DnFPlane For Efficient and High-Quality 4D Reconstruction of Deformable Tissues

Ran Bu[1], Chenwei Xu[1], Jiwei Shan[2], Hao Li[2], Guangming Wang[3], Yanzi Miao[1(✉)], and Hesheng Wang[2(✉)]

[1] China University of Mining and Technology, Xuzhou, China
myz@cumt.edu.cn
[2] Shanghai Jiao Tong University, Shanghai, China
wanghesheng@sjtu.edu.cn
[3] University of Cambridge, United Kingdom

Abstract. Reconstruction of deformable tissues in robotic surgery from endoscopic stereo videos holds great significance for a variety of clinical applications. Existing methods primarily focus on enhancing inference speed, overlooking depth distortion issues in reconstruction results, particularly in regions occluded by surgical instruments. This may lead to misdiagnosis and surgical misguidance. In this paper, we propose an efficient algorithm designed to address the reconstruction challenges arising from depth distortion in complex scenarios. Unlike previous methods that treat each feature plane equally in the dynamic and static field, our framework guides the static field with the dynamic field, generating a dynamic-mask to filter features at the time level. This allows the network to focus on more active dynamic features, reducing depth distortion. In addition, we design a module to address dynamic blurring. Using the dynamic-mask as a guidance, we iteratively refine color values through Gated Recurrent Units (GRU), improving the clarity of tissues detail in the reconstructed results. Experiments on a public endoscope dataset demonstrate that our method outperforms existing state-of-the-art methods without compromising training time. Furthermore, our approach shows outstanding reconstruction performance in occluded regions, making it a more reliable solution in medical scenarios. Code is available: https://github.com/CUMT-IRSI/DnFPlane.git.

Keywords: Depth Reliable 3D Reconstruction · Neural Rendering · Robotic Surgery

1 Introduction

Reconstructing deformable tissue structures accurately and efficiently from binocular endoscopic videos is currently a popular research topic in the field of medical image computing [19,13,20]. This technology holds the potential to advance the development of Robot-Assisted Minimally Invasive Surgery (RAMIS), primarily

R. Bu and C. Xu—Equal contribution.

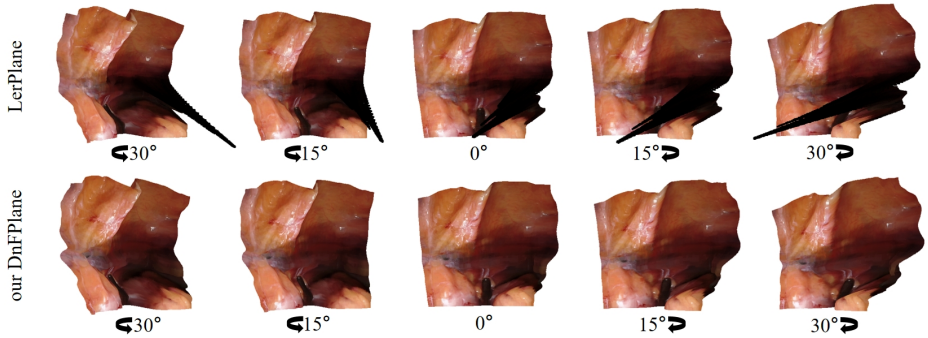

Fig. 1. Performance from different angles at the same time.We show the results of LerPlane (top) and DnFPlane (bottom) at the same time. Reconstruction of DnFPlane is more reliable in occluded regions.

providing learning data for surgical robots and creating realistic virtual surgical environments for AR/VR surgical training [18,14,4]. Furthermore, the application of fast reconstruction techniques extends to actual surgeries, enabling the construction of complete surgical scenes [16,11]. This allows surgeons to navigate surgeries in a more comprehensive and precise manner, improving the safety and success rates of surgeries.

Neural Radiance Fields (NeRF) [12] have significantly advanced the development of 3D reconstruction in endoscopic imaging. EndoNeRF [21] is the first to leverage the capabilities of NeRF for implicit geometric modeling of endoscopic scenes. It introduces a dual neural field method to simulate tissue deformation and typical density, achieving dynamic scene rendering and alleviating the impact of instrument occlusion in endoscopic-assisted surgeries. Building upon this foundation, EndoSurf [23] further employed signed distance functions to simulate tissue surfaces, imposing explicit self-consistency constraints on the neural field, thereby enhancing surface reconstruction quality. To address the challenges of fast dynamic reconstruction, LerPlane [22] separately constructs dynamic and static fields to build a four-dimensional space. This approach significantly alleviates computational burdens, striking a better balance between reconstruction quality and training time.

Although the previous methods have demonstrated remarkable results in reconstruction, they still encounter challenges in the following issues: Firstly, the obstruction of surgical instruments in soft tissues can easily result in depth distortions in the reconstruction results, thereby affecting the accuracy of the surgical scene reconstruction. Secondly, during the surgical procedure, inevitable contact between instruments and soft tissues induces deformation in the soft tissue. This deformation causes dynamic blurring during the reconstruction process, thereby impacting the clarity of the reconstruction.

We propose a novel method named DnFPlane (Dynamic Filter Plane), which enables efficient and high-quality reconstruction of deformable tissues.

Specifically, for the depth distortions problem, we design a combination of dynamic-mask and Dynamic Features Enhancement. By generating dynamic-mask from dynamic and static fields, we effectively filter features at time level.

Additionally, for the dynamic blurring problem, we design a combination of dynamic-mask and Color Iterative Refinement. By using dynamic-mask to identify more active dynamic features and guiding the color refinement process through GRU [5], we effectively reduce dynamic blurring. Our method facilitates more efficient utilization of feature planes during training, making higher-quality and stable results.

Our contributions can be summarized as follows:

1. An efficient and high-quality deformable tissue reconstruction method is developed to address depth distortion issues caused by instrument occlusion, while without compromising training time.
2. A color iterative refinement strategy based on GRU is designed to address the issue of dynamic blurring in the reconstruction process.
3. Compared to previous methods, our DnFPlane outperforms existing state-of-the-art methods on the public endoscope dataset. In particular, our method demonstrates outstanding performance in tissue reconstruction of occluded regions, providing a more reliable visual aid in medical scenarios.

2 Method

2.1 Overview

Our framework employs dynamic and static fields to reconstruct surgical scenes. It utilizes dynamic fields to guide static fields to generate a dynamic-mask(Sec. 2.3), allowing the network to focus on learning more active dynamic features (Sec. 2.4). Subsequently, utilizing GRU and guided by the dynamic-mask, an iterative refinement is applied to color values, diminishing dynamic blurring caused by tissue deformation (Sec. 2.5). Finally, volume rendering is employed to predict color and depth values for each selected ray. Rendering constraints, calculated against ground truth, are then utilized to optimize the overall framework (Sec. 2.6). Our framework is illustrated in Fig. 2.

2.2 Preliminaries

To enhance the efficiency of training and rendering, inspired by LerPlane [22], representing the surgical process as a 4D volume. The surgical scene can be represented by three static fields (XY, YZ, XZ) and three dynamic fields (XT, YT, ZT). We then sample the spatiotemporal points based on the direction of the ray $r(s)$. The coordinates are projected onto the dynamic and static fields. After bilinear interpolation, we obtain dynamic feature planes (F_{XT}, F_{YT}, F_{ZT}) and static feature planes (F_{XY}, F_{YZ}, F_{XZ}). The time dimension is (N, 1), and each feature plane has dimensions (N, C), where N represents the number of features, and C represents the number of feature channels. Then, the fused

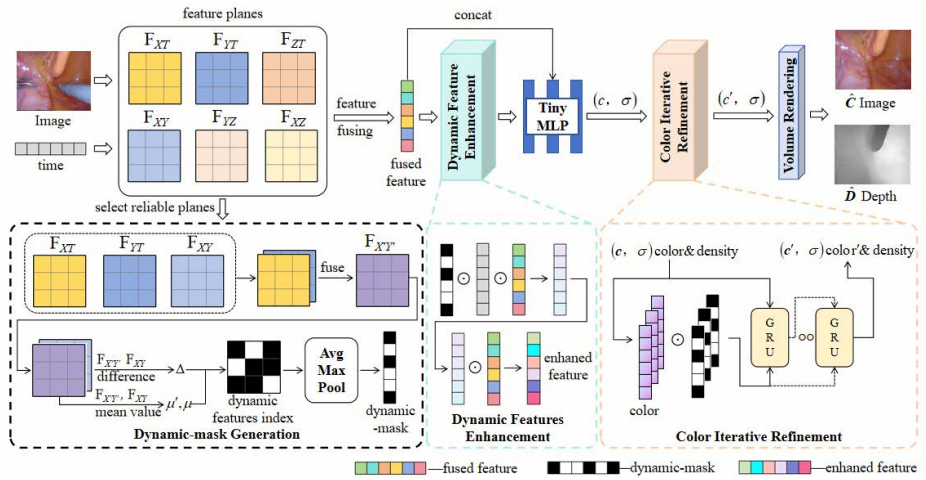

Fig. 2. Illustration of our proposed DnFPlane framework. $F_{X'Y'}$ is a dynamic representation about XY.

feature f can be represented by Eq. 1. f is sent to a tiny MLP Θ to predict the color $\boldsymbol{c}$ and density σ of the points. Finally, utilizing volume rendering[16], we obtain the predicted color $\hat{C}(r(s))$ and predicted depth $\hat{D}(r(s))$ for rapid 4D reconstruction of deformable tissues.

$$f = F_{XY} \odot F_{YZ} \odot F_{XZ} \odot F_{XT} \odot F_{YT} \odot F_{ZT} \tag{1}$$

where the $\odot$ represents element-wise multiplication.

2.3 Dynamic-mask Generation

For the task of endoscopic reconstruction, treating dynamic and static fields equally at the feature level provides rich scene information for reconstruction. However, at the time level, dynamic fields are more crucial than static fields, as they offer more depth information. Unfortunately, Lerplane treats dynamic and static fields equally at the time level, causing depth distortion, particularly in regions where soft tissue is occluded during surgery. Therefore, we guide static fields with dynamic fields to emphasize the importance of dynamic features. This process generates a dynamic-mask, filtering the features at time level to enable the network to focus on learning more active dynamic features, thereby reducing depth distortion.

Select Reliable Planes Observing soft tissues from multiple viewpoints in endoscopic surgical scenes is impractical. Additionally, the presence of surgical instruments poses challenges for the model in predicting the color and density. In order to more accurately reconstruct challenging scenes and reduce the uncertainty introduced by the depth z, we overlook the depth features in the feature

planes, retaining only the dynamic feature planes F_{XT}, F_{YT}, and static feature planes F_{XY}. This provides a reliable guide for subsequent dynamic features enhancement.

Dynamic Threshold Generation To effectively enhance dynamic features without changing the composition of the fused feature, it is necessary to find the more active dynamic part in the fused feature. For this purpose, we focus on identifying the differences between dynamic features and static features.

Specifically, we construct a dynamic representation $F_{X'Y'} = F_{XT} \oplus F_{YT}$ about the static feature plane F_{XY} by fusing dynamic feature planes F_{XT} and F_{YT}. Then, we evaluate the difference Δ between this dynamic representation and the static feature plane using $\Delta = F_{X'Y'} \ominus F_{XY}$. Next, we perform summation along the channel dimension and calculate the average variation μ' across all points for the dynamic representation. Similarly, the average variation μ is for the static feature plane F_{XY}. $\oplus$ and $\ominus$ represent element-wise addition and subtraction.

Finally, according to Eq. 2, the difference between the two average variation values serves as the threshold ε for identifying more activate dynamic features at the time level.

$$\mu' = \frac{\sum_{\mathrm{i}=0}^{\mathrm{n}} \sum_{\mathrm{j}=0}^{\mathrm{c}} F_{X'Y'}\,[\mathrm{i}]\,[\mathrm{j}]}{\mathrm{n}}, \mu = \frac{\sum_{\mathrm{i}=0}^{\mathrm{n}} \sum_{\mathrm{j}=0}^{\mathrm{c}} F_{XY}\,[\mathrm{i}]\,[\mathrm{j}]}{\mathrm{n}}, \varepsilon = \frac{\mu'}{\alpha} - \mu \tag{2}$$

Dynamic features Index Comparing the dynamic-static difference of each value $\Delta\,[i]\,[j]$ with the threshold ε, the dynamic features index is generated. When the difference surpasses ε, the index value is set to 1. Conversely, the index value is set to 0. The dynamic features index is obtained by filtering all features in the evaluation with ε. The dynamic features index undergoes average pooling and max pooling to generate a dynamic-mask with the same dimension as the time.

2.4 Dynamic Features Enhancement

Dynamic-mask is used to filter features at the time level. The more active dynamic feature values in the fused feature can be identified through the dynamic-mask. Specifically, the enhanced feature which should be attended to, is obtained by performing element-wise multiplication between the fused features and the dynamic-mask at the time level. Furthermore, concatenating the enhanced feature with the fused feature allows for a more comprehensive representation of the deformation information within the tissues.

2.5 Color Iterative Refinement

During the reconstruction process, regions that exhibit more deformations are more prone to dynamic blurring. This challenge arises from deformations caused by inherent changes in tissues or contact with surgical instruments. The difficulty lies in determining the color values that need to be refined.

Considering this challenge and drawing inspiration from [24,10], we use the iterative refinement of color values in regions with significant dynamic changes in tissues. The dynamic-mask is used to perform element-wise multiplication with color values, generating a guidance h, while the initial color $c_{initial}$ serves as the initial value for iterative refinement. Utilizing GRU, we update the initial value to obtain the residual of the color values Δc. Finally, adding the residual to the color values $c' = c + \Delta c$ to complete the iterative refinement process in regions with obvious dynamic changes in soft tissues.

2.6 Optimization

Volume Rendering By tracing rays $r(s) = o + s\varphi$ from the camera center to pixels in the captured image, where o is the ray origin and φ is the pixel's viewing direction. The predicted color $\hat{C}(r(s))$ and predicted depth $\hat{D}(r(s))$ of pixels in the camera-captured image can be computed through classical volume rendering techniques [8], as illustrated in Eq. 3. w_m represents the integration weight, c_m represents color, s_m represents sample points, δ_m represents the interval between adjacent samples.

$$\hat{C}(r(s)) = \sum_{m=1}^{M} w_m c_m, \hat{D}(r(s)) = \sum_{m=1}^{M} w_m s_m,$$
$$w_m = (1 - \exp(-\sigma_m \delta_m)) \exp\left(-\sum_{k=1}^{M} \sigma_k \delta_k\right), \delta_m = s_{m+1} - s_m. \tag{3}$$

Loss In order to optimize the rendering performance, we supervise the model not only through color loss $\mathcal{L}_{color}$ and depth loss $\mathcal{L}_{depth}$ but also introduce total variation (TV) loss [7,15] $\mathcal{L}_{tv}$ and time smoothness loss [6] $\mathcal{L}_{ts}$. This enables the model to reconstruct deformable tissues from a limited view robustly. Additionally, we incorporate histogram loss [3] $\mathcal{L}_h$ to train the sampling network, aiming to enhance the rendering quality by improving the accuracy of sampling points. The overall loss $\mathcal{L}_{total}$ is formulated as shown in Eq. 4.

$$\mathcal{L}_{total} = \mathcal{L}_{color} + \mathcal{L}_{depth} + \lambda_{tv}\mathcal{L}_{tv} + \lambda_{ts}\mathcal{L}_{ts} + \mathcal{L}_h \tag{4}$$

3 Experiments

3.1 Dataset and Evaluation Metrics

We evaluate our proposed method on the ENDONERF [21], which consists of cases captured using stereo cameras from a single viewpoint. These cases consist of challenging scenes that involve non-rigid deformation (changes in shape or structure) and instrument occlusion (obstruction caused by surgical instruments). We employed widely recognized evaluation metrics to assess the rendering results. We evaluated the rendering results using well-established metrics,

including PSNR, SSIM, and LPIPS, to measure the quality of the rendered images. Additionally, we employed the FLIP metric [1,2] to assess the consistency of the underlying 3D scene. This quantitative analysis enabled us to evaluate both visual fidelity and scene coherence in a comprehensive manner.

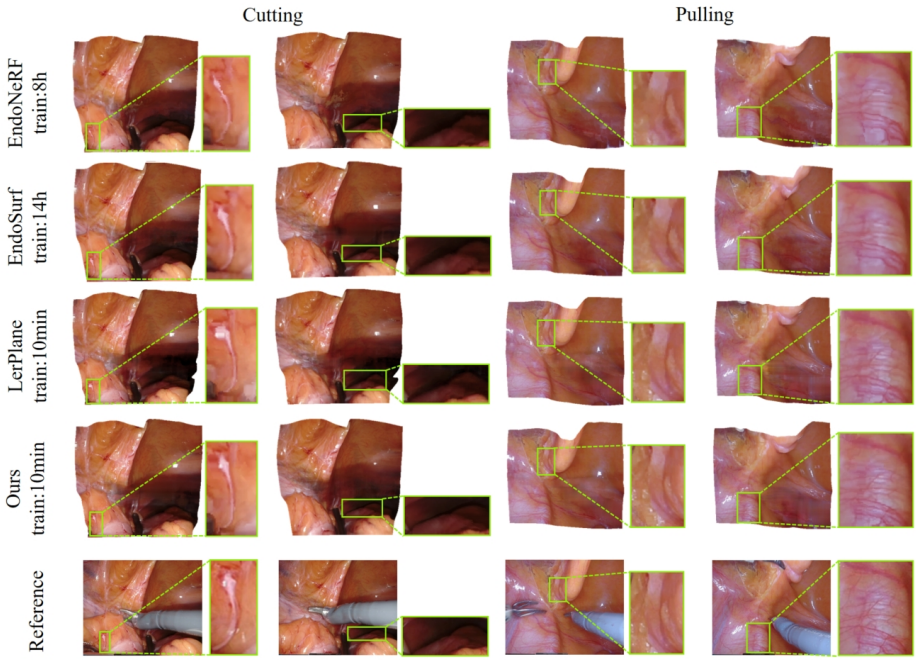

Fig. 3. Illustration of the rendered images of previous works and ours.

3.2 Implementation Details

The adjustable weight $\alpha = 3$ in Eq. 2. We use GRU for 4 iterations on the color values. For the total variation (TV) loss across all experiments, we set $\lambda_{tv} = 0.0001$, and for the time smoothness loss, we set $\lambda_{ts} = 0.03$. The Adam optimizer [9] is employed with an initial learning rate of 0.02. A cosine annealing schedule with a warm-up stage of 512 iterations is implemented. We train all scenes with 9k and 32k iterations on an RTX 4090 GPU running Ubuntu 22.04. Our DnFPlane is implemented using Python and PyTorch [17].

3.3 Qualitative and Quantitative Results

For qualitative evaluation, Fig. 3 illustrates a visual comparison of the rendered images between our method and several SOTA methods. EndoNeRF and EndoSurf are capable of high-fidelity reconstruction of deformable tissues, but the

Table 1. Performance comparison on the ENDONERF. PSNR, SSIM, LPIPS, and FLIP are employed to evaluate the result of dynamic reconstruction.

Methods	ENDONERF-Cutting				ENDONERF-Pulling				Time
	PNSR↑	SSIM↑	LPIPS↓	FLIP↓	PNSR↑	SSIM↑	LPIPS↓	FLIP↓	
EndoNeRF [21]	29.001	0.929	0.078	-	27.081	0.902	0.106	-	8h
EndoSurf [23]	34.555	**0.951**	0.125	-	36.656	0.954	0.121	-	14h
Lerplane-9k [22]	33.762	0.901	0.113	0.088	36.287	0.936	0.084	0.068	3min
ours-9k	34.908	0.917	0.099	0.078	36.862	0.941	0.078	0.064	3min
Lerplane-32k [22]	<u>36.538</u>	0.933	<u>0.065</u>	<u>0.067</u>	<u>39.725</u>	<u>0.960</u>	<u>0.046</u>	<u>0.049</u>	10min
ours-32k	**37.312**	<u>0.942</u>	**0.059**	**0.063**	**40.338**	**0.964**	**0.039**	**0.046**	10min

high computational costs limit their intraoperative use. LerPlane outperforms both of them within 10 minutes based on evaluation metrics. However, when we change the viewpoint to a side view, the result deteriorates. As shown in Fig. 1, the reconstruction performance of LerPlane in the z-axis suffers from severe distortion in regions where the instruments are occluded. DnFPlane addresses this issue by employing Dynamic Features Enhancement(DFE), allowing it to perform well in regions occluded by instruments. On the other hand, DnFPlane successfully recovers rendered maps with smoother shapes and richer details.

Table 2. Ablation studies on different components on the ENDONERF. Please refer to Sec. 3.3 for explanations of different methods.

Methods	ENDONERF-Cutting				ENDONERF-Pulling			
	PNSR↑	SSIM↑	LPIPS↓	FLIP↓	PNSR↑	SSIM↑	LPIPS↓	FLIP↓
ours w/o DFE	36.863	0.935	0.063	0.066	39.892	0.961	0.044	0.049
ours w/o CIR	37.059	0.939	0.061	0.064	40.077	0.963	0.042	0.047
ours	**37.312**	**0.942**	**0.059**	**0.063**	**40.338**	**0.964**	**0.039**	**0.046**

The quantitative results are presented in Table 1. Bold and underlined numbers denote the best and the second best respectively. DnFPlane achieves better visual quality at a low time cost. Note that, our work outperforms related approaches in most metrics and produces more reliable visual results. In Table 2, we present a quantitative ablation study on DnFPlane to understand its key components and demonstrate their effectiveness. The inclusion of the DFE module (w/o CIR) has improved the reconstruction quality of the algorithm, resulting in better performance across all evaluation metrics. Additionally, this module makes a significant contribution to mitigating depth distortion. The Color Iterative Refinement module (w/o DFE), on the other hand, helps refine the surface and plays a crucial role in improving dynamic blur.

4 Conclusion

This paper presents an efficient and high-quality deformable tissue reconstruction method. By rethinking the value of dynamic and static fields, we use dynamic fields to guide static fields at the time level to address depth distortions at regions of instrument occlusion without compromising training time. Additionally, we propose a color iteration refinement method based on GRU to enhance resolution and improve dynamic blurring during the reconstruction process. Compared to previous methods, our DnFPlane outperforms existing state-of-the-art methods on common endoscopic datasets. We hope that DnFPlane will positively impact robotic surgical scene understanding.

Acknowledgments. Research supported by the General Program of National Natural Science Foundation of China(Grant No.61976218, 62361166632, 62225309, 62073222 and U21A20480), Key R&D Plan of Xuzhou City (Social Development) Project - Medical and Health (Project Number: KC22111) and Youth Innovation Technology Project of Xuzhou Municipal Health Commission (Project Number: XWKYHT20220073).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Andersson, P., Nilsson, J., Akenine-Möller, T., Oskarsson, M., Åström, K., Fairchild, M.D.: Flip: A difference evaluator for alternating images. Proc. ACM Comput. Graph. Interact. Tech. **3**(2), 15–1 (2020)
2. Andersson, P., Nilsson, J., Shirley, P., Akenine-Möller, T.: Visualizing errors in rendered high dynamic range images (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
4. Chong, N., Si, Y., Zhao, W., Zhang, Q., Yin, B., Zhao, Y.: Virtual reality application for laparoscope in clinical surgery based on siamese network and census transformation. In: Proceedings of 2021 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2021) Medical Imaging and Computer-Aided Diagnosis. pp. 59–70. Springer (2022)
5. Dey, R., Salem, F.M.: Gate-variants of gated recurrent unit (gru) neural networks. In: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS). pp. 1597–1600. IEEE (2017)
6. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479–12488 (2023)
7. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5501–5510 (2022)

8. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. ACM SIGGRAPH computer graphics **18**(3), 165–174 (1984)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Lee, B., Lee, H., Ali, U., Park, E.: Sharp-nerf: Grid-based fast deblurring neural radiance fields using sharpness prior. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3709–3718 (2024)
11. Mahmoud, N., Collins, T., Hostettler, A., Soler, L., Doignon, C., Montiel, J.M.M.: Live tracking and dense reconstruction for handheld monocular endoscopy. IEEE transactions on medical imaging **38**(1), 79–89 (2018)
12. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
13. Muthusamy, V.R., Lightdale, J.R., Acosta, R.D., Chandrasekhara, V., Chathadi, K.V., Eloubeidi, M.A., Fanelli, R.D., Fonkalsrud, L., Faulx, A.L., Khashab, M.A., et al.: The role of endoscopy in the management of gerd. Gastrointestinal endoscopy **81**(6), 1305–1310 (2015)
14. Nicolau, S., Soler, L., Mutter, D., Marescaux, J.: Augmented reality in laparoscopic surgical oncology. Surgical oncology **20**(3), 189–201 (2011)
15. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)
16. Overley, S.C., Cho, S.K., Mehta, A.I., Arnold, P.M.: Navigation and robotics in spinal surgery: where are we now? Neurosurgery **80**(3S), S86–S99 (2017)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. Peng, H., Yang, X., Su, Y.H., Hannaford, B.: Real-time data driven precision estimator for raven-ii surgical robot end effector position. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 350–356. IEEE (2020)
19. Pimentel-Nunes, P., Dinis-Ribeiro, M., Ponchon, T., Repici, A., Vieth, M., De Ceglie, A., Amato, A., Berr, F., Bhandari, P., Bialek, A., et al.: Endoscopic submucosal dissection: European society of gastrointestinal endoscopy (esge) guideline. Endoscopy **47**(09), 829–854 (2015)
20. Valdastri, P., Simi, M., Webster III, R.J.: Advanced technologies for gastrointestinal endoscopy. Annual review of biomedical engineering **14**, 397–429 (2012)
21. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 431–441. Springer (2022)
22. Yang, C., Wang, K., Wang, Y., Yang, X., Shen, W.: Neural lerplane representations for fast 4d reconstruction of deformable tissues. arXiv preprint arXiv:2305.19906 (2023)
23. Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2023)

24. Zhang, X., Bi, S., Sunkavalli, K., Su, H., Xu, Z.: Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5449–5458 (2022)