2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

DDS-SLAM: Dense Semantic Neural SLAM for Deformable Endoscopic Scenes

Jiwei Shan[1*] Yirui Li[2*] Lujia Yang[2] Qiyu Feng[1] Lijun Han[1] Hesheng Wang[1]

***Abstract*— Estimating camera motion and continuously reconstructing dense scenes in deformable environments presents a complex and open challenge. Many existing approaches tend to rely on assumptions about the scene's topology or the nature of deformable motion. However, these assumptions do not hold true in medical endoscopy applications. To address these challenges, we introduce DDS-SLAM, a novel dense deformable semantic neural SLAM that achieves accurate camera tracking, continuous dense scene reconstruction, and high-quality image rendering in deformable scenes. First, we propose a novel hybrid neural scene representation method capable of capturing both natural and artificial deformations. Additionally, by leveraging the 2D semantic information of the scene, we introduce a semantic loss function based on semantic distance fields. This approach guides network optimization at a higher level, thereby enhancing system performance. Furthermore, we validate our method through a series of experiments conducted on several representative medical datasets, demonstrating its superiority over other state-of-the-art approaches. The code is available at: https://github.com/IRMVLab/DDS-SLAM.**

I. INTRODUCTION

Accurately estimating the camera's pose and reconstructing dense anatomical tissue from endoscopic videos are vital tasks in medical applications. For instance, in minimally invasive surgery, surgeons face challenges due to the endoscope's limited view and restricted movement [1], which hinder the surgeon's ability to observe and comprehend the target tissue, thereby affecting the safety and success of the procedure. Precise camera pose estimation allows for identifying the spatial relationships between surgical tools and critical tissue structures [2]. This enables surgeons to avoid vital structures such as nerves, significantly minimizing surgical risks. Furthermore, dense reconstructions enhance the surgeon's understanding of the anatomy and provide precise scene geometry for tasks like intra-operative and preoperative registration [2]. Dense visual SLAM offers a cost-effective and efficient solution to meet these needs.

Visual SLAM fundamentally assumes a rigid environment, which is insufficient in medical endoscopy because the environment deforms over time. Although recent research has significantly advanced the accuracy of visual SLAM in deformable scenes [3]–[6], several key issues remain unaddressed. First, there is the challenge of finding a reasonable and universally applicable way to represent deformable scenes. Some studies employ triangular meshes to model deformed surfaces [5], but they rely on the assumption that the scene's surface has a planar topology, which is unsuitable for tubular structures, such as colons. Other methods use sparse point clouds [3], which do not adequately capture the scene's geometry in detail [7], [8]. Second, many existing methods presume specific types of scene motion, such as isometric or quasi-isometric deformation. However, these assumptions often do not hold in real-world endoscopic procedures, particularly when surgical instruments interact with soft tissues. Lastly, challenges such as lighting variations and weak textures in endoscopic scenes significantly degrade the performance of the visual SLAM data association module, affecting the algorithm's accuracy and robustness.

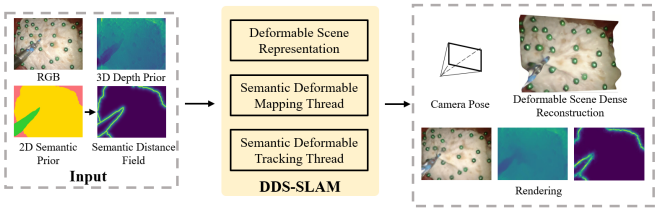

Fig. 1. A demonstration of DDS-SLAM. Our method processes RGB images, 3D depth priors, and semantic distance fields derived from 2D semantic segmentation images as inputs. Through DDS-SLAM, we can accurately track camera poses, perform dense reconstruction of deformed scenes, and provide high-quality rendering results.

In this paper, we introduce DDS-SLAM, a **d**ense, **d**eformable-aware, **s**emantic neural SLAM designed for endoscopic videos that addresses above challenges. First, for deformable scenes, we propose a novel hybrid neural scene representation. It combines an explicit hash grid to model the 3D canonical space with an implicit multi-layer perceptron for the 4D deformation field. Unlike traditional methods, DDS-SLAM does not rely on assumptions regarding the scene's topology or the nature of its deformation. This flexibility allows it to capture various types of deformations, from natural physiological movements, such as breathing and heartbeat, to artificial alterations caused by surgical instruments interacting with soft tissues. DDS-SLAM provides a universal representation of deformable scenes and delivers continuous and detailed reconstructions by leveraging dense geometric priors. Furthermore, DDS-SLAM constructs losses through differentiable rendering, eliminating the reliance on explicit data association. To enhance system performance

*The first two authors contributed equally. This work was supported in part by the Natural Science Foundation of China under Grant 62361166632, 62225309, 62073222, and U21A20480; was partially supported by a grant from the NSFC/RGC Joint Research Scheme sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China and the National Natural Science Foundation of China (Project No. N_CUHK410/23). Corresponding Author: Hesheng Wang, Lijun Han.

[1]Department of Automation, Key Laboratory of System Control and Information Processing of Ministry of Education, Key Laboratory of Marine Intelligent Equipment and System of Ministry of Education, Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai Jiao Tong University, Shanghai 200240, China

[2]School of mechanical engineering, Shanghai Jiao Tong University, Shanghai 200240, China

further, we incorporate semantic information through a novel loss based on semantic distance fields. Extensive experiments on datasets featuring both artificial and natural deformations demonstrate that DDS-SLAM surpasses existing state-of-the-art algorithms in camera tracking accuracy and image rendering quality. Our contributions are summarized as follows:

1) We present DDS-SLAM, a dense, deformable-aware, semantic neural SLAM designed for medical endoscopic videos. It achieves high-accuracy pose tracking, continuous dense scene reconstruction, and superior image rendering quality.
2) We introduce a novel hybrid neural representation for deformable scenes, modeling the 3D canonical space and the 4D deformation field with an explicit hash grid and an implicit multi-layer perceptron, respectively.
3) We propose a semantic distance loss to guide the network optimization at a higher level and achieve superior scene reconstruction results.
4) We conduct extensive evaluations on two challenging datasets with various types of deformations to demonstrate the superior performance of our method in comparison to existing ones. Ablation studies further validate the effectiveness of our key contributions.

II. RELATED WORKS

SLAM in Endoscopy. Over the past decades, many SLAM systems specifically developed for medical endoscopy scenarios have emerged. To address the pose estimation inaccuracies caused by weak textures and illumination changes, works utilizing hardware-based approaches [9] or algorithmic solutions [2], [10] have been proposed. Additionally, previous researchers have attempted to introduce semantic information into endoscopic SLAM to identify or eliminate dynamic points and enhance data association [11]. However, visual SLAM relying on rigid assumptions tends to degrade in deformable scenes. While efforts have been made to overcome this issue [3]–[5], a universally applicable representation of deformable scenes has yet to be established. In this study, we propose a versatile deformable scene representation method capable of addressing both artificial and natural deformations in endoscopic scenes.

Neural Implicit SLAM. Neural implicit representation is an emerging method for 3D scene modeling that utilizes neural networks to capture both geometry and appearance. Initially, iMAP [12] developed an RGB-D SLAM system based on this neural implicit representation. Since then, numerous neural implicit SLAM systems [13]–[16] have emerged, addressing issues such as system forgetfulness, memory usage, and slow optimization. While these developments are mainly tailored for static indoor scenes, the potential of neural implicit representation in medical endoscopic SLAM remains largely untapped [17]. This study introduces a dense deformable semantic neural SLAM, tailored for endoscopic deformable scenarios, offering a novel solution to the challenges faced in endoscopic SLAM.

III. METHODS

In Fig. 2, we illustrate the DDS-SLAM pipeline, which comprises three key components: the deformable semantic tracking thread, the deformable semantic mapping thread, and the deformable scene representation module. Given an input endoscopic image sequence $\{I_i\}_{i=1}^{M}$ with known camera intrinsics K, we perform dense deformable semantic mapping and real-time tracking by jointly optimizing the scene representation and camera poses by leveraging the depth priors $\{D_i\}_{i=1}^{M}$ and semantic informations $\{S_i\}_{i=1}^{M}$. Sec. III-A describes our hybrid neural deformable scene representation approach, explaining how we model deformable scenes using 3D canonical space and a 4D deformation field. Sec. III-B details the rendering process, which transforms the raw outputs of the deformable scene representation into pixel-based colors, depths, and semantics. Sec. III-C introduces the loss functions. Sec. III-D provides the details of the localization and reconstruction of our SLAM system.

A. Hybrid Neural Deformable Scene Representation

In deformable scenes, the camera captures images from only one viewpoint at each specific time instant. To tackle the challenge of observation sparsity, maintaining intrinsic correlations across different time steps and effectively sharing relevant information is crucial. Inspired by recent advancements in dynamic neural radiance fields [1], [18], we model the deformable scene as a combination of a 3D canonical space and a 4D deformation field. The canonical space encompasses the scene's 3D geometry, appearance, and semantic distance information. The deformation field represents how the canonical space deforms over time.

Deformation Network. The deformation network D_θ establishes a deformation field linking the scene at a specific time to its representation in the canonical space. For any sampling point x in the 3D space at time t, the deformation network D_θ estimates the position offset Δx between x and its corresponding point in canonical space. Without loss of generality, we set the scene at $t = 0$ as the canonical scene. Formally, the deformation network can be expressed as:

$$D_\theta(x, t) = \begin{cases} \Delta x, & \text{if } t \neq 0 \\ 0, & \text{if } t = 0 \end{cases} \tag{1}$$

To enhance the network's detail capture, we transform low-dimensional coordinates x and time t into a high-dimensional frequency domain using position encoding [19]:

$$\gamma(p) = \langle \sin(2^l \pi p), \cos(2^l \pi p) \rangle_0^L. \tag{2}$$

where p is t and each component of x, with L being 10 for x and 4 for t.

Canonical Network. Through the deformation network, scenes at different time steps are interconnected via a common canonical space [18]. This setup facilitates reliable information sharing and mitigates the observation sparsity issue in deformable scenes. The canonical network aims to encode the scene's 3D geometry, appearance, and semantic information in a canonical configuration. NeRF [19] and dynamic NeRF [1], [18] usually use MLP to represent

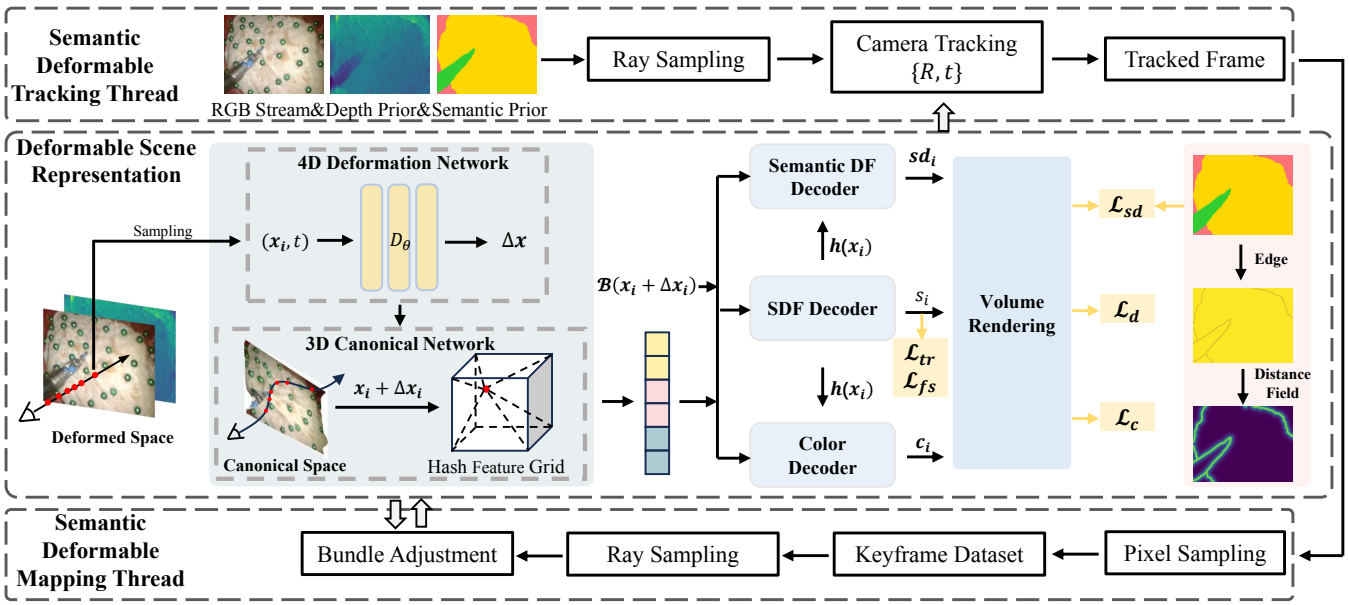

Fig. 2. Overview of DDS-SLAM. It processes a sequence of RGB images from a deformable scene, along with associated depth and semantic priors. DDS-SLAM comprises three main components: 1) Deformable Scene Representation: We adopt a hybrid neural deformable scene representation to represent deformable scenes. Through the 4D implicit deformation field and the 3D explicit hash feature grids, feature vectors of the input 4D points are extracted and mapped to RGB, SDF, and semantic distance values via three decoders. These components are updated online by minimizing our carefully designed loss functions. 2) Semantic Deformable Tracking Thread: This thread focuses on optimizing per-frame camera poses by minimizing losses. 3) Semantic Deformable Mapping Thread: This thread employs global bundle adjustment to jointly optimize the deformable scene representation, including parameters of the 4D implicit deformation field and the 3D canonical network, as well as the poses of keyframes. These two threads run with alternating optimization.

canonical networks, but these methods have problems such as slow convergence. Inspired by [14], [20], we adopt sparse parameter encoding to enhance training efficiency.

Specifically, we utilize L resolution hash feature grids $V_\alpha = \{V_\alpha^l\}_{l=1}^L$ to represent the entire scene. For a 3D sample point x and its viewing direction d, we first identify the eight nearest feature vertices at each resolution, then use trilinear interpolation to query the feature vectors. To ensure system smoothness, we employ One-blob encoding [21] for the input point x. The feature vector $V_\alpha(x)$ and One-blob encoded spatial coordinate $\beta(x)$ are input into the SDF decoder f_ν to predict the TSDF value s and a feature vector z_h:

$$f_\nu\left(\beta(x), V_\alpha(x)\right) \mapsto (z_h, s). \tag{3}$$

The feature vectors z_h and the position-encoded viewing direction $\gamma(d)$ are input into the color decoder f_μ and semantic distance decoder f_τ to predict the color c and semantic distance value m:

$$f_\mu\left(\beta(x), z_h, \gamma(d)\right) \mapsto c, \quad f_\tau\left(\beta(x), z_h, \gamma(d)\right) \mapsto m. \tag{4}$$

Here, $\phi = \{\alpha, \nu, \mu, \tau\}$ represent the learnable parameters associated with the canonical network.

B. Color, Depth, and Semantic Distance Rendering

To establish the relationship between the deformable scene representation and the system's inputs, we employ differentiable rendering [12]–[15] to integrate the predictions from Sec III-A to render pixel color, depth, and semantic distance field value. Specifically, given the camera origin o and the ray direction d, we sample R points along the ray, denoted as $x_i = o + h_i d$. The deformation network and canonical network provide the predicted color, TSDF, and semantic distance value at each sample point. Using the SDF-based rendering technique [22], we convert the TSDF values into volume densities:

$$\sigma(x_i) = \kappa \cdot \text{Sigmoid}(-\kappa \cdot s_i(x_i)). \tag{5}$$

where κ is a learnable parameter. These volume densities $\sigma(x_i)$ are then used to render the associated color, depth, and semantic distance values for each ray:

$$\hat{c} = \sum_{r=1}^{R} w_i c_i \quad \hat{d} = \sum_{r=1}^{R} w_i h_i \quad \hat{m} = \sum_{r=1}^{R} w_i m_i \tag{6}$$

where w_i represents the weight of each sample point:

$$w_i = \exp\left(-\sum_{k=1}^{i-1} \sigma(x_k)\right)\left(1 - \exp\left(-\sigma(x_i)\right)\right) \tag{7}$$

C. Loss Functions

To optimize the parameters of deformable scene representation and camera poses, we design five loss functions.

Color and Depth Loss. We construct the color and depth losses by comparing the differences between the rendered and the ground truth color and depth values:

$$L_c = \frac{1}{|P|} \sum_{p \in P} \left\| C(p) - \hat{C}(p) \right\|_2. \tag{8}$$

$$L_d = \frac{1}{|P|} \sum_{p \in P} \left\| d(p) - \hat{d}(p) \right\|_2. \tag{9}$$

Here, P denotes the set of rays/pixels with valid depth measurements, p represents a sampling ray/pixel, and $C(p)$ and $d(p)$ denote the GT values.

TSDF Free Space and Truncation Region Loss. According to the TSDF definition, in a batch of rays P with valid depth, for sampling points $q \in S_p^{fs}$ far away from the surface (outside the truncation distance), we apply a TSDF free-space loss [14], [15] to ensure the SDF values predicted by the network are equal to the truncation distance T:

$$L_{fs} = \frac{1}{|P|} \sum_{p \in P} \frac{1}{|S_p^{fs}|} \sum_{q \in S_p^{fs}} \left(f_\nu(q) - T \right)^2 . \quad (10)$$

For sampling points $q \in S_p^{tr}$ located within the truncation distance, we follow [14] for the loss function:

$$L_{tr} = \frac{1}{|P|} \sum_{p \in P} \frac{1}{|S_p^{tr}|} \sum_{q \in S_p^{tr}} \left(f_\nu(q) + h(q) - D(p) \right)^2 . \quad (11)$$

where $D(p)$ is the depth value of the ray p obtained from the input depth map, and $h(q)$ represents the depth of the sampling point q relative to the camera's pose.

Semantic Distance Loss. To improve the robustness and precision of our system, we employ semantic information to guide network optimization. Different from existing NeRF-based semantic SLAM systems [23], we utilize the semantic edge distance field to capture local relationships among various semantic categories and guide the network optimization at a higher level. Specifically, we apply the Canny algorithm [24] to identify edges between semantic categories from the input semantic map. We then compute each pixel's distance to the nearest edge, thereby establishing a semantic distance field as the ground truth. We define the semantic distance loss by comparing the discrepancy between the rendered semantic distances and the ground truth:

$$L_m = \frac{1}{|P|} \sum_{p \in P} \| m(p) - \hat{m}(p) \|_2 . \quad (12)$$

The global loss function for our method is formulated as:

$$L = \lambda_c L_c + \lambda_d L_d + \lambda_{fs} L_{fs} + \lambda_{tr} L_{tr} + \lambda_m L_m \quad (13)$$

Here, $\{\lambda_c, \lambda_d, \lambda_{fs}, \lambda_{tr}, \lambda_m\}$ are the respective weighting coefficients. The objective remains consistent for both mapping and tracking phases in our framework.

D. Mapping and Tracking

When the system starts running, the first input frame is used to optimize the deformation and canonical network parameters to accurately represent the first frame. For each subsequent frame, the camera pose is estimated in the tracking thread. Concurrently, for every k frames, we select one keyframe to add to the keyframe database and optimize both all scene parameters and the poses of the keyframes simultaneously. Following [14], we only use a subset of pixels (approximately 5%) to represent each keyframe.

Tracking. For tracking, we calculate the camera-to-world transformation matrix $\mathbf{T}_{wc} = \exp(\xi_t^\wedge) \in SE(3)$ for each new frame, utilizing a constant velocity motion model for initial pose estimation. We then sample Q pixels from the current frame and refine the camera parameters ξ_t by minimizing the global loss function. During this process, we keep all scene parameters unchanged.

Mapping. In the mapping stage, we select Q_g rays randomly from the keyframe dataset. Following [14], we first optimize the scene parameters for k_m steps and then update the camera poses of the chosen keyframes using the accumulated gradient on camera parameters $\{\xi_t\}$. We then sample Q_t pixels from the current image and optimize all scene parameters while the camera pose remains fixed.

IV. EXPERIMENTS

Datasets. We evaluate DDS-SLAM on a variety of scenes from two different datasets. The first is the **Semantic-SuPer Dataset** [11]. This dataset primarily captures artificial deformations caused by surgical tools. It comprises four trials, each containing 150 images at a resolution of 640×480, referred to as Lab 1, 2, 3, and 4 in subsequent sections. The second dataset, **StereoMIS** [6], is recorded with the da Vinci Xi surgical robot. Ground-truth camera poses are generated using the endoscope forward kinematics. We select one open-source sequence, named P2_1, and select 4000 consecutive images from this sequence, at a resolution of 640×512. P2_1 captures the natural deformations from breathing motion.

Metrics. For the StereoMIS dataset [6], we evaluate camera tracking accuracy using the ATE RMSE. For the Semantic-SuPer Dataset [11], where camera pose ground truth is unavailable, we estimate reprojection error using the positioning of green pins, as described in [11]. We also evaluate image rendering quality using Peak Signal-to-Noise Ratio (PSNR), SSIM [26], and LPIPS [27].

Baselines. The baselines fall into two categories. The first category comprises neural SLAM methods, for which we select four state-of-the-art approaches: iMAP [12], NICE-SLAM [13], ESLAM [15], and Co-SLAM [14]. We utilized the iMAP* model in our experiments, which is a reimplementation referenced in [13]. The second category includes SLAM methods tailored for deformable endoscopy scenarios. We include three state-of-the-art algorithms: DefSLAM [25], Semantic-SuPer [11], and RECP [6].

Implementation Details. We run DDS-SLAM on a desktop PC equipped with an NVIDIA RTX 2080 GPU. In the mapping thread, we set the keyframe frequency to $k = 5$ with 10 iterations for the long-sequence StereoMIS dataset, and to $k = 1$ with 200 iterations for the short-sequence Semantic-SuPer dataset. In the tracking thread, we perform 10 iterations for the StereoMIS dataset and 20 iterations for the Semantic-SuPer dataset. We sample $Q = 1024$ pixels for tracking, $Q_g = 2048$ pixels globally, and $Q_t = 2048$ pixels locally for mapping. On each ray, we uniformly sample 32 points and an additional 16 points near the estimated depth. All experiments were conducted adhering to the default parameter settings: $\lambda_c = 5$, $\lambda_d = 0.1$, $\lambda_{fs} = 10$, $\lambda_{tr} = 1000$, $\lambda_m = 0.5$, $k_m = 5$, $T = 0.1$, and $L = 16$.

In order to obtain deep priors, we utilize the pre-trained depth estimation models provided by each dataset on both the Semantic-SuPer [11] and StereoMIS [6] datasets. Additionally, to capture the scene's semantic information, we

TABLE I

QUANTITATIVE EVALUATION ON THE SEMANTIC-SUPER DATASET [11]. THE BOLD FONT INDICATES THE BEST RESULTS.

Methods	Lab1				Lab2			
	Rep.Err.↓	PSNR↑	SSIM↑	LPIPS↓	Rep.Err.↓	PSNR↑	SSIM↑	LPIPS↓
DefSLAM [25]	16.5(12.5)	-	-	-	14.5(13.2)	-	-	-
Semantic-SuPer [11]	7.5(6.1)	-	-	-	8.6(7.6)	-	-	-
iMAP [12]	4.7(1.3)	22.875	0.744	0.367	4.4(0.9)	22.668	0.750	0.350
NICE-SLAM [13]	5.8(1.0)	20.800	0.692	0.382	5.6(1.1)	21.027	0.701	0.381
ESLAM [15]	5.0(1.5)	22.018	0.724	0.339	4.8(4.8)	22.195	0.720	0.344
Co-SLAM [14]	4.9(1.6)	20.754	0.513	0.594	4.6(1.9)	21.713	0.593	0.481
DDS-SLAM(ours)	**3.3(0.4)**	**28.649**	**0.797**	**0.231**	**3.0(0.5)**	**29.678**	**0.828**	**0.175**
Methods	Lab3				Lab4			
	Rep.Err.↓	PSNR↑	SSIM↑	LPIPS↓	Rep.Err.↓	PSNR↑	SSIM↑	LPIPS↓
DefSLAM [25]	12.8(8.8)	-	-	-	7.0(5.2)	-	-	-
Semantic-SuPer [11]	6.0(4.9)	-	-	-	4.3(3.8)	-	-	-
iMAP [12]	5.7(1.3)	20.675	0.617	0.548	3.2(1.2)	20.500	0.597	0.424
NICE-SLAM [13]	5.7(1.3)	21.652	0.666	0.401	4.5(1.6)	20.284	0.624	0.464
ESLAM [15]	4.5(1.6)	23.241	0.707	0.326	3.1(0.7)	25.592	0.707	0.273
Co-SLAM [14]	3.5(1.4)	23.996	0.696	0.354	2.5(0.5)	24.182	0.688	0.298
DDS-SLAM(ours)	**2.4(0.4)**	**27.230**	**0.782**	**0.195**	**2.0(0.2)**	**27.340**	**0.734**	**0.210**

* Following [11], Rep.Err. represents the average reprojection errors across all points, formatted as "mean (standard deviation)". The symbol ↑ denotes that higher values indicate higher accuracy, and vice versa. The results of DefSLAM [5] and Semantic-SuPer [11] are both from [11]. " – " means that the methods do not support image rendering and therefore cannot be evaluated for that specific metric.

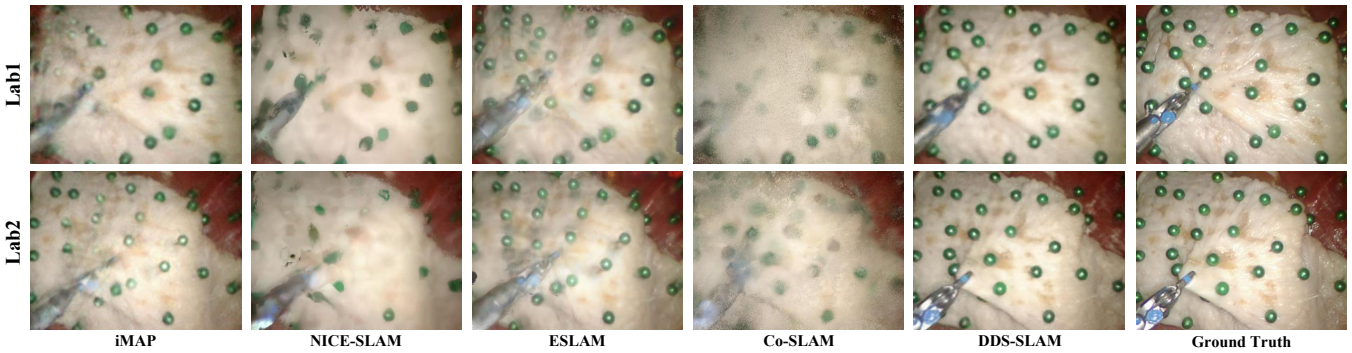

Fig. 3. Rendering results on the Semantic-SuPer dataset. Compared to baselines, our approach achieves high-quality RGB image renderings.

employ the pre-trained DeepLabv3+ model [28] provided by each dataset for semantic segmentation.

A. Experimental Results

Evaluation on the Semantic-SuPer Dataset [11]. We present a quantitative analysis of the experimental results for four sequences from the Semantic-SuPer dataset [11] in Table I. As shown in Table I, our approach significantly outperforms the baselines in both pose tracking accuracy and image rendering quality. Furthermore, our method exhibits lower variance, indicating greater stability and robustness compared to existing methods. A qualitative analysis of the Semantic-SuPer dataset [11] is provided in Fig. 3, revealing that our DDS-SLAM can render higher-quality RGB images with reduced noise and fewer artifacts in deforming scenes.

Evaluation on the StereoMIS Dataset [6]. We further evaluate our method on the StereoMIS dataset [6]. Unlike the Semantic-SuPer dataset [11], the sequence in this dataset comprises a larger number of images (4000) and includes natural deformation. The quantitative results, summarized in Table II, demonstrate that our method outperforms the baselines in tracking accuracy. Visualizations of camera tracking performance are presented in Fig. 4. Our method aligns more closely with the ground truth, reinforcing its superior accuracy. Table II also highlights that our approach excels over the baseline methods in image rendering quality.

TABLE II

QUANTITATIVE EVALUATION ON THE STEREOMIS DATASET [11]. THE BOLD FONT INDICATES THE BEST RESULTS.

Methods	Localization	Rendering		
	ATE(mm)↓	PSNR↑	SSIM↑	LPIPS↓
RECP [6]	28.791	-	-	-
iMAP [12]	36.849	14.047	0.465	0.681
NICE-SLAM [13]	37.365	14.037	0.531	0.576
ESLAM [15]	14.833	20.803	0.585	0.533
Co-SLAM [14]	19.412	22.029	0.579	0.526
DDS-SLAM(ours)	**8.261**	**22.513**	**0.592**	**0.496**

B. Ablation Study

To verify the contribution of key components of the DDS-SLAM, we conduct a series of ablation studies on the StereoMIS dataset [6], with the results presented in Table III.

Effect of Hybrid Neural Deformable Scene Representation (HNDSR). The term 'w/o HNDSR' refers to a variant from which the deformation network is removed, leaving only the canonical network in the hybrid neural deformable scene representation. As illustrated in Table III, our full model demonstrates superior localization accuracy and image

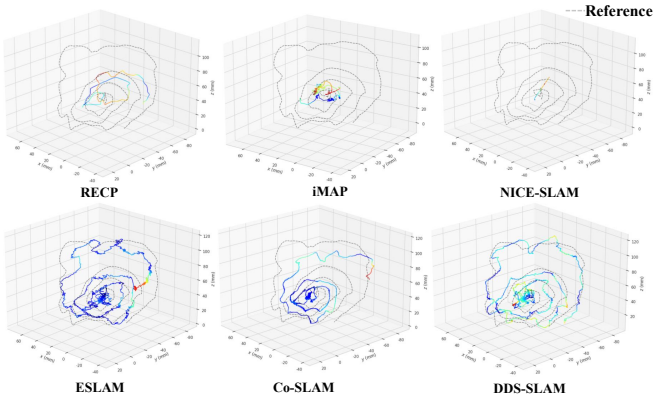

Fig. 4. Visualization of camera tracking results on the StereoMIS dataset.

TABLE III
THE ABLATION STUDY RESULTS FOR KEY COMPONENTS OF DDS-SLAM ON THE STEREOMIS DATASET [6].

Name	ATE(mm)↓	PSNR↑	SSIM↑	LPIPS↓
w/o HNDSR	9.834	21.906	0.580	0.529
w/o SDL	8.628	22.036	0.591	0.492
Full	**8.261**	**22.513**	**0.592**	**0.496**

rendering quality compared to the variant without the deformation network. This finding confirms the effectiveness of the hybrid neural deformation scene representation.

Effect of Semantic Distance Loss (SDL). The notation 'w/o SDL' indicates a variant without semantic distance loss. The quantitative results in Table III reveal that incorporating SDL enhances both camera tracking accuracy and image rendering quality. SDL captures the local interrelationships among different semantic categories and guides the optimization of the network at a higher level.

V. CONCLUSION AND DISCUSSION

In this paper, we introduce DDS-SLAM, a novel dense semantic neural SLAM for deforming endoscopic scenes. Our main design philosophy is twofold. First, we develop a novel hybrid neural representation to model deformable scenes, applicable to medical endoscopic scenarios with both artificial and natural deformations. Second, we propose a semantic distance loss based on the semantic distance field to guide the network's optimization from a higher level. Extensive experimental results confirm that DDS-SLAM excels in deformable endoscopic scenarios, offering superior camera pose tracking and image rendering quality. We believe that DDS-SLAM holds promising potential for applications in medical endoscopy, such as surgical interventions.

REFERENCES

[1] Y. Wang, Y. Long, S. H. Fan, and Q. Dou, "Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery," in *MICCAI*. Springer, 2022, pp. 431–441.

[2] X. Liu, Z. Li, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, "Sage: slam with appearance and geometry prior for endoscopy," in *ICRA*. IEEE, 2022, pp. 5587–5593.

[3] J. Lamarca, J. J. G. Rodríguez, J. D. Tardos, and J. M. Montiel, "Direct and sparse deformable tracking," *RAL*, vol. 7, no. 4, pp. 11 450–11 457, 2022.

[4] J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Tracking monocular camera pose and deformation for slam inside the human body," in *IROS*. IEEE, 2022, pp. 5278–5285.

[5] J. Lamarca, S. Parashar, A. Bartoli, and J. Montiel, "Defslam: Tracking and mapping of deforming scenes from monocular sequences," *TRO*, vol. 37, no. 1, pp. 291–303, 2020.

[6] M. Hayoz, C. Hahne, M. Gallardo, D. Candinas, T. Kurmann, M. Allan, and R. Sznitman, "Learning how to robustly estimate camera pose in endoscopic videos," *IJCARS*, pp. 1–8, 2023.

[7] J. Liu, G. Wang, Z. Liu, C. Jiang, M. Pollefeys, and H. Wang, "Regformer: An efficient projection-aware transformer network for large-scale point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8451–8460.

[8] J. Liu, D. Zhuo, Z. Feng, S. Zhu, C. Peng, Z. Liu, and H. Wang, "Dvlo: Deep visual-lidar odometry with local-to-global feature fusion and bi-directional structure alignment," *arXiv preprint arXiv:2403.18274*, 2024.

[9] L. Qiu and H. Ren, "Endoscope navigation and 3d reconstruction of oral cavity by visual slam with mitigated data scarcity," in *CVPRW*, 2018, pp. 2197–2204.

[10] R. Ma, R. Wang, Y. Zhang, S. Pizer, S. K. McGill, J. Rosenman, and J.-M. Frahm, "Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy," *MIA*, vol. 72, p. 102100, 2021.

[11] S. Lin, A. J. Miao, J. Lu, S. Yu, Z.-Y. Chiu, F. Richter, and M. C. Yip, "Semantic-super: a semantic-aware surgical perception framework for endoscopic tissue identification, reconstruction, and tracking," in *ICRA*. IEEE, 2023, pp. 4739–4746.

[12] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *ICCV*, 2021, pp. 6229–6238.

[13] 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 *CVPR*, 2022, pp. 12 786–12 796.

[14] H. Wang, J. Wang, and L. Agapito, "Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam," in *CVPR*, 2023, pp. 13 293–13 302.

[15] M. M. Johari, C. Carta, and F. Fleuret, "Eslam: Efficient dense slam system based on hybrid representation of signed distance fields," in *CVPR*, 2023, pp. 17 408–17 419.

[16] T. Deng, G. Shen, T. Qin, J. Wang, W. Zhao, J. Wang, D. Wang, and W. Chen, "Plgslam: Progressive neural scene represenation with local to global bundle adjustment," in *CVPR*, June 2024, pp. 19 657–19 666.

[17] J. Shan, Y. Li, T. Xie, and H. Wang, "Enerf-slam: A dense endoscopic slam with neural implicit representation," *IEEE Transactions on Medical Robotics and Bionics*, 2024.

[18] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *CVPR*, 2021, pp. 10 318–10 327.

[19] 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," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.

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

[21] T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák, "Neural importance sampling," *ToG*, vol. 38, no. 5, pp. 1–19, 2019.

[22] R. Or-El, X. Luo, M. Shan, E. Shechtman, J. J. Park, and I. Kemelmacher-Shlizerman, "Stylesdf: High-resolution 3d-consistent image and geometry generation," in *CVPR*, 2022, pp. 13 503–13 513.

[23] F. Tosi, Y. Zhang, Z. Gong, E. Sandström, S. Mattoccia, M. R. Oswald, and M. Poggi, "How nerfs and 3d gaussian splatting are reshaping slam: a survey," *arXiv preprint arXiv:2402.13255*, 2024.

[24] J. Canny, "A computational approach to edge detection," *TPAMI*, no. 6, pp. 679–698, 1986.

[25] J. Lamarca, S. Parashar, A. Bartoli, and J. Montiel, "Defslam: Tracking and mapping of deforming scenes from monocular sequences," *TRO*, vol. 37, no. 1, pp. 291–303, 2020.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 801–818.