

# Novel View Synthesis for Unseen Surgery Recordings

Mana Masuda<sup>1</sup>, Hideo Saito<sup>1</sup>, Yoshifumi Takatsume<sup>3</sup> and Hiroki Kajita<sup>2</sup>

<sup>1</sup>Department of Information and Computer Science, Keio University, Yokohama, Japan

<sup>2</sup>Department of Plastic and Reconstructive Surgery, Keio University School of Medicine, Shinjuku-ku, Tokyo, Japan

<sup>3</sup>Department of Anatomy, Keio University School of Medicine, Shinjuku-ku, Tokyo, Japan

**Keywords:** Medical Image Application, Novel View Synthesis.

**Abstract:** Recording surgery in operating rooms is a crucial task for both medical education and evaluation of medical treatment. In this paper, we propose a method for visualizing surgical areas that are occluded by the heads or hands of medical professionals in various surgical scenes. To recover the occluded surgical areas, we utilize a surgery recording system equipped with multiple cameras embedded in the surgical lamp, with the aim of ensuring that at least one camera can capture the surgical area without occlusion. We propose the application of a transformer-based Neural Radiance Field (NeRF) model, originally proposed for normal scenes, to surgery scenes, and demonstrate through experimentation that it is feasible to generate occluded surgical areas. We believe this research has the potential to make our multi-camera recording system practical and useful for physicians.

## 1 INTRODUCTION

The recording of surgeries in an operating room with cameras has proven to be an indispensable task for a variety of purposes, including education, the sharing of surgical technologies and techniques, the performance of case studies of diseases, and the evaluation of medical treatment (Masuda et al., 2022; Hachiuma et al., 2020; Matsumoto et al., 2013; Sadri et al., 2013; Shimizu et al., 2020). The targets that depict the surgery, such as the surgical field, surgeon's hands, and surgical tools, should be captured in the recordings of surgeries for these purposes.

It is challenging, however, to continuously record these targets without any occlusion. The most straightforward method for recording surgery is to attach a camera to the operating room environment, but this may result in occlusion of the surgical field by surgeons, nurses, or surgical machinery. Another option is to attach the camera to the head of the surgeon and record from a first-person perspective, but this video is often affected by motion blur due to head movements and the surgeon may not always be looking at the surgical field. Additionally, the camera attached to the surgeon's head may interfere with the surgery itself. As a result, cameras attached to the op-

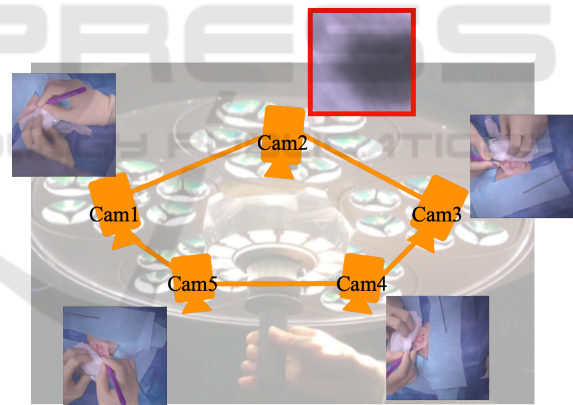




Figure 1: We expect that at least one of the cameras attached to the surgical lamp will capture the surgical field (Cam1, 3, 4, 5), but not all of the cameras will for most of the cases (Cam2).

erating room environment or the surgeon's head are not effective solutions for recording surgeries.

Shimizu *et al.* (Shimizu et al., 2020) proposed a novel surgical lamp system with multiple embedded cameras to record surgeries. A generic surgical lamp, commonly used in open surgeries, has multiple light bulbs that illuminate the surgical field from multiple directions in order to reduce shadows caused by surgeons. Shimizu *et al.* expect that at least one of the multiple light bulbs will almost always illuminate the surgical field. They embedded the cameras into each

<sup>a</sup> <https://orcid.org/0000-0002-9050-5306>

<sup>b</sup> <https://orcid.org/0000-0002-2421-9862>

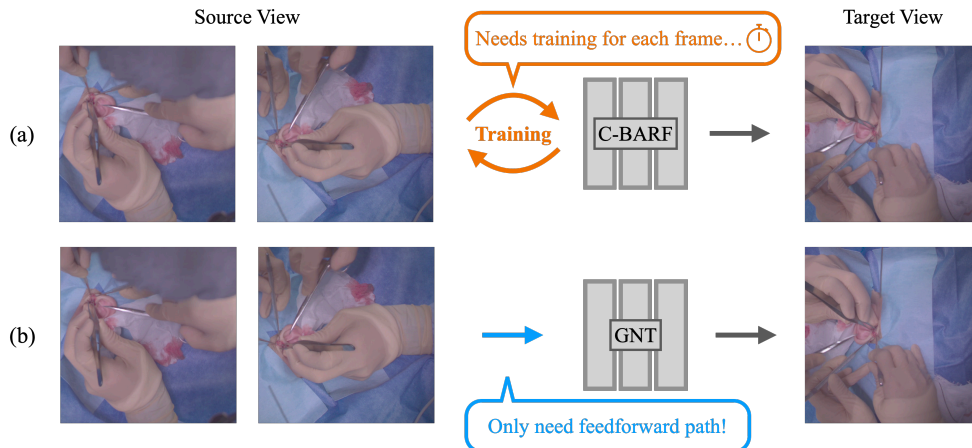


Figure 2: (a) Conditional-BARF (C-BARF) (Masuda et al., 2022) requires training the network for each frame, which takes a tremendous amount of time to create a video from the beginning to the end of the surgery. (b) We propose using Generalizable NeRF Transformer (GNT) (Wang et al., 2022) for surgery scenes. This method does not require training even when rendering unseen scenes, so it is possible to create the video in significantly less time than C-BARF.

light blub, ensuring that there is always one camera that captures the surgical field as long as the surgical field is illuminated.

As not all cameras are capable of adequately capturing the surgical field, as depicted in Fig. 1, post-processing is often required to create reviewable videos from multi-view video. Various methods have been proposed for this post-processing (Masuda et al., 2022; Hachiuma et al., 2020; Shimizu et al., 2020). Shimizu *et al.* (Shimizu et al., 2020) and Hachiuma *et al.* (Hachiuma et al., 2020) have proposed models to select the best view frames, while Masuda *et al.* (Masuda et al., 2022) have proposed the use of Conditional-BARF (C-BARF; see Fig.2-(a)) to synthesize occlusion-free image, with the aim of generating videos with smooth camera pose transitions. The C-BARF utilizes the relative position of the camera as a condition for improved camera pose estimation, as COLMAP (Schönberger and Frahm, 2016) is unable to accurately determine camera position from only five images. However, this approach requires the training of a network for every frame in order to generate novel view images, leading to a significant computational time when creating a video of an entire surgery. Additionally, it may struggle to render images if the target camera pose is not closely aligned with the source view’s camera pose.

Recently, transformer-based Neural Radiance Field (NeRF) has been attracting much attention. Wang *et al.* proposed IBNet (Wang et al., 2021), which introduces the Ray Transformer, enabling the learning of a generic view interpolation that generalizes to novel scenes. Varma *et al.* proposed the Generalizable NeRF Transformer (GNT) (Wang et al., 2022), which introduced the View Transformer as the

first stage. This transformer-based architecture allows for the generation of unseen scenes from source-view images.

In this paper, we propose the use of GNT for synthesizing surgical scenes (Fig. 2-(b)) with the aim of synthesizing novel view images of new surgical scenes without training the network using the GNT’s features, which can be pre-trained using previous surgical scenes. Our experimental results demonstrate that GNT can generalize surgical scenes using real-world surgical data, even though surgical data is typically less numerous than everyday scenes and includes complex shapes (e.g., faces, legs, surgeon’s hands, and surgical instruments). We record all data at Keio University School of Medicine using the recording system proposed by Shimizu *et al.* (Shimizu et al., 2020).

## 2 RELATED WORK

### 2.1 Surgical Recording Systems

Recording surgeries and generating videos for reviewing surgeries or teaching skills to future generations is an important task for doctors. Surgeries performed through an endoscope camera, such as laparoscopic surgery, can be easily recorded. However, surgeries in which the surgeon directly visualizes the surgical field, such as open surgery, are difficult to record due to spatial restrictions, as the head or hands of the surgeons and medical equipment may occluded the important field of the surgery.

Many attempts have been made to record the surgical field (Shimizu et al., 2020; Matsumoto et al.,

2013; Murala et al., 2010; Kumar and Pal, 2004; Byrd et al., 2003). Kumar *et al.* (Kumar and Pal, 2004) and Byrd *et al.* (Byrd et al., 2003) proposed recording systems that place a camera in the surgery room environment. However, the view is easily occluded by the surgeon's head or body. Observing the surgical field with a single camera without any occlusion is a difficult task. Matsumoto *et al.* and Murala *et al.* proposed recording systems that ask surgeons to wear cameras. However, this system is not only limited by hardware in its ability to produce high-quality videos, but it is also uncomfortable for surgeons to wear.

To solve this problem, Shimizu *et al.* proposed a new system that embeds cameras on a surgical lamp. This system not only allows one of the cameras to record the surgical field while one of the light bulbs illuminates it but also does not interfere with the surgeons during surgeries.

## 2.2 Camera Switching System

As the cameras obtain multiple videos of a single surgery, Shimizu *et al.* proposed a method for automatically selecting the image with the best view of the surgical field at each moment using Dijkstra's algorithm based on the size of the surgical field to generate a single video. Hachiuma *et al.* (Hachiuma et al., 2020) proposed Deep Selection, which selects the camera with the best view of the surgery using a fully supervised deep neural network. However, a problem with these methods is that the video quality is often low due to frequent changes in the viewing direction.

## 2.3 Novel View Synthesis

Novel view synthesis is one of the fundamental functionality and long-standing problem in computer vision (Gortler et al., 1996; Levoy and Hanrahan, 1996; Davis et al., 2012).

*For Non-medical Images:* Recently, novel view synthesis methods for everyday scenes have made significant progress by using neural networks. Mildenhall *et al.* (Mildenhall et al., 2020) proposed NeRF, a method for synthesizing novel views of static, complex scenes from a set of input images with known camera poses. Wang *et al.* proposed IBRNet (Wang et al., 2021), which introduced the Ray Transformer. This method enables learning a generic view interpolation function that generalizes to novel scenes, unlike previous neural scene representation work that optimized per-scene functions for rendering. Varma T. *et al.* proposed the Generalizable NeRF Transformer (GNT) (Wang et al., 2022), which introduced the

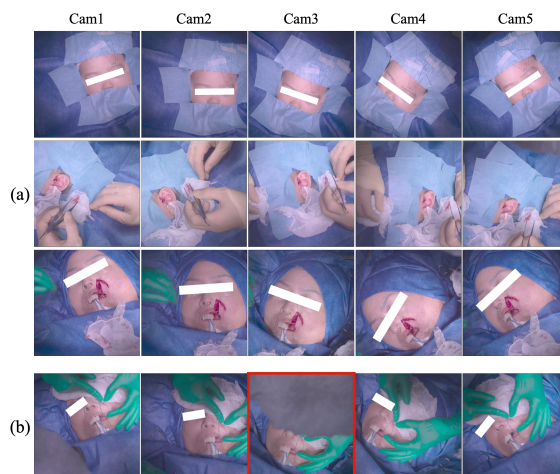


Figure 3: Examples of the data captured by the recording system proposed by Shimizu *et al.* (Shimizu et al., 2020). (a) The examples we used for the training, in which no cameras were obstructed by the surgeon's head. (b) A bad example in which the surgical area was hidden by the surgeon's head (Cam3).

View Transformer as the first stage. They also demonstrated that depth and occlusion could be inferred from the learned attention maps, which implies that the pure attention mechanism is capable of learning a physically-grounded rendering process. Furthermore, this transformer-based architecture makes it possible to generate unseen scenes from source-view images.

*For Medical Images:* The novel view synthesis method is begging to be used for medical scenes. Masuda *et al.* (Masuda et al., 2022) proposed C-BARF, which can synthesize novel view images of surgical scenes. They used the relative position of the camera to more accurately estimate the camera poses. They achieved a novel view synthesis method for surgical videos that consists of a small number of images.

## 3 METHOD

Our objective is to generate the surgical areas occluded by the surgeons' or nurse's head in order to create a comprehensive video for reviewing the surgical procedure. Our methodology can be divided into five distinct steps.

The first step is to prepare some sets of multi-view training videos using the camera recording system proposed by Shimizu *et al.* (Shimizu et al., 2020). In order to train GNT with a diversity of surgeries and surgical procedures, we recorded the surgeries with multiple types and for various areas. Presently, there are numerous frames with occlusions as depicted in Fig. 3-(b).

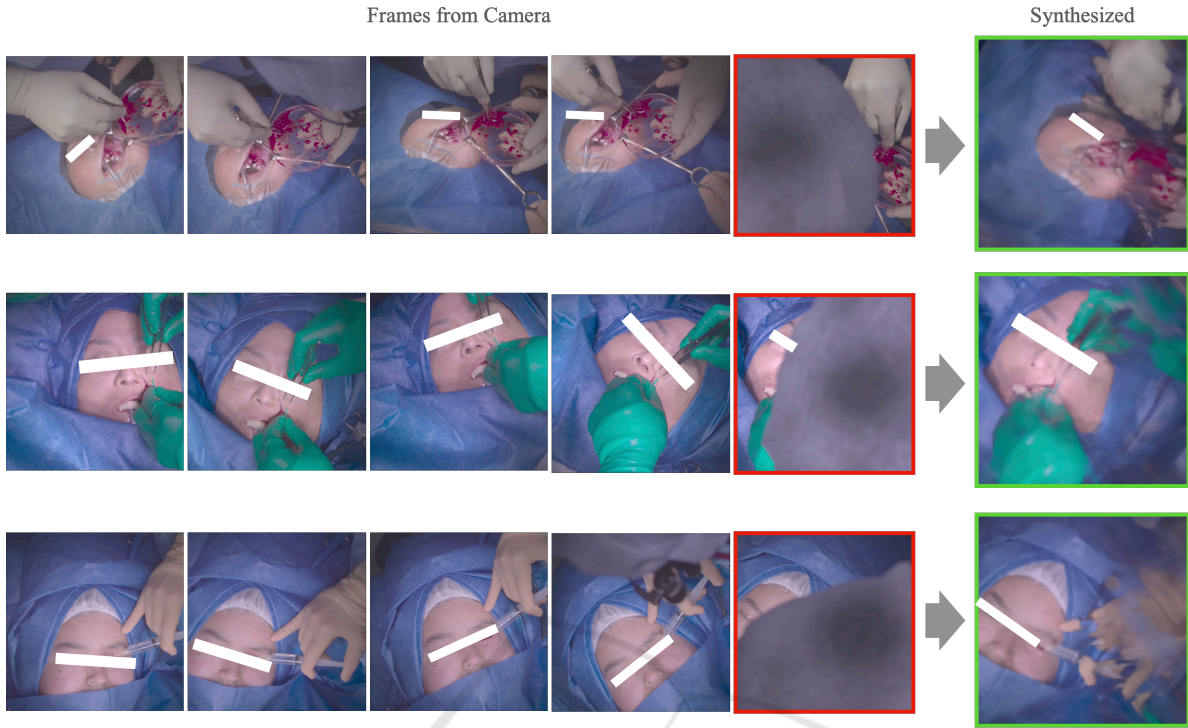


Figure 4: Novel view synthesis result for frames not included in the training set. The GNT has the capability to generate the surgical area occluded by the surgeon’s head. The images framed in red depict the frames that are surgical areas occluded by the surgeon’s head. The images framed in green illustrated the generated result for the same camera positions as the red-framed images.

The second step is to extract the frames from the training videos. In this step, we only extract the frames that are not occluded by the doctors or nurses, as shown in Fig. 3-(a).

The third step is to run COLMAP (Schönberger and Frahm, 2016) to obtain the camera parameters for the extracted frames. This step also removes the frames in which COLMAP failed to map feature points from the dataset. This third step finalizes the creation of the training dataset.

The fourth step is to train the GNT using the dataset we create in Steps 1 through 3. Through training in many types of surgical scene frames, GNT learns a neural representation of the surgical scenes.

The fifth step is generating occluded areas using the new multi-view images. In this step, we also use COLMAP to obtain the camera parameters.

## 4 EXPERIMENTS

### 4.1 Real Surgical Scene Dataset

As no available dataset contains surgery recordings with multiple cameras, we use the system proposed

by Shimizu *et al.* (Shimizu et al., 2020) to create our dataset. All surgeries were recorded at Keio University School of Medicine. Video recording of the patients were approved by the Keio University School of Medicine Ethics Committee, and written informed consent is obtained from all patients or their legal guardians. We recorded nine different types of surgeries with five cameras attached to the surgical lamp in a regular pentagon manner, as depicted in Fig. 1. We estimated the camera position using the COLMAP structure-from-motion package (Schönberger and Frahm, 2016) for both training and evaluation data.

### 4.2 Training Data

We used six different types of surgeries as the training data (three of these examples are shown in Fig. 3-(a)). For the training data, we extracted frames from the videos at regular intervals. We excluded some time stamps where the doctor’s head obscured the operative area in some cameras, as shown in Fig. 3-(b).

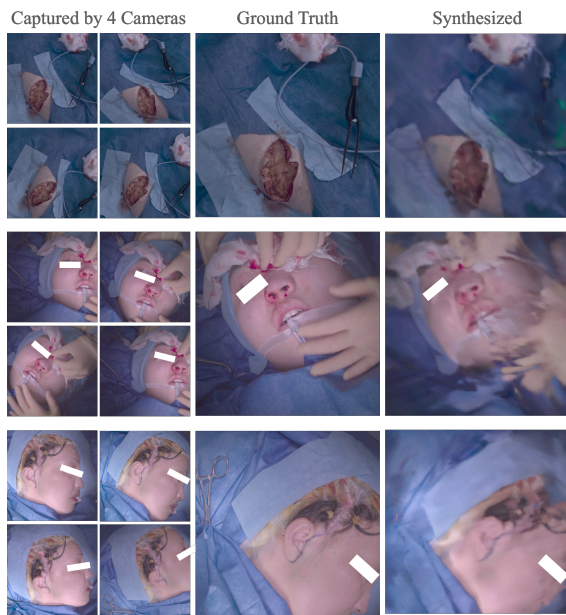


Figure 5: Novel view synthesis result for the multi-view videos not in the training set. The GNT could generate the occluded areas for three different kinds of surgical procedures that are not in the training set.

### 4.3 Evaluation Data

For the evaluation, we used two different types of datasets. The first one is the unknown “frames” that are excluded from the training data because they have some frames that do not capture the surgical areas hidden by the doctor’s head. The second one is an unknown “surgical scene”. These data are not included in the training set and are different from the training data for both the surgical type and the patient.

### 4.4 Training/Inference Setup

Following the official implementation of the GNT network, we train both the feature extraction network and the GNT end-to-end on datasets of multi-view posed images using the Adam (Kingma and Ba, 2014) optimizer to minimize the mean-squared error between predicted and ground truth RGB pixel values. For all our experiments, we trained for 250,000 steps with 512 rays sampled in each iteration. We followed the GNT’s multi-scene training setup for other experimental parameters.

### 4.5 Result

*On the Unknown Frames:* The result of the unknown frame dataset is shown in Fig. 4. The synthesized frames are bordered with green and the frames with

the target camera pose are bordered with red. We can see that the surgical area completely hidden by the head is properly synthesized.

*On the Unknown Surgical Scenes:* The result of the unknown surgical scenes dataset is shown in Fig. 5. We can see that the surgical area is properly synthesized even though the GNT does not train in these surgical scenes.

## 5 CONCLUSION

In this paper, we propose the utilization of a transformer-based NeRF network called GNT (Wang et al., 2022) for the purpose of generation of occluded areas of surgical scenes. The aim of this approach is to generate the occluded areas for new surgical scenes without the need for a training network during inference, using the GNT’s feature which can be pre-trained by the previous surgical scenes. Our experiments demonstrated that the GNT can effectively learn real-world surgical scenarios, and can also generate the occluded surgical areas not only for unknown frames as well as for unknown surgical scenarios. As a future endeavor, we intend to devise a model that can accurately determine the optimal camera position for rendering videos, to make it easier to create review videos from the acquired multi-view videos.

## ACKNOWLEDGEMENT

This work was supported by MHLW Health, Labour, and Welfare Sciences Research Grants Research on Medical ICT and Artificial Intelligence Program Grant Number 20AC1004, the MIC/SCOPE #201603003, and JSPS KAKENHI Grant Number 22H03617.

## REFERENCES

- Byrd, R. J., Ujji, V. M., Kongchan, S. S., and Reed, H. D. (2003). Surgical lighting system with integrated digital video camera. US Patent 6,633,328.
- Davis, A., Levoy, M., and Durand, F. (2012). Unstructured light fields. In *Computer Graphics Forum*, volume 31, pages 305–314. Wiley Online Library.
- Gortler, S. J., Grzeszczuk, R., Szeliski, R., and Cohen, M. F. (1996). The lumigraph. In *Proceedings of the 23rd Annual Conference on Computer Graphics and interactive techniques*, pages 43–54.
- Hachiuma, R., Shimizu, T., Saito, H., Kajita, H., and Takatsume, Y. (2020). Deep selection: A fully supervised

- camera selection network for surgery recordings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 419–428. Springer.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kumar, A. S. and Pal, H. (2004). Digital video recording of cardiac surgical procedures. *The Annals of thoracic surgery*, 77(3):1063–1065.
- Levoy, M. and Hanrahan, P. (1996). Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42.
- Masuda, M., Saito, H., Takatsume, Y., and Kajita, H. (2022). Novel view synthesis for surgical recording. In *MICCAI Workshop on Deep Generative Models*, pages 67–76. Springer.
- Matsumoto, S., Sekine, K., Yamazaki, M., Funabiki, T., Orita, T., Shimizu, M., and Kitano, M. (2013). Digital video recording in trauma surgery using commercially available equipment. *Scandinavian journal of trauma, resuscitation and emergency medicine*, 21(1):1–5.
- Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer.
- Murala, J. S., Singappuli, K., Swain, S. K., and Nunn, G. R. (2010). Digital video recording of congenital heart operations with "surgical eye". *The Annals of thoracic surgery*, 90(4):1377–1378.
- Sadri, A., Hunt, D., Rhobaye, S., and Juma, A. (2013). Video recording of surgery to improve training in plastic surgery. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, 66(4):e122–e123.
- Schönberger, J. L. and Frahm, J.-M. (2016). Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Shimizu, T., Oishi, K., Hachiuma, R., Kajita, H., Takatsume, Y., and Saito, H. (2020). Surgery recording without occlusions by multi-view surgical videos. In *VISIGRAPP (5: VISAPP)*, pages 837–844.
- Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z., et al. (2022). Is attention all nerf needs? *arXiv preprint arXiv:2207.13298*.
- Wang, Q., Wang, Z., Genova, K., Srinivasan, P., Zhou, H., Barron, J. T., Martin-Brualla, R., Snavely, N., and Funkhouser, T. (2021). Ibrnet: Learning multi-view image-based rendering. In *CVPR*.