Realistic Endoscopic Illumination Modeling for NeRF-Based Data Generation

Dimitrios Psychogyios(✉), Francisco Vasconcelos, and Danail Stoyanov

University College London, London, UK
{dimitris.psychogyios.19,f.vasconcelos,danail.stoyanov}@ucl.ac.uk

Abstract. Expanding training and evaluation data is a major step towards building and deploying reliable localization and 3D reconstruction techniques during colonoscopy screenings. However, training and evaluating pose and depth models in colonoscopy is hard as available datasets are limited in size. This paper proposes a method for generating new pose and depth datasets by fitting NeRFs in already available colonoscopy datasets. Given a set of images, their associated depth maps and pose information, we train a novel light source location-conditioned NeRF to encapsulate the 3D and color information of a colon sequence. Then, we leverage the trained networks to render images from previously unobserved camera poses and simulate different camera systems, effectively expanding the source dataset. Our experiments show that our model is able to generate RGB images and depth maps of a colonoscopy sequence from previously unobserved poses with high accuracy. Code and trained networks can be accessed at https://github.com/surgical-vision/REIM-NeRF.

Keywords: Surgical Data Science · Surgical AI · Data generation · Neural Rendering · Colonoscopy

1 Introduction

During colonoscopy screenings, localizing the camera and reconstructing the colon directly from the video feed could improve the detection of polyps and help with navigation. Such tasks can be either treated individually using depth [3,6,15,16] and pose estimation approaches [1,14] or jointly, using structure from motion (SfM) and visual simultaneous localization and mapping (VSLAM) algorithms [5,9,10]. However, the limited data availability present in surgery often makes evaluation and supervision of learning-based approaches difficult.

To address the lack of data in surgery, previous work has explored both synthetic pose and depth data generation [2,15], and real data acquisition [2,4,12].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_51.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 535–544, 2023.
https://doi.org/10.1007/978-3-031-43996-4_51

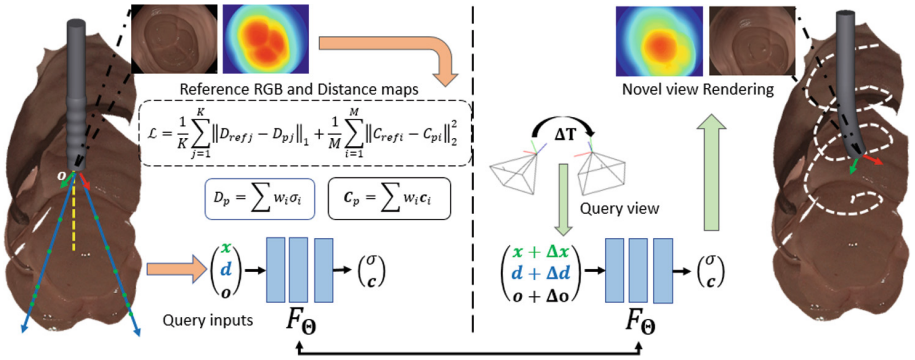


Fig. 1. (Left) The reference depth and RGB images (yellow trajectory) are used to learn an implicit representation of a scene F_θ. (Right) After training, the user can define a new trajectory (white) and extent the original dataset. (Color figure online)

Generating datasets from endoscopic sequences using game engines is scalable and noise-free but often cannot replicate the material properties and lighting conditions of the target environment. In contrast, capturing real data is a laborious process that often introduces sources of error.

Neural Radiance Field (NeRF) [11] networks aim to learn an implicit 3D representation of a 3D scene from a set of images captured from known poses, enabling image synthesis from previously unseen viewpoints. NeRF models render 3D geometry and color, including view-dependent reflections, enabling the rendering of photo-realistic images and geometrically consistent depth maps. EndoNeRF [20] applied NeRF techniques for the first time on surgical video. The method fits a dynamic NeRF [13] on laparoscopic videos, showing tools manipulating tissue from a fixed viewpoint. After training, the video sequences were rendered again without the tools obstructing the tissue. However, directly applying similar techniques in colonoscopy, is challenging because NeRF assumes fixed illumination. As soon as the endoscope moves, changes in tissue illumination result in color ambiguities.

This paper aims to mitigate the depth and pose data scarcity in colonoscopy. Inspired by work in data generation using NeRF [18], we present an extension of NeRF which makes it more suitable for use in endoscopic scenes. Our approach aims to expand colonoscopy VSLAM datasets [4] by rendering views from novel trajectories while allowing simulation of different camera models Fig. 1. Our approach addresses the scalability issues of real data generation techniques while reproducing realistic images. Our main contributions are: 1)The introduction of a depth-supervised NeRF variant conditioned on the location of the endoscope's light source. This extension is important for modeling variation in tissue illumination while the endoscope moves. 2) We evaluate our model design choices on the C3VD dataset and present renditions of the dataset from previously unseen viewpoints in addition to simulating different camera systems.

2 Method

Our method requires a set of images with known intrinsic and extrinsic camera parameters and sparse depth maps of a colonoscopy sequence. This information is already available in high-quality VSLAM [2,4,12] datasets which we wish to expand or can be extracted by running an SfM pipeline such as COLMAP [17] on prerecorded endoscopic sequences. Our pipeline involves first optimizing a special version of NeRF modified to model the unique lighting conditions present in endoscopy. The resulting network is used to render views and their associated dense depth maps from user-defined camera trajectories while allowing to specify of the camera model used during rendering. Images rendered from our models closely resemble the characteristics of the training samples. Similarly, depth maps are geometrically consistent as they share a commonly learned 3D representation. Those properties make our method appealing as it makes realistic data generation easy, configurable, and scalable.

2.1 Neural Radiance Fields

A NeRF [11] implicitly encodes the geometry and radiance of a scene as a continuous volumetric function $F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma))$. The inputs of F_Θ are the 3D location of a point in space $\mathbf{x} = (x, y, z)$ and the 2D direction from which the point is observed $\mathbf{d} = (\phi, \theta)$. The outputs are the red-green-blue (RGB) color $\mathbf{c} = (r, g, b)$ and opacity σ. F_θ is learned and stored in the weights of two cascaded multi-layer perceptron (MLP) networks. The first, f_σ, is responsible for encoding the opacity σ of a point, based only on $\mathbf{x}$. The second MLP, f_c, is responsible for encoding the point's corresponding $\mathbf{c}$ based on the output of the f_σ and $\mathbf{d}$. NeRFs are learned using differentiable rendering techniques given a set of images of scenes and their associated poses. Optimization is achieved by minimizing the L2 loss between the predicted and reference color of all pixels in the training images. To predict the color of a pixel, a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is defined in space starting for the origin of the corresponding image o and heading towards $\mathbf{d}$ which is the direction from where light projects to the corresponding pixel. N points $\mathbf{r}(t_i), i \in [n, f]$ are sampled along the ray between a near t_n and a far t_f range to query NeRF for both σ and $\mathbf{c}$. The opacity values σ of all points along the $\mathbf{r}$ can be used to approximate the accumulated transmittance T_i as defined in Eq. (1), which describes the probability of light traveling from $\mathbf{o}$ to $\mathbf{r}(t_i)$. T_i together with the color output of f_σ for every point along the ray, can be used to compute the pixel color C_p using alpha composition as defined in Eq. (2)

$$T_i = \exp\left(-\sum_{j=0}^{i-1} \sigma_j \Delta_j\right), \Delta_k = t_{k+1} - t_k \tag{1}$$

$$C_p = \sum_{i=1}^{N} w_i \mathbf{c}_i, \text{ where } w_i = T_i(1 - \exp(-\Delta_i \sigma_i)) \tag{2}$$

Similarly, the expected ray termination distance D_p can be computed from Eq. (3), which is an estimate of how far a ray travels from the camera until it hits solid geometry. D_p can be converted to z-depth by knowing the uv coordinates of the corresponding pixel and camera model.

$$D_p = \sum_{i=1}^{N} w_i t_i \tag{3}$$

In practice, NeRF uses two pairs of MLPs. Initially, a coarse NeRF $F_{\Theta c}$ is evaluated on N_c samples along a ray. The opacity output of the coarse network is used to re-sample rays with more dense samples where opacity is higher. The new N_f ray samples are used to query a fine NeRF $F_{\Theta f}$. During both training and inference, both networks are working in parallel. Lastly, to enable NeRF to encapsulate high-frequency geometry and color details, every input of F_Θ is processed by a hand-crafted positional encoding module $\gamma(\cdot)$, using Fourier features [19].

2.2 Extending NeRF for Endoscopy

Light-Source Location Aware MLP. During a colonoscopy, the light source always moves together with the camera. Light source movement results in illumination changes on the tissue surface as a function of both viewing direction (specularities), camera location (exposure changes), and distance between the tissue and the light source (falloff) Fig. 2. NeRF [11], only models radiance as a function of viewing direction as this is enough when the scene is lit uniformly from a fixed light source and the camera exposure is fixed. To model changes in illumination as a function of light source location, we extend the original NeRF formulation by conditioning f_c on both the 2D ray direction $\gamma(\mathbf{d})$ and also the location of the light source $\mathbf{o}$. For simplicity, throughout this work, we assume a single light source co-located with the camera. This parameterization allows the network to learn how light decays as it travels away from the camera and adjusts the scene brightness accordingly.

Depth Supervision. NeRF achieves good 3D reconstruction of scenes using images captured from poses distributed in a hemisphere [11]. This imposes geometric constraints during the optimization because consistent geometry would result in a 3D point projecting in correct pixel locations across different views. Training a NeRF on colonoscopy sequences is hard because the camera moves along a narrow tube-like structure and the colon wall is often texture-less. Supervising depth together with color can guide NeRF to learn a good 3D representation even when pose distribution is sub-optimal [7]. In this work, we compute the distance between the camera and tissue D_{ref} from the reference depth maps and we sample K out of M pixel Eq. (5) to optimize both color and depth as

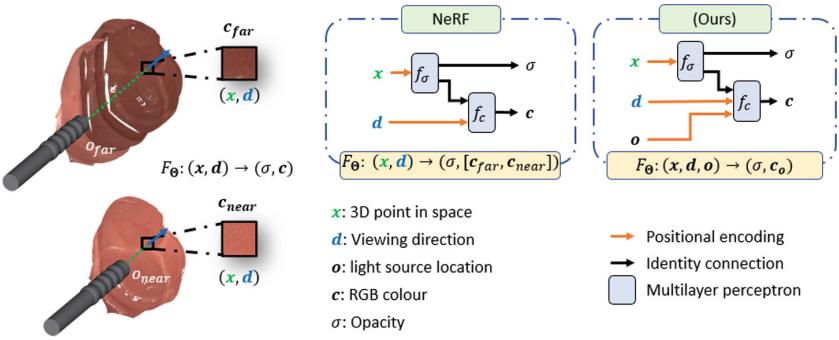


Fig. 2. (left) The color of a point as seen from the colonoscope, from two different distances due to changes in illumination. (center) original NeRF formulation is not able to assign different colors for the same point and viewing direction. (right) ours is conditioned on the location of the light source and the point, modeling light decay.

described in Eq. (4).

$$\mathcal{L} = \frac{1}{K}\sum_{j=1}^{K}\|D_{refj} - D_{pj}\|_1 + \frac{1}{M}\sum_{i=1}^{M}\|C_{refi} - C_{pi}\|_2^2 \tag{4}$$

$$D_{refj} \sim \mathcal{U}[D_{ref1}, D_{refM}], i \in K \leq |M| \tag{5}$$

$\|\cdot\|_1$ is the L1 loss, $\|\cdot\|_2^2$ is the L2 loss, $\mathcal{U}$ denotes uniform sampling.

3 Experiments and Results

3.1 Dataset

We train and evaluate our method on C3VD [4], which provides 22 small video sequences captured from a real wide-angle colonoscopy at 1350×1080 resolution, moving inside 4 different colon phantoms. The videos include sequences from the colon cecum, descending, sigmoid, and transcending. Videos range from 61 to 1142 frames adding to 10.015 in total. We use the per-frame camera poses and set $K/M = 0.03$ in Eq. (5). For each scene, we construct a training set using one out of every 5 frames. We further remove and allocate one out of every 5 poses from the training set for evaluation. Frames not present in either the train or evaluation set are used for testing. We choose to sample both poses and depth information to allow our networks to interpolate more easily between potentially noisy labels and also create a dataset that resembles the sparse output of SfM algorithms.

3.2 Implementation Details

Before training, we spatially re-scale and shift the 3D scene from each video sequence such that every point is enclosed within a cube with a length of two,

centered at (0,0,0). Prescaling is important for the positional encoding module $\gamma(\mathbf{x})$ to work properly. We configure positional encoding modules to compute 10 frequencies for each component of $\mathbf{x}$ and 4 frequencies for each component of $\mathbf{d}$ and $\mathbf{o}$. We train models on images of 270×216 resolution to ignore both depth and RGB information outside a circle with a radius of 130 pixels centered at a principal point to avoid noise due to inaccuracies of the calibration model. We used Adam optimizer [8] with a batch size of 1024 for about 140K iterations for all sequences, with an initial learning rate of 5e-4 which we later multiply by 0.5 at 50% and 75% of training. We set the number of samples along the ray to $N_c = 64$ and $N_f = 64$. We configure positional encoding modules to compute 10 frequencies for each component of $\mathbf{x}$ and 4 frequencies for each component of $\mathbf{d}$ and $\mathbf{o}$. Each model from this work is trained for around 30 min on 4 graphics cards from an NVIDIA DGX-A100.

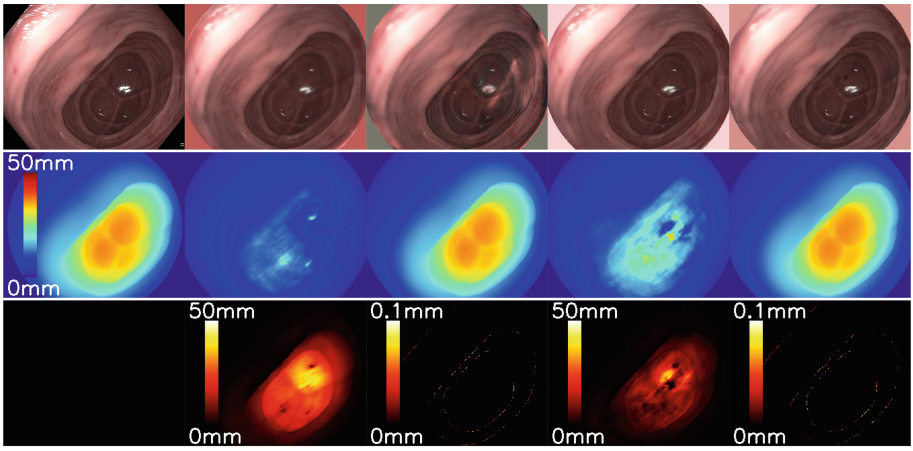
3.3 Model Ablation Study

We ablate our model showing the effects of conditioning NeRF on the light source location and supervising depth. To assess RGB reconstruction, we measure the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) between reconstructed and reference images at 270×216 resolution. We evaluate depth using the mean squared error (MSE) in the original dataset scale. For each metric, we report the average across all sequences together with the average standard deviation in Table 1. Conditioning NeRF based on the light source location (this work) produces better or equally good results compared to vanilla NeRF for all metrics. Both depth-supervised models, learn practically the same 3D representation but our model achieves better image quality metrics.

Table 1. Mean and standard deviation metrics of every method aggregated across all sequences of C3VD dataset.

Model	PSNR ↑	SSIM ↑	Depth MSE (mm)↓
NeRF	32.097 ± (1.173)	0.811 ± (0.021)	4.263 ± (1.178)
NeRF + ls_loc(Ours)	32.489 ± (1.128)	0.820 ± (0.018)	1.866 ± (0.594)
NeRF + depth	30.751 ± (1.163)	0.788 ± (0.022)	0.015 ± (0.016)
Full model (Ours)	31.662 ± (1.082)	0.797 ± (0.020)	0.013 ± (0.018)

Figure 3 shows renditions of each model of the ablation study for the same frame. Both non-depth-supervised models failed to capture correct geometry but were able to reconstruct accurate RGB information. Since non-depth-supervised networks were optimized only on color with weak geometric constraints, they learn floating artifacts in space which when viewed from a specific viewpoint, closely approximate the training samples. In contrast, depth-supervised networks learned a good representation of (3D) geometry while being able to reconstruct



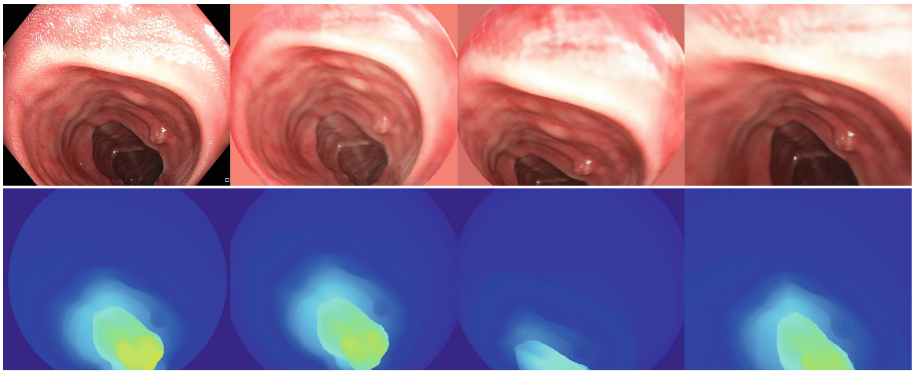
(a) Reference (b) NeRF (c) NerF+Depth (d) NeRF+ls_loc (e) Full Model

Fig. 3. Renditions from all model ablations compared to a reference RGB image and depth map. Reference and reconstructed images (top row), depth maps (middle row), and MSE depth difference between the reference and each model prediction in different scales (bottom row).

RGB images accurately. The depth-supervised NeRF model produces flare artifacts in the RGB image. That is because, during optimization, points are viewed from the same direction but at different distances from the light source. Our Full model is able to cope with illumination changes resulting in artifact-free images and accurate depth. Notably, most of the errors in depth for the depth-supervised approaches are located around sharp depth transitions. Such errors in depth may be a result of inaccuracies in calibration or imperfect camera pose information. Nevertheless, we argue that using RGB images and depth maps produced from our approach can be considered error-free because during inference the learned 3D geometry is fixed and consistent across all rendered views.

3.4 Data Generation

We directly use our proposed model of the d4v2 C3VD scene from the ablation study to render novel views and show results in Fig. 4. In the second column, we show an image rendered from a previously unseen viewpoint, radially offset from the original camera path. Geometry is consistent and the RGB image exhibits the same photo-realistic properties observed in the training set. In the third column, we render a view by rotating a pose from the training trajectory. In the rotated view, the tissue is illuminated in a realistic way even though the camera never pointed in this direction in the training set. In the fourth column, we show an image rendered using a pinhole camera model whilst only fisheye images were used during training. This is possible because NeRF has captured a good representation of the underlying scene and image formation is done by



(a) Reference (b) Translation (c) Rotation (d) Pinhole

Fig. 4. Data generated using our work. The top row shows RGB images and the second row shows depth maps. a)Reference view from frame 60 of the d4v2 C3VD sequence. b) Translating (a) along all axis. c) Rotating (a) around both the x and y-axis d) Simulating a pinhole model in (a).

projecting rays in space based on user-defined camera parameters. All the above demonstrate the ability of our method to render images from new, user-defined, trajectories and camera systems similar to synthetic data generation while producing photo-realistic images.

4 Conclusion

We presented an approach for expanding existing VSLAM datasets by rendering RGB images and their associated depth maps from user-defined camera poses and models. To achieve this task, we propose a novel variant of NeRF, conditioned on the location of the light source in 3D space and incorporating sparse depth supervision. We evaluate the effects of our contributions on phantom datasets and show that our work effectively adapts NeRF techniques to the lighting conditions present in endoscopy. We further demonstrate the efficacy of our method by showing RGB images and their associated depth maps rendered from novel views of the target endoscopic scene. 3D information and conditioning NeRF based on the light source location made NeRF suitable for use in Endoscopy. Currently, our method assumes a static environment and requires accurate camera intrinsic and extrinsic information. Subsequent work can incorporate mechanisms to represent deformable scenes [13] and refine camera parameters during training [21]. Further research can investigate adopting the proposed model for data generation in other endoscopic scenes or implementing approaches to perform label propagation for categorical data [22]. We hope this work will mitigate the data scarcity issue currently present in the surgical domain and inspire the community to leverage and improve neural rendering techniques for data generation.

Aknowledgements. This research was funded, in whole, by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145/Z/16/Z]; the Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P012841/1]; the Royal Academy of Engineering Chair in Emerging Technologies Scheme, and Horizon 2020 FET (863146). For the purpose of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

References

1. Armin, M.A., Barnes, N., Alvarez, J., Li, H., Grimpen, F., Salvado, O.: Learning camera pose from optical colonoscopy frames through deep convolutional neural network (CNN). In: Cardoso, M.J., et al. (eds.) CARE/CLIP -2017. LNCS, vol. 10550, pp. 50–59. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67543-5_5
2. Azagra, P., et al.: Endomapper dataset of complete calibrated endoscopy procedures. arXiv preprint arXiv:2204.14240 (2022)
3. Batlle, V.M., Montiel, J.M., Tardós, J.D.: Photometric single-view dense 3D reconstruction in endoscopy. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4904–4910. IEEE (2022)
4. Bobrow, T.L., Golhar, M., Vijayan, R., Akshintala, V.S., Garcia, J.R., Durr, N.J.: Colonoscopy 3D video dataset with paired depth from 2D-3D registration. arXiv preprint arXiv:2206.08903 (2022)
5. Chen, R.J., Bobrow, T.L., Athey, T., Mahmood, F., Durr, N.J.: SLAM endoscopy enhanced by adversarial depth prediction. arXiv preprint arXiv:1907.00283 (2019)
6. Cheng, K., Ma, Y., Sun, B., Li, Y., Chen, X.: Depth estimation for colonoscopy images with self-supervised learning from videos. In: de Bruijne, M., et al. (eds.) MICCAI 2021, Part VI. LNCS, vol. 12906, pp. 119–128. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_12
7. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised NeRF: fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12882–12891 (2022)
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Lamarca, J., Parashar, S., Bartoli, A., Montiel, J.: DefSLAM: tracking and mapping of deforming scenes from monocular sequences. IEEE Trans. Rob. **37**(1), 291–303 (2020)
10. Ma, R., et al.: RNNSLAM: reconstructing the 3D colon to visualize missing regions during a colonoscopy. Med. Image Anal. **72**, 102100 (2021)
11. 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: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 405–421. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_24
12. Ozyoruk, K.B., et al.: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Med. Image Anal. **71**, 102058 (2021)
13. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10318–10327 (2021)

14. Rau, A., Bhattarai, B., Agapito, L., Stoyanov, D.: Bimodal camera pose prediction for endoscopy. arXiv preprint arXiv:2204.04968 (2022)
15. Rau, A., et al.: Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int. J. Comput. Assist. Radiol. Surg. **14**(7), 1167–1176 (2019). https://doi.org/10.1007/s11548-019-01962-w
16. Rodriguez-Puigvert, J., Recasens, D., Civera, J., Martinez-Cantin, R.: On the uncertain single-view depths in colonoscopies. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part III. LNCS, vol. 13433, pp. 130–140. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_13
17. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
18. Tancik, M., et al.: Block-NeRF: scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8248–8258 (2022)
19. Tancik, M., et al.: Fourier features let networks learn high frequency functions in low dimensional domains. Adv. Neural. Inf. Process. Syst. **33**, 7537–7547 (2020)
20. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022, Part VII. LNCS, vol. 13437, pp. 431–441. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_41
21. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: NeRF–: neural radiance fields without known camera parameters. arXiv preprint arXiv:2102.07064 (2021)
22. Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J.: In-place scene labelling and understanding with implicit scene representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15838–15847 (2021)