.85% and 81.40% also outperform all other methods. Even in room0, where baseline methods perform well, our method maintains a high score of 86.75%, indicating strong robustness and consistency across environments. As shown in Fig. 5, the semantic reconstruction of the floor in room0 is more complete. In room1, large objects such as the door and sofa are well reconstructed in semantic structure. The continuity at the junction between the window and the wall in room0, together with the semantic details of the desk lamp, is handled more precisely. The vase in room1 also exhibits better detail preservation. These results demonstrate the advantage of our method in reconstructing small objects and structural boundaries with higher semantic fidelity. Fig. 6 presents qualitative semantic segmentation results on the Replica dataset, further highlighting the superiority of our method in semantic accuracy and structural fidelity. Compared with SNI-SLAM, our approach produces more accurate segmentation of fine-grained structures including door frames and TV edges, as well as small objects such as desk lamps and items on the desk, with clearer boundaries and improved semantic consistency.

Method	MTFF	FRS	mIoU(%) ↑
w/o MTFF + w/o FRS			78.63
w/o MTFF		✓	88.65
w/o FRS	✓		82.16
Ours	✓	✓	90.21

TABLE VI: Quantitative Ablation Study on Replica(office1).

E. Runtime Analysis

We compare our method with NeRF-SLAM approaches without semantic modeling and semantic NeRF-SLAM methods. As shown in the Tab. V, while our method is not the fastest among all NeRF-SLAM methods in terms of average per-frame processing time, it performs competitively within the category of semantic NeRF-SLAM approaches. Compared to the baselines, our method achieves faster mapping speed, thereby improving overall runtime efficiency.

F. Ablation Study

Tab. VI presents a series of experiments on the office1 scene from the Replica dataset to validate the effectiveness of the multi-tier feature fusion mechanism and the feature redundancy suppressor in our method.

a) Multi-Tier Feature Fusion (MTFF): After applying the multi-tier feature fusion module, the semantic segmentation results in Fig. 7 show more continuous, clearer door-frame edges and more complete, structurally accurate ceiling lights, highlighting the model's improved ability to capture fine details and preserve spatial consistency in complex boundary regions.

b) Feature Redundancy Suppressor (FRS): Both the quantitative results in Tab. VI and the qualitative comparisons in Fig. 6 consistently demonstrate a substantial improvement in segmentation accuracy when the Feature Redundancy Suppressor is incorporated, elevating the accuracy from 78.63% to 82.16%. The ceiling light, previously prone to misclassification or fragmentation, is now correctly labeled, demonstrating the module's effectiveness in refining details and reducing ambiguity in complex regions.

V. Conclusion

We propose MTE-SLAM, a novel semantic SLAM system that achieves both efficiency and high-fidelity scene understanding. It introduces a multi-tier semantic fusion strategy and a dynamic feature redundancy suppression module to better exploit semantic cues, resolve structural ambiguity and reduce computational overhead. The framework combines global semantic reasoning with local spatial refinement, enabling accurate, compact, and coherent scene reconstructions. Extensive experiments demonstrate that MTE-SLAM consistently outperforms existing methods in both tracking and mapping accuracy while significantly reducing runtime.

REFERENCES

[1] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[2] Chi-Ming Chung, Yang-Che Tseng, Ya-Ching Hsu, Xiang Qian Shi, Yun-Hung Hua, Jia-Fong Yeh, Wen-Chin Chen, Yi-Ting Chen, and Winston H. Hsu. Orbeez-SLAM: A real-time monocular visual SLAM with ORB features and nerf-realized mapping. In *IEEE International Conference on Robotics and Automation*, pages 9400–9406, London, UK, May–June 2023. IEEE.

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2432–2443. IEEE Computer Society, 2017.

[4] Calvin Galagain, Martyna Poreba, and François Goulette. Is semantic SLAM ready for embedded systems? a comparative survey. *arXiv preprint*, abs/2505.12384, May 2025.

[5] Yasaman Haghighi, Suryansh Kumar, Jean-Philippe Thiran, and Luc Van Gool. Neural implicit dense semantic SLAM. *arXiv preprint*, abs/2304.14560, 2023.

[6] Sijia Jiang, Jing Hua, and Zhizhong Han. Query quantized neural SLAM. In Toby Walsh, Julie Shah, and Zico Kolter, editors, *Association for the Advancement of Artificial Intelligence*, pages 4057–4065. AAAI Press, 2025.

[7] Mohammad Mahdi Johari, Camilla Carta, and François Fleuret. ESLAM: Efficient dense SLAM system based on hybrid representation of signed distance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17408–17419, Vancouver, BC, Canada, June 2023. IEEE.

[8] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J. Davison. vmap: Vectorised object mapping for neural field SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 952–961, Vancouver, BC, Canada, June 2023. IEEE.

[9] J. Li, Z. Chen, J. Chen, and Q. Lin. Diversity-sensitive generative adversarial network for terrain mapping under limited human intervention. *IEEE Trans. Cybern.*, 51(12):6029–6040, 2021.

[10] J. Li, Y. Zhang, Z. Chen, J. Wang, M. Fang, C. Luo, and H. Wang. A novel edge-enabled SLAM solution using projected depth image information. *Neural Comput. Appl.*, 32(19):15369–15381, 2020.

[11] Kunyi Li, Michael Niemeyer, Nassir Navab, and Federico Tombari. DNS-SLAM: Dense neural semantic-informed SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 7839–7846. IEEE, 2024.

[12] Jiuming Liu, Guangming Wang, Zhe Liu, Chaokang Jiang, Marc Pollefeys, and Hesheng Wang. Regformer: An efficient projection-aware transformer network for large-scale point cloud registration. In *IEEE/CVF International Conference on Computer Vision*, pages 8417–8426. IEEE, 2023.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 3431–3440. IEEE Computer Society, 2015.

[14] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7210–7219. Computer Vision Foundation / IEEE, 2021.

[15] John McCormac, Ankur Handa, Andrew J. Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *IEEE International Conference on Robotics and Automation*, pages 4628–4635. IEEE, 2017.

[16] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision European Conference*, volume 12346 of *Lecture Notes in Computer Science*, pages 405–421. Springer, 2020.

[17] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[18] Raul Mur-Artal and Juan D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[19] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[20] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *IEEE International Conference on Robotics and Automation*, pages 1689–1696. IEEE, 2020.

[21] Erik Sandström, Yue Li, Luc Van Gool, and Martin R. Oswald. Point-slam: Dense neural point cloud-based SLAM. In *IEEE/CVF International Conference on Computer Vision, Paris, France, October 1-6, 2023*, pages 18387–18398. IEEE, 2023.

[22] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Yuheng Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard A. Newcombe. The Replica Dataset: A digital replica of indoor spaces. *arXiv preprint*, abs/1906.05797, 2019.

[23] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580. IEEE, 2012.

[24] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J. Davison. imap: Implicit mapping and positioning in real-time. In *IEEE/CVF International Conference on Computer Vision*, pages 6209–6218. IEEE, 2021.

[25] Zachary Teed and Jia Deng. DROID-SLAM: deep visual SLAM for monocular, stereo, and RGB-D cameras. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, pages 16558–16569, 2021.

[26] Yulun Tian, Yun Chang, Fernando Herrera Arias, Carlos Nieto-Granda, Jonathan P. How, and Luca Carlone. Kimera-Multi: Robust, distributed, dense metric-semantic SLAM for multi-robot systems. *IEEE Transactions on Robotics*, 38(4):2022–2038, 2022.

[27] Guangming Wang, Xinrui Wu, Shuyang Jiang, Zhe Liu, and Hesheng Wang. Efficient 3d deep lidar odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5749–5765, 2023.

[28] Guangming Wang, Xinrui Wu, Zhe Liu, and Hesheng Wang. Hierarchical attention learning of scene flow in 3d point clouds. *IEEE Transactions on Image Processing*, 30:5168–5181, 2021.

[29] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-SLAM: Joint coordinate and sparse parametric encodings for neural real-time SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13293–13302, Vancouver, BC, Canada, June 2023. IEEE.

[30] Xingrui Yang, Hai Li, Hongjia Zhai, Yuhang Ming, Yuqian Liu, and Guofeng Zhang. Vox-fusion: Dense tracking and mapping with voxel-based neural implicit representation. In Henry B. L. Duh, Ian Williams, Jens Grubert, J. Adam Jones, and Jianmin Zheng, editors, *IEEE International Symposium on Mixed and Augmented Reality*, pages 499–507. IEEE, 2022.

[31] Siting Zhu, Guangming Wang, Hermann Blum, Jiuming Liu, Liang Song, Marc Pollefeys, and Hesheng Wang. SNI-SLAM: Semantic neural implicit SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21167–21177. IEEE, 2024.

[32] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. NICE-SLAM: neural implicit scalable encoding for SLAM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12776–12786. IEEE, 2022.