

Neural Rendering for Stereo 3D Reconstruction of Deformable Tissues in Robotic Surgery

Yuehao Wang¹, Yonghao Long¹, Siu Hin Fan², and Qi Dou¹(✉)

¹ Dept. of Computer Science and Engineering, The Chinese University of Hong Kong

² Dept. of Biomedical Engineering, The Chinese University of Hong Kong

Abstract. Reconstruction of the soft tissues in robotic surgery from endoscopic stereo videos is important for many applications such as intra-operative navigation and image-guided robotic surgery automation. Previous works on this task mainly rely on SLAM-based approaches, which struggle to handle complex surgical scenes. Inspired by recent progress in neural rendering, we present a novel framework for deformable tissue reconstruction from binocular captures in robotic surgery under the single-viewpoint setting. Our framework adopts dynamic neural radiance fields to represent deformable surgical scenes in MLPs and optimize shapes and deformations in a learning-based manner. In addition to non-rigid deformations, tool occlusion and poor 3D clues from a single viewpoint are also particular challenges in soft tissue reconstruction. To overcome these difficulties, we present a series of strategies of tool mask-guided ray casting, stereo depth-cueing ray marching and stereo depth-supervised optimization. With experiments on DaVinci robotic surgery videos, our method significantly outperforms the current state-of-the-art reconstruction method for handling various complex non-rigid deformations. To our best knowledge, this is the first work leveraging neural rendering for surgical scene 3D reconstruction with remarkable potential demonstrated. Code is available at: <https://github.com/med-air/EndoNeRF>.

Keywords: 3D Reconstruction · Neural Rendering · Robotic Surgery.

1 Introduction

Surgical scene reconstruction from endoscope stereo video is an important but difficult task in robotic minimally invasive surgery. It is a prerequisite for many downstream clinical applications, including intra-operative navigation and augmented reality, surgical environment simulation, immersive education, and robotic surgery automation [2,12,20,25]. Despite much recent progress [10,22,28,29,30,33], several key challenges still remain unsolved. First, surgical scenes are deformable with significant topology changes, requiring dynamic reconstruction to capture a high degree of non-rigidity. Second, endoscopic videos show sparse viewpoints due to constrained camera movement in confined space, resulting in limited 3D clues of soft tissues. Third, the surgical instruments always occlude part of the soft tissues, which affects the completeness of surgical scene reconstruction.

Previous works [1,13] explored the effectiveness of surgical scene reconstruction via depth estimation. Since most of the endoscopes are equipped with stereo cameras, depth can be estimated from binocular vision. Follow-up SLAM-based methods [23,31,32] fuse depth maps in 3D space to reconstruct surgical scenes under more complex settings. Nevertheless, these methods either hypothesize scenes as static or surgical tools not present, limiting their practical use in real scenarios. Recent work SuPer [8] and E-DSSR [11] present frameworks consisting of tool masking, stereo depth estimation and SurfelWarp [4] to perform single-view 3D reconstruction of deformable tissues. However, all these methods track deformation based on a sparse warp field [16], which degenerates when deformations are significantly beyond the scope of non-topological changes.

As an emerging technology, neural rendering [6,27,26] is recently developed to break through the limited performance of traditional 3D reconstruction by leveraging differentiable rendering and neural networks. In particular, neural radiance fields (NeRF) [15], a popular pioneering work of neural rendering, proposes to use *neural implicit field* for continuous scene representations and achieves great success in producing high-quality view synthesis and 3D reconstruction on diverse scenarios [14,15,17]. Meanwhile, recent variants of NeRF [18,19,21] targeting dynamic scenes have managed to track deformations through various neural representations on non-rigid objects.

In this paper, we endeavor to reconstruct highly deformable surgical scenes captured from single-viewpoint stereo endoscopes. We embark on adapting the emerging neural rendering framework to the regime of deformable surgical scene reconstruction. We summarize our contributions as follows: 1) To accommodate a wide range of geometry and deformation representations on soft tissues, we leverage neural implicit fields to represent dynamic surgical scenes. 2) To address the particular tool occlusion problem in surgical scenes, we design a new mask-guided ray casting strategy for resolving tool occlusion. 3) We incorporate a depth-cueing ray marching and depth-supervised optimization scheme, using stereo prior to enable neural implicit field reconstruction for single-viewpoint input. To the best of our knowledge, this is the first work introducing cutting-edge neural rendering to surgical scene reconstruction. We evaluate our method on 6 typical in-vivo surgical scenes of robotic prostatectomy. Compared with previous methods, our results exhibit great performance gain, both quantitatively and qualitatively, on 3D reconstruction and deformation tracking of surgical scenes.

2 Method

2.1 Overview of the Neural Rendering-based Framework

Given a single-viewpoint stereo video of a dynamic surgical scene, we aim to reconstruct 3D structures and textures of surgical scenes without occlusion of surgical instruments. We denote $\{(\mathbf{I}_i^l, \mathbf{I}_i^r)\}_{i=1}^T$ as a sequence of input stereo video frames, where T is the total number of frames and $(\mathbf{I}_i^l, \mathbf{I}_i^r)$ is the pair of left and right images at the i -th frame. The video duration is normalized to $[0, 1]$. Thus, time of the i -th frame is i/T . We also extract binary tool masks $\{\mathbf{M}_i\}_{i=1}^T$ for the

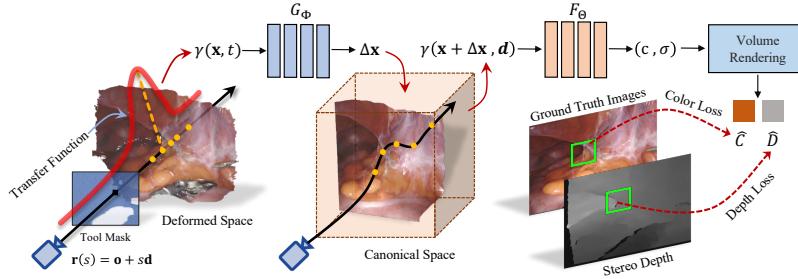


Fig. 1: Illustration of our proposed novel approach of neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery.

left views to identify the region of surgical instruments. To utilize stereo clues, we estimate coarse depth maps $\{\mathbf{D}_i\}_{i=1}^T$ for the left views from the binocular captures. We follow the modeling in D-NeRF [21] and represent deformable surgical scenes as a canonical neural radiance field along with a time-dependent neural displacement field (cf. Sec. 2.2). In our pipeline, each training iteration consists of the following six stages: i) randomly pick a frame for training, ii) run tool-guided ray casting (cf. Sec. 2.3) to shoot camera rays into the scene, iii) sample points along each camera ray via depth-cueing ray marching (cf. Sec. 2.4), iv) send sampled points to networks to obtain color and space occupancy of each point, v) evaluate volume rendering integral on sampled points to produce rendering results, vi) optimize the rendering loss plus depth loss to reconstruct shapes, colors and deformations of the surgical scene (cf. Sec. 2.5). The overview of key components in our approach is illustrated in Fig. 1. We will describe the detailed methods in the following subsections.

2.2 Deformable Surgical Scene Representations

We represent a surgical scene as a canonical radiance field and a time-dependent displacement field. Accordingly, each frame of the surgical scene can be regarded as a deformation of the canonical field. The canonical field, denoted as $F_\Theta(\mathbf{x}, \mathbf{d})$, is an 8-layer MLP with network parameter Θ , mapping coordinates $\mathbf{x} \in \mathbb{R}^3$ and unit view-in directions $\mathbf{d} \in \mathbb{R}^3$ to RGB colors $c(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^3$ and space occupancy $\sigma(\mathbf{x}) \in \mathbb{R}$. The time-dependent displacement field $G_\Phi(\mathbf{x}, t)$ is encoded in another 8-layer MLP with network parameters Φ and maps input space-time coordinates (\mathbf{x}, t) into displacement between the point \mathbf{x} at time t and the corresponding point in the canonical field. For any time t , the color and occupancy at \mathbf{x} can be retrieved as $F_\Theta(\mathbf{x} + G_\Phi(\mathbf{x}, t), \mathbf{d})$. Compared with other dynamic modeling approaches [18,19], a displacement field is sufficient to explicitly and physically express all tissue deformations. To capture high-frequency details, we use positional encoding $\gamma(\cdot)$ to map the input coordinates and time into Fourier features [24] before feeding them to the networks.

2.3 Tool Mask-Guided Ray Casting

With scene representations, we further leverage the differentiable volume rendering used in NeRF to yield renderings for supervision. The differentiable volume rendering begins with shooting a batch of camera rays into the surgical scene from a fixed viewpoint at an arbitrary time t . Every ray is formulated as $\mathbf{r}(s) = \mathbf{o} + s\mathbf{d}$, where \mathbf{o} is a fixed origin of the ray, \mathbf{d} is the pointing direction of the ray and s is the ray parameter. In the original NeRF, rays are shot towards a batch of randomly selected pixels on the entire image plane. However, there are many pixels of surgical tools on the captured images, while our goal is to reconstruct underlying tissues. Thus, training on these tool pixels is unexpected. Our main idea for solving this issue is to bypass those rays traveling through tool pixels over the training stage. We utilize binary tool masks $\{\mathbf{M}_i\}_{i=1}^T$, where 0 stands for tissue pixels and 1 stands for tool pixels, to inform which rays should be neglected. In this regard, we create importance maps $\{\mathbf{V}_i\}_{i=1}^T$ according to \mathbf{M}_i and perform importance sampling to avoid shooting rays for those pixels of surgical tools. Eq. (1) exhibits the construction of importance maps, where \otimes is element-wise multiplication, $\|\cdot\|_F$ is Frobenius norm and $\mathbf{1}$ is a matrix with the same shape as \mathbf{M}_i while filled with ones:

$$\mathbf{V}_i = \mathbf{A} \otimes (\mathbf{1} - \mathbf{M}_i), \quad \mathbf{A} = \left(\mathbf{1} + \sum_{j=1}^T \mathbf{M}_j \middle/ \left\| \sum_{j=1}^T \mathbf{M}_j \right\|_F \right). \quad (1)$$

The $\mathbf{1} - \mathbf{M}_i$ term initializes the importance of tissue pixels to 1 and the importance of tool pixels to 0. To balance the sampling rate of occluded pixels across frames, the scaling term \mathbf{A} specifies higher importance scaling for those tissue areas with higher occlusion frequencies. Normalizing each importance map as $\hat{\mathbf{V}}_i = \mathbf{V}_i / \|\mathbf{V}_i\|_F$ will yield a probability mass function over the image plane. During our ray casting stage for the i -th frame, we sample pixels from the distribution $\hat{\mathbf{V}}_i$ using inverse transform sampling and cast rays towards these sampled pixels. In this way, the probability of shooting rays for tool pixels is guaranteed to be zero as the importance of tool pixels is constantly zero.

2.4 Stereo Depth-Cueing Ray Marching

After shooting camera rays over tool occlusion, we proceed ray marching to sample points in the space. Specifically, we discretize each camera ray $\mathbf{r}(s)$ into batch of points $\{\mathbf{x}_j | \mathbf{x}_j = \mathbf{r}(s_j)\}_{j=1}^m$ by sampling a sequence of ray steps $s_1 \leq s_2 \leq \dots \leq s_m$. The original NeRF proposes hierarchical stratified sampling to obtain $\{s_j\}_{j=1}^m$. However, this sampling strategy hardly exploits accurate 3D structures when NeRF models are trained on single-view input. Drawing inspiration from early work in iso-surface rendering [7], we create Gaussian transfer functions with stereo depth to guide point sampling near tissue surfaces. For the i -th frame, the transfer function for a ray $\mathbf{r}(s)$ shooting towards pixel (u, v) is formulated as:

$$\delta(s; u, v, i) = \exp(-(s - \mathbf{D}_i[u, v])^2 / 2\xi^2). \quad (2)$$

The transfer function $\delta(s; u, v, i)$ depicts an impulse distribution that continuously allocates sampling weights for every location on $\mathbf{r}(s)$. The impulse is centered at $\mathbf{D}_i[u, v]$, i.e., the depth at the (u, v) pixel. The width of the impulse is controlled by the hyperparameter ξ , which is set to a small value to mimic Dirac delta impulse. In our ray marching, $s_1 \leq s_2 \leq \dots \leq s_m$ are drawn from the normalized impulse distribution $\frac{1}{\xi\sqrt{2\pi}}\delta(s; u, v, i)$. By this means, sampled points are concentrated around tissue surfaces, imposing stereo prior in rendering.

2.5 Optimization for Deformable Radiance Fields

Once we obtain the sampled points in the space, the emitted color $\widehat{\mathcal{C}}$ and optical depth \widehat{D} of a camera ray $\mathbf{r}(s)$ can be evaluated by volume rendering [5] as:

$$\begin{aligned}\widehat{\mathcal{C}}(\mathbf{r}(s)) &= \sum_{j=1}^{m-1} w_j c(\mathbf{x}_j, \mathbf{d}), \quad \widehat{D}(\mathbf{r}(s)) = \sum_{j=1}^{m-1} w_j s_j, \\ w_j &= (1 - \exp(-\sigma(\mathbf{x}_j)\Delta s_j)) \exp(-\sum_{k=1}^{j-1} \sigma(\mathbf{x}_k)\Delta s_k), \quad \Delta s_j = s_{j+1} - s_j.\end{aligned}\tag{3}$$

To reconstruct the canonical and displacement fields from single-view captures, we optimize the network parameters Θ and Φ by jointly supervising the rendered color and optical depth [3]. Specifically, the loss function for training the networks is defined as:

$$\mathcal{L}(\mathbf{r}(s)) = \|\widehat{\mathcal{C}}(\mathbf{r}(s)) - \mathbf{I}_i[u, v]\|_2^2 + \lambda |\widehat{D}(\mathbf{r}(s)) - \mathbf{D}_i[u, v]|,\tag{4}$$

where (u, v) is the location of the pixel that $\mathbf{r}(s)$ shoots towards, λ is a hyper-parameter weighting the depth loss.

Last but not least, we conduct statistical depth refinement to handle corrupt stereo depth caused by fuzzy pixels and specular highlights on the images of surgical scenes. Direct supervision on the estimated depth will overfit corrupt depth in the end, leading to abrupt artifacts in reconstruction results (Fig. 3). Our preliminary findings reveal that our model at the early training stage would produce smoother results both in color and depth since the underfitting model tends to average learned colors and occupancy. Thus, minority corrupt depth is smoothed by majority normal depth. Based on this observation, we propose to patch the corrupt depth with the output from underfitting radiance fields. Denoting $\widehat{\mathbf{D}}_i^K$ as the underfitting output depth maps for the i -th frame after K iterations of training, we firstly find residual maps through $\epsilon_i = |\widehat{\mathbf{D}}_i^K - \mathbf{D}_i|$, then we compute a probabilistic distribution over the residual maps. After that, we set a small number $\alpha \in [0, 1]$ and locate those pixels with the last α -quantile residuals. Since those located pixels statistically correspond to large residuals, we can identify them as occurrences of corrupt depth. Finally, we replace those identified corrupt depth pixels with smoother depth pixels in $\widehat{\mathbf{D}}_i^K$. After this refinement procedure, the radiance fields are optimized on the patched depth maps in the subsequent training iterations, alleviating corrupt depth fitting.

3 Experiments

3.1 Dataset and Evaluation Metrics

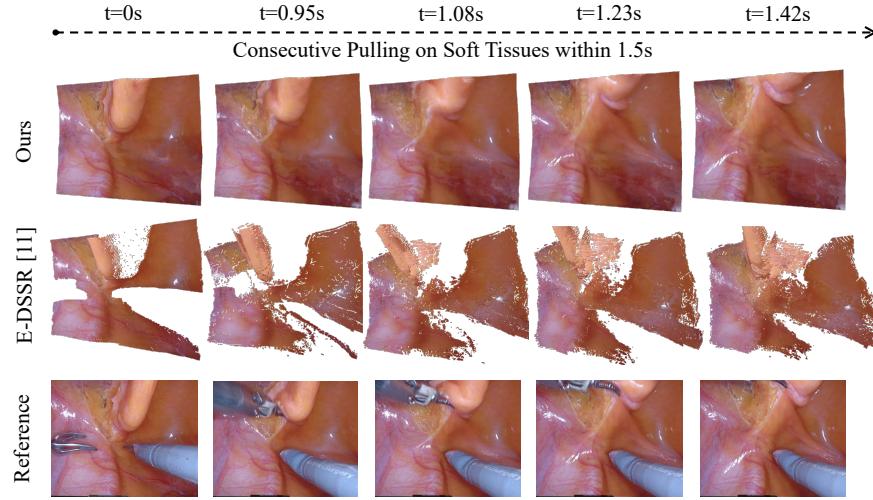
We evaluate our proposed method on typical robotic surgery stereo videos from 6 cases of our in-house DaVinci robotic prostatectomy data. We totally extracted 6 clips with a total of 807 frames. Each clip lasts for 4~8s with 15*fps*. Each case is captured from stereo cameras at a single viewpoint and encompasses challenging scenes with non-rigid deformation and tool occlusion. Among the selected 6 cases, 2 cases contain traction on thin structures such as fascia, 2 cases contain significant pushing and pulling of tissue, and 2 cases contain tissue cutting, which altogether present the typical soft tissue situations in robotic surgery. For comparison, we take the most recent state-of-the-art surgical scene reconstruction method of E-DSSR [11] as a strong comparison. For qualitative evaluation, We exhibit our reconstructed point clouds and compare textural and geometric details obtained by different methods. We also conduct an ablation study on our depth-related modules through qualitative comparison. Due to clinical regulation in practice, it is impossible to collect ground truth depth for numerical evaluation on 3D structures. Following the evaluation method in [11] and wide literature in neural rendering, we alternatively use photometric errors, including PSNR, SSIM and LPIPS, as evaluation metrics for quantitative comparisons.

3.2 Implementation Details

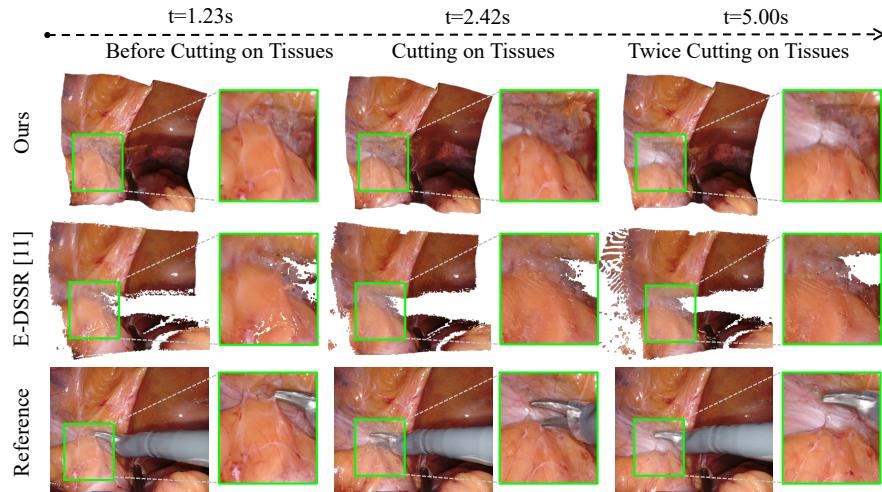
In our implementation, we empirically set the width of the transfer function $\xi = 1$, the weight of depth loss $\lambda = 1$, depth refinement iteration $K = 4000$ and $\alpha = 0.1$. Other training hyper-parameters follow the settings in the state-of-the-art D-NeRF [21]. We calibrate the endoscope in advance to acquire its intrinsics. In all of our experiments, tool masks are obtained by manually labeling and coarse stereo depth maps are generated by STTR-light [9] pretrained on Scene Flow. We optimize each model over 100*K* iterations on a single case. To recover explicit geometry from implicit fields, we render optimized radiance fields to RGBD maps, smooth rendered depth maps via bilateral filtering, and back-project RGBD into point clouds based on the endoscope intrinsics.

3.3 Qualitative and Quantitative Results

For qualitative evaluation, Fig. 2 illustrates the reconstruction results of our approach and the comparison method, along with a reference to the original video. In the test case of Fig. 2(a), the tissues are pulled by surgical instruments, yielding relatively large deformations. Benefiting from the underlying continuous scene representations, our method can reconstruct water-tight tissues without being affected by the tool occlusion. More importantly, per-frame deformations are captured continuously, achieving stable results over the episode of consecutive pulling. In contrast, the current state-of-the-art method [11] could not fully track these large deformations and its reconstruction results include holes



(a) Results on the case “pulling tissues”, where soft tissues are drastically pulled within 2s. We exhibit 5 reconstruction results of our method and E-DSSR over time.



(b) Results on the case “cutting tissues twice”. We show 3 frames corresponding to no deformation, cutting once and cutting twice, respectively. The close-ups of the cutting areas display reconstructed tissue details before and after cutting.

Fig. 2: Qualitative comparisons of 2 cases, demonstrating reconstruction of soft tissues with large deformations and topology changes.

and noisy points under such a challenging situation. We further demonstrate a more difficult case in Fig. 2(b) which includes soft tissue cutting with topology changes. From the reconstruction results, it is observed that our method manages to track the detailed cutting procedures, owing to the powerful neural

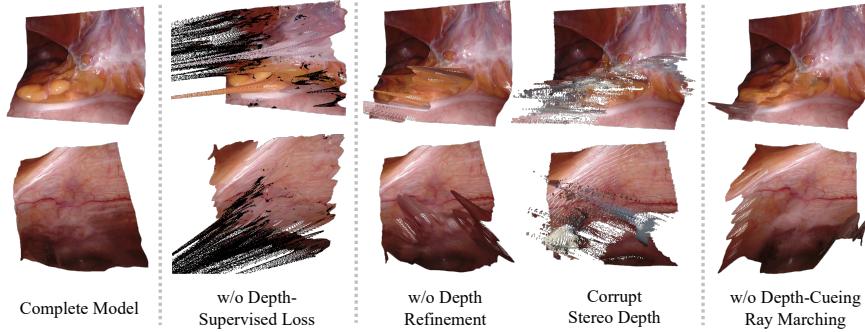


Fig. 3: Ablation study on our depth-related modules, i.e., depth-supervised loss, depth refinement and depth-cueing ray marching.

Table 1: Quantitative evaluation on photometric errors of the dynamic reconstruction on metrics of PSNR, SSIM and LPIPS.

Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
E-DSSR [11]	13.398 \pm 1.387	0.630 \pm 0.057	0.423 \pm 0.047
Ours w/o D	24.088 \pm 2.567	0.849 \pm 0.023	0.230 \pm 0.023
Ours	29.831 \pm 2.208	0.925 \pm 0.020	0.081 \pm 0.022

representation of displacement fields. In addition, it can bypass the issue of tool occlusion and recover the hidden tissues, which is cooperatively achieved by our mask-guided ray casting and the interpolation property of neural implicit fields. On the other hand, the comparison method is not able to capture these small changes on soft tissues nor patch all the tool-occluded areas. Table 1 summarizes our quantitative experiments, showing overall performance on the dataset. Our method dramatically outperforms E-DSSR by \uparrow 16.433 PSNR, \uparrow 0.295 SSIM and \downarrow 0.342 LPIPS. To assess the contribution of the dynamics modeling, we also evaluate our model without neural displacement field (Ours w/o D). As expected, removing this component leads to a noticeable performance drop, which reflects the effectiveness of the displacement modeling.

We present a qualitative ablation study on our depth-related modules in Fig. 3. Without depth-supervision loss, we observe that the pipeline is not capable of learning correct geometry from single-viewpoint input. Moreover, when depth refinement is disabled, abrupt artifacts occur on the reconstruction results due to corruption in stereo depth estimation. Our depth-cueing ray marching can further diminish artifacts on 3D structures, especially for boundary points.

4 Conclusion

This paper presents a novel neural rendering-based framework for dynamic surgical scene reconstruction from single-viewpoint binocular images, as well as

addressing complex tissue deformation and tool occlusion. We adopt the cutting-edge dynamic neural radiance field method to represent surgical scenes. In addition, we propose mask-guided ray casting to handle tool occlusion and impose stereo depth prior upon the single-viewpoint situation. Our approach has achieved superior performance on various scenarios in robotic surgery data such as large elastic deformations and tissue cutting. We hope the emerging NeRF-based 3D reconstruction techniques could inspire new pathways for robotic surgery scene understanding, and empower various down-stream clinical-oriented tasks.

Acknowledgements. This work was supported in part by CUHK Shun Hing Institute of Advanced Engineering (project MMT-p5-20), in part by Shenzhen-HK Collaborative Development Zone, and in part by Multi-Scale Medical Robotics Centre InnoHK.

References

1. Brandao, P., Psychogios, D., Mazomenos, E., Stoyanov, D., Janatka, M.: Hapnet: hierarchically aggregated pyramid network for real-time stereo matching. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **9**(3), 219–224 (2021) [2](#)
2. Chen, L., Tang, W., John, N.W., Wan, T.R., Zhang, J.J.: Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. Computer methods and programs in biomedicine **158**, 135–146 (2018) [1](#)
3. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. arXiv preprint arXiv:2107.02791 (2021) [5](#)
4. Gao, W., Tedrake, R.: Surfelwarp: Efficient non-volumetric single view dynamic reconstruction. In: Robotics: Science and Systems XIV (2019) [2](#)
5. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. ACM SIGGRAPH computer graphics **18**(3), 165–174 (1984) [5](#)
6. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: CVPR. pp. 3907–3916 (2018) [2](#)
7. Kniss, J., Ikits, M., Lefohn, A., Hansen, C., Praun, E., et al.: Gaussian transfer functions for multi-field volume visualization. In: IEEE Visualization, 2003. VIS 2003. pp. 497–504. IEEE (2003) [4](#)
8. Li, Y., Richter, F., Lu, J., Funk, E.K., Orosco, R.K., Zhu, J., Yip, M.C.: Super: A surgical perception framework for endoscopic tissue manipulation with surgical robotics. IEEE Robotics and Automation Letters **5**(2), 2294–2301 (2020) [2](#)
9. Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F.X., Taylor, R.H., Unberath, M.: Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In: ICCV. pp. 6197–6206 (2021) [6](#)
10. Liu, X., Stiber, M., Huang, J., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Reconstructing sinus anatomy from endoscopic video—towards a radiation-free approach for quantitative longitudinal assessment. In: MICCAI. pp. 3–13. Springer (2020) [1](#)
11. Long, Y., Li, Z., Yee, C.H., Ng, C.F., Taylor, R.H., Unberath, M., Dou, Q.: E-dssr: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. In: MICCAI. pp. 415–425. Springer (2021) [2, 6, 8](#)

12. Lu, J., Jayakumari, A., Richter, F., Li, Y., Yip, M.C.: Super deep: A surgical perception framework for robotic tissue manipulation using deep learning for feature extraction. In: ICRA. pp. 4783–4789. IEEE (2021) [1](#)
13. Luo, H., Wang, C., Duan, X., Liu, H., Wang, P., Hu, Q., Jia, F.: Unsupervised learning of depth estimation from imperfect rectified stereo laparoscopic images. Computers in biology and medicine **140**, 105109 (2022) [2](#)
14. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., Barron, J.T., Dosovitskiy, A., Duckworth, D.: Nerf in the wild: Neural radiance fields for unconstrained photo collections. In: CVPR. pp. 7210–7219 (2021) [2](#)
15. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV. pp. 405–421. Springer (2020) [2](#)
16. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: CVPR. pp. 343–352 (2015) [2](#)
17. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: CVPR. pp. 11453–11464 (2021) [2](#)
18. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: CVPR. pp. 5865–5874 (2021) [2, 3](#)
19. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. ACM Transactions on Graphics (TOG) **40**(6), 1–12 (2021) [2, 3](#)
20. Penza, V., De Momi, E., Enayati, N., Chupin, T., Ortiz, J., Mattos, L.S.: envisors: enhanced vision system for robotic surgery. a user-defined safety volume tracking to minimize the risk of intraoperative bleeding. Frontiers in Robotics and AI **4**, 15 (2017) [1](#)
21. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR. pp. 10318–10327 (2021) [2, 3, 6](#)
22. Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J., Civera, J.: Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. IEEE Robotics and Automation Letters **6**(4), 7225–7232 (2021) [1](#)
23. Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G.: Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. IEEE Robotics and Automation Letters **3**(1), 155–162 (2017) [2](#)
24. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singh, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. NeurIPS **33**, 7537–7547 (2020) [3](#)
25. Tang, R., Ma, L.F., Rong, Z.X., Li, M.D., Zeng, J.P., Wang, X.D., Liao, H.E., Dong, J.H.: Augmented reality technology for preoperative planning and intraoperative navigation during hepatobiliary surgery: a review of current methods. Hepatobiliary & Pancreatic Diseases International **17**(2), 101–112 (2018) [1](#)
26. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Xu, Z., Simon, T., Nießner, M., Tretschk, E., Liu, L., et al.: Advances in neural rendering. In: ACM SIGGRAPH 2021 Courses, pp. 1–320 (2021) [2](#)
27. Tewari, A., Fried, O., Thies, J., Sitzmann, V., Lombardi, S., Sunkavalli, K., Martin-Brualla, R., Simon, T., Saragih, J., Nießner, M., et al.: State of the art on neural rendering. In: Computer Graphics Forum. vol. 39, pp. 701–727. Wiley Online Library (2020) [2](#)

28. Tukra, S., Marcus, H.J., Giannarou, S.: See-through vision with unsupervised scene occlusion reconstruction. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (01), 1–1 (2021) [1](#)
29. Wei, G., Yang, H., Shi, W., Jiang, Z., Chen, T., Wang, Y.: Laparoscopic scene reconstruction based on multiscale feature patch tracking method. In: 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS). pp. 588–592. IEEE (2021) [1](#)
30. Wei, R., Li, B., Mo, H., Lu, B., Long, Y., Yang, B., Dou, Q., Liu, Y., Sun, D.: Stereo dense scene reconstruction and accurate laparoscope localization for learning-based navigation in robot-assisted surgery. *arXiv preprint arXiv:2110.03912* (2021) [1](#)
31. Zhou, H., Jagadeesan, J.: Real-time dense reconstruction of tissue surface from stereo optical video. *IEEE transactions on medical imaging* **39**(2), 400–412 (2019) [2](#)
32. Zhou, H., Jayender, J.: Emdq-slam: Real-time high-resolution reconstruction of soft tissue surface from stereo laparoscopy videos. In: MICCAI. pp. 331–340. Springer (2021) [2](#)
33. Zhou, H., Jayender, J.: Real-time nonrigid mosaicking of laparoscopy images. *IEEE Transactions on Medical Imaging* **40**(6), 1726–1736 (2021) [1](#)