applied sciences

MDPI

Communication

Multi-View Surgical Camera Calibration with None-Feature-Rich Video Frames: Toward 3D Surgery Playback

Mizuki Obayashi [1,*], Shohei Mori [1,2,*], Hideo Saito [1,*], Hiroki Kajita [3] and Yoshifumi Takatsume [3]

[1] Graduate School of Science and Technology, Keio University, Yokohama 223-8852, Japan
[2] Institute of Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria
[3] Department of Plastic and Reconstructive Surgery, Keio University School of Medicine, Shinjuku-ku, Tokyo 160-8582, Japan
* Correspondence: mizuki_obayashi@keio.jp (M.O.); s.mori.jp@ieee.org (S.M.); hs@keio.jp (H.S.); Tel.: +81-45-563-1141 (ext. 43230) (H.S.)

Abstract: Mounting multi-view cameras within a surgical light is a practical choice since some cameras are expected to observe surgery with few occlusions. Such multi-view videos must be reassembled for easy reference. A typical way is to reconstruct the surgery in 3D. However, the geometrical relationship among cameras is changed because each camera independently moves every time the lighting is reconfigured (i.e., every time surgeons touch the surgical light). Moreover, feature matching between surgical images is potentially challenging because of missing rich features. To address the challenge, we propose a feature-matching strategy that enables robust calibration of the multi-view camera system by collecting a set of a small number of matches over time while the cameras stay stationary. Our approach would enable conversion from multi-view videos to a 3D video. However, surgical videos are long and, thus, the cost of the conversion rapidly grows. Therefore, we implement a video player where only selected frames are converted to minimize time and data until playbacks. We demonstrate that sufficient calibration quality with real surgical videos can lead to a promising 3D mesh and a recently emerged 3D multi-layer representation. We reviewed comments from surgeons to discuss the differences between those 3D representations on an autostereoscopic display with respect to medical usage.

Keywords: multi-camera shadowless lamp; multi-frame multi-view calibration; 3D view synthesis

Citation: Obayashi, M.; Mori, S.; Saito, H.; Kajita, H.; Takatsume, Y. Multi-View Surgical Camera Calibration with None-Feature-Rich Video Frames: Toward 3D Surgery Playback. *Appl. Sci.* **2023**, *13*, 2447. https://doi.org/10.3390/app13042447

Academic Editors: Jiann-Der Lee and Jong-Chih Chien

Received: 20 January 2023
Revised: 5 February 2023
Accepted: 11 February 2023
Published: 14 February 2023

1. Introduction

1.1. Motivation

Recording surgeries has a potential demand for educational and archiving purposes. Deeper analyses with video examples encourage better understanding and avoid future mistakes. Affordable, small form-factor cameras make recordings more casual and multi-view. Aiming for solid capturing of patients under surgery, mounting multi-view cameras within a surgical light (aka. shadowless lamp, hereafter referred to as a multi-camera shadowless lamp: *McSL*) is a reasonable choice. A shadowless lamp has multiple light sources verged at a point. At least one of the light sources lights an area; hence, fewer shadows appear, even with interferences by surgeons' heads, hands, and tools. The analogy applies to viewpoints. At least one of the cameras prevents uncollected data capture.

Once the surgery is recorded with a McSL, separated videos must be reassembled into an easy-to-refer-to format. Shimizu et al. proposed a video analysis approach that can select the most non-occluded view over frames [1]. Reconstructing the surgery in 3D would be another option, given that a 3D scene allows viewers to navigate the scene, avoiding occlusions. However, one McSL video tends to be long (in several tens of minutes to hours), making frame-to-3D model conversion time consuming. Moreover, a major challenge lies in frame alignment because of two reasons. First, the geometrical relationship among the cameras is changed because each camera independently moves every time when the

lighting is reconfigured (i.e., every time surgeons touch the surgical light). Second, surgery videos may contain highly reflective tissues and tools, uniform blue covers, and significantly occluding substances. Therefore, finding feature correspondences over cameras becomes difficult and 3D point localization becomes fragile. Manual calibration is always accessible [1], but the video post-processing becomes further cumbersome.

We address the above issues with a video player that allows a user to access a McSL video hierarchically. First, the user highlights important frames in a McSL video and selects shorter videos around the frames. Then, our system performs feature matching and frame alignment algorithms that utilize the shorter video frames to find feature correspondences robustly. Therefore, only the highlighted frames are converted on-demand into a 3D mesh or multi-layer scene representation, which renders the frame in 3D and thus in an autostereoscopic display.

1.2. Background and Related Work

Here, we provide an overview of surgical video recording and 3D view synthesis approaches and the challenges in their application to non-feature-rich frames of surgeries.

Surgical Video Recording. Recording surgery is needed for various purposes, such as preserving case studies, educating trainees, and passing on skills [1–4]. Contrary to laparoscopic surgery [5,6], recording in open surgeries, in which physicians directly view the affected areas, is challenging because of possible interference with the surgery, spatial limitations, and occlusions by surgeons.

There are two major ways to install cameras to record surgeries: cameras on the surgeon's head [2,7–9] and arms [10], and in the operating room [11]. Knowing possible cable management, image noises (e.g., motion blurs), and occlusions, McSL is a reasonable solution for multi-view observations [1].

Displaying non-aligned multi-view images makes the analysis difficult as it requires mental image warping. Mathematically aligning such images requires calibrating the cameras. However, objects in surgical videos are often featureless, which makes calibration challenging. Calibration before video recording is useless as cameras in McSL move when surgeons manipulate it. To address these issues, we implement a feature-matching scheme that utilizes multiple frames over time to collect sufficient feature points for calibration.

3D View Synthesis. To seek system requirements, we reviewed comments from two surgeons who have experience using a McSL recording system at Keio University School of Medicine. One is a Prof. Dr. of the Department of Plastic and Reconstructive Surgery and the other is a Dr. of the Department of Anatomy. The surgeons requested to see and compare several frames in 3D each at the beginning and end of the surgery, especially for plastic surgery. Therefore, we implement a video player that allows medical doctors to seek and select frames so that the system can instantly reconstruct the view in 3D at the selected frames only. From this background, we did not use view synthesis approaches that require per-frame extensive training [12,13].

Another approach is to convert all frames into a 3D video, such as neural videos [14–17] and multi-layer mesh and texture atlas videos [18,19]. However, neural videos are typically in low resolution and are computationally demanding. A multi-layer mesh provides faster rendering at higher resolutions, although conversion is nonetheless computationally expensive [18]. We instead provide on-demand 3D views only at selected frames.

Disocclusion Rendering. Real-time disocclusion rendering is often referred to as diminished reality [20]. Multi-view approaches bring observed background pixels from cameras to the current view to disocclude foreground objects [21–23]. These approaches either rely on plane geometry proxy [21] or range sensors [22,23] for millisecond-order rendering, which prevents us from applying them to our application. Another approach relies on image inpainting, in which the recovered area is completely synthetic (i.e., pixels are hallucinated by a collection of pixels within the field of view) [24–26]. Therefore, inpainting approaches are not suitable for our application.

Rather than relying on the above approaches, we utilize a recently developed offline multi-view approach [27] and soft 3D representation [28,29] (i.e., multi-layer scene representation) that is known to be more robust than explicit 3D representations. Although 3D rendering allows users to peep at occluded backgrounds, we further extend the soft 3D representation for disocclusion rendering.

1.3. Contributions

In summary, we contribute to medical imaging and display in the following ways:

- We propose a video player that allows the user to selectively convert a portion of a long-shot McSL video into free viewpoint images to minimize time and data until playbacks.
- We propose a multi-frame multi-view feature matching strategy to estimate intrinsic and extrinsic camera parameters from a set of McSL video frames. This process is always required after the independent movement of internal cameras (e.g., when surgeons touch the surgical light during the operation). We also analyze the number of frames to achieve stable calibration results using real surgical videos.
- With the robustly estimated camera parameters, we demonstrate a 3D mesh and a recently emerged 3D multi-layer reconstruction. The latter enables disocclusion rendering to remove foregrounds in the generated 3D scene representation for better surgical field visibility.
- We reviewed comments from surgeons to discuss the differences between those 3D representations with respect to medical usage.

2. Materials and Methods

The proposed workflow consists of the following four steps (Figure 1).

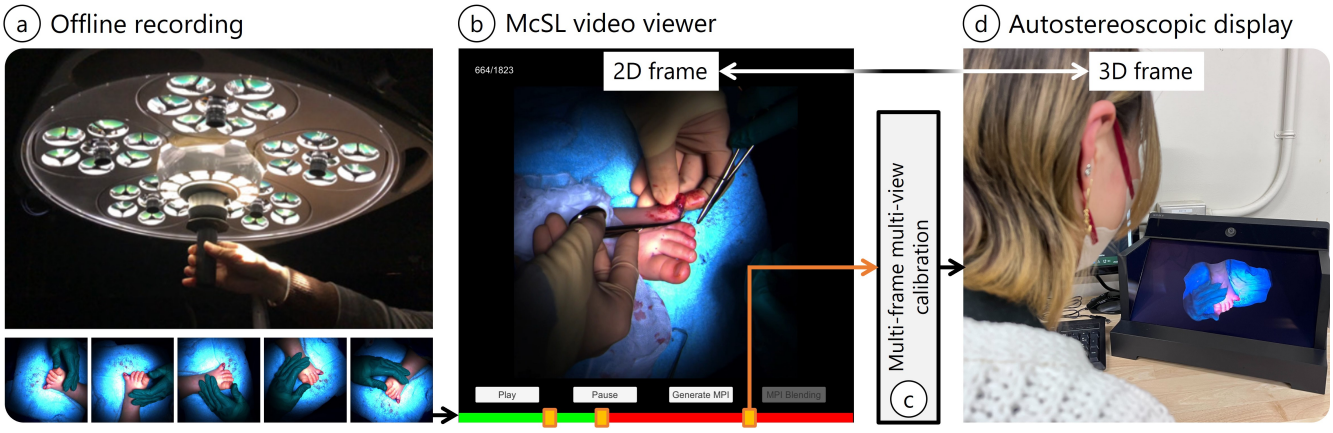


Figure 1. Pipeline of our system. Our system consists of four steps, including (**a**) on-site McSL video recording, (**b**) playback with our McSL video player, (**c**) McSL calibration, and (**d**) view synthesis on an autostereoscopic display. The view synthesis is available only after a successful calibration, which is challenging with non-feature-rich multi-view images in surgical videos. We, therefore, implement a calibration strategy that utilizes multiple frames over time.

1. On-site McSL video recording: We record a surgery case with a McSL (Figure 1a). At this point, no additional effort is required (e.g., no calibration pattern recording is necessary).
2. Playback with our McSL video player: The user (e.g., medical trainees) plays the video and selects a frame to be reconstructed in 3D (Figure 1b, Section 2.3).
3. McSL calibration: Upon frame selection, we run our calibration algorithm to calculate the cameras' intrinsic and extrinsic parameters (Figure 1c, Section 2.1).
4. View synthesis: The 3D frame data are generated and saved (Section 2.2), and the selected frame is highlighted in the player. Therefore, the user can switch between the

2D and 3D viewers. We provide ways to avoid occlusions in surgical fields. Given that the frame is in 3D representation, we can display the frame on an autostereoscopic display (Figure 1d).

Our McSL consists of the following components.

- The one used to capture the polysyndactyly surgery and the cleft lip surgery: stand-alone shadowless lamp (DAI-ICHI SHOMEI CO., LTD., LEDXⅡ 5S) + camera (LUCID Vision Labs, PHX032S-CC) × 5 units;
- The one used to capture the cleft lip surgery and the accessory auricle surgery: stand-alone shadowless lamp (DAI-ICHI SHOMEI CO., LTD., LEDXIV 5S) + HDR camera (LUCID Vision Labs, TritonHDR) × 5 units.

2.1. Calibration with a McSL Video

The relative poses of multi-cameras in McSL change every time a surgeon manipulates the McSL position. This mechanism prevents photographing a calibration target and calibrating the cameras before the surgery. Instead, we propose to use surgical video frames for calibration. However, surgical video frames are not feature-rich (see an example in Figure 1a). Therefore, to collect more features, we utilize feature points and their matching results that appear on every available time frame.

The key idea is to treat the scene geometry that changes over time as a whole structure observed at a time since such structure changes are irrelevant from the inter-camera poses. Assuming that M consecutive frames are recorded under a fixed McSL location, we randomly select N ($\leq$M) multi-camera frames. At a time frame, we extract feature points [30] at every camera image and run matching [31] between McSL cameras. We obtain N such sets of inter-camera feature correspondences and use all correspondences as if all features were detected at a time frame (Figure 2). Finally, we perform structure from motion (SfM), followed by bundle adjustment, for which we rely on the COLMAP framework [32,33].

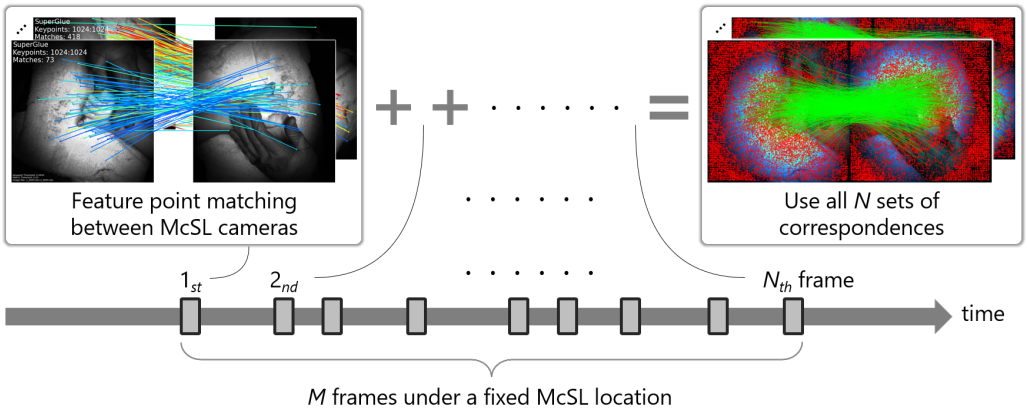


Figure 2. Our multi-frame multi-view feature matching strategy. We randomly select N ($\leq$M) frames from M consecutive frames with a fixed McSL location. We then extract feature points from the camera images in each frame and run matching between the McSL cameras. The resulting correspondences between all N sets of inter-camera features are used to determine the geometric relationship between the cameras.

2.2. Three-Dimensional Frame Generation

Users can choose to generate either a 3D mesh or a recently emerged multi-plane image (MPI) for an in-depth investigation of the surgery. We discuss the advantages and disadvantages of those different types of 3D frames through a comparison on an autostereoscopic display (see Section 3).

Mesh Generation. We create a 3D mesh from a selected multi-view frame. Given the intrinsic and extrinsic parameters from our calibration procedure, a 3D mesh is constructed from a 3D point cloud from COLMAP using a multi-view stereo library, OpenMVS [27].

MPI Generation and Disocclusion Rendering. Given the same multi-camera data for the mesh generation, we can generate an MPI using local light field fusion (LLFF) [29]. LLFF is a 3D convolutional neural network that can infer multi-layer scene representation (aka. MPI) [34] or soft 3D representation [28] from calibrated multi-view images. MPI consists of multiple RGB+α layers spaced along the camera forward direction in inverse depth for efficiency. LLFF blends multi-view MPI depending on per-camera weights for quality:

$$c = \frac{\sum_{i=1}^{K} w_i^{\mathrm{C}} \alpha_i c_i}{\sum_{i=1}^{K} w_i^{\mathrm{C}} \alpha_i + t}, \tag{1}$$

where w_i^{C} represents a blending weight (scalar) of i_{-th} rendered MPI (among MPIs of individual McSL cameras). K means the total number of cameras (i.e., $K = 5$ in our case of McSL. See Figure 1a). We calculate Euclid distances between the rendered view and the five McSL cameras so that we blend closer cameras with more weights and vice versa. c_i is a pixel color (RGB values) of the i_{th} rendered MPI and α_i is the corresponding opacity. t is a small value to avoid zero division.

We apply the LLFF inference to a selected McSL frame to render the frame in 3D. To avoid occlusion, users can control the viewpoint until the rendered view disoccludes the obstacle. Furthermore, given a camera occlusion metric, we can down-weight cameras under occlusion:

$$c = \frac{\sum_{i=1}^{K} w_i^{\mathrm{O}} w_i^{\mathrm{C}} \alpha_i c_i}{\sum_{i=1}^{K} w_i^{\mathrm{O}} w_i^{\mathrm{C}} \alpha_i + t}, \tag{2}$$

where w_i^{O} is a per-camera value that suggests disocclusion significance in the camera view (e.g., the camera score [1]).

2.3. McSL Video Player

Our video player allows the user to switch between the 2D video and 3D scene viewers (Figure 1b,d). The 2D video player performs similarly to a conventional video player of a selected camera view or arranged multi-camera views in a grid. Upon a request from the user (a button click), MPI or 3D mesh at the time frame is generated in the background. The user can choose which 3D representation to generate. Once the 3D representation generation is completed, a highlight appears over the sequence bar. Clicking the highlight toggles the 3D scene viewer. The 3D scene is displayed in an autostereoscopic display for a glasses-free experience. Users can adjust the position, orientation, and scale of the 3D representation with mouse dragging and scrolling.

MPI takes several seconds for generation (with our setup using NVIDIA GeForce RTX 3080 GPU, MPI generation for five viewpoints takes approximately 20 s) and increases the data amount, per camera, by a factor of L (=32 by default) MPI layers, each of which adds an additional α channel. The 3D mesh likewise requires an amount of time and data to generate. As such, we do not generate 3D representation at every frame, minimizing the time and storage.

3. Results

We evaluate our calibration system in two experiments using real surgical videos. We first present a quantitative evaluation of the McSL camera calibration (Section 2.1). Then, we demonstrate disocclusion view synthesis using MPI (Section 2.2).

Note that other 3D surgical video reconstruction works than this study are for laparoscopic surgery and these approaches assume a single-view moving video input with fewer significant occlusions in the input [5,6]. In our McSL setups, one or a few cameras are significantly or entirely occluded. Namely, [5,6] and we explore distinct problems and, thus, direct comparison is difficult.

3.1. McSL Calibration Accuracy

We prepared McSL videos with five cameras of three different types of surgeries: polysyndactyly surgery, cleft lip surgery, and accessory auricle surgery (Figure 3). To investigate the impact of frame counts used in calibration, we randomly selected $N = 1, 10, 20, \ldots, 100$ frames 10 times in each video and performed the calibration. We calculated the mean and standard deviation values of reprojection errors for each number of frames and the success ratios. For the success ratios, we calculated the number of fails of COLMAP bundle adjustment over all trials.

Figure 3. Examples of none-feature-rich frames. Each shows one of five view images in McSL in three different surgeries: polysyndactyly, cleft lip, and accessory auricle surgery (from **left** to **right**).

Figure 4 summarizes the results. A one-frame input is not enough for the calibration because of the lack of features in the frames. Increasing the number of inputs increases the success ratio. However, no improvement is observed over the 20-frame input. The use of more than 70 frames shows a slight decrease in success ratio, perhaps due to the increased interference of erratic feature matching. The reprojection errors stay very similar regardless of the number of images.

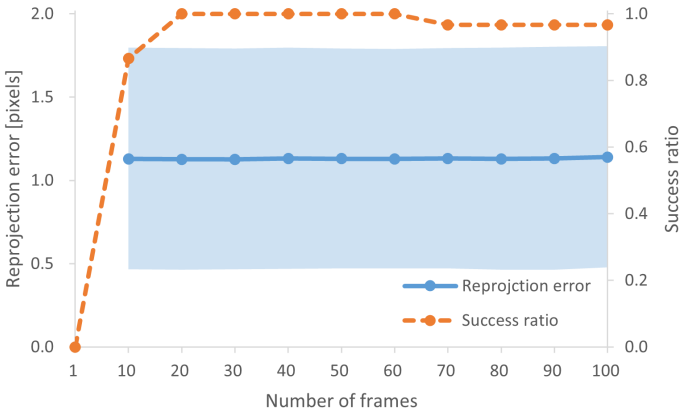


Figure 4. Reprojection errors and success ratios over various numbers of images for McSL calibration. A 1-frame input is apparently not enough for calibrating the system. Almost identical mean errors and standard deviations are observed. However, the success ratio reaches 1.0 in the 20-frame input. Notably, with more than 70 frames, the success ratio is slightly lower, which suggests that including a significantly larger number of images would introduce erratic and less beneficial images for calibration.

In 3D reconstruction, in general, COLMAP [32,33] is the de facto standard and state-of-the-art method. Although the approach is extended with a graph-based neural network approach (i.e., SuperPoint [30] and SuperGlue [31]) instead of SIFT features [35] and heuristic matching, the result is summarized as 1-frame (the leftmost result) in Figure 4 in our experiment.

Figure 5 presents a qualitative result of our calibration method with a 20-frame input. The red cones represent the camera view frustum and their heights indicate the focal lengths. The camera views are posed with extrinsic parameters. The colored points represent a 3D point cloud reconstruction. These results show our method provides qualitatively reasonable performance.

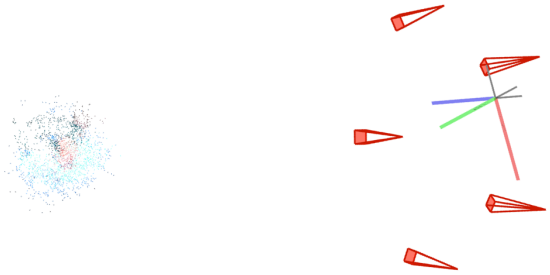
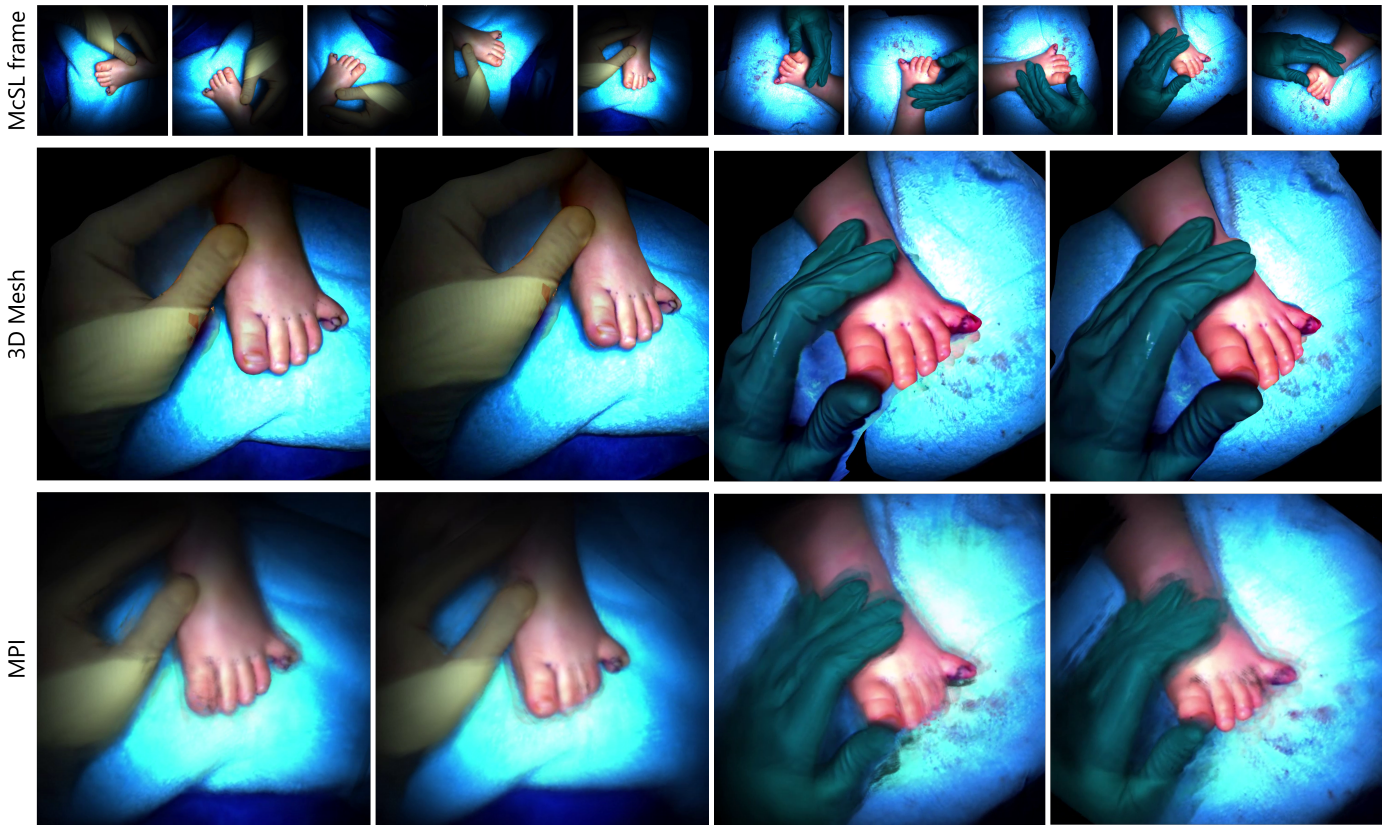


Figure 5. Qualitative results of our calibration method with a 20-frame input. The red cones represent the camera view frustum and their heights indicate the focal lengths. The cone directions and positions represent the extrinsic parameters. The colored points are a 3D point cloud. The results are qualitatively convincing, compared to the camera poses attached to the shadowless lamp (Figure 1a).

3.2. Three-Dimensional Mesh Rendering

We show 3D mesh generation results. From the results in Section 3.1, we used 20 randomly selected frames for calibration in all results. Figure 6 (2nd row) shows the 3D meshes at the beginning and ending in the same surgical video and frames to compare the differences.



(a) Beginning of the surgery **(b)** Ending of the surgery

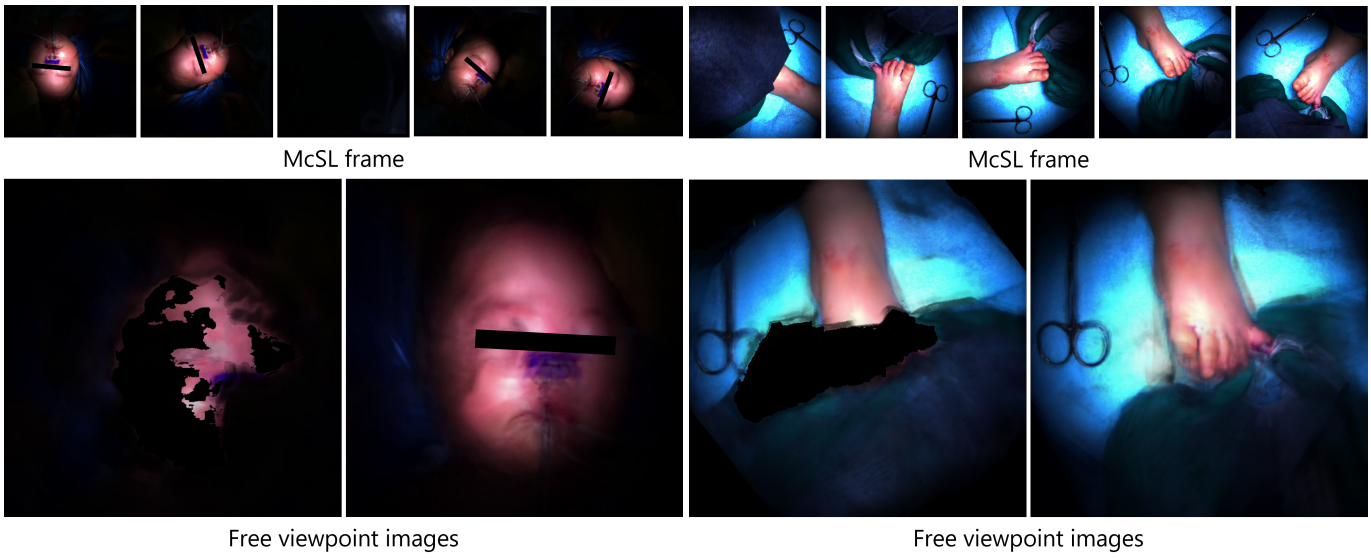
Figure 6. Comparing (**a**) beginning and (**b**) ending of a polysyndactyly surgery in 3D mesh and MPI at the selected McSL frames. Notice the difference in the suture of the little toe.

3.3. Occlusion-Free View Synthesis Using MPI

We show view synthesis results. As in Section 3.2, we used 20 randomly selected frames for the calibration.

Figure 6 (3rd row) shows the free-viewpoint images at the beginning and ending of a polysyndactyly surgery to compare the differences. We emphasize that such a comparison in 3D rendering is achieved because of our successful calibration.

Figure 7 shows the free-viewpoint images generated with two different disocclusion rendering strategies, as discussed in Section 2.2. Navigating the viewpoint to where no occlusion occurs simply clarifies the view (Figure 7a). With our disocclusion rendering (Equation (2)), the occlusion is virtually diminished at any viewpoint. We prioritized the second MPI by setting weights manually depending on the interruption degrees of the surgeon's head to demonstrate our disocclusion rendering conceptually (Figure 7b).



(a) Changing viewpoint in a cleft lip surgery **(b)** Disocclusion rendering in a polysyndactyly surgery

Figure 7. Occlusion-free rendering using two techniques. These frames contain a viewpoint where the operation field is occluded by the surgeon's head (first row). Our free-viewpoint rendering techniques (**a**) by changing the viewpoint in 3D or (**b**) by disocclusion rendering (Equation (2)) can reveal the occluded area.

3.4. Experts' Comments

Using the camera's intrinsic and extrinsic parameters calculated by the calibration method in this paper, we performed 3D scene recovery in two ways: 3D mesh and MPI. Our main contributions are the pipeline and calibration system (the first and second item in Section 1.3) that enable 3D rendering (the third item in Section 1.3) in the medical imaging context. As such, we consider that asking medical doctors to provide their comments is the best way to investigate how well and in what way the generated 3D rendering performs in medical usage. We asked two medical doctors from Keio University School of Medicine to observe and provide comments on each rendering result displayed on the spatial reality display (SRD) from Sony Corporation [36], an autostereoscopic display that reproduces 3D spatial images using real-time gaze recognition and dedicated image generation algorithms. One of the medical doctors is a Prof. Dr. of the Department of Plastic and Reconstructive Surgery and the other is a Dr. of the Department of Anatomy.

Figure 8 shows the 3D mesh models and MPIs on SRD shown to the doctors. The following summarizes the positives (P) and negatives (N) from the medical doctors.

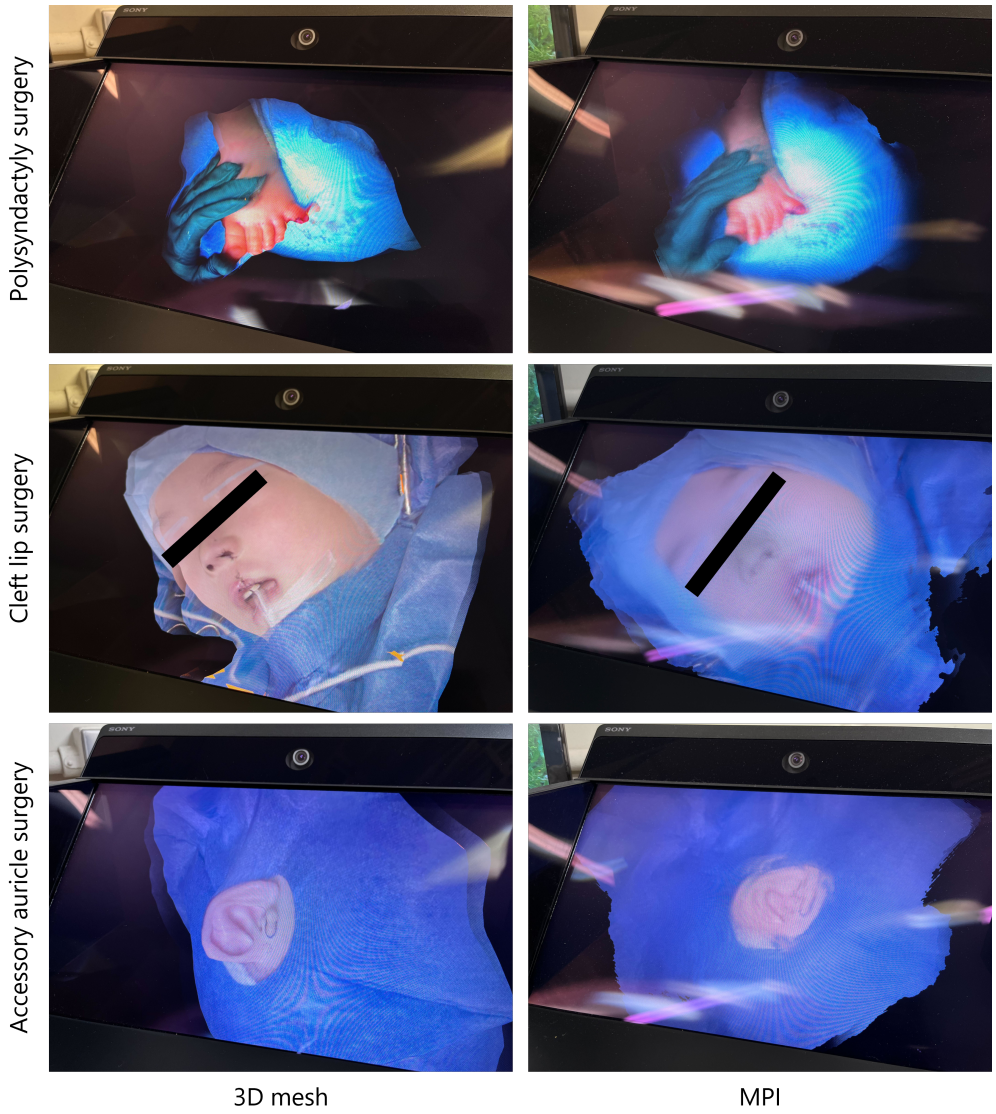


Figure 8. 3D mesh and MPI on Sony SRD. We showed polysyndactyly, cleft lip, and accessory auricle surgery images to medical doctors to review their comments.

3D mesh.

- (P1) Very close impression to what we see in actual surgeries regarding appearance and shape;
- (N1) The blind spots on the sides are sharply vertical, unlike the actual shape;
- (N2) The scale should be adjusted.

MPI.

- (P1) Feeling of being there, however, worse than 3D mesh;
- (P2) Interesting to see "images" in 3D;
- (N1) Blurry and unclear;
- (N2) Only rough shapes can be grasped;
- (N3) Planes are obvious from a steep angle (i.e., stack-of-cards artifacts);
- (N4) The scale should be adjusted.

Although MPI enables new disocclusion rendering, as demonstrated in Section 3.3, there are clear limitations. MPI has more disadvantages and one of the main reasons would be wider baselines between the cameras in McSL. One medical doctor commented that "it is difficult to list advantages from a clinical point of view" due to the limitations. Nonetheless, we should note that MPI can be extended to videos rather more straightforwardly than

the 3D mesh [18]. As a common issue, the medical doctors must have adjusted the scale manually for individual scenes, given that the scene scale is unknown and thus different in each scene.

4. Discussion

Our approach successfully estimates the intrinsic and extrinsic parameters of multi-view cameras in McSL using non-feature-rich multi-view image sequences. The generated multi-layer scene enables disoccluding visual disturbances, such as a surgeon's head, for better surgical field visibility. However, we found several limitations to be addressed that indicate future research directions.

Shorter video extraction. The current calibration procedure assumes that the user extracts a shorter video, in which McSL does not move, from a long video. Although such changes in McSL videos are obvious and easy to detect visually, automation of the process is a reasonable future extension. In the previous approach [1], experts need to align images from McSL by hand, although the resultant alignment is image-to-image registration in 2D. Once a frame chunk is specified, our approach can automate the process, and the resultant alignment is 3D registration and 3D view synthesis.

Blurry artifacts. We observe blurry artifacts in MPI rendering results, especially when there are significant occlusions. The LLFF network, which we rely on for MPI generation, implicitly finds color matches between images. Therefore, ambiguous matches introduced by occlusion pixels can lead to blurred pixels. Denser camera arrangements can provide appropriate plenoptic sampling [29]. However, further quality improvements in MPI inference using non-feature-rich images exceeds our focus and remains for future work.

Scale ambiguity. As pointed out by the medical doctors, scale ambiguity needs to be manually corrected in our system. Camera baselines or scene objects with a known scale can potentially resolve this issue. Whereas the former varies depending on McSL video frames, the latter can be estimated from known faces, hands, and tools in the field of view.

5. Conclusions

This paper presents a system to calibrate McSL cameras using non-feature-rich multi-view frames and recover a 3D scene representation. Given the long-shot surgical videos from McSL, the system provides a way to seek video frames where the user wishes to recover the 3D structure. Given that McSL camera configurations can change every time surgeons move McSL, calibration must be performed with a selected shorter clip, in which McSL is static. Within the shorter clip, our calibration algorithm randomly selects multiple frames to robustly estimate the intrinsic and extrinsic parameters. Owing to the robust calibration, the system can generate a 3D mesh or a multi-layer scene representation to provide 3D frames. The evaluation results using real surgical videos revealed the advantages and the reasonable number of frames for the calibration, and the future challenges in automation and rendering quality improvements.

Author Contributions: Conceptualization, H.S. and H.K.; methodology, M.O., S.M. and H.S.; software, M.O. and S.M.; validation, M.O.; formal analysis, M.O.; investigation, M.O.; resources, H.K. and Y.T.; data curation, H.K. and Y.T.; writing—original draft preparation, M.O.; writing—review and editing, M.O., S.M. and H.S.; visualization, M.O.; supervision, H.S.; project administration, H.S. and H.K.; funding acquisition, H.K. All authors have read and agreed to the published version of the manuscript.

Funding: 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.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of Keio University School of Medicine.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

McSL	Multi-camera shadowless lamp
SfM	Structure from motion
LLFF	Local light field fusion
MPI	Multi-plane image
SRD	Spatial reality display

References

1. Shimizu, T.; Oishi, K.; Hachiuma, R.; Kajita, H.; Saito, H.; Takatsume, Y. Surgery recording without occlusions by multi-view surgical videos. In Proceedings of the International Conference on Computer Vision Theory and Applications, Valletta, Malta, 27–29 February 2020.
2. Matsumoto, S.; Sekine, K.; Yamazaki, M.; Funabiki, T.; Orita, T.; Shimizu, M.; Kitano, M. Digital video recording in trauma surgery using commercially available equipment. *Scand. J. Trauma Resusc. Emerg. Med.* **2013**, *21*, 27. [CrossRef] [PubMed]
3. Sadri, A.; Hunt, D.; Rhobaye, S.; Juma, A. Video recording of surgery to improve training in plastic surgery. *J. Plast. Reconstr. Aesthet. Surg.* **2013**, *66*, e122–e123. [CrossRef]
4. Hachiuma, R.; Shimizu, T.; Saito, H.; Kajita, H.; Takatsume, Y. Deep selection: A fully supervised camera selection network for surgery recordings. In Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI), Lima, Peru, 4–8 October 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 419–428.
5. Hu, M.; Penney, G.; Figl, M.; Edwards, P.; Bello, F.; Casula, R.; Rueckert, D.; Hawkes, D. Reconstruction of a 3D surface from video that is robust to missing data and outliers: Application to minimally invasive surgery using stereo and mono endoscopes. *Med. Image Anal.* **2012**, *16*, 597–611. [CrossRef] [PubMed]
6. Cano González, A.M.; Sánchez-González, P.; Sánchez-Margallo, F.M.; Oropesa, I.; del Pozo, F.; Gómez, E.J. Video-endoscopic image analysis for 3D reconstruction of the surgical scene. In Proceedings of the International Federation for Medical and Biological Engineering (IFMBE), Miami, FL, USA, 15–17 May 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 923–926.
7. Murala, J.S.K.; Singappuli, K.; Swain, S.K.; Nunn, G.R. Digital video recording of congenital heart operations with "surgical eye". *Ann. Thorac. Surg.* **2010**, *90*, 1377–1378. [CrossRef] [PubMed]
8. Nair, A.G.; Kamal, S.; Dave, T.V.; Mishra, K.; Reddy, H.S.; Della Rocca, D.; Della Rocca, R.C.; Andron, A.; Jain, V. Surgeon point-of-view recording: Using a high-definition head-mounted video camera in the operating room. *Indian J. Ophthalmol.* **2015**, *63*, 771–774. [CrossRef] [PubMed]
9. Graves, S.N.; Shenaq, D.S.; Langerman, A.J.; Song, D.H. Video capture of plastic surgery procedures using the GoPro HERO 3+. *Plast. Reconstr. Surg. Glob. Open* **2015**, *3*, e312. [CrossRef] [PubMed]
10. Kumar, A.S.; Pal, H. Digital video recording of cardiac surgical procedures. *Ann. Thorac. Surg.* **2004**, *77*, 1063–1065; discussion 1065. [CrossRef]
11. Byrd, R.J.; Ujjin, V.M.; Kongchan, S.S.; Reed, H.D. Surgical Lighting System with Integrated Digital Video Camera. U.S. Patent 6,633,328, 14 October 2003.
12. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. In Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 405–421.
13. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph. (TOG)* **2022**, *41*, 102:1–102:15. [CrossRef]
14. Tretschk, E.; Tewari, A.; Golyanik, V.; Zollhöfer, M.; Lassner, C.; Theobalt, C. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 12959–12970.
15. Li, Z.; Niklaus, S.; Snavely, N.; Wang, O. Neural scene flow fields for space-time view synthesis of dynamic scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 6498–6508.
16. Zhang, J.; Liu, X.; Ye, X.; Zhao, F.; Zhang, Y.; Wu, M.; Zhang, Y.; Xu, L.; Yu, J. Editable free-viewpoint video using a layered neural representation. *ACM Trans. Graph. (TOG)* **2021**, *40*, 1–18. [CrossRef]
17. Wang, L.; Wang, Z.; Lin, P.; Jiang, Y.; Suo, X.; Wu, M.; Xu, L.; Yu, J. IButter: Neural interactive bullet time generator for human free-viewpoint rendering. In Proceedings of the ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021.
18. Broxton, M.; Flynn, J.; Overbeck, R.; Erickson, D.; Hedman, P.; Duvall, M.; Dourgarian, J.; Busch, J.; Whalen, M.; Debevec, P. Immersive light field video with a layered mesh representation. *ACM Trans. Graph. (TOG)* **2020**, *39*, 86:1–86:15. [CrossRef]

19. DuVall, M.; Flynn, J.; Broxton, M.; Debevec, P. Compositing light field video using multiplane images. In Proceedings of the 12th ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia, Los Angeles, CA, USA, 28 July–1 August 2019.
20. Mori, S.; Ikeda, S.; Saito, H. A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Trans. Comput. Vis. Appl. (CVA)* **2017**, *9*. [CrossRef]
21. Barnum, P.; Sheikh, Y.; Datta, A.; Kanade, T. Dynamic seethroughs: Synthesizing hidden views of moving objects. In Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR), Orlando, FL, USA, 19–22 October 2009; pp. 111–114.
22. Meerits, S.; Saito, H. Real-time diminished reality for dynamic scenes. In Proceedings of the International Symposium on Mixed and Augmented Reality Workshop (ISMAR-Workshop), Fukuoka, Japan, 29 September–3 October 2015; pp. 53–59.
23. Ienaga, N.; Bork, F.; Meerits, S.; Mori, S.; Fallavollita, P.; Navab, N.; Saito, H. First deployment of diminished reality for anatomy education. In Proceedings of the International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Merida, Mexico, 19–23 September 2016; pp. 294–296.
24. Kawai, N.; Sato, T.; Yokoya, N. Diminished Reality Based on Image Inpainting Considering Background Geometry. *IEEE Trans. Vis. Comput. Graph. (TVCG)* **2016**, *22*, 1236–1247. [CrossRef] [PubMed]
25. Mori, S.; Herling, J.; Broll, W.; Kawai, N.; Saito, H.; Schmalstieg, D.; Kalkofen, D. 3d pixmix: Image inpainting in 3d environments. In Proceedings of the International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Bari, Italy, 4–8 October 2021; pp. 1–2.
26. Mori, S.; Erat, O.; Broll, W.; Saito, H.; Schmalstieg, D.; Kalkofen, D. InpaintFusion: Incremental RGB-D inpainting for 3D scenes. *IEEE Trans. Vis. Comput. Graph. (TVCG)* **2020**, *26*, 2994–3007. [CrossRef] [PubMed]
27. Cernea, D. OpenMVS: Multi-View Stereo Reconstruction Library. 2020. Available online: https://cdcseacave.github.io/openMVS (accessed on 19 January 2023).
28. Penner, E.; Zhang, L. Soft 3D reconstruction for view synthesis. *ACM Trans. Graph. (TOG)* **2017**, *36*, 1–11. [CrossRef]
29. 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 Trans. Graph. (TOG)* **2019**, *38*, 1–14. [CrossRef]
30. DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperPoint: Self-supervised interest point detection and description. In Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018.
31. Sarlin, P.E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning feature matching with graph neural networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
32. Schönberger, J.L.; Frahm, J.M. Structure-from-motion revisited. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
33. Schönberger, J.L.; Zheng, E.; Frahm, J.M.; Pollefeys, M. Pixelwise view selection for unstructured multi-view stereo. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9907, pp. 501–518.
34. Flynn, J.; Broxton, M.; Debevec, P.; DuVall, M.; Fyffe, G.; Overbeck, R.; Snavely, N.; Tucker, R. Deepview: View synthesis with learned gradient descent. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2367–2376.
35. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the International Conference on Computer Vision, Kerkyra, Greece, 20–25 September 1999; Volume 2, pp. 1150–1157.
36. Sony Corporation. SPATIAL REALITY DISPLAY | White Paper. 2022. Available online: https://www.sony.net/Products/Developer-Spatial-Reality-display/en/develop/WhitePaper.html (accessed on 19 January 2023).