This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

BASED: Bundle-Adjusting Surgical Endoscopic Dynamic Video Reconstruction using Neural Radiance Fields

Shreya Saha[1] Zekai Liang[1] Shan Lin[1] Jingpei Lu[1] Michael Yip[1] Sainan Liu[2]
[1]University of California, San Diego, USA [2]Intel Labs, USA
{ssaha, z9liang, shl102, jil360, yip}@ucsd.edu sainan.liu@intel.com

Abstract

Reconstruction of deformable scenes from endoscopic videos is important for many applications such as intraoperative navigation, surgical visual perception, and robotic surgery. It is a foundational requirement for realizing autonomous robotic interventions for minimally invasive surgery. However, previous approaches in this domain have been limited by their modular nature and are confined to specific camera and scene settings. Our work adopts the Neural Radiance Fields (NeRF) approach to learning 3D implicit representations of scenes that are both dynamic and deformable over time, and furthermore with unknown camera poses. This work removes the constraints of known camera poses and overcomes the drawbacks of the state-of-the-art unstructured dynamic scene reconstruction technique, which relies on the static part of the scene for accurate reconstruction. Through several experimental datasets, we demonstrate the versatility of our proposed model to adapt to diverse camera and scene settings, and show its promise for both current and future robotic surgical systems.

1. Introduction

Surgical scene reconstruction from endoscopic videos is an important and challenging task in robotic-assisted minimally invasive surgery [50]. It also has widespread use in many downstream clinical applications such as surgical workflow analysis, image-guided robotic surgery, automation, surgical environment simulation, and augmented reality to train surgeons to be familiar with preoperative assessment and planning. Many learning and automation algorithms for robotic surgery are reliant on accurate 3D reconstruction of the surgical scenes [15, 17, 35].

The majority of current approaches for reconstructing surgical scenes from endoscopic videos adopt a modular sequence of tasks, each tackled by independent models [8,13,22,24,27,29,37,43,44]. This fragmented methodology is susceptible to accumulating errors. For instance, a significant number of existing methods either estimate or rely on ground truth camera poses before proceeding with scene reconstruction. Errors in pose estimation often lead to faulty scene reconstructions. Camera poses obtained through structure from motion (SfM) or simultaneous localization and mapping (SLAM) [7, 12, 19, 31, 36] often exhibit noise. When the camera is fixed to the end effector of a robotic arm, camera pose estimation can be accomplished through forward kinematics. But it's prone to errors stemming from in joint angle measurements inaccuracies attributed to instrument sag, cable stretch, positioning biases, and reading drift [30]. Furthermore, in addition to being modular, these methods will not generalize from static to deformable scene settings.

Neural Radiance Fields (NeRF) [26] has shown tremendous performance in the domain of 3D reconstruction. They have been applied recently to stereo-endoscopic videos for deformable surgery scene reconstruction [42, 51]. However, some constraints in these methods are that the camera should be static with known stereo camera information. Alternatively, some works have extended NeRF with camera pose optimization (e.g., [2, 18]). However, these methods generally assume the scene to be static.

More recently, 3D Gaussian Splatting (3DGS) [10] represents the scene with a cloud of 3D Gaussians and estimate the Gaussian parameters based on a efficient rendering algorithm. Similar to NeRF, this technique has been applied to deformable surgical scenes with fixed or known camera poses [20, 53], as well as simultaneous estimation of static scene representations and unknown camera poses [6]. Although 3DGS has achieved state-of-the-art visual rendering quality, it does not capture the 3D geometric information of the scene accurately enough since surgical videos are typically highly under-constrained where the camera movement is relatively small. Moreover, since training of 3DGS primarily focuses on minimizing the difference between rendered and actual images rather than learning accurate 3D geometry of the scene, the geometric accuracy of the Gaussians highly depends on the quality of the point cloud and

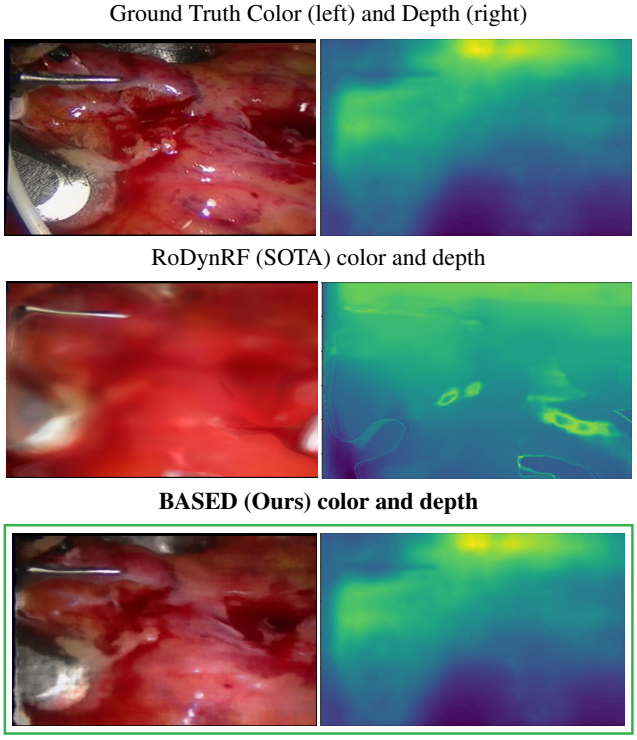


Figure 1. BASED is a novel NeRF-based method that can be used in dynamic and deformable scenes with unknown camera poses. It can produce novel viewpoint renderings with robust color (left) and depth (right) reconstructions, even from monocular untracked camera images. Comparisons with state-of-the-art methods show that it has leading performance in scene reconstruction. Depth in the ground truth row is the reference depth, estimated from [29].

camera poses initialized using off-the-shelf SfM methods like COLMAP [33, 34], which still struggle to handle certain surgical videos. Therefore, we develop our method based on NeRF and demonstrate its improved ability to capture scene geometry through several experiments.

In summary, we present the first NeRF-based method that can reconstruct deformable surgical scenes captured from untracked monocular endoscopic videos. The overall method combines a learnable pose layer that parameterizes camera poses, a deformation module that learns the deformation of a 3D point at a given time step with respect to a canonical position and a rendering module that takes the canonical coordinates of a 3D point at every time step along with the 2D camera directions and outputs the volumetric density and color information. We can also capture more accurate geometric information without compromising on rendering capability for deforming scenes, even in instances where the camera motion is unknown. This makes the method applicable to a wide range of robotic systems and semi-autonomous applications. The major contributions of this paper are thus:

- A NeRF approach for *dynamic and deformable* scenes from data captured at *unknown camera poses*,
- A method for describing multi-view correspondence loss, specifically catered to a dynamic and deformable setting, that enables establishing correspondence across spatial temporal drift in tissues, and
- A framework that offers a generic, unrestricted, and versatile solution to robotic surgical perception that is validated on several surgical scene reconstruction datasets.

2. Related Work

2.1. Traditional Scene Reconstruction Techniques

Structure from Motion (SfM) utilizes a series of two-dimensional images to construct a three-dimensional model of a scene or an object, and it is widely used in the domain of scene/organ/tissue reconstruction with endoscopic images [8, 24, 27, 37, 43, 44]. However, SfM techniques like COLMAP [33, 34] often fail to estimate camera poses for the entire length of sequences. SLAM-based approaches [7, 12, 19, 31, 36] have been employed in the field of endoscopic scene reconstructions to model sparse and dense tissue surfaces. The scene is usually assumed to be static, and the estimated deformations do not always look natural.

2.2. Static and Dynamic NeRF

NeRFs [26] are implicit volumetric representations that encode the appearance and geometry of a 3D scene. It learns a continuous representation of the scene, enabling novel view renderings, unlike SLAM/SfM techniques that usually learn a discrete representation. It takes a 3D point coordinate and a 2D camera viewing direction and outputs RGB color information and the density of the point. The original NeRF model strictly catered to static scenes. D-NeRF [28] is the first work that expanded the NeRF model to learn deformable scenes. It consists of two separate models that progressively learn the deformation and the volumetric scene reconstruction simultaneously. Robust Dynamic NeRF, or RoDynRF [21], tries to reconstruct dynamic scenes by estimating the camera poses using only the static part of the scene, which is not suitable for surgical scene reconstruction given that the scenes are highly deformable. EndoNeRF [42] builds on D-NeRF for endoscopic scenes by adding a tool-guided ray casting and by sampling points along a ray closer to the tissue surfaces using a Gaussian transfer function but relying on stereo depth. EndoSurf [51] further improved upon EndoNeRF [42] by employing three networks to learn the deformation, signed distance function (SDF), and color of a viewpoint. However, as seen in Figure 1, these methods strictly cater to cases where the camera is static. If extended to moving camera sequences with camera poses estimated by COLMAP, the rendering quality degrades significantly.

2.3. Static and Dynamic Gaussian Splatting

3DGS is a fast rendering technique for novel view synthesis. It uses 3D Gaussians for scene representation and a novel tile-based rasterizer for fast 3D Gaussian rendering. Given a set of camera poses and their corresponding 2D views, a cloud of 3D Gaussian representation can be learned end-to-end. While the original work focused on static scenes, dynamic and deformable scenes have also been explored in recent months.

Dynamic 3D Gaussians [23] builds a static scene using 3DGS for the first frame, then incrementally learns the offset for each 3D Gaussian. More recently, 3DGStream [38] has improved the efficiency of offset cache storage and enabled new Gaussian spawning. However, these models can only work with multiview images and rely heavily on the information from the first frame.

Yang *et al.* [47] and Wu *et al.* [45] both use a deformation field to model offsets to a canonical space of explicit 3D Gaussians, whereas [48] has increased GS dimension into 4D. However, these methods assume ground truth camera poses are known. Our model does not have this assumption.

2.4. Joint Pose Estimation and Scene Reconstruction

Bundle adjusting neural radiance fields (BARF) [16] first proposed an end-to-end NeRF-based framework that can jointly estimate camera extrinsics and reconstruct the 3D scene. It implements a coarse-to-fine technique to gradually activate the higher frequency components of the positional encoding scheme that aids in pose refinement, and only uses the photometric loss for backpropagation through the network. There have been numerous follow-up works ([1–4, 9, 18, 25, 41, 46]) that improved BARF further. However, all these methods specifically deal with static scenes. RoDynRF has firstly implemented estimating camera poses while simultaneously reconstructing dynamic scenes. However, RoDynRF requires the dynamic pixels to be masked out prior to training. This is hard to acquire as the entire surgical scene may be deformable and textureless. Hence, as seen in Figure 1, it cannot be generalized for endoscopic videos. Our proposed model overcomes the limitations of EndoNeRF and EndoSurf by expanding to moving cameras with unknown poses. It does not constrain part of the scene to be static for it to be able to learn deformations, thus also overcoming the limitations of RoDynRF.

3. Methodology

Figure 2 gives a high-level overview of our method. The overall model consists of three parts: (1) the pose estimation module (see 3.1), (2) the deformable module (see 3.2), and (3) a canonical NeRF module (see 3.3).

3.1. Camera Pose Module

We use a camera pose module to model the appearance of the rigid-body motion of the scene as is shown in Figure 2 (A). We follow the convention of previous work [16] and model the camera pose with $P_i = [R_i|t_i] \in SE(3)$ representing camera-to-world transform of camera i, where $R_i \in SO(3)$ denotes the rotation and $t_i \in \mathbb{R}^3$ denotes the translation. We use $K \in \mathbb{R}^{3\times3}$ to represent the intrinsic matrix.

In terms of implementation details, the poses are parameterized as a learnable parameter $P_i \in \mathbb{R}^9$ for $i \in [1, N]$ for N frames. The first six columns of the matrix refer to the first two rows of the camera rotation matrix, and the last three values refer to the camera translation vector. The third row of the rotation matrix can be estimated by taking a cross-product of the other two rows. The pose layer is initialized to identity and updated for the first 200 iterations, after which the layer is frozen.

3.2. Deformation Module

After the camera pose module explains away the rigid-body motion, the deformation module will model the non-rigid-body deformation of the scene from a canonical state at t_0 as is shown in Figure 2 (B). The deformation module $\psi_\Delta(x, t)$ takes as input 3D spatial coordinates $x \in \mathbb{R}^3$ and a time step t and learns the deformation Δx of the coordinate x from time step zero:

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

If x_0 is the coordinate of a point at time $t = 0$ and it deforms to x_t at time t, then we can say that $\psi_\Delta(x_t, t) + x_t = x_0$.

3.3. Canonical NeRF Module

This module will capture the implicit 3D representation of the canonical state of the scene. The canonical module ψ_x can be assumed to be a vanilla NeRF model, which takes as input a spatial 3D location and 2D camera orientations, and outputs color (RGB values) and the volumetric density of the location. x_0 calculated above is the input to the canonical model along with camera viewing directions d. The rendered output of the canonical model (color c, and density σ) is corresponding to the deformed point x_t at time t.

$$\psi_x(x_t + \psi_\Delta(x_t, t), d) = c, \sigma \tag{2}$$

We apply conventional positional encoding [28] to all the inputs, $< (sin(2^l\pi p), cos(2^l\pi p)) >_0^L$, with $L = 10$ for 3D point coordinates x and time frame t, and $L = 4$ for camera directions d.

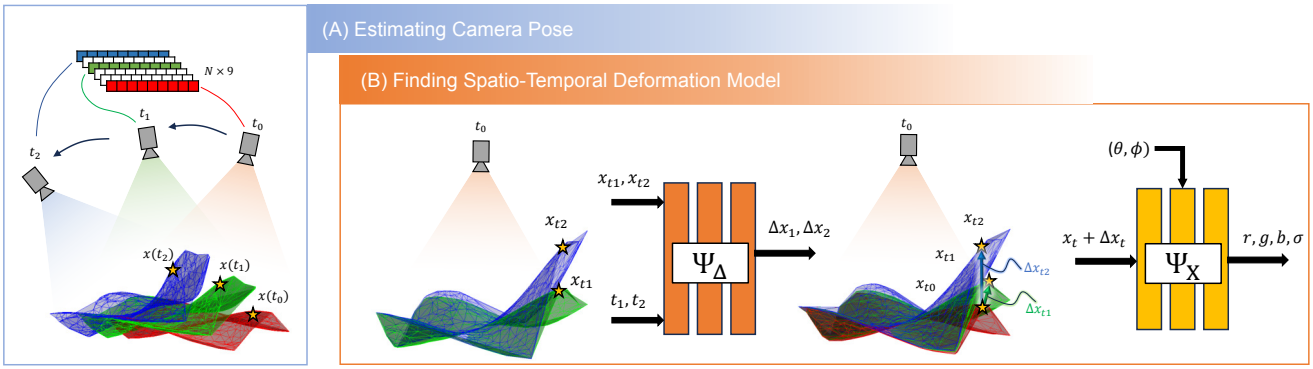


Figure 2. Method overview: (A) The first part of the diagram shows a tissue that appears to be dynamic and deformable at three timestamps t_0, t_1, and t_2. Camera poses are estimated by a $N \times 9$ learnable matrix through the camera pose estimation layer. $x(t_0), x(t_1)$ and $x(t_2)$ shows the trajectory taken by the same point x on the tissue at these three different timesteps. (B) x_{t_1} and x_{t_2} represents non-rigid deformation trajectory of x at time t_1 and t_2. The deformation model denoted byψ_Δ takes in x_{t_1} and x_{t_2} and predicts their displacement from the canonical configuration of the scene. The canonical model ψ_x then takes in the mapped canonical position of the point $x_{t_0} = x_{t_1} + \Delta x_{t_1} = x_{t_2} + \Delta x_{t_2}$ along with 2D camera directions, and predicts the color and density information.

3.4. Tool mask-guided ray casting and Stereo Depth Cue Ray Marching

Surgical tools occlude part of the scene in most of our sequences. Instead of shooting rays randomly all through the image, we adopt a tool mask-guided ray casting strategy [42]. Using binary tool masks, an importance map is created which highlights the regions containing tool pixels. During training, instead of random ray sampling throughout the image, we only sample pixels from this distribution using inverse transform sampling. Secondly, we sample pixels close to the tissue surface using a Gaussian transfer function and use a depth refinement strategy to get rid of corrupt artifacts in the estimated depths.

3.5. Losses

Our model uses photometric loss, correspondence loss, and estimated depth Loss for monocular dynamic camera deformable scene reconstruction.

RGB Photometric Loss: We use photometric loss based on conventional setup in NeRF models [26]. At every training iteration, we compare the rendered color of the sampled pixels with the ground truth colors using the L_2 loss, and this loss is backpropagated through the entire network with all three modules.

$$L_{pho} = \frac{1}{N}\Sigma_{i=1}^{N}||\hat{C}_i(p,t) - C_i(p,t)||_2^2 \tag{3}$$

where $\hat{C}_i$ is the rendered color and C_i is the ground truth color of the pixel p at time t with respect to frame i, $i = 0...N$.

Dynamic Multi-View Correspondence Loss: We introduce a novel 3D multi-view correspondence loss for dynamic scenes to offer extra constraints on pose estimation. Figure 3 gives a high-level idea of this loss. For a given pair of images, I_i^a (viewpoint P_i at time t_a) and I_j^b (viewpoint P_j at time t_b) with matching pixels $p^a \in I_i^a$ and $q^b \in I_j^b$, both pixels p^a and q^b should project to the same point in 3D space in the canonical space. z_p^a and z_q^b are the estimated depths of the pixels p^a and q^b respectively. Let f_P^{-1} be the function that maps the 3D coordinate of a pixel given its estimated depth and its respective camera viewpoint. Let $x_p^a \in \mathbb{R}^3$ be the 3D coordinate of pixel p^a at time t_a and $x_q^b \in \mathbb{R}^3$ be the 3D coordinate of pixel q^b at time t_b. Then, the estimated coordinates are

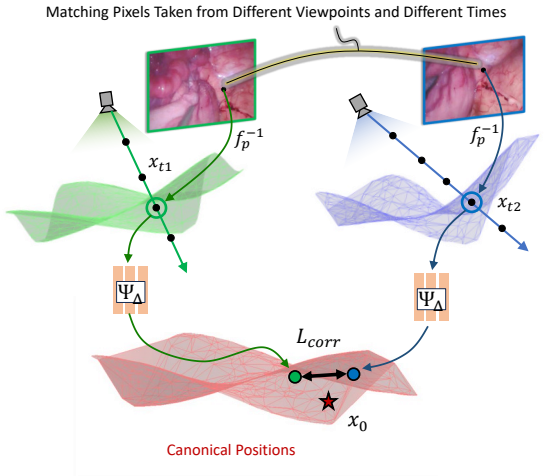


Figure 3. Overview of the correspondence loss: Matching pixels are chosen from a random pair of images (having correspondence above a certain threshold) using PDC-Net [39]. The matching pixels are projected from 2D pixel space to their 3D positions in the world frame, and then further passed through the deformation model to get their canonical 3D positions. The difference between the mapped 3D points is used to calculate the flow correspondence loss. The loss is designed to pull the outputs of correspondence points returned by the deformation model closer together towards the same canonical point denoted as x_0.

$$x_p^a = f_{P_i}^{-1}(p^a, z_p^a) \tag{4}$$

Table 1. Ablation studies on Hamlyn Rectified18-1 dataset with dynamic motions. ✓ indicates loss usage. BA stands for bundle-adjusting. EndoNeRF experiment is using COLMAP for camera pose estimation.

	BA	L_d	L_{corr}	PSNR↑	SSIM↑	LPIPS↓
EndoNeRF [42]	COLMAP	✓	-	26.97	0.895	0.164
BASED w/o $L_d L_{corr}$	✓	-	-	30.236	0.925	0.124
BASED w/o L_d	✓	-	✓	31.415	0.935	0.108
BASED w/o L_{corr}	✓	✓	-	31.403	0.937	**0.102**
BASED (final)	✓	✓	✓	**32.227**	**0.944**	0.106

$$x_q^b = f_{P_j}^{-1}(q^b, z_q^b) \tag{5}$$

If there was no deformation, x_p^a and x_q^b should be the same point in 3D space. However, since this is a deformable scene, we need to map both the points back to their canonical state in order to calculate the loss:

$$\hat{x}_p^a = \psi_\Delta(x_p^a, t_a) + x_p^a \tag{6}$$

$$\hat{x}_q^b = \psi_\Delta(x_q^b, t_a) + x_q^b \tag{7}$$

Then, the loss becomes:

$$L_{corr} = w(L_\delta(\hat{x}_p^a, \hat{x}_q^b)) \tag{8}$$

where w is loss weight term and L_δ is the Huber loss function.

Estimated Depth Loss: In spite of only having 2D images, we can use an *estimated* depth loss to further assist better reconstruction when the camera pose estimation is optimized. The depth estimates can be found via an off-the-shelf image-to-depth estimator. This loss is only used to optimize the deformable and canonical modules after the pose layer is frozen. Applying the estimated depth loss before the camera pose matrix is optimized leads to abrupt artifacts. We collect depths estimated from RGB images via existing pre-trained methods [14], [29] during training. We compare the rendered depth values of the sampled pixels with the corresponding estimated depth values using a L_2 loss. The gradient from this loss is only backpropagated through the deformable NerRF modules, after the pose layer is frozen.

$$L_d = \frac{1}{N} \Sigma_{i=1}^{N} ||\hat{D}_i(p,t) - D_i(p,t)||_2^2 \tag{9}$$

where $\hat{D}_i(p,t) = \frac{1}{\sigma_i(p,t)}$ is the rendered depth, σ_i is the rendered density and $D_i(p,t)$ is the reference depth of the pixel p at time t with respect to frame $i \in [1, N]$ for N frames. Note that if a hardware solution to depth is available (e.g., stereo cameras or RGB-D sensor, the ground truth depth can be used as well to calculate an estimated depth loss.

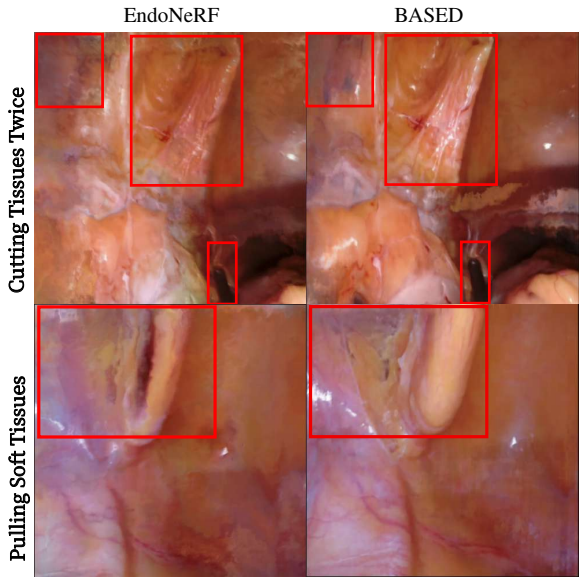


Figure 4. Qualitative analysis of our proposed model BASED and EndoNeRF on "Cutting Tissues Twice" dataset for novel view renderings. Highlighted sections show how BASED is able to avoid different artifacts in the images, resulting in a sharper image.

4. Experiments

4.1. Implementation Details

All of the experiments have been performed on a single Nvidia RTX-A6000 GPU. The weights of the NeRF network and the pose estimation network are simultaneously updated with a learning rate of 5e-4. After 200 iterations, the pose layer is fixed, and only the NeRF model is updated for another 100K iterations. In the first stage, one batch consists of all of the training images, because it requires more views of the scene for better pose estimation. In the second stage, one batch consists of a single training image for faster training. Tool masks are manually created for the Hamlyn dataset sequences, partly using Segment Anything Model [11]. We use an off-the-shelf dense pixel correspondence network, PDCNet [40], to obtain corresponding matching pixels between a pair of images for calculating the multi-view correspondence loss L_{corr}.

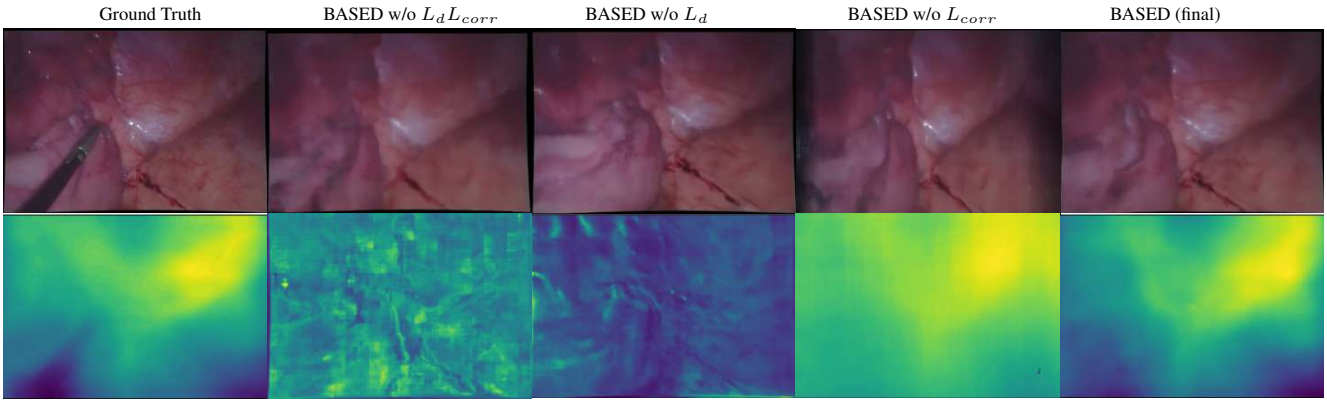


Figure 5. Ablation analysis on Hamlyn Rectified 18-1 dataset shows how different losses bring about an improvement in the rendered results from the final model. Please note that the depth in the Ground Truth column is the reference depth.

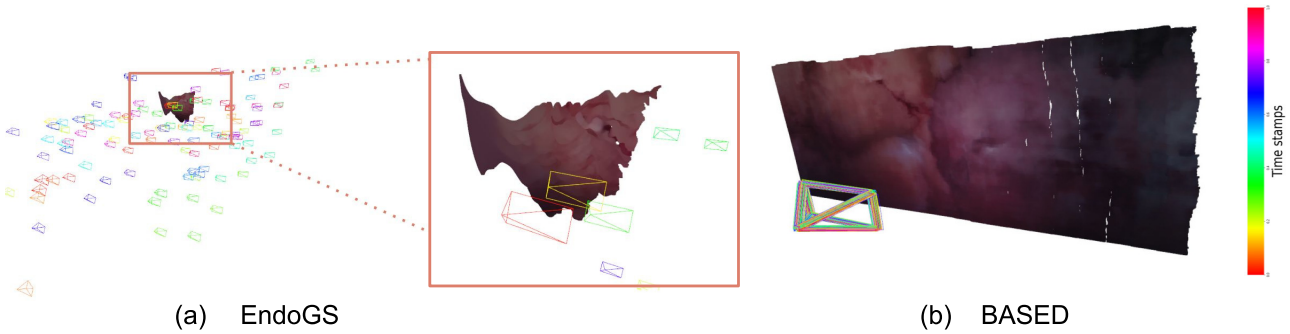


Figure 6. Camera poses and point cloud visualization of (a) EndoGS and (b) BASED on hamlyn dataset Rectified18-2. The colors assigned to the cameras correspond to their timestamps in the video.

4.2. Datasets

We evaluated BASED on the following datasets:

Hamlyn Dataset *[49]*: We choose three videos from the Hamlyn dataset of robotic surgery with deformable scenes. Hamlyn dataset contains weak textures, and the scene is often occluded by blood and surgical instruments. Rectified04 contains 158 frames, and it appears to have slight camera motion. The camera in Rectified04 is mostly static, whereas Rectified18 exhibits slightly more camera motion, so two subsets of Rectified18 are chosen from the whole video with 113 frames in each. The frames are sampled at regular intervals (every 10 frames), as the original videos were too large. 10 percent of the total images were used for testing. We use [29] to generate reference depth needed for calculating the estimated depth loss.

EndoNeRF da Vinci dataset *[42]*: We used two datasets from the EndoNeRF paper that are publicly available. They last for 4-8 sec with 15 fps. Each of the videos is taken from a single static viewpoint setting. One of the videos demonstrates soft tissues being drastically pulled, while the other video demonstrates soft tissue cutting with challenging topological changes. As mentioned in the original paper, the depth maps are obtained using STTR-LIGHT [14] from stereo images.

4.3. Metrics

The quality of image rendering is evaluated through commonly used metrics, including (1) Peak Signal-to-Noise Ratio (PSNR), (2) Learned Perceptual Image Patch Similarity (LPIPS) [52], and (3) Structural Similarity Index (SSIM) [52]. We also quantitatively evaluate the 3D reconstruction performance by comparing predicted depth maps with the reference depth (refer to Section 4.2 for details), using standard metrics for depth evaluation [5]. With this, we are treating the reference as pseudo ground truth. It should be noted that while the quality this pseudo ground truth may be slightly inferior than that of true depth, using it is still a common strategy in endoscopic video analysis as true ground truth can be hard to obtain in a surgical environment [32].

4.4. Results

4.4.1 Ablation Study

Table 1 and Figure 5 provide both quantitative and qualitative analysis of the effect of the various losses. As shown in

Table 2. Quantitative analysis of BASED and State-of-the-Art on the Hamlyn Dataset.

	PSNR ↑				SSIM ↑				LPIPS ↓			
	EndoNeRF	RoDynRF	EndoGS	Ours	EndoNeRF	RoDynRF	EndoGS	Ours	EndoNeRF	RoDynRF	EndoGS	Ours
Rectified04-1 (scds)	-	<u>14.902</u>	-	**20.947**	-	<u>0.468</u>	-	**0.621**	-	<u>0.577</u>	-	**0.269**
Rectified18-1 (mcds)	22.838	25.700	<u>30.821</u>	**32.277**	0.817	0.874	<u>0.928</u>	**0.944**	0.273	0.192	**0.100**	<u>0.106</u>
Rectified18-2 (mcds)	29.572	24.690	**35.969**	<u>33.017</u>	0.869	0.875	**0.956**	<u>0.936</u>	0.235	0.201	**0.079**	<u>0.103</u>

'-' denotes that the model fails to complete the task.
Best results are **bolded**; second best are <u>underlined</u>.

Table 3. Quantitative analysis of BASED and state-of-the-art on the EndoNeRF da Vinci dataset.

	PSNR ↑				SSIM ↑				LPIPS ↓			
	EndoNeRF	RoDynRF	EndoGS	Ours	EndoNeRF	RoDynRF	EndoGS	Ours	EndoNeRF	RoDynRF	EndoGS	Ours
pulling soft tissues	<u>35.832</u>	27.502	**36.750**	35.097	**0.941**	0.900	0.925	<u>0.936</u>	**0.058**	0.158	0.135	<u>0.059</u>
cutting tissues twice	<u>35.720</u>	21.105	**37.372**	35.706	<u>0.935</u>	0.700	**0.964**	0.934	<u>0.063</u>	0.280	**0.050**	0.073

Best results are **bolded**; second best are <u>underlined</u>.

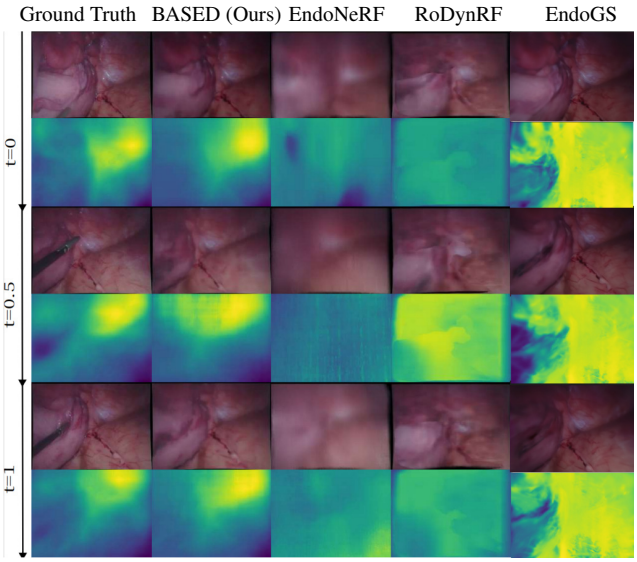


Figure 7. Qualitative analysis of our proposed model BASED compared with RoDynRF, EndoNeRF and EndoGS (with COLMAP estimated poses) for sequences obtained from the public Hamlyn dataset. Please note that the depth in the Ground Truth column is the reference depth.

Table 1, bundle-adjusting leads to significant improvement in image rendering quality compared with EndoNeRF [42]. Also, training with estimated depth guidance L_d and correspondence loss L_{corr} improves rendering quality, respectively. Figure 5 shows that estimated depth guidance clearly improves the reconstructed textures and depths. Additionally, correspondence loss L_{corr} leads to better results, as shown quantitatively by image rendering results and qualitatively by predicted depth quality, indicating that the correspondence loss may have improved the estimated camera poses by instilling multi-view consistency. However, the sequences are still severely underconstrained (camera movement/viewing angles of the scene are still limited in those sequences, which is common in surgical scenes), hence the correspondence loss only showed a relatively slight improvement in depth reconstruction. The tool masks for sequence - Rectified04-1 from the Hamlyn dataset does not always completely mask out the tools, which is why the resultant rendered images partially have tools in them (Figure 1), and not a transparent overlay like we see for the other rendered results.

4.4.2 Comparison on Hamlyn Dataset

The main dataset we are evaluating is the Hamlyn dataset. We compare BASED with state-of-the-art methods including EndoNeRF, RoDynRF and EndoGS, for deformable scene reconstruction with obvious camera motions. The image rendering are quantitatively and qualitatively compared in Table 2 and Figure 7, respectively. The 3D scene reconstruction performance is evaluated through predicted depth maps, as shown in Figure 7. Our model outperforms both EndoNeRF and RoDynRF by a significant margin in both image rendering and 3D reconstruction.

On the other hand, BASED is still competitive in image rendering compared with EndoGS, which is based on 3DGS. 3DGS has inherent advantages in image rendering, as the color and position of the Gaussians can be easily updated to fit the camera view. However, the depth estimated by EndoGS may not accurately represent the true geometric shape of the environment. To demonstrate the superiority of our method in capturing the 3D geometric information of the scene, beyond Figure 7, we also quantitatively measure

Table 4. Quantitative analysis of BASED and EndoGS depth estimation on the Hamlyn Rectified18-2 dataset.

	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE Log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
EndoGS	3.834	1.149	2.938	4.225	3.356	6.908	10.930
BASED	**1.124**	**0.228**	**1.790**	**1.341**	**10.227**	**11.000**	**11.000**

Table 5. Quantitative analysis of BASED and EndoGS depth estimation on the EndoNeRF "Cutting Tissues Twice" dataset.

	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE Log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
EndoGS	4.644×10^6	1.643×10^6	**54.342**	622.943	52.230	69.472	85.756
BASED	**47.931**	**27.107**	65.810	**140.360**	**91.750**	**92.898**	**93.772**

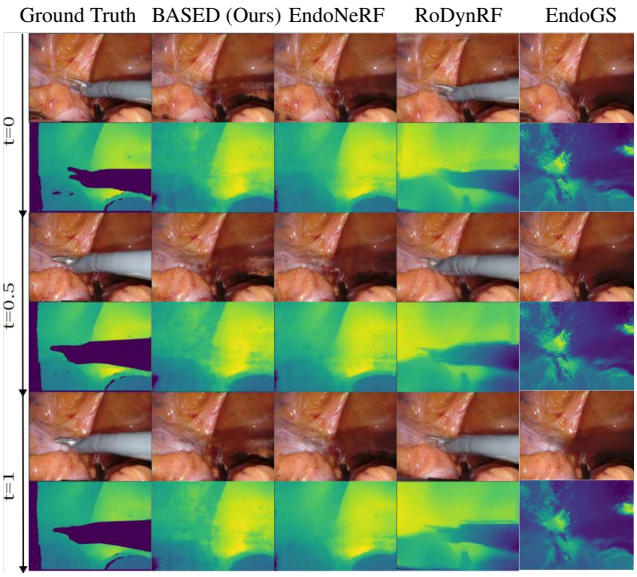


Figure 8. Qualitative analysis of our proposed model BASED with RoDynRF and EndoNeRF for two publicly available sequences - "Cutting Tissues Twice" [42] guided by stereo depth.

the quality of predicted depth based on the reference depth, as shown in Table 4. BASED provides better depth maps both quantitatively and qualitatively. Furthermore, we visualize the camera pose along with the point cloud estimated by EndoGS and BASED in Figure 6. Given the small camera motion in this dataset, BASED aligns more closely with this pattern compared to EndoGS. A major drawback of EndoNeRF and EndoGS is that they have only been tested with single-view static cameras, which are severely underconstrained. In dynamic scenes, both models use COLMAP to estimate camera poses.

Thus, these two models fail when COLMAP fails to predict camera poses as seen for Rectified 04-1 in Table 2. This clearly demonstrates the superiority of our method. Also, even in cases where COLMAP is able to estimate camera poses, our model, BASED, more accurately captures the dynamic and deformable aspects of the scene, as shown by the quantitative and qualitative results provided.

4.4.3 Comparison on EndoNeRF da Vinci Dataset

We further show that our model can generalize toward deformable surgical scenes with stationary cameras on the sequences: Cutting tissues twice and Pulling soft tissues as shown in Table 3 and Figure 8. We can see that our model is on par, if not better than EndoNeRF, and performs significantly better than RoDynRF. In Figure 4, our model can generalize better novel view renderings compared to EndoNeRF which has been strictly trained with static identity camera poses. Hence, our model retains EndoNeRF's capacity for static deformable monocular videos.

Furthermore, we still compare our model with EndoGS on static surgical scenes. From Table 5 and Figure 8, both qualitative and quantitative results show although EndoGS shows its inherent advantage in image rendering, it still almost fails in geometric shape capturing even on static camera scenes, which is very crucial for real-world applications.

5. Conclusion

We propose BASED, a network that simultaneously estimates dynamic scene motions and reconstructs deformable surgical scenes from monocular videos. This is the first work that tries to reconstruct scenes from video sequences that contain mixed static and dynamic parts with unknown monocular camera poses. Through our experiments, we have shown significant improvement over the existing methods, and also potential for future research. Limitations of the presented model include slow optimization time (similar to EndoNeRF at several hours), though many recent strategies are drastically cutting this time by orders of magnitude (including ours [18]). Some potential future directions could be removing the dependency of BASED on reference depths, reducing optimization time, and automatic tool mask generation.

References

[1] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. NoPe-NeRF: Optimising neural radiance field with no pose prior. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 4160–4169, 2023.

[2] Yu Chen and Gim Hee Lee. DBARF: Deep bundle-adjusting generalizable neural radiance fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 24–34, June 2023.

[3] Zezhou Cheng, Carlos Esteves, Varun Jampani, Abhishek Kar, Subhransu Maji, and Ameesh Makadia. LU-NeRF: Scene and pose estimation by synchronizing local unposed NeRFs, 2023.

[4] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *Eur. Conf. Comput. Vis.*, 2022.

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Adv. Neural Inf. Process. Syst.*, 27, 2014.

[6] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. COLMAP-Free 3D gaussian splatting. *arXiv preprint arXiv:2312.07504*, 2023.

[7] Oscar Garcia-Grasa, Ernesto Bernal, Santiago Casado, Ismael Gil, and J. Montiel. Visual SLAM for handheld monocular endoscope. *IEEE Trans. Med. Imag.*, 33:135—-146, 09 2013.

[8] Mingxing Hu, Graeme Penney, Michael Figl, Philip Edwards, Fernando Bello, Roberto Casula, Daniel Rueckert, and David Hawkes. Reconstruction of a 3D surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes. *Med. Image Anal.*, 2012.

[9] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Animashree Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Int. Conf. Comput. Vis.*, 2021.

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):1–14, 2023.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv*, 2023.

[12] José Lamarca, Shaifali Parashar, Adrien Bartoli, and José M. M. Montiel. DefSLAM: Tracking and mapping of deforming scenes from monocular sequences. *IEEE Trans. Robot.*, 37(1):291–303, February 2021.

[13] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Int. Conf. Comput. Vis.*, pages 6197–6206, October 2021.

[14] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021.

[15] Xiao Liang, Fei Liu, Yutong Zhang, Yuelei Li, Shan Lin, and Michael Yip. Real-to-sim deformable object manipulation: Optimizing physics models with residual mappings for robotic surgery. *arXiv preprint arXiv:2309.11656*, 2023.

[16] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In *Int. Conf. Comput. Vis.*, 2021.

[17] Fei Liu, Zihan Li, Yunhai Han, Jingpei Lu, Florian Richter, and Michael C Yip. Real-to-sim registration of deformable soft tissue with position-based dynamics for surgical robot autonomy. In *IEEE Int. Conf. Robot. Autom.*, pages 12328–12334. IEEE, 2021.

[18] Sainan Liu, Shan Lin, Jingpei Lu, Shreya Saha, Alexey Supikov, and Michael Yip. BAA-NGP: Bundle-adjusting accelerated neural graphics primitives. *arXiv*, 2023.

[19] Xingtong Liu, Zhaoshuo Li, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, and Mathias Unberath. SAGE: SLAM with appearance and geometry prior for endoscopy. In *IEEE Int. Conf. Robot. Autom.*, pages 5587–5593, 2022.

[20] Yifan Liu, Chenxin Li, Chen Yang, and Yixuan Yuan. EndoGaussian: Gaussian splatting for deformable surgical scene reconstruction. *arXiv preprint arXiv:2401.12561*, 2024.

[21] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[22] Yonghao Long, Zhaoshuo Li, Chi Hang Yee, Chi Fai Ng, Russell H Taylor, Mathias Unberath, and Qi Dou. E-DSSR: Efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 415–425. Springer, 2021.

[23] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024.

[24] Kristen L. Lurie, Roland Angst, Dimitar V. Zlatev, Joseph C. Liao, and Audrey K. Ellerbee Bowden. 3D reconstruction of cystoscopy videos for comprehensive bladder records. *Biomed. Opt. Expr.*, pages 2106–2123, 2017.

[25] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Eur. Conf. Comput. Vis.*, 2020.

[27] Tan-Binh Phan, Dinh-Hoan Trinh, Didier Wolf, and Christian Daul. Optical flow-based structure-from-motion for the reconstruction of epithelial surfaces. *Pattern Recognition*, 105:107391, 2020.

[28] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NERF: Neural radiance fields

for dynamic scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10318–10327, 2021.

[29] David Recasens, José Lamarca, José M Fácil, JMM Montiel, and Javier Civera. Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. *IEEE Int. Conf. Robot. Autom. Letters*, 6(4):7225–7232, 2021.

[30] Florian Richter, Jingpei Lu, Ryan K Orosco, and Michael C Yip. Robotic tool tracking under partially visible kinematic chain: A unified approach. *IEEE Trans. Robot.*, 38(3):1653–1670, 2021.

[31] Juan J Gómez Rodríguez, José Lamarca, Javier Morlana, Juan D Tardós, and José MM Montiel. SD-DefSLAM: Semi-direct monocular slam for deformable and intracorporeal scenes. In *IEEE Int. Conf. Robot. Autom.*, pages 5170–5177, 2021.

[32] Adam Schmidt, Omid Mohareri, Simon DiMaio, Michael C Yip, and Septimiu E Salcudean. Tracking and mapping in medical computer vision: A review. *Med. Image Anal.*, page 103131, 2024.

[33] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[34] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proc. European Conf. Comput. Vis.*, 2016.

[35] Nikhil Uday Shinde, Xiao Liang, Fei Liu, Yutong Zhang, Florian Richter, Sylvia Herbert, and Michael C Yip. Jiggle: An active sensing framework for boundary parameters estimation in deformable surgical environments. *arXiv preprint arXiv:2405.09743*, 2024.

[36] Jingwei Song, Jun Wang, Liang Zhao, Shoudong Huang, and Gamini Dissanayake. MIS-SLAM: Real-time large scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing. *IEEE Int. Conf. Robot. Autom. Letters*, PP:4068–4075, 03 2018.

[37] Timothy D. Soper, Michael P. Porter, and Eric J. Seibel. Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance. *IEEE Trans. Biomed. Engi.*, pages 1670–1680, 2012.

[38] Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 3DGStream: On-the-fly training of 3D gaussians for efficient streaming of photo-realistic free-viewpoint videos. *arXiv preprint arXiv:2403.01444*, 2024.

[39] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 5714–5724, 2021.

[40] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 5714–5724, 2021.

[41] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. SPARF: Neural radiance fields from sparse and noisy poses. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[42] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. In *Med. Imag. Comput. Comput. Assis. Interv.*, pages 431–441. Springer, 2022.

[43] Aji Resindra Widya, Yusuke Monno, Kosuke Imahori, Masatoshi Okutomi, Sho Suzuki, Takuji Gotoda, and Kenji Miki. 3d reconstruction of whole stomach from endoscope video using structure-from-motion. In *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., 2019, pp. 3900-3904.*

[44] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki. Whole stomach 3D reconstruction and frame localization from monocular endoscope video. *IEEE J. Transl. Eng. Health Med.*, 7:1–10, 2019.

[45] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023.

[46] Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool. SiNeRF: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction. In *Proc. 33rd Brit. Mach. Vis. Conf., London, UK, Nov. 21-24, 2022*.

[47] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D Gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023.

[48] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4D gaussian splatting. In *Int. Conf. Learn. Represent.*, 2024.

[49] Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang-Zhong Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *arXiv preprint arXiv:1705.08260*, 2017.

[50] Michael Yip, Septimiu Salcudean, Ken Goldberg, Kaspar Althoefer, Arianna Menciassi, Justin D Opfermann, Axel Krieger, Krithika Swaminathan, Conor J Walsh, He Huang, et al. Artificial intelligence meets medical robotics. *Science*, 381(6654):141–146, 2023.

[51] Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge. Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In *Med. Imag. Comput. Comput. Assis. Interv.*, pages 13–23. Springer, 2023.

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 586–595, 2018.

[53] Lingting Zhu, Zhao Wang, Jiahao Cui, Zhenchao Jin, Guying Lin, and Lequan Yu. EndoGS: Deformable endoscopic tissues reconstruction with gaussian splatting.