2025 IEEE International Conference on Robotics and Automation (ICRA)
May 19-23, 2025. Atlanta, USA

NeuraLoc: Visual Localization in Neural Implicit Map with Dual Complementary Features

Hongjia Zhai[1], Boming Zhao[1], Hai Li[2], Xiaokun Pan[1], Yijia He[2], Zhaopeng Cui[1], Hujun Bao[1], Guofeng Zhang[1*]
[1]State Key Lab of CAD&CG, Zhejiang University [2]RayNeo

***Abstract*— Recently, neural radiance fields (NeRF) have gained significant attention in the field of visual localization. However, existing NeRF-based approaches either lack geometric constraints or require extensive storage for feature matching, limiting their practical applications. To address these challenges, we propose an efficient and novel visual localization approach based on the neural implicit map with complementary features. Specifically, to enforce geometric constraints and reduce storage requirements, we implicitly learn a 3D keypoint descriptor field, avoiding the need to explicitly store point-wise features. To further address the semantic ambiguity of descriptors, we introduce additional semantic contextual feature fields, which enhance the quality and reliability of 2D-3D correspondences. Besides, we propose descriptor similarity distribution alignment to minimize the domain gap between 2D and 3D feature spaces during matching. Finally, we construct the matching graph using both complementary descriptors and contextual features to establish accurate 2D-3D correspondences for 6-DoF pose estimation. Compared with the recent NeRF-based approaches, our method achieves a 3× faster training speed and a 45× reduction in model storage. Extensive experiments on two widely used datasets demonstrate that our approach outperforms or is highly competitive with other state-of-the-art NeRF-based visual localization methods. Project page: https://zju3dv.github.io/neuraloc**

I. INTRODUCTION

Visual localization plays a crucial role in many robotics applications [1], [2], [3], aiming to estimate the 6-Degree-of-Freedom (6-DoF) camera pose of a query image with respect to a pre-built 3D map.

With the advancement of deep learning, convolutional neural networks (CNNs) have shown great potential in extracting high-level contextual features, significantly promoting the development of visual localization. Current visual localization approaches can be broadly categorized into feature-based and regression-based methods. Feature-based methods [4], [5] typically represent the scene using point clouds. These methods rely on keypoint detection and matching techniques [6], [7], [8], [9] to reconstruct a 3D point map. Once 2D-3D correspondences between the query image and the point-cloud map are established, the 6-DoF pose can be estimated via the Perspective-n-Point (PnP) algorithm [10]. However, their performance is limited by the repeatability and discriminative power of the extracted or learned keypoint descriptors. Furthermore, feature-based methods require substantial storage to retain map information, such as 3D points, descriptors, and covisibility relationships, which can be storage-intensive. On the other hand, regression-based approaches [11], [12], [13] employ CNNs to directly regress the pose from features extracted from a single image or to estimate the 3D coordinates of pixels in the camera's view. While these methods avoid the need for storing point-wise descriptors and demand less storage space, their generalization performance in large, complex scenarios is relatively poor. Additionally, regression-based methods typically require large amounts of 3D data to optimize the regression network. As a result, while these methods are more storage-efficient, they are generally inferior to feature-based methods in terms of localization accuracy and generalization performance.

Recently, NeRF [14] has emerged as a new paradigm for implicit scene representation. The neural implicit representation uses multi-layer-perceptions (MLPs) and additional parametric encoding [15], [16], [17], [18], [19] to model the scene property and achieve impressive results in novel view synthesis and surface reconstruction. Benefiting from differentiable volume rendering [20], NeRF-based approaches enable end-to-end parameter optimization without the need of 3D supervision. Several recent works [3], [21], [22], [23], [24] have explored using neural implicit representations to support 6-DoF pose estimation and refinement. For example, iNeRF [21] was the first method to use a pre-trained NeRF model for estimating the pose of an unknown query image. However, it is limited to object-level localization and requires an initial pose estimate. In visual localization, LENS [25] and NeRF-SCR [23] are regression-based approaches that use NeRF to synthesize virtual camera views to aid scene coordinate regression networks. However, similar to other regression-based methods, these approaches suffer from unsatisfactory localization results due to the lack of geometric constraints. To introduce geometric constraints, PNeRFLoc [22] uses point-based neural representation [26] to represent the scene map and performs 2D-3D feature matching for pose estimation. This was the first work to combine feature-based methods with NeRF-based representations for visual localization. While PNeRFLoc achieves satisfactory results across various scenes, it requires explicit storage of features for each point, leading to significant storage requirements for the reconstructed scene model.

To address the issues mentioned above, we propose an efficient and novel visual localization approach, NeuraLoc, based on a learned neural implicit map. Specifically, to

* Corresponding author: Guofeng Zhang (zhangguofeng@zju.edu.cn)
This work was partially supported by the NSF of China (No.6242500063).

2025 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3315-4139-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICRA55743.2025.11127718

Authorized licensed use limited to: Zhejiang University. Downloaded on September 21,2025 at 12:01:10 UTC from IEEE Xplore. Restrictions apply.

reduce the size of the scene model, we avoid explicitly storing point cloud descriptors. Instead, we implicitly learn a descriptor field from an MLP decoder, distilled from a 2D keypoint detection model [6]. Additionally, to address the semantic ambiguity of the learned descriptors and filter outliers, we distill a semantic contextual feature field, which helps construct the accurate matching graph. Furthermore, to reduce the domain gap between 2D and 3D descriptor feature spaces, we introduce a similarity alignment loss to align 2D-3D and 2D-2D similarity distributions. Finally, we construct a matching graph using the dual complementary features to establish 2D-3D correspondences for 6-DoF pose estimation.

In summary, the major contributions of our proposed approach are summarized as follows:

- We propose an efficient visual localization approach based on a reconstructed neural implicit map, achieving accurate localization with fewer scene model parameters.
- To maintain localization performance with a compact scene model, we introduce dual complementary descriptors and semantic contextual features distilled from 2D foundation models for accurate 6-DoF pose estimation.
- We reduce the domain gap between 2D and 3D feature spaces during matching by applying alignment loss on descriptor similarity distributions.

II. RELATED WORKS

A. Visual Localization

Visual localization approaches can be broadly divided into regression-based [11], [12], [27], and feature-based methods [4], [22], [28]. Regression-based approaches often rely on CNNs to directly regress either the camera pose [11] or the scene coordinates [29]. One of the key limitations of these methods is their heavy dependence on large amounts of 3D ground truth data for supervision. Additionally, they generally lack interpretability, which limits their performance in terms of both localization accuracy and generalization. On the other hand, feature-based methods [4], [5], which are the most commonly used for visual localization, utilize sparse or dense feature matching to estimate camera poses by establishing 2D-3D correspondences. Thanks to advanced keypoint detection [6], [8] and matching techniques [7], [30], feature-based approaches can reliably compute 2D-3D correspondences between query images and a prebuilt 3D map. For pose estimation, these methods typically apply Perspective-n-Point [10] and RANSAC [31] algorithms to register query images. Regression-based methods tend to have fewer parameters but suffer from lower generalization ability and localization accuracy compared to feature-based methods.

B. Neural Radiance Fileds

Recent advances in implicit representations have shown great potential in various 3D computer vision tasks, such as surface reconstruction [32], [33], [34], and neural implicit SLAM [2], [35], [36], [37], [38], [39]. For example, NeuS [32] employs a signed distance function (SDF) to represent object geometry and optimizes it using volumetric rendering techniques [20] with an unbiased weighting function. Similarly, iMAP [35] is the first neural SLAM approach to use MLPs to represent the scene, performing bundle adjustments to optimize the tracked RGB-D camera trajectory. However, MLPs have limited representation capacity when applied to large-scale reconstructions, and additional parametric encoding features are proposed, such as voxel [17], [40], point [26], and tri-plane [18]. To extend implicit representations to outdoor scenes, Tao recently proposed SILVR [41], the first neural Lidar SLAM system, capable of handling large-scale, outdoor environments.

C. NeRF-based Visual Localization

Recently, several approaches [3], [22], [23], [24], [25], [42], [43] have explored the use of differentiable representation for pose estimation and visual localization, benefiting from their high-quality novel view synthesis capabilities. For instance, iNeRF [21] was the first to estimate the camera pose of a query image by leveraging the differentiability of a pre-trained NeRF. However, iNeRF requires a coarse initial camera pose and is primarily limited to object-level pose estimation. LENS [25] employs NeRF-W [44] to reconstruct scenes and synthesize virtual camera views, aiding the training of a scene coordinate network. To address the need for geometric constraints, methods such as [3], [22] utilize robust keypoint detection [8] and matching networks [30] to improve pose estimation. Zhao introduces PNeRFLoc [22], which adapts point-based neural representations to facilitate both feature matching and rendering, further enhancing the localization process.

III. METHOD

Given posed RGB-D images $\{I_i \in \mathbb{R}^3, D_i \in \mathbb{R}, \}$, we first extract 2D feature maps with powerful vision foundation models [6], [45] for RGB images. Then, we reconstruct the consistent appearance/geometry property and dual complementary descriptor/contextual feature fields for the entire scene (Sec. III-A). With the reconstructed neural implicit semantic map, we can estimate the 6-DoF pose of the query image based on the matching graph (Sec. III-B). The whole reconstruction and localization pipeline is shown in Fig. 1.

A. Reconstruction Process

In this part, we introduce the geometry and semantic reconstruction process of our neural implicit map.

Scene Representation and Volume Rendering. As noted in [35], a single MLP struggles with representing large scenes. Therefore, to improve the MLP's representation capacity, we employ the multi-resolution hash encoding [15] to capture high-frequency information (*e.g.*, color, contextual feature). As illustrated in the top part of Fig. 1, we utilize separate hash parametric encodings for different branches, $\mathcal{T}_{geo} = \{\mathcal{T}^l_{geo}\}^L_{l=1}$ for geometry branch and $\mathcal{T}_{sem} = \{\mathcal{T}^l_{sem}\}^L_{l=1}$ for semantic branch, L is the level of different resolutions. For each sampled point, $p \in \mathbb{R}^3$, along the

Authorized licensed use limited to: Zhejiang University. Downloaded on September 21,2025 at 12:01:10 UTC from IEEE Xplore. Restrictions apply.

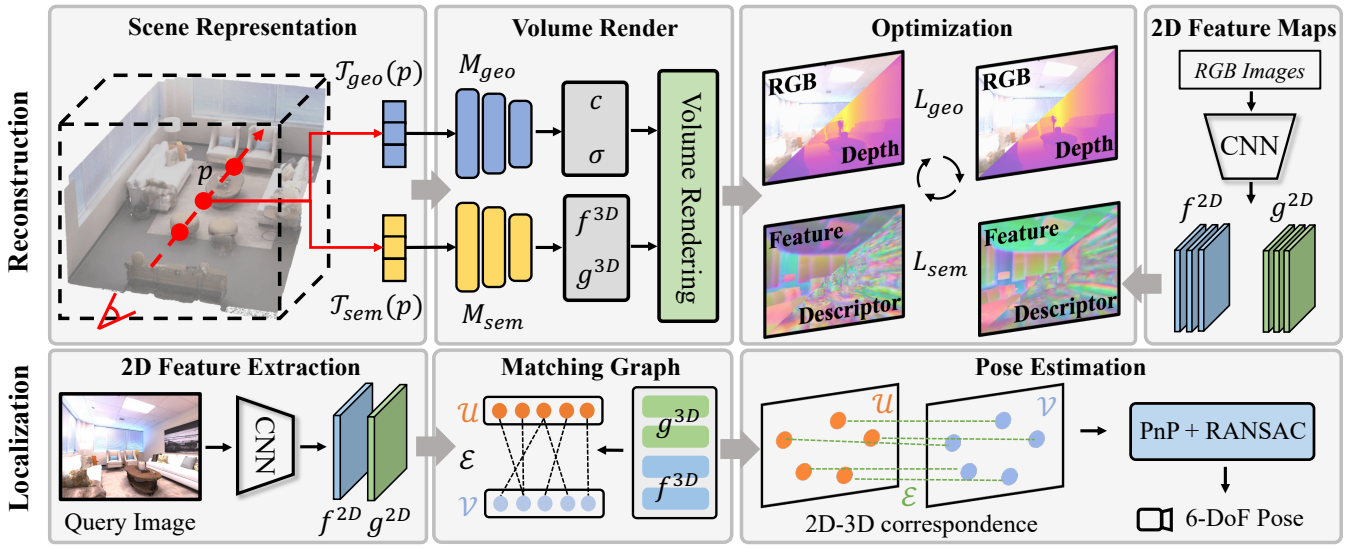

Fig. 1. The whole pipeline of our system. (1) Reconstruction: We employ different parametric encodings ($\mathcal{T}_{geo}$ and $\mathcal{T}_{sem}$) for geometry and semantic branches. Scene properties, including color c, SDF σ, semantic contextual feature f^{3D}, and keypoint descriptor g^{3D} are produced by separated shadow decoders ($\mathcal{M}_{geo}$ and $\mathcal{M}_{sem}$). We use pre-trained CNN models (SuperPoint [6] and SAM [45]) to generate 2D feature maps for the optimization of the semantic branch. (2) Localization: We extract 2D descriptors and semantic contextual features for the query image to build the matching graph between 3D points. Then, we estimate the 6-DoF pose based on the 2D-3D correspondence.

camera ray, its encoding feature is obtained via trilinear interpolation from the multi-resolution feature volume. To decode the scene properties, we adopt two separate shallow MLPs:

$$(c, \sigma), (f^{3D}, g^{3D}) = \mathcal{M}_{geo}(\mathcal{T}_{geo}(p)), \mathcal{M}_{sem}(\mathcal{T}_{sem}(p)), \tag{1}$$

Here, $\mathcal{M}_{geo}$ and $\mathcal{M}_{sem}$ are the MLP decoders for each branch. where c, σ are the color, SDF, and f^{3D}, g^{3D} are the learned 3D semantic contextual feature and descriptors.

Instead of using occupancy or density-based rendering [14], [44], we adopt SDF-based volume rendering [32], [34], [36] for superior scene geometry reconstruction. Following [34], we adopt a simple bell-shaped rendering weight function that can transform the signed distance function into rendering weight:

$$w_i = \text{Sigmoid}(\frac{\sigma_i}{tr}) \cdot \text{Sigmoid}(-\frac{\sigma_i}{tr}), \tag{2}$$

where tr is the truncation distance for supervising σ.

The scene and semantic properties of each ray are then obtained through volume rendering [20] with the following equation:

$$c = \sum_{i=1}^{M} \widetilde{w}_i c_i, \quad d = \sum_{i=1}^{M} \widetilde{w}_i d_i, \tag{3}$$

$$f^{3D} = \sum_{i=1}^{M} \widetilde{w}_i f_i^{3D}, \quad g^{3D} = \sum_{i=1}^{M} \widetilde{w}_i g_i^{3D}, \tag{4}$$

where M is the number of sampled pixels along each ray, and $\widetilde{w}_i$ is the normalized weight, calculated as $\widetilde{w}_i = w_i / \sum_j^M w_j$.

Scene Reconstruction. To reconstruct the appearance and geometry of the scene, we apply the loss functions like [34]: RGB loss (L_c), Depth loss (L_d), SDF loss ($\mathcal{L}_{sdf}$), and free space loss ($\mathcal{L}_{fs}$). Those losses are defined as follows:

$$\mathcal{L}_c = \sum_{r \in R} (c(r) - I(r))^2, \mathcal{L}_d = \sum_{r \in R} (d(r) - D(r))^2 \tag{5}$$

where $\mathcal{L}_c$ and $\mathcal{L}_d$ are performed for each sampled ray r. For more accurate geometry learning, $\mathcal{L}_{fs}$ and $\mathcal{L}_{sdf}$ are used to optimze the signed distance value, σ:

$$\mathcal{L}_{fs} = \sum_{r \in R} \frac{1}{|P_r^{fs}|} \sum_{p \in P_r^{fs}} (\sigma - 1)^2, \tag{6}$$

$$\mathcal{L}_{sdf} = \sum_{r \in R} \frac{1}{|P_r^{tr}|} \sum_{p \in P_r^{tr}} (d(p) + \sigma \cdot tr - D(r))^2, \tag{7}$$

where P_r^{fs} and P_r^{tr} are the point set which outside the truncation region, $|d(p) - D(r)| > tr$, and inside the truncation region, $|d(p) - D(r)| < tr$.

The final reconstruction loss for the geometry branch is expressed as:

$$\mathcal{L}_{geo} = \lambda_c \cdot \mathcal{L}_c + \lambda_d \cdot \mathcal{L}_d + \lambda_{sdf} \cdot \mathcal{L}_{sdf} + \lambda_{fs} \cdot \mathcal{L}_{fs}, \tag{8}$$

where λ_i are the weights for different components, which are set to 0.5, 1.0, 5000, and 10 respectively.

Complementary Feature and Descriptor Distillation. To reduce the storage requirements of keypoint descriptors, we learn a discriminative descriptor field in the semantic branch through distillation from the pre-trained 2D keypoint detection model, SuperPoint [6]. The learned descriptor can be used to estimate 2D-3D correspondence. However, due to the semantic ambiguity of similar structures, a single descriptor is not effective enough for localization, which may lead to incorrect correspondence. To address this, we distill an additional complementary semantic feature field in the

Authorized licensed use limited to: Zhejiang University. Downloaded on September 21,2025 at 12:01:10 UTC from IEEE Xplore. Restrictions apply.

semantic branch, focusing on contextual information around pixels. We use the segmentation foundation model, SAM [45] as the supervision, leveraging its robust capability to identify semantically similar pixels. To learn 3D consistent context semantic features and keypoint descriptor fields, we first extract the 2D feature map of database images through [6], [45]. For each sampled pixel during volume rendering, we obtain its feature in the 2D feature map, f_i^{2D}, g_i^{2D}. We align the 3D feature rendered from neural fields to the feature maps with the following equation:

$$\mathcal{L}_{dis} = \sum_i [1 - cos(f_i^{2D}, f_i^{3D})] + [1 - cos(g_i^{2D}, g_i^{3D})], \quad (9)$$

where $cos(\cdot)$ is the cosine similarity function between two feature vectors, f_i^{3D} and g_i^{3D} are the rendered 3D complementary feature via Eq. (4).

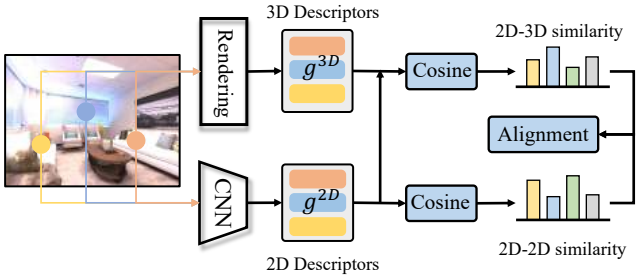

Fig. 2. Descriptor similarity alignment. To reduce the domain gap, we perform the similarity distribution alignment between the 2D-2D and 2D-3D similarity distribution for better optimization.

Descriptor Similarity Distribution Alignment. We utilize the similarity between 2D and 3D descriptors for feature matching and correspondence estimation. However, due to the inherent domain gap between 2D and 3D feature spaces, this can result in incorrect matches. As shown in Fig. 2, we address this issue by imposing alignment constraints on the 2D-3D similarity distributions.

During each iteration, we randomly sample M keypoints from a selected database frame. First, we obtain the L2-normalized 2D and 3D descriptors, g^{2D} and g^{3D}. For the i-th keypoint, we compute its 2D-2D similarity distribution relative to other keypoints in the same image, denoted as $p_i^{2D} = \{cos(g_i^{2D}, g_j^{2D}\}_{j=1}^M$ Similarly, we calculate the 2D-3D similarity distribution for its 3D descriptor, $p_i^{3D} = \{cos(g_i^{3D}, g_j^{2D}\}_{j=1}^M$. Since the relationship between different descriptors is crucial during the matching process, we optimize the 3D descriptor field by aligning the learned 2D-3D descriptor similarity distribution with the original 2D-2D similarity distribution. To reduce the gap, we align the similarity distribution with the following equation:

$$\mathcal{L}_{kl} = \sum_i KL(\texttt{Softmax}(p_i^{2D}) || \texttt{Softmax}(p_i^{3D})), \quad (10)$$

where $\texttt{KL}(\cdot)$ is the Kullback-Leibler divergence between two descriptor similarity distributions. Following the approach [46], we apply temperature scaling to focus the optimization on positive matches.

Finally, the reconstruction loss for the semantic branch is formulated as:

$$\mathcal{L}_{sem} = \lambda_{dis} \cdot \mathcal{L}_{dis} + \lambda_{kl} \cdot \mathcal{L}_{kl}, \quad (11)$$

where λ_i are the weights for different components, which are set to 1.0 and 0.01, respectively.

B. Localization Process

In this part, we show the process of localization for query images. The whole process is shown at the bottom of Fig. 1.
Problem Formulation. To obtain the 6-DoF pose of the query image, we need to establish 2D-3D correspondence between the query image and the 3D neural implicit map. So, this localization problem can be modeled as constructing and solving a bipartite graph matching problem. The bipartite graph is denoted as $\mathcal{G} = \{\mathcal{U}, \mathcal{V}, \mathcal{E}\}$, where $\mathcal{U} = \{u_1, u_2, \cdots, u_m\} \in \mathbb{R}^{m\times 2}$ is the 2D keypoints in the query image, and $\mathcal{V} = \{v_1, v_2, \cdots, v_n\} \in \mathbb{R}^{n\times 3}$ is the 3D points in the implicit map. The edge set $\mathcal{E} = \{e_{i,j}\} \in \mathbb{R}^{m\times n}$, where $e_{i,j} = e(u_i, v_j)$ represent the probability that u_i and v_j are correspond. The goal of the localization process is to solve the bipartite graph matching problem and obtain the assignment matrix, $\mathcal{S}$, which represents the true correspondences.
Matching Graph Construction. To construct the graph, we first extract the 2D keypoints $\mathcal{U}$, 2D semantic contextual features $\{f_i^{2D}\}$, and descriptors $\{g_i^{2D}\}$ for each query image using [6], [45]. For the 3D points $\mathcal{V}$, we consider those within the frustum of the retrieved reference image [47] to build the matching graph $\mathcal{G}$. Their 3D semantic contextual feature $\{f_i^{3D}\}$, and descriptor $\{g_i^{3D}\}$ via volume rendering (Eq. (4)). Since there is a domain gap between 2D and 3D feature spaces, directly matching keypoints with descriptors can often lead to incorrect results. To address this, we incorporate the semantic contextual features and compute the 2D-3D matching scores using the following equation:

$$e_{ij} = \cos(g_i^{2D}, g_i^{3D}) + \alpha \cdot \cos(f_i^{2D}, f_j^{3D}). \quad (12)$$

Additionally, to reduce the number of candidate matches in the graph, we filter out matches whose semantic contextual feature similarity falls below a specified threshold.
Pose Estimation. Inspired by [5], [48], after constructing the graph $\mathcal{G} = \{\mathcal{U}, \mathcal{V}, \mathcal{E}\}$, we apply the Hungarian algorithm [49] to solve the bipartite matching problem and obtain the final matched correspondences.

$$\mathcal{S} = \texttt{Hungarian}(\mathcal{G} = \{\mathcal{U}, \mathcal{V}, \mathcal{E}\}), \quad (13)$$

where $\mathcal{S} = \{s_1, s_2, \cdots, s_k\}$ is the assignment matrix, with $s_i \in \{0, 1\}$ represents whether the i-th edge is contained in the maximum-weight matching. Using the estimated 2D-3D correspondences, we apply the RANSAC and PnP algorithms [31], [10] to estimate the 6-DoF pose of the query image.

IV. EXPERIMENTS

A. Dataset, Baselines, and Evaluation Protocol

We evaluate our approach on two commonly used datasets: Replica [51] and 12-Scenes [52]. The Replica dataset, which

Authorized licensed use limited to: Zhejiang University. Downloaded on September 21,2025 at 12:01:10 UTC from IEEE Xplore. Restrictions apply.

TABLE I

VISUAL LOCALIZATION RESULTS ON REPLICA DATASET. WE REPORT MEDIAN TRANSLATION AND ROTATION ERRORS (CM, DEGREE).

Method	Room 0	Room 1	Room 2	Office 0	Office 1	Office 2	Office 3	Office 4
SCRNet [13]	2.05 / 0.33	1.84 / 0.34	1.31 / 0.26	1.69 / 0.34	2.10 / 0.52	2.21 / 0.41	2.13 / 0.37	2.25 / 0.43
SCRNet-ID [50]	2.33 / 0.28	1.83 / 0.35	1.78 / 0.29	1.79 / 0.37	1.65 / 0.42	2.07 / 0.37	1.79 / 0.28	2.42 / 0.35
SRC [29]	2.78 / 0.54	1.92 / 0.35	2.97 / 0.63	1.45 / 0.30	2.07 / 0.53	2.53 / 0.51	3.44 / 0.63	4.84 / 0.90
NeRF-SCR [23]	1.53 / 0.24	1.96 / 0.31	1.34 / 0.22	1.61 / 0.35	1.54 / 0.44	1.69 / 0.33	2.40 / 0.38	1.69 / 0.32
PNeRFLoc [22]	1.00 / 0.21	1.32 / 0.28	1.43 / 0.29	0.72 / 0.15	1.08 / 0.28	1.71 / 0.37	2.39 / 0.30	1.63 / 0.32
Ours	**0.51 / 0.08**	**1.06 / 0.20**	**1.11 / 0.22**	**0.39 / 0.08**	**0.82 / 0.21**	**1.18 / 0.22**	**1.32 / 0.21**	**1.05 / 0.17**

TABLE II

VISUAL LOCALIZATION RESULTS ON 12-SCENES DATASET. WE REPORT MEDIAN TRANSLATION AND ROTATION ERRORS (CM, DEGREE).

Scenes	Apartment 1		Apartment 2				Office 1				Office 2	
Method	kitchen	living	bed	kitchen	living	luke	gates362	gates381	lounge	manolis	5a	5b
SCRNet [13]	2.3 / 1.3	2.4 / 0.8	3.3 / 1.5	2.1 / 1.0	4.2 / 1.8	4.4 / 1.4	2.6 / 0.8	3.4 / 1.4	2.7 / 0.9	1.8 / 1.0	3.6 / 1.5	3.4 / 1.2
SCRNet-ID [50]	2.6 / 1.4	2.0 / 0.8	2.0 / 0.8	1.8 / 0.9	3.0 / 1.2	3.7 / 1.3	2.1 / 1.0	2.9 / 1.2	3.4 / 1.1	2.6 / 1.2	3.3 / 1.2	3.8 / 1.3
NeRF-SCR [23]	**0.9 / 0.5**	2.1 / 0.6	1.6 / 0.7	1.2 / 0.5	2.0 / 0.8	2.6 / 1.0	2.0 / 0.8	2.7 / 1.2	1.8 / **0.6**	1.6 / 0.7	2.5 / 0.9	2.6 / 0.8
PNeRFLoc [22]	1.0 / 0.6	1.5 / 0.5	**1.2 / 0.5**	**0.8 / 0.4**	1.4 / **0.5**	8.1 / 3.3	1.6 / 0.7	8.7 / 3.2	2.3 / 0.8	1.1 / **0.5**	X	2.8 / 0.9
Ours	**0.9 / 0.5**	**1.1 / 0.4**	1.3 / 0.6	1.0 / 0.6	**1.2 / 0.5**	**1.4 / 0.7**	**1.1 / 0.5**	**1.1 / 0.5**	**1.7 / 0.6**	**1.0 / 0.5**	**1.3 / 0.6**	**1.5 / 0.5**

contains high-quality RGB-D sequences, is widely used in recent NeRF-based works. Following the setup in [23], we use 8 scenes provided by [53] for evaluation. For each scene, the first sequence is used for the training set, and the second sequence is used for testing. For the 12-Scenes dataset, we follow the common setting [52], where the first sequence is used for testing, and the remaining sequences are used for training. Instead of using all training images, we sample one frame every five frames as the training data.

We compared our approach with recent regression-based and NeRF-based approaches, including SCRNet [13], SCRNet-ID [50], SRC [29], NeRF-SCR [23], and PNeRFLoc [22]. To measure localization accuracy, we use the commonly adopted relative rotation error and relative translation error as metrics:

$$\Delta R = \arccos((\mathrm{Tr}(R^T \hat{R}) - 1)/2), \quad \Delta t = ||t - \hat{t}||_2, \quad (14)$$

where $\hat{t}$ and $\hat{R}$ are the ground-truth translation and rotation, respectively, and t and R are the estimated ones.

B. Implementation Details

For each scene, we train the scene geometry, descriptor, and semantic feature branches over 10,000 iterations. The multi-resolution hash encodings consist of 16 levels of detail, with each level containing a 2-dimensional feature vector, and the finest resolution is set to 2 cm. In the geometry branch, both the geometry and appearance decoders are 2-layer MLPs with 32 hidden units each. In the semantic branch, we use two separate 2-layer MLPs with 128 hidden units for learning descriptors and contextual features. For SAM [45] and SuperPoint [6], we use the 'vit_h' model and 'inloc' configuration, respectively, resulting in 2D feature maps with 256 output dimensions.

C. Localization Results

In this part, we evaluate the localization performance of our proposed approach, presenting both quantitative and qualitative results.

Quantitative and Qualitative Results. We compare our approach with other baselines on Replica and 12-Scenes datasets. Quantitative results for both datasets are shown in Tab. I and Tab. II, respectively. The best localization results are highlighted in bold. For the Replica dataset, our approach achieves state-of-the-art performance across all scenes. Since the dataset contains high-quality depth data, our SDF-based representation models the geometric information of the scene more accurately. For the 12-Scenes dataset, our approach achieves the best localization results on 10 scenes and the second-best on the remaining two. Real-world scenes typically feature complex lighting and material conditions. While PNeRFLoc [22] failed (denoted as "X") on challenging scenes like "Ofiice2/5a", our approach performs visual localization robustly.

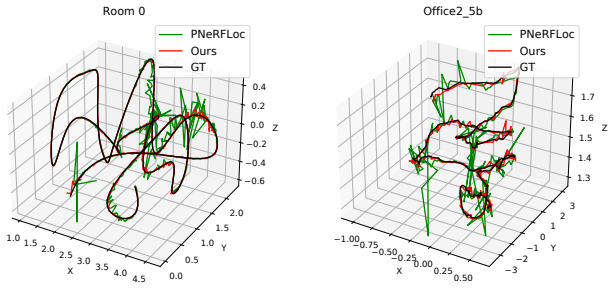

Fig. 3. Trajectory visualization of two selected scenes.

We present the estimated camera trajectory of two select scenes ("Room 0" from Replica and "Office2/5b" from 12-Scenes) in Fig. 3. Compared to PNeRFLoc [22], our method demonstrates more stable pose estimation with fewer outliers. Additionally, visual matching results for both datasets are shown in Fig. 4. For visualization, we project the 3D map points into the reference image. As illustrated, our approach

Authorized licensed use limited to: Zhejiang University. Downloaded on September 21,2025 at 12:01:10 UTC from IEEE Xplore. Restrictions apply.

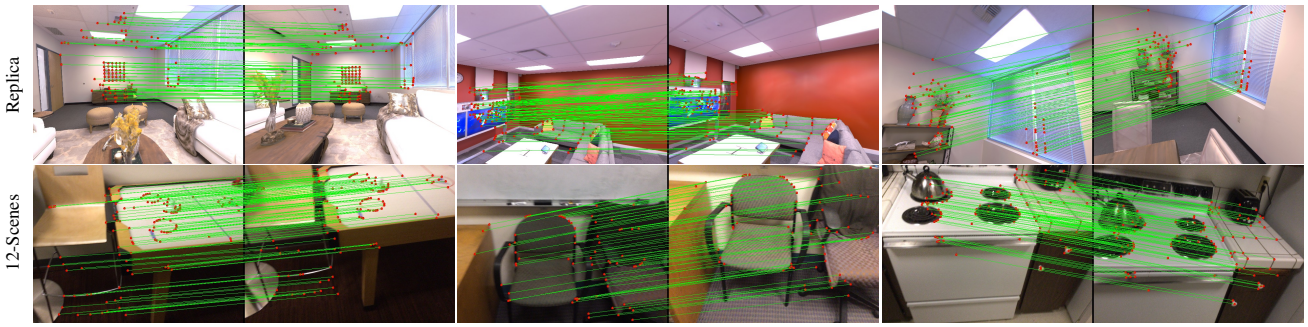

Fig. 4. Qualitative results of feature matching. We show some matching results of our method on Replica [51] and 12-Scenes [52] dataset.

accurately estimates correspondences despite changes in viewpoints.

TABLE III
TRAINING TIME AND MEMORY USAGE OF DIFFERENT METHODS.

Method	SCRNet	NeRF-SCR	PNeRFLoc	Ours
Training Time	2 days	16 hours	1 hours	20 mins
Memory	165 MB	-	788 MB	15.48 MB

Training Time and Memory Usage. In Tab. III, we show the training time and memory usage of different methods in scene "Office2/manolis". Both PNeRFLoc and our approach were tested on an AMD Ryzen 9 7950X 16-core CPU and an RTX 4090 24GB GPU, while the results of other methods are taken from their respective papers. As shown, regression-based approaches generally require fewer parameters but take longer to train and are less effective for localization compared to feature-based methods. Our method strikes a balance, achieving better localization performance with fewer parameters and faster convergence.

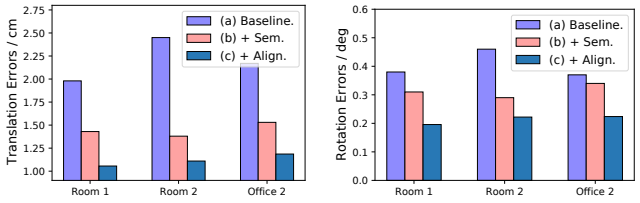

Fig. 5. Median localization errors (cm, degree) of using different features.

D. Ablation Studies

In this part, we perform ablation studies to investigate the effect of components and designs in our system.

Effects of Different Features and Alignment. Here, we show the localization performance using different features and alignment loss for localization. (a) Directly matching learned descriptors with nearest neighbor search. (b) Using semantic contextual features to construct the matching graph. (c) Using similarity alignment loss to reduce the gap between 2D and 3D feature space. The comparison results of different settings are shown in Fig. 5. As shown, incorporating semantic information in the matching graph effectively filters out keypoints with semantic ambiguity during feature matching. Additionally, using similarity alignment constraints further minimizes the domain gap between the 2D and 3D descriptor feature spaces.

TABLE IV
MEDIAN LOCALIZATION ERRORS (CM, DEGREE).

Case	Description	Room 0	Office2/5b
#1	w/o Separate Encoding	5.34 / 2.77	3.6 / 2.1
#2	w/o Match Candidate Filter	0.78 / 0.20	1.8 / 0.7
#3	w/o Graph Match	1.02 / 0.69	2.9 / 1.1
#4	Ours Full	**0.51 / 0.08**	**1.5 / 0.5**

Effects of Design Choices. In Tab. IV, we show the localization performance of various design choices. The results in the table have validated the effectiveness of our system's design. #1 shows that without additional parametric encoding for the semantic branch, it can lead to significant performance degradation. This is due to that a single MLP can not fit the high-frequency feature very well for large indoor scenes. #2 shows that using semantic contextual features to filter outlier matches can lead to accurate 2D-3D correspondence estimation. #3 shows that using descriptor and semantic contextual features to construct the matching graph in Sec. III-B is also effective.

V. CONCLUSIONS

In this paper, we presented an efficient and novel visual localization approach based on our reconstructed neural implicit map. Specifically, to enforce geometric constraints while minimizing storage requirements, we learn a 3D descriptor field instead of storing descriptors for individual 3D points. Additionally, we learned a complementary semantic contextual feature field for more robust matching graph reconstruction. Besides, to reduce the domain gap between the 2D and 3D feature spaces, we employed similarity distribution alignment, which enhances the estimation of 2D-3D correspondences. Currently, a key limitation of our approach is its inability to scale for large-scale scene localization. Despite this, our approach offers an efficient and novel solution to visual localization, paving the way for future research and improvements in this area.

Authorized licensed use limited to: Zhejiang University. Downloaded on September 21,2025 at 12:01:10 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] Y. Ming, X. Yang, W. Wang, Z. Chen, J. Feng, Y. Xing, and G. Zhang, "Benchmarking neural radiance fields for autonomous robots: An overview," *Eng. Appl. Artif. Intell.*, vol. 140, p. 109685, 2025.

[2] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "NICE-SLAM: neural implicit scalable encoding for SLAM," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12786–12796, 2022.

[3] J. Liu, Q. Nie, Y. Liu, and C. Wang, "NeRF-Loc: Visual localization with conditional neural radiance field," in *IEEE International Conference on Robotics and Automation*, pp. 9385–9392, 2023.

[4] P. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From Coarse to Fine: Robust hierarchical localization at large scale," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12716–12725, 2019.

[5] H. Yu, W. Ye, Y. Feng, H. Bao, and G. Zhang, "Learning bipartite graph matching for robust visual localization," in *IEEE International Symposium on Mixed and Augmented Reality*, pp. 146–155, 2020.

[6] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

[7] P. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4937–4946, 2020.

[8] J. Revaud, C. R. de Souza, M. Humenberger, and P. Weinzaepfel, "R2D2: reliable and repeatable detector and descriptor," in *Advances in Neural Information Processing Systems*, pp. 12405–12415, 2019.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate $O(n)$ solution to the pnp problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.

[11] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2616–2625, 2018.

[12] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-dof camera relocalization," in *IEEE International Conference on Computer Vision*, pp. 2938–2946, 2015.

[13] X. Li, S. Wang, Y. Zhao, J. Verbeek, and J. Kannala, "Hierarchical scene coordinate classification and regression for visual localization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11980–11989, 2020.

[14] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision*, 2020.

[15] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, 2022.

[16] J. Kulhánek and T. Sattler, "Tetra-NeRF: Representing neural radiance fields using tetrahedra," in *IEEE/CVF International Conference on Computer Vision*, pp. 18412–18423, 2023.

[17] H. Li, X. Yang, H. Zhai, Y. Liu, H. Bao, and G. Zhang, "Vox-Surf: Voxel-based implicit surface representation," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–12, 2022.

[18] M. M. Johari, C. Carta, and F. Fleuret, "ESLAM: Efficient dense slam system based on hybrid representation of signed distance fields," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[19] H. Li, H. Zhai, X. Yang, Z. Wu, Y. Zheng, H. Wang, J. Wu, H. Bao, and G. Zhang, "ImTooth: Neural implicit tooth for dental augmented reality," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 5, pp. 2837–2846, 2023.

[20] J. T. Kajiya and B. V. Herzen, "Ray tracing volume densities," in *SIGGRAPH*, pp. 165–174, 1984.

[21] Y. Lin, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T. Lin, "iNeRF: Inverting neural radiance fields for pose estimation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1323–1330, IEEE, 2021.

[22] B. Zhao, L. Yang, M. Mao, H. Bao, and Z. Cui, "PNeRFLoc: Visual localization with point-based neural radiance fields," in *Conference on Artificial Intelligence*, pp. 7450–7459, 2024.

[23] L. Chen, W. Chen, R. Wang, and M. Pollefeys, "Leveraging neural radiance fields for uncertainty-aware visual localization," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[24] J. Sun, Y. Xu, M. Ding, H. Yi, C. Wang, J. Wang, L. Zhang, and M. Schwager, "NeRF-Loc: Transformer-based object localization within neural radiance fields," *IEEE Robotics Autom. Lett.*, vol. 8, no. 8, pp. 5244–5250, 2023.

[25] A. Moreau, N. Piasco, D. Tsishkou, B. Stanciulescu, and A. de La Fortelle, "LENS: localization enhanced by nerf synthesis," in *Conference on Robot Learning*, vol. 164, pp. 1347–1356, 2021.

[26] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann, "Point-NeRF: Point-based neural radiance fields," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5438–5448, 2022.

[27] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "DSAC - differentiable RANSAC for camera localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2492–2500, 2017.

[28] H. Li, T. Fan, H. Zhai, Z. Cui, H. Bao, and G. Zhang, "BDLoc: Global localization from 2.5D building map," in *IEEE International Symposium on Mixed and Augmented Reality*, pp. 80–89, 2021.

[29] S. Dong, S. Wang, Y. Zhuang, J. Kannala, M. Pollefeys, and B. Chen, "Visual localization via few-shot scene region classification," in *International Conference on 3D Vision*, pp. 393–402, 2022.

[30] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021.

[31] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[32] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," in *Annual Conference on Neural Information Processing Systems 2021*, pp. 27171–27183, 2021.

[33] M. Oechsle, S. Peng, and A. Geiger, "UNISURF: unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *IEEE/CVF International Conference on Computer Vision*, pp. 5569–5579, 2021.

[34] D. Azinović, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies, "Neural RGB-D surface reconstruction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6290–6301, 2022.

[35] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "iMAP: Implicit mapping and positioning in real-time," in *IEEE/CVF International Conference on Computer Vision*, pp. 6209–6218, IEEE, 2021.

[36] X. Yang, H. Li, H. Zhai, Y. Ming, Y. Liu, and G. Zhang, "Vox-Fusion: Dense tracking and mapping with voxel-based neural implicit representation," in *IEEE International Symposium on Mixed and Augmented Reality*, pp. 499–507, 2022.

[37] H. Wang, J. Wang, and L. Agapito, "Co-SLAM: Joint coordinate and sparse parametric encodings for neural real-time SLAM," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13293–13302, 2023.

[38] H. Zhai, G. Huang, Q. Hu, G. Li, H. Bao, and G. Zhang, "NIS-SLAM: Neural implicit semantic RGB-D SLAM for 3D consistent scene understanding," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 11, pp. 7129–7139, 2024.

[39] H. Zhai, H. Li, X. Yang, G. Huang, Y. Ming, H. Bao, and G. Zhang, "Vox-fusion++: Voxel-based neural implicit dense tracking and mapping with multi-maps," *arXiv preprint arXiv:2403.12536*, 2024.

[40] L. Liu, J. Gu, K. Z. Lin, T. Chua, and C. Theobalt, "Neural sparse voxel fields," in *Annual Conference on Neural Information Processing Systems*, 2020.

[41] Y. Tao, Y. Bhalgat, L. F. T. Fu, M. Mattamala, N. Chebrolu, and M. F. Fallon, "SiLVR: Scalable lidar-visual reconstruction with neural radiance fields for robotic inspection," in *IEEE International Conference on Robotics and Automation*, pp. 17983–17989, 2024.

[42] D. Maggio, M. Abate, J. Shi, C. Mario, and L. Carlone, "Loc-NeRF: Monte carlo localization using neural radiance fields," in *IEEE International Conference on Robotics and Automation*, pp. 4018–4025, 2023.

[43] H. Zhai, X. Zhang, Z. Boming, H. Li, Y. He, Z. Cui, H. Bao, and G. Zhang, "SplatLoc: 3D Gaussian splatting-based visual localization for augmented reality," *arXiv preprint arXiv:2409.14067*, 2024.

Authorized licensed use limited to: Zhejiang University. Downloaded on September 21,2025 at 12:01:10 UTC from IEEE Xplore. Restrictions apply.

[44] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the wild: Neural radiance fields for unconstrained photo collections," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7210–7219, 2021.
[45] 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," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.
[46] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.
[47] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297–5307, 2016.
[48] Y. Liu, M. Zhu, H. Li, H. Chen, X. Wang, and C. Shen, "Matcher: Segment anything with one shot using all-purpose feature matching," in *International Conference on Learning Representations*, 2024.
[49] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
[50] T. Ng, A. Lopez-Rodriguez, V. Balntas, and K. Mikolajczyk, "Reassessing the limitations of cnn methods for camera pose regression," in *International Conference on 3D Vision*, 2021.
[51] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. D. Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The Replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.
[52] J. P. C. Valentin, A. Dai, M. Nießner, P. Kohli, P. H. S. Torr, S. Izadi, and C. Keskin, "Learning to navigate the energy landscape," in *International Conference on 3D Vision*, pp. 323–332, 2016.
[53] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation," in *IEEE/CVF International Conference on Computer Vision*, pp. 15838–15847, 2021.

Authorized licensed use limited to: Zhejiang University. Downloaded on September 21,2025 at 12:01:10 UTC from IEEE Xplore. Restrictions apply.