Improving Neural Rendering for Stereo 3D Reconstruction via Image Preprocessing and Edge Guided Ray Sampling

Report for the Advanced Medical Learning Seminar SoSe - 2023

submitted by

Franziska Hradilak

Potsdam, Germany September 1, 2023

supervised by Sumit Shekhar at the Digital Health - Machine Learning
— Hasso Plattner Institute

1 INTRODUCTION

Reconstruction of soft tissues in robotic surgery using endoscopic videos is important for a wide range of medical applications. Not only can they help doctors get a better understanding of medical conditions and procedures but also improve and enable further processing by algorithms e.g. for autonomous surgery applications. Previous approaches often struggled with limitations and disadvantages such as not being able to:

- handle complex surgical scenes
- apply to non-rigid deformations
- work with tool occlusion
- result in good results with poor 3D clues from a single viewpoint

1.1 The Existing Method

EndoNeRF, a neural rendering-based framework, is presented to overcome those issues using a variety of strategies as

- Tool mask guided ray casting: A method to only select pixels holding relevant information for reconstruction.
- Stereo depth Cueing ray marching: a sampling method to improve the selection of the sampled points from a ray for a single-view input.
- Optimization techniques: methods to improve the reconstruction and to handle corrupt stereo depth that would lead to abrupt artifacts in the final output.

1.2 Limitations and Challenges

Failure Cases for EndoNeRF are similar to those in NeRF. Poor camera poses with e.g. a too high occlusion level can limit the reconstruction quality. Also, deformations between two neighboring scenes can be too large resulting in the model not converging to a consistent deformation. Another limitation of the application is, that stereo camera data is needed or camera data with depth maps for each frame. Also, camera calibration information has to be available such as the focal length to train the model.

1.3 Problem Statement

Given a set of images of a dynamic endoscopic scene captured with a stereo camera, we aim to improve the implicit encoding and reconstruction ability of the Endo-NeRF pipeline. Or, in other words, enable an improvement of the Mapping M that given a 3D point x =(x,y,z), a view direction $d = (\theta, \phi)$) and a time instant t, outputs the emitted color c =(r,g,b) of that point and its volume density.

1.4 Contributions

Our contributions to the work are based on established image preprocessing and alterations in the EndoNeRF pipeline at the ray sampling stage and the volume rendering equation.:

(1) Preprocessing: Specularity removal on the input images.
(2) Adjustment of the volume rendering using other functions:
    - gaussian
    - x^2
(3) Adjustment of the ray sampling method using masks to take edges into account and shift the importance to the edges areas.

2 BACKGROUND

2.1 Neural Rendering

"Neural Radiance Fields" or short NeRF is a method to create realistic scenes using neural rendering. It was introduced in 2020 by Ben Mildenhall and his colleagues. Here a continuous scene is represented as a vector function. Using a fully connected deep network with a 5D coordinate (spatial location (x, y, z) and viewing direction $d = (\theta, \phi)$) as input, the volume density σ and color c, as viewed from the direction, are returned for each input coordinate. To achieve a better prediction of the color c both location and viewing direction are taken into account whereas for the density only the location is used ensuring multiview consistency. NeRF needs multiple views of the same scene to create coherent predictions across different camera positions, handle occlusions, address artifacts and noise, and predict the scene from novel viewpoints. The basis of the volume rendering is formed by radiance fields using rays that simulate the paths of light rays interacting with surfaces in a scene. Where the volume density is interpreted as the differential probability of a ray terminating at location x (with an infinitesimal distance), the color is processed as the integral of a function $T(t)$ from t near to t far ($[t_n, t_f]$). $T(t)$ represents the accumulated transmittance along a ray from its near and far bound. This can also be interpreted as the probability that the ray travels from t near to t far without hitting any other particle. To estimate this integral $[t_n, t_f]$ is split into N evenly-spaced bins, and samples are drawn at random uniformly within each bin to allow a higher resolution. Since this procedure of N query points per camera ray results in free space and occluded regions two networks are optimized. A "coarse" network uses samples as described to then use its output summing sampled colors along rays to get a more relevancy-biased distribution for the "fine" network.

Compared to former methods this approach does not need a great number of input pictures and skips the necessity of 3D modeling, which leads to reduced memory requirements. Those 3D models are furthermore expensive and often difficult to retrieve. Additionally with other techniques complicated scenes were difficult to render.

2.2 Neural Rendering for dynamic Scenes

D-NeRF is a NeRF-based approach to implicitly represent and create a dynamic scene while taking the dimension t into account. This means that the equation from a 5D space is now enriched to $M : (x, d, t) \rightarrow (\mathrm{c}, \sigma)$. D-NeRF introduced the novel technique to split the model into a canonical and a deformation network outperforming the intuitive approaches of directly learning the transformation from the 6D (x,d, and time) space to the 4D space of color and density. Also, with this approach only one single view is required per instant instead of multiple views of one rigid scene. The networks work as follows. The canonical network has to find a representation of the scene including and encoding all values (density and color) per corresponding points in all images. For an input i containing a point x viewed from the camera viewing direction this model is optimized to learn color and density for the point in the canonical space. The deformation network is then optimized to represent, as the name suggests, the deformation field between the canonical scene and the scene at time instant (where the time instant 0 represents the canonical scene). The volume rendering from NeRF is now adapted such that density and color are predicted by the canonical network. The canonical network receives the output from the deformation network which again receives a ray and time instant as input. To learn both models the mean squared error considering the RGB images of the scene with camera poses is minimized simultaneously.

This method introduced further improvements. It only requires a single view per time instant and only a sparse set of images making rendering applicable for even more scenarios. And of course, the ability to model fine geometric details while handling several dynamic input types, like human

motion and asynchronous motion, presented a great advantage compared to the former approaches. Limitations are inter alia poor camera poses and large deformations between two neighbor scenes. With this, the model possibly can't converge to a consistent deformation.

2.3 Endoscopic Reconstruction

As stated before the reconstruction of soft tissues in robotic surgery using endoscopic videos is important for a wide range of medical applications. Previous approaches often struggled with limitations and disadvantages such as not being able to:

- handle complex surgical scenes
- apply to non-rigid deformations
- work with tool occlusion
- result in good results with poor 3D clues from a single viewpoint

EndoNeRF is presented to overcome those issues using a variety of strategies. The method is based on D-NeRF/ NeRF using neural rendering with a canonical (mapping time and coordinates to density and color) and deformation network (mapping x,t to displacement between point x at time t and the corresponding point in the canonical field). We consider stereo surgical dynamic scenes that only have one viewpoint. The goal is now to reconstruct the complex 3D structures and textures while also removing the tool occlusion. To achieve this tool masks have to be acquired as well as depth maps from the binocular recordings and following alterations and additions are added to the concept. Tool mask-guided ray casting describes the curated pixel sampling for the rays shot to the image plane. Since many pixels represent surgical tools that hold misleading information for tissue reconstruction, rays that travel through those pixels are filtered out during training using the binary tools masks. In addition, it is desirable to emphasize the sampling on pixels that have higher tool occlusion frequencies as information on those areas is scarcer and holds thus a higher information density than those that are available at many time instances.

Since NeRFs sampling strategy does not apply to the current single view-point setting an iso-surface rendering inspired sampling using Gaussian transfer functions is introduced. Furthermore, to sample points preferably close to the tissue surface an impulse distribution around the depth of a pixel is used to draw points from the normalized impulse distribution. Following the volume rendering concept color and density are calculated. As an additional step, the optical depth is also evaluated by volume rendering. The loss function for training the networks now not only includes supervising the rendered color but also the optical depth. As a final addition statistical depth refinement is conducted to react to corrupt stereo depth induced by specular highlights or fuzzy pixels. Since overfitting can be beheld when directly supervising the estimated depth (resulting in artifacts in the reconstruction results) Endo-NeRF takes advantage of an early underfitting model averaging learned colors and densities leading to smoother colors and depth information. Using residual maps to find corrupt depth information, pixels can be replaced with depth values from earlier iterations.

3 IMPLEMENTATION DETAILS

3.1 Components of Endo NeRF Pipeline

The EndoNeRF Pipeline consists of the following stages that each training iteration will run through.

(1) A frame is randomly picked for training.
(2) Tool-guided ray casting will run to shoot rays into the scene ignoring those that would go through occluded pixels
(3) and points along the rays are sampled with depth cueing ray marching.

(4) Thus points are handed to the networks that return for each, color and space occupancy.
(5) The volume rendering integral on our sampled points is evaluated
(6) and finally the rendering loss and depth loss are optimized to achieve a reconstruction of the surgical scene.

3.2 Modifications

After researching Endo-NeRFs methods and the scientific work it is built up on we proceeded with a Design-thinking inspired brainstorming session on how and where to improve the pipeline. Here we found two general approaches. One is to focus on preprocessing the input images to enable the networks to work better. The second one is to alter techniques or parameter values used in the existing pipeline or to add new ones.

3.2.1 Preprocessing. I started with the first approach with the task to reduce the specularity in the images as specular highlights are known to pose challenges in image processing and machine learning tasks. We used the SHIQ [1] repository based on work from Fu et Al. that utilizes a multi-task network for joint highlight detection and removal. We followed the instructions from the README using the pre-trained model.

3.2.2 Volume Rendering. The next idea was to use other functions than exp() in equation 3 for the color and depth estimation via volume rendering as described in section 2.5. The functions considered were:

- x^2
- gaussian: e^{-x^2}

In the run_endonerf.py file from the Endonerf repository the function raw2outputs() returns the model's predictions in a semantically meaningful way. We can find the evaluation of the optical depth and the emitted color in here. We can alter the function calculating the weights from equation 3 to the functions we choose.

3.2.3 Edgeguided Ray Sampling. Another modification was to take the edges of the images into account when shooting the camera rays. Therefore we used the phycv repository [7]. It contains various image processing functionalities like color and light enhancement and different edge detections. We used the so-called Phase-Stretch Transform algorithm. Therefore, again following the README, we used run_pst.py to translate each input image to a mask where edges are detected and translated to white pixels. The other areas are colored black.

One approach was to set the sampling importance of the pixels that lay on edges to the highest level. Meaning that now not only emphasis is put on whether a point in the scene is often occluded. Also, points that lay on edges but represent a tool are still not included in the sampling. As edges often hold important geometric information, e.g. regions of high detail where materials change, it is important to ensure they are included in the sampling. Edges can also tend to be in regions of rapidly changing structures, that are thus important to capture for the reconstruction. Therefore we looked at the run_endonerf_helpers.py where we added new sampling functions, altering the ray_sampling_importance_from_masks() function. To set the sampling importance of the edges we use the importance masks calculation(taking the frequency of visibility into account) from the repository and the edges masks. Remember that in the importance mask the highest priority is marked as a 1.0 and the lowest (e.g. tool occlusion pixels) as 0.0. Joining the masks via taking the maximum we achieve the prioritization of the edges. Now we have to filter out the tool occlusion values again by generating a new tool mask where the areas with tool occlusion are represented with a 1.0 and all other values are 0.0. We now multiply our joint mask with the inverse of the new tool mask.

Another radical approach was tested to shoot rays only to the edges. Therefore we again retrieved the occlusion mask but now multiplying its inverse with the edges mask. Therefore only edges that do belong to the soft tissue are represented as a 1.0 all other areas get ignored due to their 0.0 representation. A possible next steps would be to set the sampling probability of points lying on edges higher than the most occluded ones.

3.2.4 Including new endoscopic video material. To include new stereo endoscopic surgery videos or also possibly videos from other domains we wrote a script creating the poses_bounds.npy file in the LLFF format necessary for training the networks on images. The necessary steps therefore are explained at the LLFF repository [2] in the section: Using your own poses without running COLMAP. The focal length, height, width, and close/far depth bounds have to be known or retrieved before computing. As done for the provided videos we retrieve the close and far depth the depth maps, extracting the maximum and minimum value. The focal length is retrieved from the camera calibration data taking the mean from the x,y focal distance since in EndoNeRF they assume that the x,y distance is the same. It is important to note that instead of the rotation matrices we use identity matrices for the camera poses since we only have a single-viewpoint setting. This is done to avoid interferences of badly calibrated poses. Hereby we follow the procedure described in the EndoNeRF README.

3.2.5 Others. Since it would be interesting to compare the point clouds and their geometric behavior we tried different approaches to display two point clouds fulfilling this need. Approaches were to color code areas, depth, or the whole point cloud and present them overlapping. Yet since it is difficult to decide what a ground truth would look like, since the input values are stereo images this did not lead to insightful results.

4 RESULTS AND DISCUSSION

4.1 Effect on performance and quality

The preprocessing step of removing specularity resulted in improved results for the quantitative metrics as used in the paper with ↑ 2,810 PSNR, ↑ 0,19 SSIM, ↓ 0,019 LPIPS for the case cutting tissues twice and with ↑ 2,818 PSNR, ↑ 0,17 SSIM, ↓ 0,017 LPIPS. On the other hand as can be seen in [1] and [2], fine details of tissue structure get less detailed. Yet removal of strong specular effects might improve the handling by doctors due to reduced distractions.

Applying the ray to edges strategy, metric vise the prioritized edges approach can only slightly improve PSNR, and SSIM and even has a slightly worse result for LPIPS.[2]. We also noticed an increase of artifacts in the reconstruction [4]. When looking at the results for the radical method to only shoot rays towards edges we can see a decline for all metrics. This corresponds with the increased blurriness in the reconstruction [3].

When looking at the results of the alternative functions used there is no apparent qualitative difference in the reconstruction for the square function. Applying the gaussian function results in harsh edges, especially on the areas that have more artifacts due to representing where the tools have been in the original video. When using the preprocessed images with the gaussian function the results for the pulling example still have a higher artifact rate whereas for the cutting example, the metrics and the qualitative results are comparatively similar. The difference between the tests with and without preprocessing has drastically different metric results. [5] [6]

4.2 Comparative results

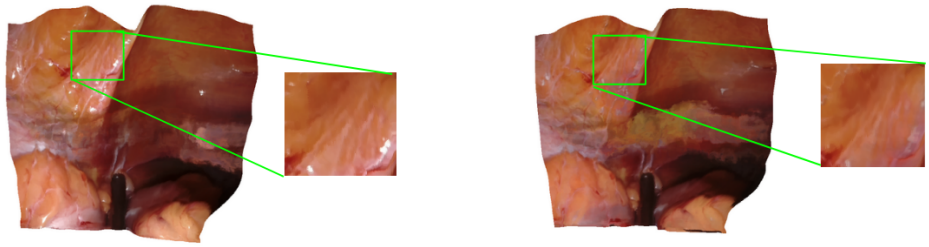

Fig. 1. Comparison between reconstruction without and with specularity preprocessing.
Video: Cutting Tissues Twice Frame 62

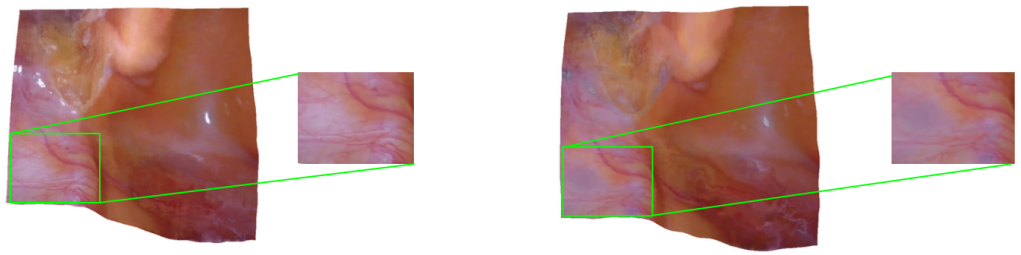

Fig. 2. Comparison between reconstruction without and with specularity preprocessing.
Video: Pulling Soft Tissue Frame 63

Method	PSNR ↑	SSIM ↑	LPIPS
w/o Adjustments	37.168	0.948	0.054
Preprocessing	39.978	0.967	0.035
Gaussian	23.596	0.913	0.106
Preprocessing and Gaussian	39.68	0.966	0.033

Table 1. Quantitative evaluation on photometric errors of the dynamic reconstruction on metrics of PSNR, SSIM, and LPIPS for video cutting tissues twice.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Adjustments	38.238	0.954	0.050
Preprocessing	41.056	0.971	0.033
Prioritized Edges	38.307	0.954	0.049
Only Edges	35.974	0.945	0.088
Square	36.559	0.943	0.048
Gaussian	23.596	0.913	0.106
Preprocessing and Gaussian	40.729	0.969	0.031

Table 2. Quantitative evaluation on photometric errors of the dynamic reconstruction on metrics of PSNR, SSIM, and LPIPS for video pulling soft tissues.

4.3 Discussion of results

Specularity removal is known to be an essential method to improve many computer vision algorithms and to be important to improve the observation and resulting observations of surgeons [3]. With less distracting extreme input values the network is possibly able to better generalize.

Even though edges present valuable information for the whole geometric structure of an image, by focusing too much on them the tissue structure in between possibly did not get sufficiently taken into account in the prioritized edges example. This could have led to the increase of artifacts in the reconstruction. The radical edges experiment might not have included enough information for generalization with not enough references for the whole image, thus leading to a blurry image. A possible explanation for the noticeable difference in the experiments with the gaussian function could be that with preprocessing the image values are already less extreme and less differing. Since aliasing effects often occur when high-frequency components from the input are not properly represented in the output and the simple gaussian function e^{-x^2} limits values.

5 CONCLUSION & OUTLOOK

5.1 Learnings

- An important learning was to first approach this research project in a very open-minded state and then later focus on a more specific task, taking the limited time into account.
- Furthermore often when we needed a processing step already existing work was very helpful.
- Many repositories could just be integrated but of course, others were not working or matching our needs. In the latter, it was important to learn not to get too attached to trying to get this one solution to work but to think more creatively and to look for other solutions.
- When starting this project it was fundamental to first get into the material and read up on the groundwork preceding the EndoNeRF paper. This way when interacting with the code

we already had a feeling of the methods and knew where we had to look for if we struggled with understanding the implementation. It was very interesting to see that when taking our first steps towards understanding the concepts everything seemed very complex and complicated. But now having read, discussed, and built upon the work and rereading the papers everything seems familiar and easily understandable.
- Finally working collaboratively on the project sharing and discussing problems and ideas regularly is essential for the learning process getting over setbacks and improving.

5.2 Future Work

Due to time constraints, many existing ideas and approaches that we looked into could not be followed yet.

- The approach to change the function in the volume rendering process can still be extended. We thought about using the tan(x) function or the function $tan(x * \pi/2)$ as their steeply increasing curve in the positive range is similar to the exponential function. The exponential function is rooted in the physical properties of the light. With the distance the light travels the density increases exponentially. This way an alteration of the steepness of the slope could be interesting.
- It would be interesting to test our adjustments on several new endoscopic videos to get further insights. There might be cases where our adjustments make a greater difference than in others or some for which the EndoNeRF method is not applicable. We were already researching for freely available endoscopic stereo videos that showed tool occlusion and presented the necessary camera calibration data. [4] presents promising sample data for this.
- In the current work, the tool masks are generated by hand. We already generated new tool masks with a library [5] based on work from Shvets and Al. . It would be interesting to check whether this makes a difference since automated tool masking would accelerate the whole process tremendously.
- Also, following up on the specularity removal more preprocessing techniques could be evaluated such as histogram equalization to enhance the visibility of features.
- The approach to adjust the sampling strategy could also lead to further improvements. A method that does include the edges could be further pursued or the current edge-favoring approach tested on different videos with different tissue structures.
- Furthermore, combining techniques such as preprocessing and edge-guided ray sampling could lead to promising results. And finally, as they already stated in their repositories README "Note that we only evaluate photometric errors due to the difficulties in collecting geometric ground truth."[6] it would be very interesting to apply or create geometric metrics to not only get photometric evaluation but also a quantitative comparative result directly linked to the point cloud reconstruction.

REFERENCES

[1] Gang Fu, Qing Zhang, Lei Zhu, Ping Li, and Chunxia Xiao. 2021. A Multi-Task Network for Joint Specular Highlight Detection and Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7752–7761.

[2] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. 2019. Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines. *ACM Transactions on Graphics (TOG)* (2019).

[3] Chao et al. Nie. 14 Jan. 2023. Specular Reflections Detection and Removal for Endoscopic Images Based on Brightness Classification.". *(Basel, Switzerland) vol. 23,2 974.* (14 Jan. 2023). https://doi.org/10.3390/s23020974

[4] David Recasens, José Lamarca, José M Fácil, JMM Montiel, and Javier Civera. 2021. Endo-Depth-and-Motion: Reconstruction and Tracking in Endoscopic Videos Using Depth Networks and Photometric Constraints. *IEEE Robotics and Automation Letters* 6, 4 (2021), 7225–7232.

[5] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. [n. d.]. Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning.

[6] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. 2022. Neural Rendering for Stereo 3D Reconstruction of Deformable Tissues in Robotic Surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 431–441.

[7] Yiming Zhou, Callen MacPhee, Madhuri Suthar, and Bahram Jalali. 2023. PhyCV: The First Physics-inspired Computer Vision Library. *arXiv preprint arXiv:2301.12531* (2023).

6 APPENDIX

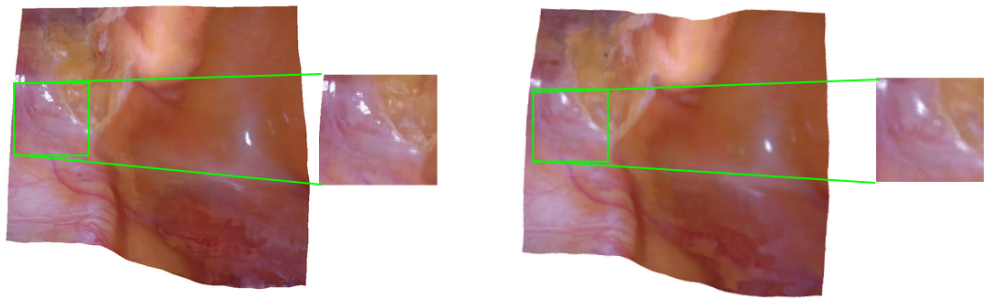

Fig. 3. Comparison between reconstruction without and with sampling only from edges.
Video: Pulling Soft Tissues Twice Frame 62

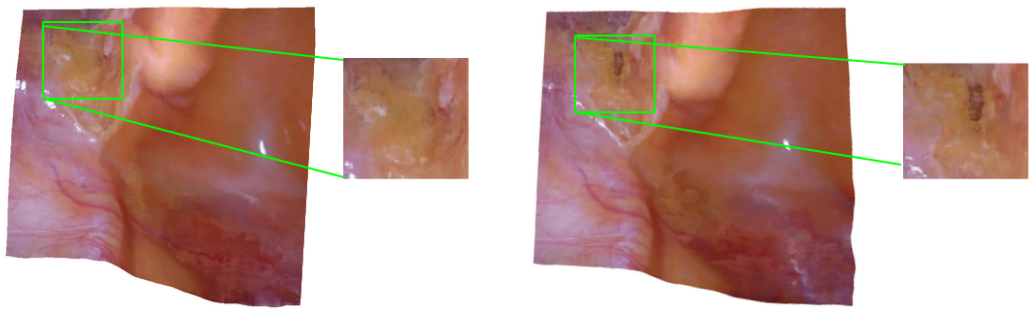

Fig. 4. Comparison between reconstruction without and with sampling with prioritized edges.
Video: Pulling Soft Tissues Twice Frame 62

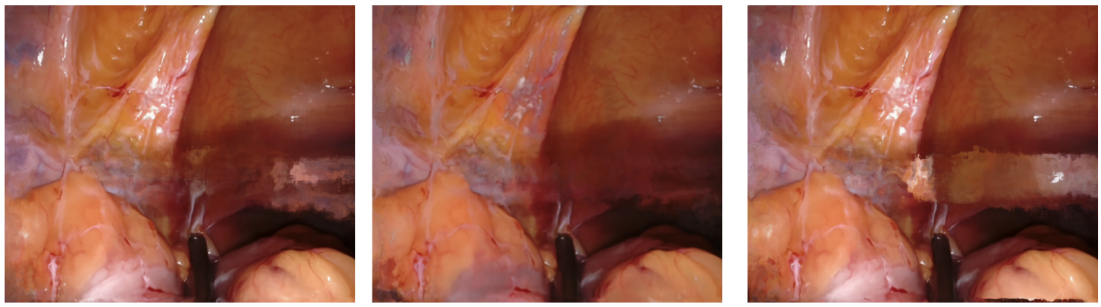

Fig. 5. Comparison between reconstructed images: without alterations, with gaussian, with preprocessing and gaussian.
Video: PCutting Tissues Twice Frame 63

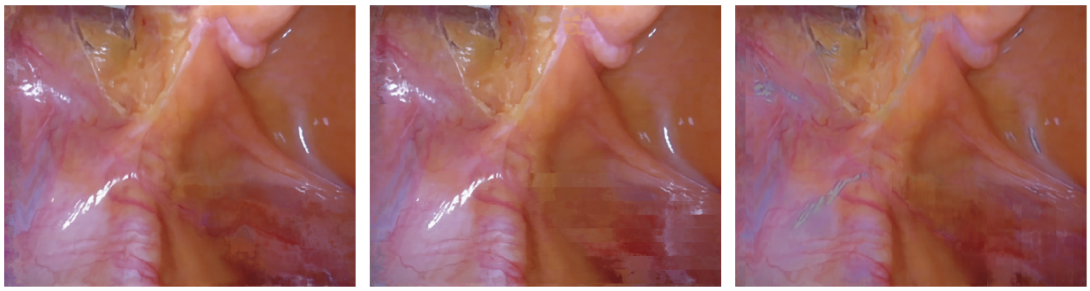

Fig. 6. Comparison between reconstructed images: without alterations, with gaussian, with preprocessing and gaussian.
Video: Pulling Soft Tissues Twice Frame 45