Wei *et al*. *Art Int Surg* 2024;4:187-98
DOI: 10.20517/ais.2024.12

Artificial Intelligence Surgery

Original Article

Open Access

Scale-aware monocular reconstruction via robot kinematics and visual data in neural radiance fields

Ruofeng Wei[1], Jiaxin Guo[2], Yiang Lu[2], Fangxun Zhong[2], Yunhui Liu[2], Dong Sun[3], Qi Dou[1]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong 000000, China.
[2]Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong 000000, China.
[3]Department of Biomedical Engineering, City University of Hong Kong, Hong Kong 000000, China.

Correspondence to: Dr. Ruofeng Wei, Department of Computer Science and Engineering, Room115, Ho Sin-Hang Engineering Building, The Chinese University of Hong Kong, Shatin, Hong Kong 000000, China. E-mail: ruofengwei@cuhk.edu.hk

How to cite this article: Wei R, Guo J, Lu Y, Zhong F, Liu Y, Sun D, Dou Q. Scale-aware monocular reconstruction via robot kinematics and visual data in neural radiance fields. *Art Int Surg* 2024;4:187-98. http://dx.doi.org/10.20517/ais.2024.12

Received: 6 Feb 2024 **First Decision:** 20 Jun 2024 **Revised:** 15 Jul 2024 **Accepted:** 31 Jul 2024 **Published:** 16 Aug 2024

Academic Editors: Andrew A. Gumbs, Eyad Elyan **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Aim: Scale-aware 3D reconstruction of the surgical scene from a monocular endoscope is important for automatic navigation systems in robot-assisted surgery. However, traditional multi-view stereo methods purely utilize monocular images, which can recover 3D structures arbitrarily scaled with the real world. Current deep learning-based approaches rely on large training data for relative depth estimation and further 3D reconstruction with no scale. Inspired by recently proposed neural radiance fields (NeRF), we present a novel pipeline, KV-EndoNeRF, which explores limited multi-modal data (i.e., robot kinematics, and monocular endoscope) for surgical scene reconstruction with absolute scale.

Methods: We first extract scale information from robot kinematics data and then integrate it into sparse depth recovered from structure from motion (SfM). Based on the sparse depth supervision, we adapt a monocular depth estimation network to the current surgical scene to obtain scene-specific coarse depth. After adjusting the scale of coarse depth, we use it to guide the optimization of NeRF, resulting in absolute depth estimation. The 3D models of the tissue surface with real scale are recovered by fusing fine depth maps.

Results: Experimental results on the Stereo Correspondence And Reconstruction of Endoscopic Data (SCARED) demonstrate that KV-EndoNeRF excels in learning an absolute scale from robot kinematics and achieves 3D reconstruction with rich details of surface texture and high accuracy, outperforming other existing reconstruction methods.

Conclusion: Combining multi-modal image data with NeRF-based optimization represents a potential approach to achieve scale-aware 3D reconstruction of monocular endoscopic scenes.

Keywords: Scale-aware reconstruction, NeRF-based optimization, multi-modal data learning, surgical navigation, robotic surgery

1. INTRODUCTION

Reconstructing scale-aware 3D structures from monocular endoscopes is a fundamental task for some emerging surgical robotic systems, such as flexible robots[1–3]. It is also a prerequisite for applications such as multi-modal image registration and automatic navigation based on real-scale 3D modeling of human anatomies[4–6]. However, relying solely on monocular images is insufficient to accurately recover 3D structures with absolute scale in the surgical scene. Several methods for scene reconstruction from monocular endoscopes have been explored. Traditional multi-view stereo methods[7] can simultaneously recover 3D point clouds and camera poses in scenes with rich features. However, these methods cannot directly reconstruct structures with real scale, requiring manual estimation of the global scale and then optimization of it using "iterative closest point" registration algorithm (ICP)[8]. Recent deep learning-based methods[8–10] have exploited large numbers of surgical images with certain requirements, such as static tissue surfaces or ground truth depth labels, to train convolutional neural networks (CNN) for relative depth estimation and further reconstruction. However, based on our experiments, these methods only predicted relative depth by large training data and computed 3D reconstruction without an accurate scale.

Surgical robotic systems provide richer information beyond images, such as robot kinematics, which describes how robotic instruments are mechanically controlled. This kinematics information can enhance the perception in a multi-modal learning style[11]. Despite much work on recognition-related tasks using robotic information[12–14], joint modeling of kinematics and visual data for monocular 3D reconstruction has been rarely studied to date due to several challenges. First, acquiring large surgical datasets with static scenes for learning-based methods is difficult. Second, generating accurate ground truth depth labels of real endoscopic images is hard. Third, for 3D reconstruction, robot kinematics and endoscopic videos represent multi-modal data, and how to efficiently integrate kinematics data into the images remains underexplored.

Neural radiance fields (NeRF) have emerged as a promising technology[15,16] for quality novel view synthesis and 3D reconstruction. These methods utilize neural implicit fields to represent continuous scenes. Several variants of NeRF[17,18] have incorporated sparse 3D points from structure from motion (SfM) techniques to guide ray termination and optimize the neural implicit field for view synthesis. However, these approaches have primarily focused on relative depth estimation in natural scenes. In the context of urban environments, urban radiance fields (URF)[19] have been introduced to apply NeRF-based view synthesis and visual reconstruction. URF leverages sparse multi-view images along with LiDAR data to reconstruct urban scenes. In the field of medicine, a recent work called EndoNeRF[20] has presented a pipeline for achieving single-view 3D reconstruction of dynamic surgical scenes. This methodology specifically addresses the challenges of reconstructing surgical scenes that involve deformable tissues.

In this paper, we propose a novel approach, KV-EndoNeRF, for reconstructing surgical scenes with an accurate scale using kinematics and visual data. Our contributions can be summarized as follows: Firstly, we introduce a NeRF-based pipeline specifically designed for scale-aware reconstruction from multi-modal data, addressing the challenging problem of reconstructing 3D scenes with scale from a monocular endoscope. Secondly, we incorporate scale information extracted from robot kinematics and coarse depth information learned from SfM into the NeRF optimization process, improving the accuracy of the reconstruction. Finally, we evaluate

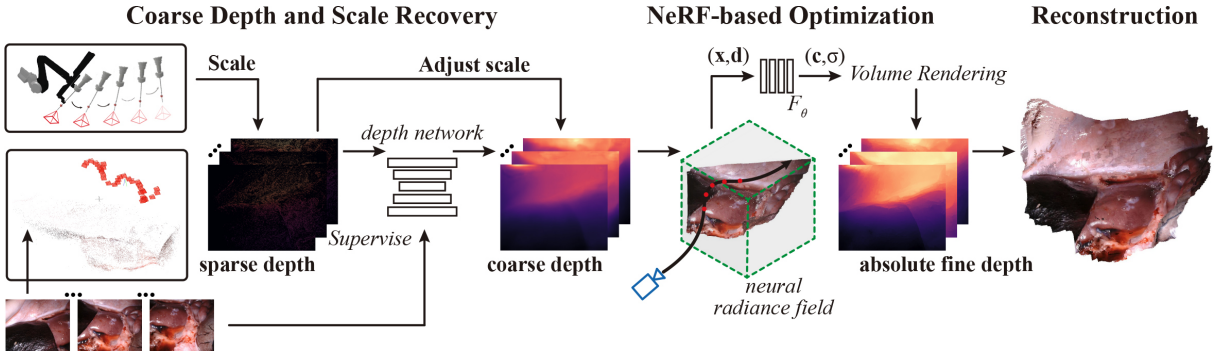

Figure 1. Illustration of our proposed KV-EndoNeRF for scale-aware monocular reconstruction from the robotic endoscope. NeFR: Neural radiance fields.

our proposed pipeline, KV-EndoNeRF, both qualitatively and quantitatively on the publicly available Stereo Correspondence And Reconstruction of Endoscopic Data (SCARED) robotic endoscope dataset. The results demonstrate that KV-EndoNeRF outperforms previous methods, showcasing its ability to achieve 3D reconstruction with accurate scale in monocular surgical scenes.

2. METHODS

2.1 Overview of NeRF-based scale-aware reconstruction

Considering robot-assisted endoscopy, the goal of our proposed pipeline, KV-EndoNeRF, is to achieve scale-aware monocular reconstruction from limited multi-modal data (i.e., kinematics, and endoscopic image sequences). It requires neither large numbers of endoscopic images for training, nor other imaging modalities, such as computed tomography (CT) and magnetic resonance image (MRI), for the ground truth labels. The key to our pipeline is to effectively incorporate the scale information from robot kinematics into NeRF-represented surgical scenes for optimization. Following the modeling in NeRF[15], we represent the surgical scene as a neural radiance field for further volume rendering (Section 2.2). As shown in Figure 1, we first extract the absolute scale from kinematics and then fuse it into sparse depth produced by SfM. Under sparse supervision, we fine-tune a monocular depth estimation network to the current endoscopic scene for scene-specific coarse depth (Section 2.3). After adjusting the scale of coarse depth estimation, we integrate it into the ray marching of NeRF and optimize the volumetric field to obtain the absolute depth (Section 2.4). Finally, the refined absolute depth maps are fused in a truncated signed distance functions (TSDF)-based volumetric representation according to the endoscopic trajectory (Section 2.5). This results in a reconstructed 3D model of the surgical scene with global-scale information.

2.2 Surgical scene representing and rendering by NeRF

NeRF has achieved impressive success in view synthesis by optimizing the neural implicit field. Our pipeline explores its potential for the optimization of depth estimates. We represent a surgical scene as a neural radiance field F_θ, which is an 8-layer multilayer perceptron (MLP) with network parameter θ. The field, $F_\theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$, maps a 3D point $\mathbf{x} \in \mathbb{R}^3$ and a viewing direction $\mathbf{d} \in \mathbb{R}^3$ to an RGB value $\mathbf{c}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^3$ and space occupancy $\sigma(\mathbf{x}) \in \mathbb{R}$. With scene representations, we further adopt the volume rendering[21] in NeRF to generate rendered images for training. The volume rendering starts with shooting a batch of endoscope rays into the surgical scene from the endoscope center $\mathbf{o}$ along the direction $\mathbf{d}$. Each ray is constructed as $\mathbf{r}(s) = \mathbf{o} + s\mathbf{d}$, where s is the ray parameter. We then proceed with ray marching to sample points in the space. Specifically, we partition each camera ray $\mathbf{r}(s)$ into a batch of points $\{\mathbf{x}_k | \mathbf{x}_k = \mathbf{r}(s_k)\}_{k=1}^m$. Then, the rendered

color image $\mathbf{C}$ and per-view depth value $\mathbf{D}$ of a camera ray $\mathbf{r}(s)$ can be computed using volume rendering as:

$$\mathbf{C}(\mathbf{r}(s)) = \sum_{k=1}^{m} w_k \cdot \mathbf{c}(\mathbf{x}_k, \mathbf{d}), \quad \mathbf{D}(\mathbf{r}(s)) = \sum_{k=1}^{m} w_k \cdot s_k \tag{1}$$

where $w_k = (1 - \exp(-\sigma(\mathbf{x}_k) \triangle s_k)) \exp\left(-\sum_{l=1}^{k-1} \sigma(\mathbf{x}_l) \triangle s_l\right)$, $\triangle s_k = s_{k+1} - s_k$.

In this way, the 3D structure of the surgical scene can be encoded as a continuous implicit function, which enables memory-efficient geometric representation with infinite resolution.

2.3 Coarse depth adaptation and scale recovery with kinematics

We use the SfM to reconstruct the surgical scene as a set of 3D points $\mathcal{X}$ and camera poses $\mathcal{T} = \{\mathbf{T}_i \in \mathbf{SE}(3) | i = 1 \cdots N\}$ for the input images extracted from the unlabeled endoscopic video. To eliminate extreme outliers in the sparse reconstruction, point cloud filtering is utilized. For each endoscopic image pair, the rigid transformation matrix ${}^{I}\mathbf{T}_i^{i+1}$ from image i to $i+1$ can be computed by the camera poses $\mathcal{T}$, where the left superscript $\{I\}$ denotes the pose described under the image coordinate. As the endoscope is attached to a robot, the camera poses under the robot coordinate system can be calculated from kinematics information, considered as a reference to recover the absolute scale. Therefore, the relative pose ${}^{R}\mathbf{T}_i^{i+1}$ under the robot base $\{R\}$ is computed. The absolute scale between the reconstructed structure and the real world can be estimated by:

$$\lambda = \exp\left(\frac{1}{N-1} \sum_{i=0}^{N-2} \log_{10}\left(\frac{\left\|{}^{R}\mathbf{t}_i^{i+1}\right\|_2}{\left\|{}^{I}\mathbf{t}_i^{i+1}\right\|_2}\right)\right), \tag{2}$$

where $\mathbf{t}_i$ is the translation vector of the camera pose $\mathbf{T}_i$. Although we can compute scale data for each frame, the noise in the kinematics data and the instability of the poses in $\mathcal{T}$ introduce severe noise to each scale. To filter the scale, we employ a logarithmic moving average with a multiplicative error model. Based on the computed scale factor, we adjust the sparse 3D structure and camera poses to match the real-world values. Afterward, the scaled 3D point cloud $\mathcal{X}'$ is projected onto each image plane with the corresponding scaled camera pose $\mathbf{T}_i'$. The re-projected z values are concatenated as the sparse depth supervision ${}^{s}\mathbf{D}_i$, where the region with no points projected onto is set to zero.

To obtain scene-specific coarse depth from the current endoscopic data, we propose adapting a depth estimation network. This network is fine-tuned using the sparse depth supervision ${}^{s}\mathbf{D}_i$. However, due to the scale ambiguity in the predicted depth map, we utilize the scale-invariant log loss[22] for training the depth network. The scale-invariant log loss is defined as:

$$\mathcal{L} = \sqrt{\frac{1}{M} \sum_{t=1}^{M} e_t^2 - \frac{\alpha}{M^2} \left(\sum_{t=1}^{M} e_t\right)^2} \tag{3}$$

where $e_t = \log y_t - \log y_t'$, y_t represents the coarse depth value predicted by the proposed depth network, and y_t' is the value of the corresponding sparse depth supervision ${}^{s}\mathbf{D}_i$. M denotes the number of pixels with valid supervision values, and α is a weighting factor.

2.4 NeRF-based optimization for absolute depth

According to Equation (2), we can determine the absolute scale between the reconstruction and real-world values using the robot kinematics information. To incorporate this calculated scale into dense monocular reconstruction, we propose guiding the NeRF sampling process with our coarse depth estimation and scale information. First, we align the scale of the coarse depth map ${}^{c}\mathbf{D}_i$ based on the depth supervision ${}^{s}\mathbf{D}_i$. Moreover, we compute the confidence map of ${}^{c}\mathbf{D}_i$ by a geometric consistency check. The depth ${}^{c}\mathbf{D}_i$ is first projected onto

all other views using the following equations:

$$\mathbf{p}_{i\to j},\ {}^{c}\mathbf{D}_{i\to j} \sim \mathbf{K}\cdot\left(\mathbf{T}_{i}^{j}\right)'\cdot {}^{c}\mathbf{D}_{i}\left(\mathbf{p}_{i}\right)\cdot\mathbf{K}^{-1}\cdot\mathbf{h}\left(\mathbf{p}_{i}\right) \tag{4}$$

$${}^{c}\mathbf{D}'_{j} = {}^{c}\mathbf{D}_{j}\left(\mathbf{p}_{i\to j}\right) \tag{5}$$

where $\mathbf{K}$ represents the endoscope intrinsic matrix, and $\mathbf{p}$ denotes a pixel in the image. Subsequently, we calculate the depth reprojection error between ${}^{c}\mathbf{D}'_{j}$ and ${}^{c}\mathbf{D}_{i\to j}$. The confidence map $\mathbf{E}_{i}$ for each view is defined as the average value of the top K minimum cross-view depth reprojection errors.

Next, during ray marching, we sample points using a Gaussian distribution guided by the prior from the scaled coarse depth. Assuming the coarse depth value for a pixel $\mathbf{p}$ to be $z_{\mathbf{p}} = {}^{c}\mathbf{D}_{i}\left(\mathbf{p}\right)$, we sample the candidates using the distribution $s_{k} \sim \mathcal{N}\left(z_{\mathbf{p}}, \delta_{\mathbf{p}}^{2}\right)$, where $\delta_{\mathbf{p}} = z_{\mathbf{p}}\cdot\mathbf{E}_{i}\left(\mathbf{p}\right)$. This sampling method ensures that the points are concentrated around tissue surfaces.

To estimate the absolute depth of endoscopic frames, we can optimize the network parameter θ by supervising the rendered color images. To be more specific, the loss function utilized to train the network is defined as follows:

$$\mathcal{L}\left(\mathbf{r}\left(s\right)\right) = \left\|\mathbf{C}\left(\mathbf{r}\left(s\right)\right) - \mathbf{I}_{i}\left(\mathbf{p}\right)\right\|_{2}^{2} \tag{6}$$

where $\mathbf{p}$ represents the location of the pixel that $\mathbf{r}\left(s\right)$ shoots toward, and $\mathbf{I}_{i}$ corresponds to the input endoscopic image.

2.5 Volumetric reconstruction on fine depth

To further refine depth accuracy, we use the view synthesis results of NeRF to calculate the per-pixel error for the predicted structure. If the rendering at a specific pixel does not match the input endoscopic image well, a high error is assigned to the depth prediction of that pixel. The error map $\mathbf{R}_{i}\left(\mathbf{p}\right)$ for the pixel $\mathbf{p}$ in the ith view is expressed as:

$$\mathbf{R}_{i}\left(\mathbf{p}\right) = \left\|\mathbf{I}_{i}\left(\mathbf{p}\right) - \mathbf{C}\left(\mathbf{p}\right)\right\|_{1} / 255 \tag{7}$$

The error map is then used to improve the estimated depth by a filter. We apply an off-the-shelf post-filtering approach [23] to obtain the fine output, which enhances absolute depth estimates, particularly in regions where the renderings are not accurate.

Afterward, these fine depth maps are fused to create a surface reconstruction. We use TSDF [24] to build a volumetric representation of the tissue surface. Since the predicted depth maps and the endoscope poses are scaled to the real world, all data are made scale-aware and -consistent before fusion. The surgical scene is represented by a discrete voxel grid, and for each of them, a weighted signed distance to the closest surface is recorded. The TSDF is updated in a straight manner, using sequential averaging for each voxel and the predicted depth for each pixel in every image. Finally, the whole 3D structure is reconstructed by the marching cubes method [25] from the volumetric representation.

3. RESULTS

3.1 Dataset and implementation details

We evaluate our scale-aware monocular reconstruction pipeline on the publicly available SCARED dataset [26]. This dataset consists of seven training datasets and two test datasets captured by a da Vinci Xi surgical robot. Each dataset is collected from a porcine model and contains four or five keyframes. Each keyframe includes a video with kinematic information about the endoscope. From each dataset, we randomly select one keyframe and extract a set of 40 to 80 images that cover the entire surgical scene. During the data collection process, the robot manipulates an endoscope to observe the interior scenes of the porcine abdominal anatomy. A projector

Figure 2. Four typical examples of the SCARED data. For every row, when the robot manipulates the endoscope to move, diversified views and corresponding robot kinematics are recorded in sequence. SCARED: Stereo Correspondence And Reconstruction of Endoscopic Data.

is used to calculate high-quality depth maps for each frame. As a result, the dataset provides endoscopic videos with ground-truth depth maps and robot kinematics. Typical examples of the SCARED data are illustrated in Figure 2. In addition, the robot kinematics information is utilized to restore the scale.

In our implementation, we used the network architecture proposed in Mannequin Challenge[27] with pre-trained weights as the monocular depth network for coarse depth adaptation. Twenty fine-tuning epochs were used in the surgical scene-specific adaptation. We set $K = 4$ for the geometric consistency check. For the NeRF-based optimization, we followed the settings in NeRF[15]. Specifically, we sampled 64 points in each ray and used a batch of 1,024 rays during the training. We added random Gaussian noise with zero mean and unit variance to the density to regularize the network. Additionally, positional encoding was utilized to capture high-frequency details. Using Adam optimizer with an initial learning rate of 5e-4, which decayed exponentially to 5e-5, we trained our NeRF on each surgical scene for 200 K iterations. All experiments were conducted on a single RTX 2080 Ti.

3.2 Performance metrics

Table 1 lists the depth evaluation metrics[28] used in our experiments, where d and d^* denote the estimated depth value and the corresponding ground truth, respectively, $\mathbf{D}$ represents the estimated depth map, and $\upsilon \in \{1.25^1, 1.25^2\}$. Additionally, since the comparison methods cannot accurately predict depth maps with an absolute scale from monocular images, we employ the ground truth median scaling method[29] to scale the predicted depth. The scaling is performed as follows:

$$\mathbf{D}_{sd} = \mathbf{D} \cdot s = \mathbf{D} \cdot \frac{median(\mathbf{G})}{median(\mathbf{D})} \tag{8}$$

where $\mathbf{D}_{sd}$ denotes the scaled predicted depth, s represents the scale information calculated by the median scaling method, and $\mathbf{G}$ is the ground truth depth.

Table 1. Depth evaluation metrics

Metrics	Definition
Abs Rel	$\frac{1}{\lvert\mathbf{D}\rvert}\sum_{d\in\mathbf{D}}\lvert d^*-d\rvert/d^*$
Sq Rel	$\frac{1}{\lvert\mathbf{D}\rvert}\sum_{d\in\mathbf{D}}\lvert d^*-d\rvert^2/d^*$
RMSE	$\sqrt{\frac{1}{\lvert\mathbf{D}\rvert}\sum_{d\in\mathbf{D}}\lvert d^*-d\rvert^2}$
RMSE_{log}	$\sqrt{\frac{1}{\lvert\mathbf{D}\rvert}\sum_{d\in\mathbf{D}}\lvert\log d^*-\log d\rvert^2}$
δ	$\frac{1}{\lvert\mathbf{D}\rvert}\{d\in\mathbf{D}\lvert\max(\frac{d^*}{d},\frac{d}{d^*}<\upsilon)\rvert\}\times 100\%$

d and d^* represent the estimated depth value and the corresponding ground truth. **D** corresponds to the estimated depth map. RMSE: Root mean square error.

3.3 Evaluation on scale-aware depth estimation

We compare the accuracy of depth estimation using the KV-EndoNeRF method with several other deep learning-based approaches and the SfM method, specifically COLMAP[7].

- COLMAP[7] is a general-purpose SfM pipeline used for reconstructing 3D point cloud reconstruction from ordered and unordered image collections. In our study, we apply it to monocular surgical scene reconstruction. The recovered points are then projected onto each image plane to obtain the sparse depth maps for evaluation.
- EndoSLAM[30] is an unsupervised relative monocular depth estimation method specifically designed for gastrointestinal tract organs. It combines residual networks with a spatial attention module to focus on highly textured tissue regions. We fine-tune the depth model using the SCARED data for comparison.
- AF-SfMLearner[10] is a novel self-supervised network for estimating monocular depth in endoscopic scenes. It is trained on the SCARED datasets, which contain severe brightness fluctuations induced by illumination variations, non-Lambertian reflections, and inter-reflections.
- DS-NeRF[17] is a general depth-supervised NeRF method that utilizes sparse reconstruction from the SfM to recover dense 3D structures. We apply DS-NeRF to estimate dense depth maps for each endoscopic image.

We present the quantitative depth comparison results on SCARED data in Table 2, which rescales the results using the ground truth median scaling method. In addition to standard depth evaluation metrics, we calculate the means and standard errors of the rescaling factors to demonstrate the scale-awareness ability. KV-EndoNeRF achieves the best up-to-scale performance with respect to five metrics and ranks the second best for the other two metrics. Notably, KV-EndoNeRF also achieves nearly perfect absolute scale estimation. These quantitative results show that our proposed method effectively extracts absolute scale information from kinematics and integrates it into NeRF for further depth optimization, resulting in accurate absolute depth estimation.

Furthermore, we select four representative images from the SCARED dataset for qualitative depth comparison. As shown in Figure 3, our method with NeRF-based optimization produces depth predictions with sharp boundaries and fine-grained details, outperforming other approaches in terms of absolute depth estimation. However, COLMAP could only recover sparse depth maps without the entire 3D geometry of the tissue surface. While EndoSLAM and AF-SfMLearner are capable of generating reasonable 3D structures of tissues, they lose many details in tissues with complex geometries and edges. Lastly, the estimated depth values from DS-NeRF contain significant noise, which could affect the surgeons' observations of complicated tissue surfaces.

Table 2. Quantitative comparisons for scale-aware depth estimation on SCARED

Method	Scale	Error ↓				Accuracy ↑	
		Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25^1$	$\delta < 1.25^2$
COLMAP [7]	4.04 ± 2.24	**0.044** ± 0.028	0.391 ± 0.435	4.766 ± 2.506	**0.065** ± 0.033	0.979 ± 0.036	0.998 ± 0.006
EndoSLAM [30]	77.77 ± 17.10	0.079 ± 0.047	0.897 ± 1.090	7.160 ± 4.818	0.099 ± 0.052	0.931 ± 0.124	0.997 ± 0.009
AF-SfMLearner [10]	2.12 ± 0.45	0.056 ± 0.028	0.437 ± 0.560	5.103 ± 3.143	0.073 ± 0.034	0.979 ± 0.047	**0.999** ± 0.005
DS-NeRF [17]	22.04 ± 9.75	0.049 ± 0.034	0.458 ± 1.012	4.866 ± 3.432	0.070 ± 0.041	0.972 ± 0.067	0.997 ± 0.012
Ours	**0.95** ± 0.07	0.048 ± 0.025	**0.347** ± 0.351	**4.583** ± 2.247	0.066 ± 0.030	**0.984** ± 0.029	**0.999** ± 0.003

The closer the scale is to 1, the better. The best result is in bold. The second best is underlined. SCARED: Stereo Correspondence And Reconstruction of Endoscopic Data; RMSE: root mean square error; NeRF: neural radiance fields.

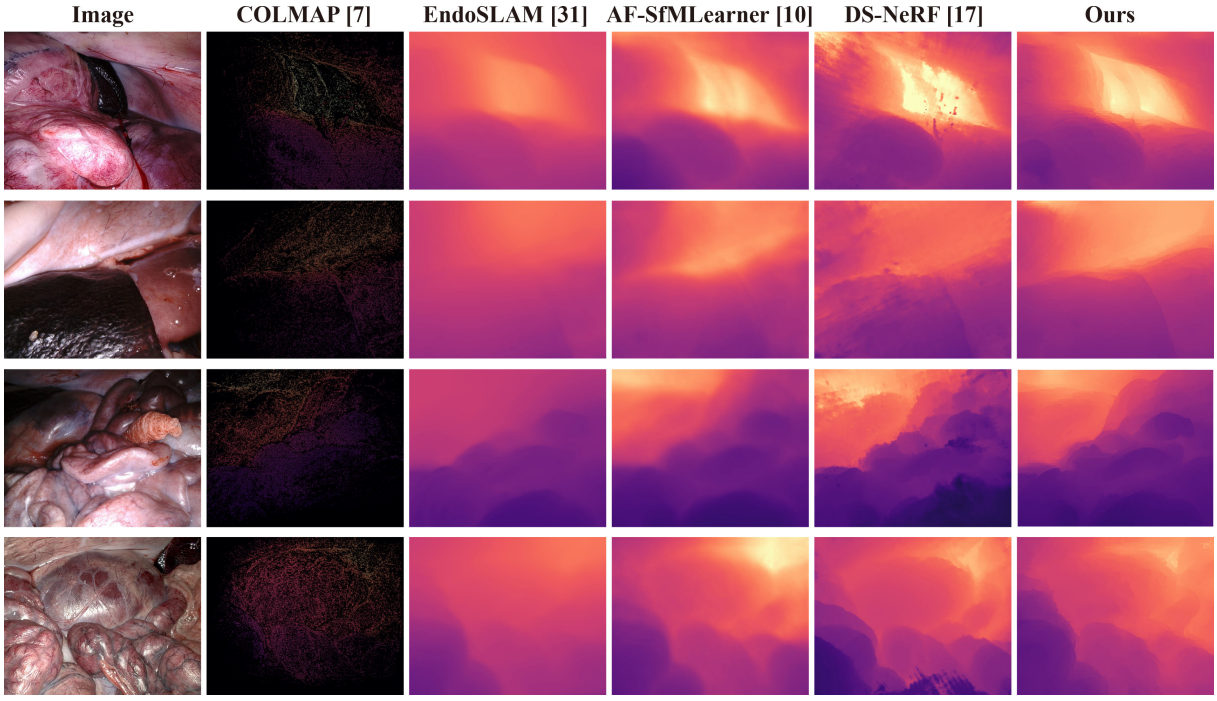

Figure 3. Qualitative comparisons on SCARED. Our method outperforms COLMAP [7], EndoSLAM [30], AF-SfMLearner [10], and DS-NeRF [17] in terms of depth quality. A large depth value is encoded with yellow, while a small depth value is encoded with purple. SCARED: Stereo Correspondence And Reconstruction of Endoscopic Data; NeRF: neural radiance fields.

3.4 Comparison with state-of-the-art methods

We compare our method with state-of-the-art approaches in terms of 3D reconstruction and view synthesis. Firstly, we quantitatively assess the reconstruction results and compare them with ground truth 3D models calculated by a structure light camera [26]. Unlike other monocular scene reconstruction methods, we do not scale the structures during evaluation, thanks to our scale-aware depth estimation. KV-EndoNeRF achieves high accuracy in 3D reconstruction, with an average root mean square error (RMSE) error of 1.259±0.257 mm across all data. Figure 4A shows a qualitative comparison of SCARED data. As shown in the figure, the ground truth models in the third column, represented by gray points, indicate that these tissues have complex surfaces. The sparse point clouds recovered by COLMAP are presented in the first column of the figure. Due to the sparsity of the 3D points, it is difficult to observe the geometric structures and the textures of the tissue surfaces. In comparison, our reconstructed meshes shown in the second column present reasonable structures and rich details of the surface. Furthermore, we register the reconstruction results with the ground truth structures, and the registration results show that our 3D reconstruction matches well with the ground truth. In summary, our method can reconstruct smooth 3D structures from a monocular endoscope with accurate scale, high accuracy, and rich details of the surface texture. Moreover, in Figure 4B, we observe that the proposed method benefits

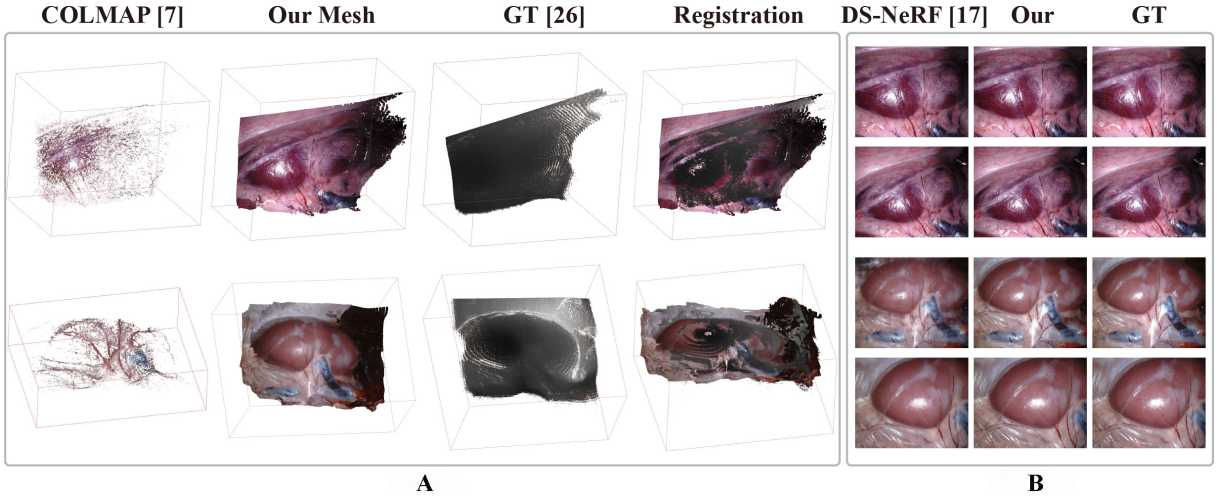

Figure 4. (A) Qualitative reconstruction comparisons on SCARED; (B) Results on view synthesis. Better viewed when zoomed in. More results are shown in the Supplementary Materials. SCARED: Stereo Correspondence And Reconstruction of Endoscopic Data.

Table 3. Ablation studies on each module of our pipeline using dataset-5/keyframe-4

Coarse Depth	NeRF	Refinement	Scale	RMSE	$\delta < 1.25^1$
✓			21.83 ± 3.64	2.864 ± 0.452	0.991 ± 0.008
✓	✓		0.89 ± 0.05	2.730 ± 0.391	0.993 ± 0.008
✓	✓	✓	**0.90** ± 0.04	**2.688** ± 0.415	**0.995** ± 0.007

The best results are in bold. NeRF: Neural radiance fields; RMSE: root mean square error.

the view synthesis quality of NeRF. With the coarse depth priors, our method improves the rendering quality for view synthesis. More 3D point cloud comparison results are illustrated in the Supplementary Materials.

3.5 Ablation studies

We perform ablation studies to validate the effectiveness of the proposed pipeline in estimating fine absolute depth using robot kinematics and NeRF-based optimization. Results in Table 3 demonstrate that each module contributes to the final depth quality. Although the coarse depth estimation is not scaled, it provides a relatively accurate depth basis for the following NeRF-based optimization, as shown in the table. After computing the scale from kinematics data, we incorporate it into NeRF to optimize depth further. We observe that the NeRF improves the depth quality and retains the absolute scale information. Additionally, the refinement operation based on the view synthesis enhances absolute depth estimates, which is beneficial to the final scale-aware reconstruction.

4. DISCUSSION

Nowadays, robotic surgery has become a valuable tool for surgeons, offering advantages such as improved precision in positioning and repetitive accuracy. However, despite these benefits, certain challenges persist, including the absence of 3D anatomical structures and a limited field of view. The accurate representation of the surgical scene in 3D, with proper scaling, is crucial for ensuring surgical safety and effectively controlling robotic systems[31]. To address these issues, we propose a novel NeRF-based method that leverages both visual information and robot kinematics to achieve scale-aware 3D reconstruction of monocular endoscopic scenes. Notably, our approach does not require labeled data or the use of CT scans for training. By incorporating robot kinematics as an additional modality, we can extract scale information that bridges the gap between the

3D reconstruction and the real world. Given the widespread adoption of robotic surgery, it is imperative to integrate robotic kinematics as a multi-modal data source in the visual reconstruction process.

In Ear-Nose-Thorat (ENT) surgery[6] or colonoscopy[32], surgeons manipulate flexible endoscopes or instruments to observe anatomies or perform specific operations. Considering the narrow space of the surgical site, it is crucial for the surgeon or the robot to have an accurate understanding of the 3D structures with real-scale representation of the environment. Therefore, our proposed method can be applied to ENT surgery and colonoscopy. When a limited number of monocular images are obtained from the endoscope, the NeRF-based method can reconstruct the 3D geometry of the tissue surface. For the kinematics data, an external tracking system, such as EM-Tracker and FBG sensors, can be embedded into the surgical robot. In this case, our proposed 3D reconstruction method seamlessly integrates into current surgical robotic systems.

While some existing methods employ external sensors, such as stereo cameras[33,34], to recover real-scale 3D structures, their practical implementation is hindered by their high cost. Additionally, in certain scenarios like ENT surgery and colonoscopy, the limited operating space poses challenges for using stereo cameras. Alternative approaches involve the use of optical tracking[35] or electromagnetic systems[36] to register the endoscope with CT/MRI data. However, these devices are typically treated as independent sources of information for multi-modal data registration. In contrast, our method integrates robotic information into a comprehensive framework, enabling the reconstruction of scale-aware structures from monocular endoscopes. Moreover, compared to learning-based monocular reconstruction approaches[37], our proposed NeRF-based method does not require large amounts of domain-specific training data and can render novel endoscopic views for surgeons to observe the surgical scenarios. Additionally, while other SLAM-based reconstruction methods[38] can only recover sparse 3D point clouds without accurate scaling, our framework can obtain dense 3D structures with an absolute scale to represent tissue surfaces.

However, our method does have some limitations that should be addressed in future work. Firstly, the current approach relies on two separate processes to extract scale data from robot kinematics and monocular images, which is complex and time-consuming. To overcome this, we aim to develop an end-to-end learning method that can efficiently distill information from different modalities. Secondly, the use of the NeRF technique to represent the 3D geometry requires significant computational resources and training time, making real-time rendering and reconstruction challenging. To tackle this issue, we plan to investigate more efficient neural representations, such as 3D gaussian, which can be integrated into our method to enhance efficiency for real-time application. Furthermore, while the kinematics information provided by rigid robots is relatively accurate and has minimal noise, flexible surgical robots can only provide rough and inaccurate kinematics data. Currently, our framework does not account for errors in robot kinematics during scale recovery. In future work, we intend to design an optimization module that can jointly utilize the translation and rotation components of the poses from robot kinematics and visual data. Additionally, we aim to collect more multi-modal data from different surgical scenes to thoroughly evaluate the performance of our method.

5. CONCLUSION

In this paper, we introduce a novel NeRF-based pipeline that enables scale-aware monocular reconstruction with limited robotic endoscope data. It neither requires large medical images nor ground truth labels for network training. We first integrate the scale information extracted from kinematics and learning-based coarse depth supervised by SfM into the optimization process of NeRF, resulting in absolute depth estimation. Then, 3D models with a real scale of tissue surfaces are reconstructed by fusing refined absolute depth maps. We also evaluate the pipeline on SCARED data to demonstrate its accuracy and efficiency. In the future, more robotic endoscope data will be collected to validate our pipeline. The reconstructed scale-aware 3D structures will be utilized for automatic navigation systems in various robotic surgeries, including ENT surgery.

DECLARATIONS

Authors' contributions

Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing - original draft: Wei R
Conceptualization, formal analysis, writing - review and editing: Guo J
Conceptualization, writing - review and editing: Lu Y, Zhong F
Conceptualization, resources, formal analysis, Writing review and editing, supervision: Liu Y, Sun D, Dou Q

Availability of data and materials

Stereo Correspondence And Reconstruction of Endoscopic Data (SCARED) is publicly available at https://endovissub2019-scared.grand-challenge.org.

Financial support and sponsorship

This research work was supported in part by Shenzhen Portion of Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone under HZQB-KCZYB-20200089, in part by Hong Kong Research Grants Council Project No. T42-409/18-R, and in part by National Natural Science Foundation of China Project No. 62322318.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

As our research does not deal with any patient data or broadly, we did not obtain an institutional review board (IRB) for this study.

Consent for publication

Not applicable.

REFERENCES

1. Diana M, Marescaux J. Robotic surgery. *J Brit Surg* 2015;102:e15–28. DOI
2. Fang G, Chow MCK, Ho JDL, et al. Soft robotic manipulator for intraoperative MRI-guided transoral laser microsurgery. *Sci Robot* 2021;6:eabg5575. DOI
3. Mo H, Wei R, Ouyang B, et al. Control of a flexible continuum manipulator for laser beam steering. *IEEE Robot Autom Lett* 2021;6:1074–81. DOI
4. Li B, Wei R, Xu J, et al. 3D perception based imitation learning under limited demonstration for laparoscope control in robotic surgery. In: 2022 International Conference on Robotics and Automation (ICRA); 2022 May 23-27; Philadelphia, USA. IEEE; 2022. pp. 7664–70. DOI
5. Wei R, Li B, Mo H, et al. Stereo dense scene reconstruction and accurate localization for learning-based navigation of laparoscope in minimally invasive surgery. *IEEE Trans Biomed Eng* 2023;70:488–500. DOI
6. Zhong F, Li P, Shi J, et al. Foot-controlled robot-enabled enDOoscope manipulator (FREEDOM) for sinus surgery: Design, control, and evaluation. *IEEE Trans Biomed Eng* 2020;67:1530–41. DOI
7. Schönberger JL, Frahm JM. Structure-from-motion revisited. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, USA. IEEE; 2016. pp. 4104–13. DOI
8. Liu X, Sinha A, Ishii M, et al. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *IEEE Trans Med Imaging* 2020;39:1438–47. DOI
9. Karaoglu MA, Brasch N, Stollenga M, et al. Adversarial domain feature adaptation for bronchoscopic depth estimation. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2021. Springer, Cham; 2021. pp. 300–10. DOI
10. Shao S, Pei Z, Chen W, et al. Self-supervised monocular depth and ego-motion estimation in endoscopy: appearance flow to the rescue. *Med Image Anal* 2022;77:102338. DOI
11. Wei R, Li B, Mo H, et al. Distilled visual and robot kinematics embeddings for metric depth estimation in monocular scene reconstruction.

In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2022 Oct 23-27; Kyoto, Japan. IEEE; 2022. pp. 8072–7. DOI

12. Colleoni E, Edwards P, Stoyanov D. Synthetic and real inputs for tool segmentation in robotic surgery. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020. Springer, Cham; 2020. pp. 700–10. DOI
13. Long Y, Wu JY, Lu B, et al. Relational graph learning on visual and kinematics embeddings for accurate gesture recognition in robotic surgery. In: 2021 IEEE International Conference on Robotics and Automation (ICRA); 2021 May 30 - Jun 05; Xi'an, China. IEEE; 2021. pp. 13346–53. DOI
14. van Amsterdam B, Funke I, Edwards E, et al. Gesture recognition in robotic surgery with multimodal attention. *IEEE Trans Med Imaging* 2022;41:1677–87. DOI
15. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun ACM* 2021;65:99–106. DOI
16. Niemeyer M, Geiger A. GIRAFFE: representing scenes as compositional generative neural feature fields. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20-25; Nashville, USA. IEEE; 2021. pp. 11448–59. DOI
17. Deng K, Liu A, Zhu JY, Ramanan D. Depth-supervised NeRF: fewer views and faster training for free. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18-24; New Orleans, USA. IEEE; 2022. pp. 12872–81. DOI
18. Wei Y, Liu S, Rao Y, Zhao W, Lu J, Zhou J. NerfingMVS: guided optimization of neural radiance fields for indoor multi-view stereo. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10-17; Montreal, Canada. IEEE; 2021. pp. 5590-9. DOI
19. Rematas K, Liu A, Srinivasan P, et al. Urban radiance fields. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18-24; New Orleans, USA. IEEE; 2022. pp. 12922–32. DOI
20. Wang Y, Long Y, Fan SH, Dou Q. Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2022. Springer, Cham; 2022. pp. 431–41. DOI
21. Kajiya JT, Von Herzen BP. Ray tracing volume densities. *ACM SIGGRAPH Comput Gr* 1984;18:165–74. DOI
22. Lee JH, Han MK, Ko DW, Suh IH. From big to small: multi-scale local planar guidance for monocular depth estimation. arXiv. [Preprint] Sep 23, 2021 [accessed on 2024 Aug 7]. Available from: https://doi.org/10.48550/arXiv.1907.10326.
23. Valentin J, Kowdle A, Barron JT, et al. Depth from motion for smartphone AR. *ACM T Graphic* 2018;37:1–19. DOI
24. Curless B, Levoy M. A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. Association for Computing Machinery; 1996. pp. 303–12. DOI
25. Lorensen WE, Cline HE. Marching cubes: a high resolution 3D surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field. Association for Computing Machinery; 1998. pp. 347–53. DOI
26. Allan M, Mcleod J, Wang C, et al. Stereo correspondence and reconstruction of endoscopic data challenge. arXiv. [Preprint] Jan 28, 2021 [accessed on 2024 Aug 7]. Available from: https://doi.org/10.48550/arXiv.2101.01133.
27. Li Z, Dekel T, Cole F, et al. Learning the depths of moving people by watching frozen people. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15-20; Long Beach, USA. IEEE; 2019. pp. 4516–25. DOI
28. Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. arXiv. [Preprint] Jun 9, 2014 [accessed on 2024 Aug 7]. Available from: https://doi.org/10.48550/arXiv.1406.2283.
29. Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised learning of depth and ego-motion from video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, USA. IEEE; 2017. pp. 6612-9. DOI
30. Ozyoruk KB, Gokceler GI, Bobrow TL, et al. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Med Image Anal* 2021;71:102058. DOI
31. Lu Y, Wei R, Li B, et al. Autonomous intelligent navigation for flexible endoscopy using monocular depth guidance and 3-D shape planning. In: 2023 IEEE international conference on robotics and automation (ICRA); 2023 May 29 - Jun 02; London, UK. IEEE; 2023. p. 1–7. DOI
32. Prendergast JM, Formosa GA, Fulton MJ, Heckman CR, Rentschler ME. A real-time state dependent region estimator for autonomous endoscope navigation. *IEEE Trans Robot* 2021;37:918–34. DOI
33. Cheng X, Zhong Y, Harandi M, Drummond T, Wang Z, Ge Z. Deep laparoscopic stereo matching with transformers. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2022. Springer, Cham; 2022. pp. 464–74. DOI
34. Zhou H, Jagadeesan J. Real-time dense reconstruction of tissue surface from stereo optical video. *IEEE Trans Med Imaging* 2020;39:400–12. DOI
35. Wang J, Suenaga H, Hoshi K, et al. Augmented reality navigation with automatic marker-free image registration using 3-D image overlay for dental surgery. *IEEE Trans Biomed Eng* 2014;61:1295–304. DOI
36. Leonard S, Sinha A, Reiter A, et al. Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on *in vivo* clinical data. *IEEE Trans Med Imaging* 2018;37:2185–95. DOI
37. Recasens D, Lamarca J, Fácil JM, Montiel JMM, Civera J. Endo-depth-and-motion: reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Robot Autom Lett* 2021;6:7225–32. DOI
38. Mahmoud N, Collins T, Hostettler A, Soler L, Doignon C, Montiel JMM. Live tracking and dense reconstruction for handheld monocular endoscopy. *IEEE Trans Med Imaging* 2019;38:79–89. DOI