Depth-Supervised NeRF for Multi-View RGB-D Operating Room Images

Beerend G.A. Gerats[1,3], Jelmer M. Wolterink[2], and Ivo A.M.J. Broeders[1,3]

[1] Centre for Artificial Intelligence, Meander Medisch Centrum, Amersfoort, The Netherlands
[2] Department of Applied Mathematics & Technical Medical Center, University of Twente, Enschede, The Netherlands
[3] Robotics and Mechatronics, University of Twente, Enschede, The Netherlands

Abstract. Neural Radiance Fields (NeRF) is a powerful novel technology for the reconstruction of 3D scenes from a set of images captured by static cameras. Renders of these reconstructions could play a role in virtual presence in the operating room (OR), e.g. for training purposes. In contrast to existing systems for virtual presence, NeRF can provide real instead of simulated surgeries. This work shows how NeRF can be used for view synthesis in the OR.

A depth-supervised NeRF (DS-NeRF) is trained with three or five synchronised cameras that capture the surgical field in knee replacement surgery videos from the 4D-OR dataset. The algorithm is trained and evaluated for images in five distinct phases before and during the surgery. With qualitative analysis, we inspect views synthesised by a virtual camera that moves in 180 degrees around the surgical field. Additionally, we quantitatively inspect view synthesis from an unseen camera position in terms of PSNR, SSIM and LPIPS for the colour channels and in terms of MAE and error percentage for the estimated depth.

DS-NeRF generates geometrically consistent views, also from interpolated camera positions. Views are generated from an unseen camera pose with an average PSNR of 17.8 and a depth estimation error of 2.10%. However, due to artefacts and missing of fine details, the synthesised views do not look photo-realistic. Our results show the potential of NeRF for view synthesis in the OR. Recent developments, such as NeRF for video synthesis and training speedups, require further exploration to reveal its full potential.

Keywords: Neural Radiance Fields · RGB-D Imaging · Operating Room

1 Introduction

The presentation of a novel approach for view synthesis called Neural Radiance Fields (NeRF) [6] caused an explosion of interest in the field of computer vision. Even though the original paper was presented recently, a large number of follow-up studies have already been published [19]. However, NeRF-based methods for clinical use remain largely unexplored. At the same time, there is increasing interest in virtual presence in operating rooms (ORs) during real or simulated surgeries. For example, the National Autonomous University of Mexico used virtual reality to simulate an OR, including distractions and interruptions, to train paediatric surgical residents during the COVID-19 pandemic [10]. Apart from surgical training with virtual reality [15], physician assistant students can greatly benefit from training in a virtual environment [2].

NeRF is a powerful technology for the reconstruction of a 3D scene from a set of images that capture the scene from various camera positions. Likewise, this technology could be used to build 3D representations of clinical environments such as the OR. NeRF requires only a set of calibrated cameras that capture the scene of interest and does not depend on any physical marker or complex sensory system. When applied to overhead cameras in the OR, NeRF could build 3D reconstructions of the surgical environment, providing the ability to render videos that virtually record surgical scenes. These videos could be used for virtual presence, helping students or clinicians to experience a surgery without the need to be actually present. The advantage over the existing systems for virtual presence in the OR, such as the ones mentioned above, is that NeRF can provide real instead of simulated surgeries.

In this paper, we show how NeRF can be used for view synthesis in the OR. We show that depth-supervision [1] helps to increase the render quality and reduces the need for many camera positions. This is particularly relevant for clinical environments, where it is generally impossible to capture the scene with tens to hundreds of cameras. In fact, we find that a depth-supervised NeRF with only three or five synchronised camera views captures the surgical field, and is able to generate images of the surgical intervention from a range of camera angles. In contrast to existing depth-supervision in NeRF, we directly optimise our model using measured RGB-D sensor data instead of estimated depth from a Structure from Motion (SfM) algorithm.

2 Related work

2.1 Neural Radiance Fields

NeRF is a method for volume rendering, based upon the *implicit representation* of a 3D scene in the weights of a neural network F_Θ [6]. This network is generally a standard multi-layer perceptron (MLP) that takes a 5D vector (x, y, z, θ, ϕ) as input and that outputs a 4D vector (RGB, σ). An input vector consists of a 3D location (x, y, z) in the captured scene and an orientation (θ, ϕ) from which this location is viewed. F_Θ returns for each vector a colour RGB and a volume density σ. With this simple setup, NeRF can reconstruct images by casting a viewing ray from each pixel, sampling points along that ray, asking the MLP to find the colours and densities for these points and to sum over these results. In this way, it is possible to use a discrete set of sampled points in the 3D scene, while representing the 3D scene in continuous form. Reconstructed images are compared with ground truth images that are taken from the same camera positions. Importantly, the rendering function that sums over the found colours and densities is differentiable such that the MLP can be optimised by stochastic gradient descent. The loss function is often a total squared error between the colours of the rendered and the ground truth images.

2.2 NeRF with Depth Priors

Although the original NeRF has the ability to synthesise photo-realistic images from unseen camera perspectives, there are some limitations. One is that NeRF does not guarantee to capture 3D geometry accurately, which appears to be problematic for poorly textured areas that often occur in indoor scenes [18]. Additionally, points are randomly sampled along a viewing ray, even though most of these points will describe empty space. It is likely that these problems will play a role when NeRF is applied to scenes in the OR, as these involve a large indoor environment.

Several solutions are proposed that enforce NeRF to find a more accurate geometry by regularisation with depth priors. Nerfing MVS [18] provides a guided optimisation scheme for NeRF, where points are sampled along a viewing ray only around depth values found earlier. A sparse set of depth values is found by applying the COLMAP SfM algorithm [14] on the multi-view images. The sparse sets are used to train a depth completion network that provides full sets of depth values. With this optimisation scheme, NeRF is able to provide more accurate depth maps and has a better understanding of the scene's 3D geometry. [8] have shown that computational costs at inference can be reduced significantly when conditioning NeRF on depth information. Their method involves another sampling strategy to include only points located around surfaces. These point locations are predicted by an oracle network.

[13] present a similar approach to leverage depth priors with depth-guided sampling. Additionally, the authors propose to add a loss term that enforces NeRF to terminate rays close to the most certain depth observations that are provided by a depth completion network. The work by [12] shows how a depth-constrained NeRF is able to reconstruct large urban areas by training the network with an additional loss term as well. The loss enforces NeRF to represent a large amount of volume density around ground truth depth values found with a LiDAR sensor and a small amount in the area along the viewing ray that is closer to the camera. The depth-supervised NeRF [1] uses an additional loss function that operates with known depth values from RGB-D data or that are obtained from the COLMAP algorithm. This loss constraints the distribution of ray terminations to be similar to the known depth distribution by minimising the KL divergence between the two. The key insight is that they consider the uncertainty of the known depth values and constrain NeRF proportionally to the uncertainty.

The work above shows that depth priors help to reduce the number of camera poses required for NeRF to synthesise high-quality images. This is particularly relevant for the OR, where it is not possible to put cameras at all locations and the overall number of available cameras is limited.

2.3 NeRF for Clinical Interventions

Although NeRF has not been widely adopted in the field of computer-assisted interventions, the technology has been used for 3D reconstruction of soft tissues in robotic surgery videos by [16]. Because the laparoscopic view is a single view, their method uses time-dependent modelling of neural radiance and displacement fields, based upon D-NeRF [11]. Using the stereoscopic camera of the surgical robot, the method finds depth maps along the coloured image frames. The depth information is used to constrain NeRF optimisation with an additional loss function.

3 Methods

3.1 Depth-Supervised NeRF

We use the depth-supervised NeRF (DS-NeRF) by [1] for building 3D reconstructions of OR scenes. This method regularises the training with an additional depth loss such that a model can be optimised with relatively few camera positions. The key idea in DS-NeRF is that most viewing rays terminate at the closest surface, which is often opaque. Therefore, most volume density should be found close to the distance of this surface along the viewing ray. DS-NeRF enforces such a distribution of volume density by minimising the KL divergence between the volume density distribution $h_i(t)$ and a normal distribution around the ground truth depth d_i of keypoint $x_i \in X$:

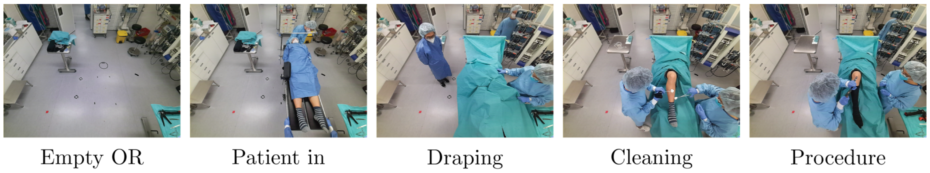
$$\mathrm{KL}[\mathbb{N}(d_i, \hat{\sigma}_i) \| h_i(t)], \tag{1}$$

where X is the set of all keypoints in an image for which the depth is known and t is the far endpoint of the viewing ray. The variance $\hat{\sigma}_i$ is set to the uncertainty of the depth estimation for keypoint x_i. When depth is estimated with COLMAP, the uncertainty is calculated by re-projecting the keypoint to and from another camera position in which the keypoint is visible. In RGB-D data, however, the depth values are measurements rather than estimations. Therefore, we set $\hat{\sigma}_i = 1.0$ for all keypoints such that each depth value is weighted equally. We sample 10^5 depth values in each ground truth image at random positions, where pixels that have a depth value equal to zero are never included.

3.2 Dataset

The 4D-OR dataset [9] from the Technical University of Munich contains RGB-D images and camera poses from ten simulated knee replacement surgeries. We use this dataset to train a DS-NeRF in the reconstruction of five distinct phases in the surgery: "empty OR", "patient in", "draping", "cleaning" and "procedure" (see Figure 1). In this way, we can evaluate the quality of synthesised views for different OR activities.

The dataset contains synchronised images from six cameras that have fixed locations in the OR. Three of these cameras are located above the surgical field,



Empty OR Patient in Draping Cleaning Procedure

Fig. 1. Five distinct phases during or before the surgery. All images are obtained with the same camera.

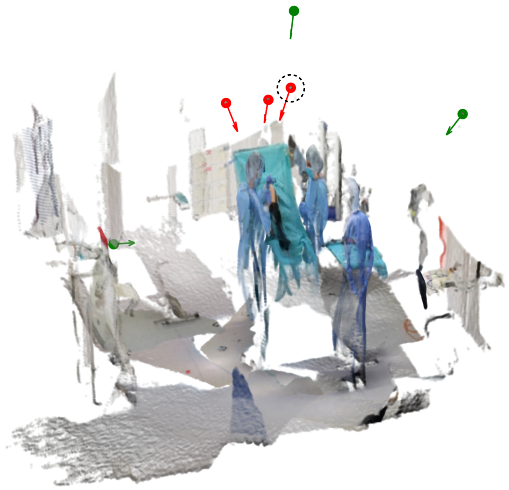


Fig. 2. Locations of the RGB-D cameras are indicated by a red or green dot. The viewing angle is directed by the arrows. Red dots indicate three cameras located above the surgical field, whereas the green dots capture the OR from very different perspectives. The scene is a coloured point cloud formed by camera projection using depth values.

each rotated approximately 90 degrees around the yaw axis (the red dots in Figure 2). Two of the three other cameras capture the OR from a wide perspective, while the sixth camera records from a position that is closer to the ground (the green dots in Figure 2).

3.3 Experimental Setup

Our experiments are separated into two steps: a qualitative and quantitative analysis. In the qualitative step, we train DS-NeRF with three images from the cameras above the surgical field (the “red” cameras). After training, we instruct the algorithm to synthesise views from these exact poses as well as from interpolated poses that together form a 180 degrees rotation around the surgical field. In the quantitative step, we train DS-NeRF with five images captured by the camera positions that are not circled in Figure 2. At inference, the algorithm synthesises views from the remaining sixth position (the circled camera pose) which is unseen during training. We compare the resulting coloured images with ground truth images in terms of PSRN, SSIM [17] and LPIPS [20]. Similarly, we evaluate the quality of depth perception in terms of mean absolute error (MAE) and in percentage of the ground truth depth value. Note that depth values equal to zero in the ground truth images are not included in this evaluation, since these values can be perceived as inaccurate measurements.

For all reconstructed scenes, DS-NeRF is trained to synthesise images with a resolution of 512×384 pixels with the following hyperparameters: 4096 selected rays per batch, 64 points sampled per ray, 5×10^4 iterations and depth loss weighting factor λ_D set to 0.1.

4 Results

4.1 Qualitative Analysis

The synthesised views for surgical phases "empty OR", "patient in" and "procedure" are given in Figure 3. When comparing the top-row synthesised views with the ground truth images, it can be seen that DS-NeRF is able to reconstruct the surgical scenes independent of surgical phase. Lighting conditions such as reflections on the floor and shadows are realistically rendered. However, the reconstructed scene looks smoothed with missing details, e.g. in the keyboard of the mobile monitor or the sterile clothing of the physician.

Synthesised views from the interpolated camera poses (images without red borders in Figure 3) correctly grasp the geometry of the scenes and realistically find the right lighting conditions. Nevertheless, the images do not look realistic due to a number of artefacts. One is the white background pixels that are rendered when DS-NeRF does not find any material density along viewing rays. These are present in the far corners of the OR. Second, a number of objects seem to be misaligned and occurring twice. For example, the tape on the floor and the instrument table are not always represented correctly. Surprisingly, this artefact seems to be less present in the "procedure" phase. Third, fine-grained details, such as the surgical instruments in the hands of the left physician, are missing to make the images convincingly realistic.

4.2 Quantitative Analysis

Results of the quantitative analysis can be found in Table 1. On average, DSNeRF is able to synthesise views from the unseen camera pose with 17.8 PSNR, 0.60 SIMM and 0.47 LPIPS. These results are comparable to the performance of the original NeRF on the NeRF Real dataset [5] when trained with 5 images: 18.2 PSNR, 0.57 SSIM and 0.50 LPIPS [1]. However, DSNeRF on the same dataset gains higher performance scores: 22.6 PSNR, 0.69 SSIM and 0.35 LPIPS. This indicates that our dataset is more challenging for the synthesis of high-quality images. The image quality of the rendered views on the 4D-OR dataset differs per phase and ranges in PSNR from 16.9 for "patient in" to 19.3 for "empty

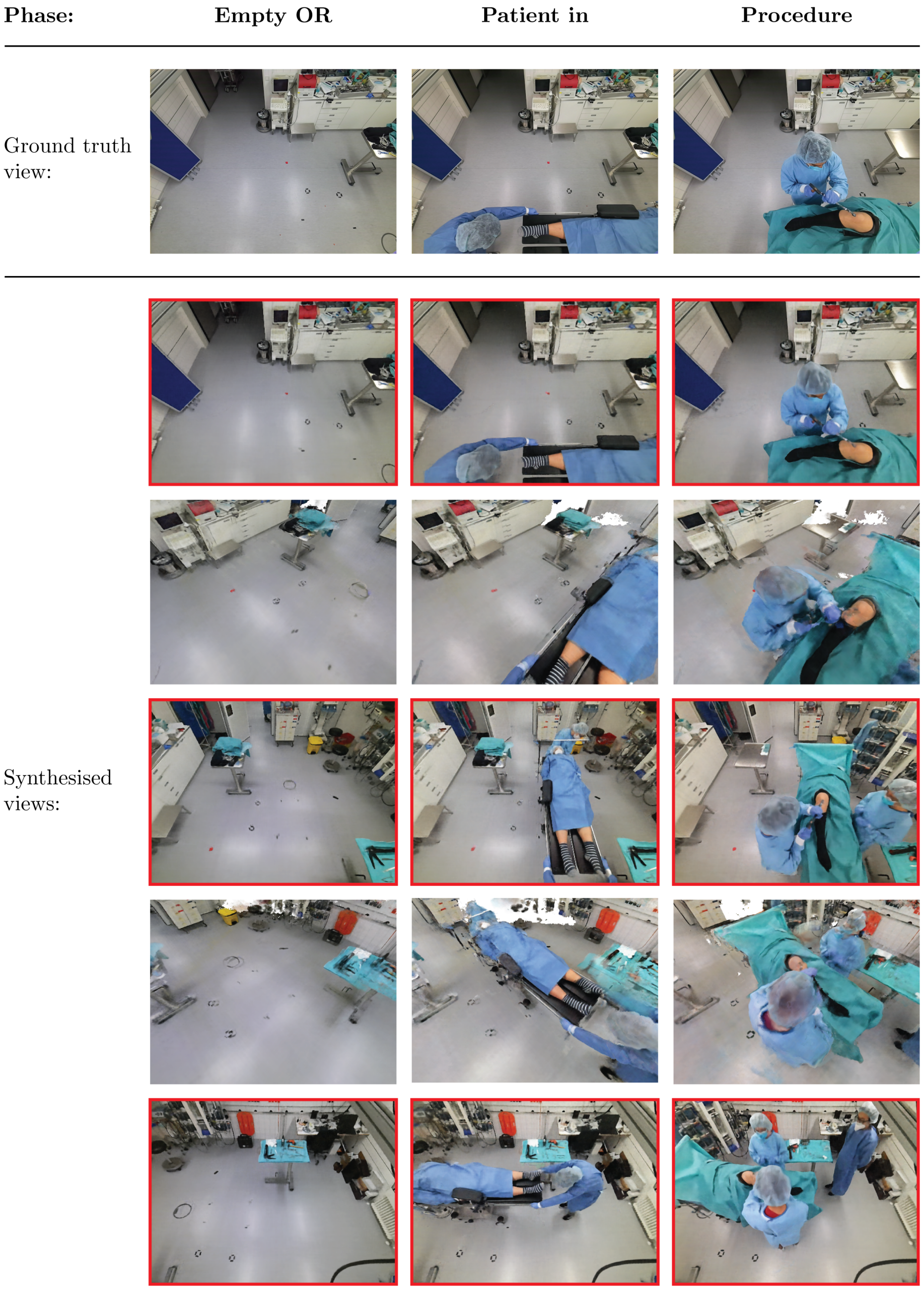


Fig. 3. DS-NeRF synthesised views for three phases in the OR where the virtual camera rotates 180 degrees around the surgical field. The top row displays the ground truth images for the starting camera pose. Views with a red border are generated from the camera poses with which the algorithm is trained, corresponding to the red camera positions in Figure 2.

Table 1. Evaluation metrics comparing DS-NeRF synthesised views with unseen ground truth RGB-D images.

	Colour Image			Depth Map	
Surgery phase	PSNR↑	SSIM↑	LPIPS↓	MAE (in cm)↓	Error (in %)↓
Empty OR	19.3	0.71	0.44	2.3	0.97
Patient in	16.9	0.57	0.48	3.4	1.48
Draping	18.3	0.60	0.46	5.3	2.11
Cleaning	17.3	0.57	0.48	4.4	2.28
Procedure	17.3	0.57	0.50	5.8	3.67
Average	**17.8**	**0.60**	**0.47**	**4.2**	**2.10**

Table 2. Evaluation metrics for NeRF synthesised views with and without depth-supervision.

	Colour Image			Depth Map	
Depth-supervision	PSNR↑	SSIM↑	LPIPS↓	MAE (in cm)↓	Error (in %)↓
✓	**17.8**	**0.60**	**0.47**	**4.2**	**2.10**
×	14.2	0.50	0.64	56.2	23.35

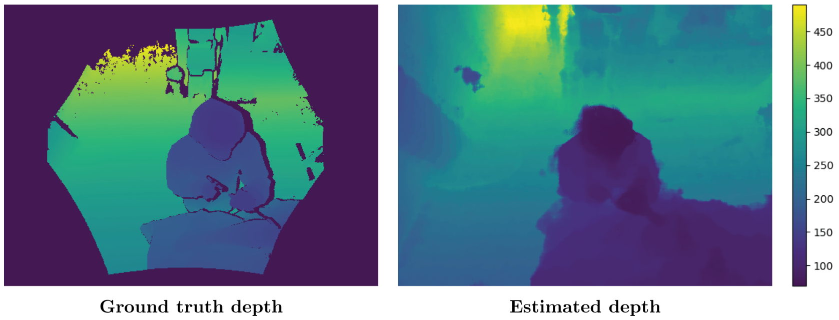
OR”. This shows that the image quality is dependent on the complexity of the surgical scene.

Table 2 displays the quality of the colour images and depth maps for NeRF with and without depth-supervision. The quality of the colour images decreases from 17.8 to 14.2 PSNR, while the error in depth estimation accuracy increases drastically from 4.2 to 56.2 cm. In terms of percentages, the depth error increases from 2.10 to 23.35 %. These results show that depth-supervision helps NeRF to reconstruct the scene’s 3D geometry accurately, resulting in a higher quality of synthesised images.

Figure 4 displays the estimated depth for an unseen camera position in the “procedure” phase in comparison with the ground truth depth channel captured from the same camera pose. It can be seen that DS-NeRF is able to grasp the geometry of the captured scene accurately. Moreover, the algorithm is able to generate depth values that are not present in the ground truth image due to the depth sensor’s hexagon shape or sensor artefacts (e.g. the zero-valued “shadows”).

5 Discussion

In this work, we explored the use of Neural Radiance Fields (NeRF) [6] for reconstructing surgical scenes in the operating room (OR) with multi-view RGB-D images from the 4D-OR dataset [9]. We showed that the application of an



Ground truth depth **Estimated depth**

Fig. 4. Example depth estimation output of DS-NeRF for the "procedure" phase (right) in comparison with the ground truth depth channel (left). Color bar displays distance in cm.

original NeRF does not provide optimal reconstruction results and that the use of depth-supervision benefits the quality of the synthesised views. Depth-supervised NeRF (DS-NeRF) [1] removes the necessity to train the algorithm with tens to hundreds of camera positions, which are difficult to obtain during a surgical procedure.

Visual inspection of synthesised views around the surgical field showed that DS-NeRF is able to generate realistic images from the camera poses that the algorithm is trained with. Additionally, it is possible to use the technology to virtually rotate around the surgical field while the synthesised images remain geometrically consistent. However, the synthesised views from unseen camera poses miss fine details and contain artefacts that make the images look unrealistic. Training with more camera positions, larger image resolutions or stronger geometric priors can potentially help the algorithm produce views with higher quality.

The quantitative evaluation of DS-NeRF in generating views from an unseen camera position finds that the algorithm can produce views with a PSNR of 17.8. In comparison, DS-NeRF could generate views with larger image quality on the NeRF Real and the Redwood-3dscan datasets [1]. The performance difference can be explained by the distinct camera poses in the 4D-OR dataset, making the dataset more challenging for view synthesis. On the contrary, we found that DS-NeRF is able to find accurate depth maps for unseen camera poses, with an average depth error of 2.10%.

We envision several potential uses for NeRF in the synthesis of OR images. The technology could be used for rendering virtual environments displaying real surgeries in 3D. With the current state of NeRF, where the rendering process is time-consuming, the most obvious use is for post-surgery rendering, where videos display the surgery from angles where no camera was originally positioned. In the long term, when NeRF rendering becomes cheaper, one could think of more interactive ways of experiencing a surgical procedure, either during or post-surgery, where the user could choose camera position, angle and zoom. The virtual renders could be static moments in a procedure, like “snapshots”, or could display the surgery over time.

Besides improving the quality of view synthesises, there are other developments that could increase the impact of NeRF for 3D reconstruction in the OR. One is the synthesis of video instead of static images, which is shown to be possible [4]. The time dimension can put an additional constraint on the training of NeRF, which is likely to benefit the rendering quality. More importantly, the synthesis of video would make NeRF more interesting for the development of virtual training applications. Another development is the drastic speedup in the training and rendering of NeRF algorithms. For example, [7] have shown that it is possible to train high-quality neural graphics in seconds rather than minutes to hours. Last, the implicit 3D representation of the surgical scene could be used for further processing. For example, methods that use 3D representations for detecting human poses [3] have a potential benefit in using NeRF representations of humans in the OR, especially when NeRF is fitted to videos. Other video processing tasks, such as background removal, object detection and segmentation could potentially benefit similarly.

In conclusion, depth-supervised NeRF is able to synthesise views of the surgical field from OR images in which the 3D geometry is captured accurately. To reveal the full potential of NeRF for OR view synthesis, there remain several developments, such as video synthesis, training speedups and the use of implicit representations for downstream video processing tasks, that require further exploration.

Declarations

Research at the Centre for Artificial Intelligence at Meander Medisch Centrum is sponsored by Johnson & Johnson, Ltd. Jelmer M. Wolterink was supported by NWO domain Applied and Engineering Sciences VENI grant (18192).

GitHub projects with the DS-NeRF code implementation [1] and instructions for downloading the 4D-OR dataset [9] are available publicly online.

References

1. 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)
2. Francis, E.R., Bernard, S., Nowak, M.L., Daniel, S., Bernard, J.A.: Operating room virtual reality immersion improves self-efficacy amongst preclinical physician assistant students. Journal of surgical education **77**(4), 947–952 (2020)
3. Gerats, B.G., Wolterink, J.M., Broeders, I.A.: 3d human pose estimation in multi-view operating room videos using differentiable camera projections. arXiv preprint arXiv:2210.11826 (2022)
4. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3d video synthesis from multi-view video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5521–5531 (2022)

5. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
6. 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 (2020)
7. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022). https://doi.org/10.1145/3528223.3530127, `https://doi.org/10.1145/3528223.3530127`
8. Neff, T., Stadlbauer, P., Parger, M., Kurz, A., Mueller, J.H., Chaitanya, C.R.A., Kaplanyan, A., Steinberger, M.: Donerf: Towards real-time rendering of compact neural radiance fields using depth oracle networks. In: Computer Graphics Forum. vol. 40, pp. 45–59. Wiley Online Library (2021)
9. Özsoy, E., Örnek, E.P., Czempiel, T., Tombari, F., Navab, N.: 4d-or: Semantic scene graphs for or domain modeling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2022)
10. Pérez-Escamirosa, F., Medina-Alvarez, D., Ruíz-Vereo, E.A., Ordorica-Flores, R.M., Minor-Martínez, A., Tapia-Jurado, J.: Immersive virtual operating room simulation for surgical resident education during covid-19. Surgical Innovation **27**(5), 549–550 (2020)
11. 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)
12. Rematas, K., Liu, A., Srinivasan, P.P., Barron, J.T., Tagliasacchi, A., Funkhouser, T., Ferrari, V.: Urban radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12932–12942 (2022)
13. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
14. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
15. Seymour, N.E.: Vr to or: a review of the evidence that virtual reality simulation improves operating room performance. World journal of surgery **32**(2), 182–188 (2008)
16. 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)
17. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
18. Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5610–5619 (2021)
19. Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., Sridhar, S.: Neural fields in visual computing and beyond. In: Computer Graphics Forum. vol. 41, pp. 641–676. Wiley Online Library (2022)

20. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)