Progressive Optimization of Camera Pose and 4D Radiance Fields for long Endoscopic Videos

Florian Stilz[*,1] Mert Karaoglu[*,1,3] Felix Tristram[*,1,2] Nassir Navab[1,2]
Benjamin Busam[1,2] Alexander Ladikos[3]
[1] Technical University Munich [2] MCML [3] Imfusion

Abstract

Reconstructing endoscopic scenes is vital for medical purposes, such as post-operative assessments and educational training. Recently, neural rendering has emerged as a promising method for reconstructing endoscopic scenes involving tissue deformation. Yet, current techniques exhibit major limitations, such as reliance on static endoscopes, limited deformation, or the need for external tracking devices to obtain camera pose data. In this paper we introduce a novel solution that can tackle these challenges posed by a moving stereo endoscope in a highly deformable setting. Our method divides the scene into multiple overlapping 4D neural radiance fields (NeRFs) and uses a progressive optimization approach via optical flow and geometry supervision for simultaneous reconstruction and camera pose estimation. Tested on videos of up to fifteen times longer than what prior work experiment on, our method greatly improves usability, extending detailed reconstruction to much longer surgical videos without external tracking. Comprehensive evaluations using the StereoMIS dataset show that our method substantially enhances novel view synthesis quality while maintaining competitive pose accuracy.

1. INTRODUCTION

Visually and geometrically accurate reconstructions of surgical scenes are crucial for various computer vision and AR/VR applications such as post-surgical longitudinal assessment [16], surgical training [14], and data generation for other learning-based computer vision and robotics applications [18]. However, endoscopic videos present a range of visual and practical challenges, including strong non-homeomorphic deformations, prolonged recording times, and the difficulty of determining camera positions. These challenges often lead to a reliance on external tools for acquisition, diminishing the ease of use and practicality of the reconstruction frameworks.

Recent methods for 4D endoscopic reconstruction and novel view synthesis [37, 39–41] assume a static camera or use forward kinematics of a robotic endoscope to acquire poses in this highly dynamic setting. This limits their applicability to real-life surgical recordings, which can exhibit substantial camera movement. Additionally, acquiring camera poses from robot kinematics can also be problematic as they are often inaccurate and require refinement [8]. To address these limitations, we propose FLex, a novel NeRF-based architecture that handles the complex setup of a moving endoscope in a dynamic surgical environment. FLex introduces an implicit scene separation into multiple overlapping 4D neural radiance fields (NeRFs) and employs a progressive optimization scheme for joint 3D reconstruction and camera pose estimation from scratch. Extensive evaluations on the StereoMIS [11] dataset demonstrate that FLex significantly improves the quality of novel view synthesis while maintaining competitive pose accuracy, showcasing its potential for practical surgical applications.

To summarize, our contributions are:

- A novel NeRF architecture for dynamic reconstruction in highly deformable endoscopic scenes without the need for camera pose information, accomplished by progressive optimization and optical flow supervision.
- An efficient scaling method that splits the scene into multiple overlapping 4D models, enabling detailed reconstruction of theoretically unlimited length dynamic surgical videos.
- Significant improvement over prior State-of-the-art in novel view synthesis with competitive accuracy in camera pose estimation on the StereoMIS dataset.

2. Related Work

2.1. Static Reconstruction

Traditionally, camera poses and scene geometry are estimated by extracting and matching features from images, then triangulating their 3D positions, as exemplified by

* The authors contributed equally.
Corresponding author: Florian Stilz (*florian.stilz@web.de*).

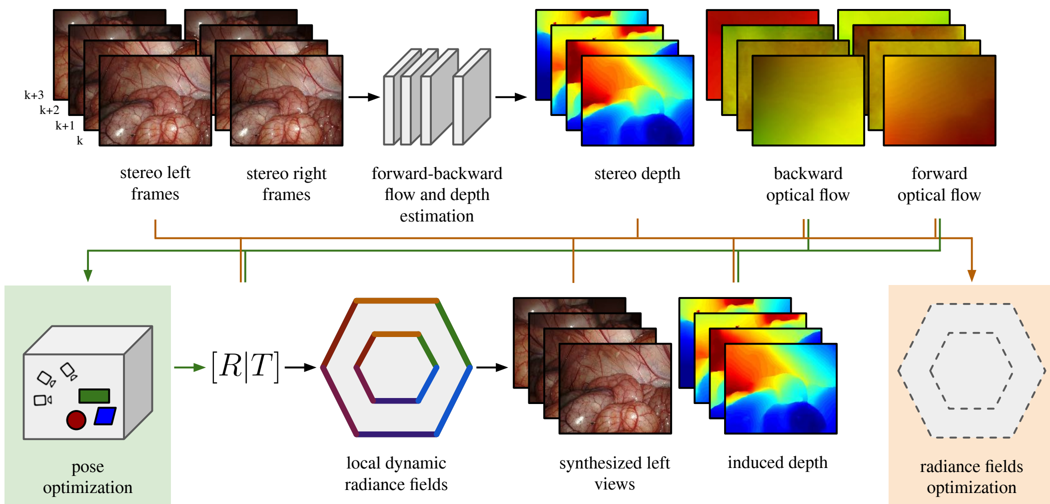

Figure 1. Overview of our proposed method. k indexes the frames along the temporal dimension. We obtain stereo depth images and forward/backward optical flow from RAFT [32] and use it during our optimization. The arrows symbolize which components are used for which optimization, where backward/forward optical flow and induced depth are used to optimize the camera poses (green) and stereo images, depth and optical flow are used to optimize the radiance fields (orange).

COLMAP [28], which uses SIFT [19] features in a sequential pipeline. Recently, methods like VGGSfM [35] and Dust3R [36] have advanced this process by using fully differentiable approaches. VGGSfM [35] recovers all cameras simultaneously based on 2D point tracks and optimizes geometry and poses globally. Dust3R [36] uses a transformer architecture to regress point maps from image pairs and aligns them in multi-view cases.

Poses and geometry from the aforementioned methods often initialize Neural Radiance Fields (NeRFs) [22], which optimize a dense 3D representation in an MLP for rendering novel viewpoints. They also initialize Gaussian Splatting methods [12]. Some NeRF approaches [23, 34, 38], like BARF [15] and LocalRF [21], jointly optimize camera poses and 3D representations without needing known poses. BARF [15] uses filtered positional encoding to smooth gradient flow, while LocalRF [21] optimizes local scene representations with additional supervision. Block-NeRF [31] also optimizes overlapping scene representations and poses but assumes only slight pose inaccuracies. All these methods assume static scenes, making them unsuitable for endoscopic reconstruction.

For our work we take inspiration from LocalRF [21] and choose to represent our endoscopic scenes as multiple local representations but change the underlying network architecture to also handle dynamic scene content.

2.2. Dynamic Reconstruction

Reconstructing dynamic scenes with non-rigid motion is challenging due to the breakdown of 3D consistency, making traditional SfM and NeRF approaches ineffective. Shape-from-Template [3, 7] and Non-Rigid-Structure-from-Motion (NRSfM) [1, 4, 33] methods attempt to address this by incorporating spatial and temporal priors, but they rely on accurate 2D point tracks or 2D-3D matches. Recently, NeRFs have been used for dynamic scenes, either by decoupling deformations from scene geometry [24] or adding time as an input [10]. More recently, some works utilize an explicit scene representation by including a learnable 4D feature volume [5, 9]. However, most methods rely on prior pose information, making them vulnerable to inaccuracies. RoDyNeRF [17] addresses this by jointly optimizing poses and reconstruction, but it assumes some static content, which is unsuitable for constantly moving environments like endoscopy.

2.3. Reconstructing Endoscopic Scenes

Prior works explore explicit representations like point clouds from visual odometry [30] and SLAM [26] for camera tracking and reconstruction, but these methods struggle with incomplete geometry when rendering new views. EndoNeRF [37] was the first to adapt dynamic NeRF [24] for endoscopic scenes, followed by EndoSurf[41], which uses a signed-distance function, and LerPlane [40]and For-Plane [39], which employ explicit data structures [5, 6, 9] for faster optimization and rendering. However, these approaches rely on external camera pose measurements, which are hard to obtain in endoscopic environments.

FLex, along with concurrent work BASED [27], is among the first to investigate joint pose optimization for dynamic endoscopic scenes. FLex also scales efficiently to long sequences, tested on surgical recordings with up

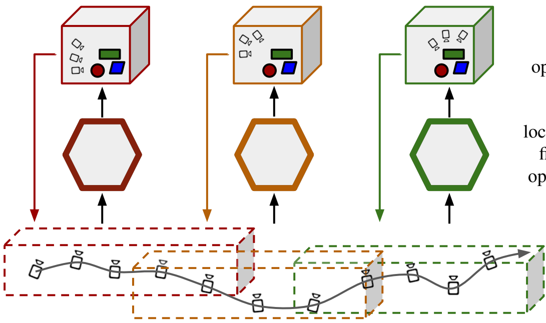

Figure 2. Joint progressive pose and local dynamic radiance fields optimization. Spatial extents clustered within the bounding boxes of different colors represent the spatio-temporal domain of the corresponding local radiance fields. The arrow on the camera trajectory shows the temporal direction.

to 5,000 frames, unlike prior works limited to 300 frames. This makes FLex a significant step toward dynamic neural rendering in real surgical setups.

3. Method

3.1. Overview

Given a rectified stereo-endoscopic video, our goal is to reconstruct the 4D scene accurately without prior camera pose information. For this, we propose a new method **FLex**, standing for **F**low-optimized **L**ocal **Hex**planes, depicted in Fig. 1, which combines advancements from recent NeRF literature to build multiple smaller dynamic models that are progressively optimized. In contrast to prior work [37, 39, 41], we do not have one unified representation of the scene but multiple smaller overlapping ones. Furthermore, we adopt a progressive optimization scheme that enables the optimization of poses from scratch. Since endoscopic environments often have textureless surfaces which make geometry optimization from photometric consistency difficult we additionally incorporate supervision through optical flow and stereo depth priors.

3.2. 4D Scene Representation

NeRFs [22] implicitly model a 3D scene utilizing differentiable volume rendering to predict pixel colors. They can be adapted to a 4D scene representation by adding the timestep k as an additional input to the model. We choose HexPlane [5] as our local model, which represents a dynamic scene using an explicit 4D feature grid paired with an implicit MLP.

3.3. Progressive Optimization

Endoscopic videos pose challenges for NeRF architectures due to their reliance on external tools for pose estimation and the potential for arbitrarily long sequences in dynamic

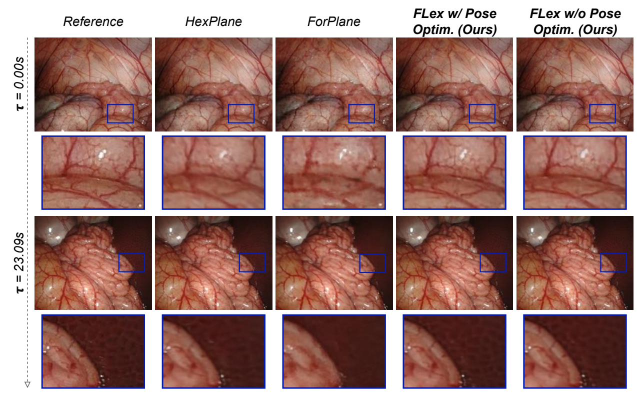

Figure 3. Qualitative results on a 1,000 frame scene with breathing deformations and camera motion. Best viewed in digital version and zoomed in.

environments. To address these, we introduce a joint pose and radiance fields optimization scheme that combines *progressive optimization* and dynamic allocation of *local HexPlane models*, inspired by LocalRF [20] and visualized in Fig. 2. We sequentially add frames, while optimizing their pose and local HexPlane model, until a certain number of frames is reached or the camera pose moves to far from the initial frame. We then instantiate a new local model, where the process starts over. During inference contributions from overlapping models are blended based on proximity.

3.4. Training Objectives

We optimize our method with a combination of photometric $\mathcal{L}_{rgb}$, depth $\mathcal{L}_z$ and optical flow losses $\mathcal{L}_f$, which are balanced by factors $\lambda_{z,f}$:

$$\mathcal{L} = \mathcal{L}_{rgb} + \lambda_z \mathcal{L}_z + \lambda_f \mathcal{L}_f \tag{1}$$

The exact loss formulations are clarified in the supplementary material.

4. Experiments

4.1. Dataset and Evaluation Metrics

We assess the efficacy of our approach using the publicly available StereoMIS [11] dataset, recorded using a stereo endoscope of a da Vinci Xi robot; ground-truth camera trajectories are measured using the forward kinematics. In total we extract five sequences for general comparison, each 1,000 frames long (∼29s). Furthermore, we create two additional longer sequences (5000 & 4000 frames) to study the method's behavior given a larger temporal and spatial extent, which are discussed in the supplementary material. We report PNSR, SSIM and LPIPS [42] (both AlexNet [13] and VGG [29]) metrics, as well as L1-Distance in mm to

Model	ATE-RMSE ↓	RPE-Trans ↓	RPE-Rot ↓
Robust-Pose Estimation [11]	$\mathbf{2.164} \pm 2.68e-1$	$\mathbf{0.073} \pm 3e-5$	$\mathbf{0.043} \pm 2e-6$
LocalRF^{†_2} [20]	7.704 ± 1.506	$0.160 \pm 8e-4$	$0.119 \pm 2e-5$
FLex w/ Pose Optim. (Ours)	$\underline{2.565} \pm 1.6e-1$	$\underline{0.127} \pm 9e-4$	$\underline{0.102} \pm 4e-6$e

Table 1. Average Pose accuracy on StereoMIS dataset. ATE-RMSE and RPE-Trans are in mm, RPE-Rot is in degrees. The best results are marked in bold, second best are underlined. Our method improves substantially over LocalRF and performs close to the fully supervised Robust-Pose-Estimation, which was trained on the Stereo-MIS Dataset.

Method	PSNR ↑	SSIM ↑	LPIPS$_a$ ↓	LPIPS$_v$ ↓	L1-Distance ↓
EndoNeRF [37]	21.99	0.590	0.496	0.514	–
EndoSurf [41]	25.18	0.622	0.528	0.529	8.105
ForPlane [39]	30.35	0.783	0.208	0.301	23.717
LocalRF^{†_2} [20]	27.41	0.781	0.245	0.288	4.576
HexPlane^{†_1} [5]	<u>30.85</u>	<u>0.819</u>	0.211	0.273	1.532
FLex w/o Pose Optim. (Ours)	**31.10**	**0.836**	<u>0.200</u>	**0.244**	<u>1.456</u>
FLex w/ Pose Optim. (Ours)	30.62	0.818	**0.179**	<u>0.245</u>	**1.273**

Table 2. View synthesis quality on StereoMIS dataset. The metrics are computed as an average for five 1,000 frame sequences. L1-Distance is computed between the synthesized and the ground truth depth images in mm. The best result for each metric is marked in bold, while second best is underlined.

evaluate geometry reconstruction. We consider the stereo-estimated depth as *ground truth* since a measured depth is not available. For evaluating camera pose accuracy we report root-mean-squared absolute trajectory error (ATE-RMSE), relative translational and rotational pose errors (RPE-Trans and RPE-Rot).

4.2. Implementation Details

We ensure equal model capacity for all methods using explicit data structures [5, 20, 39], meaning all those methods have equal feature grid dimensions spatially and proportionally to the covered image sequence for the temporal dimension. This is to make any results more comparable, since a higher capacity can achieve better results. We also make small changes to HexPlane and LocalRF to make them usable in an endoscopic setting, indicated by $\dagger_{1,2}$. Where we do not optimise for the poses as well, we use Robust-Pose Estimation [11] to estimate the camera poses. More implementation details can be found in the supplementary material.

4.3. Quantitative and Qualitative Results

We conduct a comprehensive comparison of the proposed method against the latest published state-of-the-art (SoTA) NeRF methods designed for endoscopy [37, 39, 41] and two additional baselines [5, 20] that are not specifically designed for endoscopy. The results in Table 2, summarizing the average results across all 5 scenes, demonstrate that FLex without pose optimization consistently outperforms all baselines and notably surpasses the current endoscopic SoTA, ForPlane, by 5.3 SSIM while achieving substantially better geometry reconstruction as measured by L1-Distance. These quantitative findings are substantiated by our qualitative results presented in Fig. 3, highlighting that FLex renders images with clearer high-frequency details and less blur than the most competitive baselines.

4.4. Pose Accuracy

We compare FLex against a SoTA method in visual odometry for endoscopic scenes, Robust-Pose Estimation [11], and the original LocalRF [21] on 3 sequences each with 1,000 frames. As highlighted in Table 1, FLex performs competitively achieving close results to Robust-Pose Estimation and outperforms LocalRF by a good margin. However, please note that this task is not the main focus of our work and can be improved using robust optimization and globally consistent methods in the future.

5. Conclusion

In this work, we present FLex, a novel method for reconstructing pose-free, long surgical videos with challenging tissue deformations and camera motion. Our approach successfully eliminates the reliance on prior poses by jointly optimizing for 4D reconstruction and camera trajectory via optical flow and depth supervision in a progressive manner. FLex improves upon the scalability of dynamic NeRFs for larger scenes thus becoming more applicable to boundlessly long surgical recordings, while improving over current methods on the StereoMIS dataset in terms of novel view synthesis with competitive pose accuracy. We believe that FLex can pave the way towards more easily accessible, realistic and reliable 4D endoscopy reconstructions to improve post surgical analysis and medical education.

References

[1] Antonio Agudo and Francesc Moreno-Noguer. Simultaneous pose and non-rigid shape with particle dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. pages 5470–5479, 2022. 1

[3] Adrien Bartoli, Yan Gérard, Francois Chadebecq, Toby Collins, and Daniel Pizarro. Shape-from-template. *IEEE transactions on pattern analysis and machine intelligence*, 37:2099–2118, 2015. 2

[4] Christoph Bregler and Aaron Hertzmann. Recovering non-rigid 3d shape from image streams. In *In Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 690–696, 2000. 2

[5] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2, 3, 4, 1

[6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 2

[7] Ajad Chhatkuli, Daniel Pizarro, and Adrien Bartoli. Stable template-based isometric 3d reconstruction in all imaging conditions by linear least-squares. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 708–715, 2014. 2

[8] Zejian Cui, João Cartucho, Stamatia Giannarou, and Ferdinando Rodriguez y Baena. Caveats on the first-generation da vinci research kit: Latent technical constraints and essential calibrations. *IEEE Robotics & Automation Magazine*, 2023. 1

[9] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[10] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 2

[11] Michel Hayoz, Christopher Hahne, Mathias Gallardo, Daniel Candinas, Thomas Kurmann, Maximilian Allan, and Raphael Sznitman. Learning how to robustly estimate camera pose in endoscopic videos. *International journal of computer assisted radiology and surgery*, pages 1–8, 2023. 1, 3, 4

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 2

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 3

[14] Tim Lange, Daniel J Indelicato, and Joseph M Rosen. Virtual reality in surgical training. *Surgical oncology clinics of North America*, 9(1):61–79, 2000. 1

[15] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 2

[16] Xingtong Liu, Maia Stiber, Jindan Huang, Masaru Ishii, Gregory D Hager, Russell H Taylor, and Mathias Unberath. Reconstructing sinus anatomy from endoscopic video–towards a radiation-free approach for quantitative longitudinal assessment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 3–13. Springer, 2020. 1

[17] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Yung-Yu Chuang Kopf Johannes Ayush Saraf, Changil Kim, and Jia-Bin Huang. Robust dynamic radiance fields. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023. 2

[18] Yonghao Long, Jianfeng Cao, Anton Deguet, Russell H Taylor, and Qi Dou. Integrating artificial intelligence and augmented reality in robotic surgery: An initial dvrk study using a surgical education scenario. In *2022 International Symposium on Medical Robotics (ISMR)*, pages 1–8. IEEE, 2022. 1

[19] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[20] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16539–16548, 2023. 3, 4, 1

[21] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16539–16548, 2023. 2, 4

[22] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[23] Keunhong Park, Philipp Henzler, Ben Mildenhall, Jonathan T Barron, and Ricardo Martin-Brualla. Camp: Camera preconditioning for neural radiance fields. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 2

[24] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[25] Konstantinos Rematsas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12932–12942, 2022. 1

[26] Juan J Gomez Rodriguez, JM Montiel, and Juan D Tardos. Nr-slam: Non-rigid monocular slam. *arXiv preprint arXiv:2308.04036*, 2023. 2

[27] Shreya Saha, Sainan Liu, Shan Lin, Jingpei Lu, and Michael Yip. Based: Bundle-adjusting surgical endoscopic dynamic video reconstruction using neural radiance fields. *arXiv preprint arXiv:2309.15329*, 2023. 2

[28] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[30] Jingwei Song, Jun Wang, Liang Zhao, Shoudong Huang, and Gamini Dissanayake. Dynamic reconstruction of deformable soft-tissue with stereo scope in minimal invasive surgery. *IEEE Robotics and Automation Letters*, 3(1):155–162, 2017. 2

[31] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 2

[32] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. pages 402–419, 2020. 2, 1

[33] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under ortography: a factorization method. *Internation Journal of Computer Vision*, 1992. 2

[34] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023. 2

[35] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024. 2

[36] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[37] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 431–441. Springer, 2022. 1, 2, 3, 4

[38] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[39] Chen Yang, Kailing Wang, Yuehao Wang, Qi Dou, Xiaokang Yang, and Wei Shen. Efficient deformable tissue reconstruction via orthogonal neural plane. *arXiv preprint arXiv:2312.15253*, 2023. 1, 2, 3, 4

[40] Chen Yang, Kailing Wang, Yuehao Wang, Xiaokang Yang, and Wei Shen. Neural lerplane representations for fast 4d reconstruction of deformable tissues. *arXiv preprint arXiv:2305.19906*, 2023. 2

[41] Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge. Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 13–23. Springer, 2023. 1, 2, 3, 4

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3