Segmentation and removal of surgical instruments for background scene visualization from endoscopic/ laparoscopic video

Hasan, S. M. Kamrul, Simon, Richard, Linte, Cristian

S. M. Kamrul Hasan, Richard A. Simon, Cristian A. Linte, "Segmentation and removal of surgical instruments for background scene visualization from endoscopic/laparoscopic video," Proc. SPIE 11598, Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling, 115980A (15 February 2021); doi: 10.1117/12.2580668

SPIE.

Event: SPIE Medical Imaging, 2021, Online Only

Segmentation and Removal of Surgical Instruments for Background Scene Visualization from Endoscopic / Laparoscopic Video

S. M. Kamrul Hasan[a, b] (✉), Richard A. Simon[a, c], and Cristian A. Linte[a,b,c]

[a]Biomedical Modeling, Visualization and Image-guided Navigation (BiMVisIGN) Lab , RIT
[b]Center for Imaging Science, Rochester Institute of Technology, NY, USA
[c]Biomedical Engineering, Rochester Institute of Technology, NY, USA

ABSTRACT

Surgical tool segmentation is becoming imperative to provide detailed information during intra-operative execution. These tools can obscure surgeons' dexterity control due to narrow working space and visual field-of-view, which increases the risk of complications resulting from tissue injuries (e.g. tissue scars and tears). This paper demonstrates a novel application of segmenting and removing surgical instruments from laparoscopic/endoscopic video using digital inpainting algorithms. To segment the surgical instruments, we use a modified U-Net architecture (U-NetPlus) composed of a pre-trained VGG11 or VGG16 encoder and redesigned decoder. The decoder is modified by replacing the transposed convolution operation with an up-sampling operation based on nearest-neighbor (NN) interpolation. This modification removes the artifacts generated by the transposed convolution, and, furthermore, these new interpolation weights require no learning for the upsampling operation. The tool removal algorithms use the tool segmentation mask and either the instrument-free reference frames or previous instrument-containing frames to fill-in (i.e., inpaint) the instrument segmentation mask with the background tissue underneath. We have demonstrated the performance of the proposed surgical tool segmentation/removal algorithms on a robotic instrument dataset from the MICCAI 2015 EndoVis Challenge. We also showed successful performance of the tool removal algorithm from synthetically generated surgical instruments-containing videos obtained by embedding a moving surgical tool into surgical tool-free videos. Our application successfully segments and removes the surgical tool to unveil the background tissue view otherwise obstructed by the tool, producing visually comparable results to the ground truth.

Keywords: Surgical tool segmentation, tool removal, video inpainting, non-parametric optical flow, affine parametric motion, Poisson blending

1. DESCRIPTION OF PURPOSE

Minimally invasive medical procedures based upon optical imaging systems have become increasingly common in today's healthcare practice, due to the reduced patient recovery time and mortality rate. Optical imaging enabled robotic platforms such as da Vinci surgical system (Intuitive Surgery) to be used to perform minimally invasive complex surgical procedures. However, surgical instruments used in the endoscopic surgical suturing procedures, obscure surgeons' dexterity control due to narrow working space and visual field-of-view. These hindrances in the visual field increase the risk of tissue scars and tears. Hence, removal or masking the surgical instruments transparent from the background and then inpainting the foreground masked region with background content is paramount.

Research efforts focused on surgical instrument segmentation from endoscopic/laparoscopic videos have become more widespread and have been the focus of many biomedical imaging challenges. In this paper, we present an innovative application of our neural net-based surgical tool segmentor (U-NetPlus)[1] to digitally remove surgical tools from video frames enabling the visualization of anatomy obscured by the tool. The authors know of only one other work tackling the segmentation and modification of surgical instruments in endoscopic/laparoscopic videos. Koreeda *et al.*[2] presented

Further author information:
S. M. Kamrul Hasan (E-mail: sh3190@rit.edu)
Richard A. Simon (E-mail: rasbme@rit.edu)
Cristian A. Linte (E-mail: calbme@rit.edu)

Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling,
edited by Cristian A. Linte, Jeffrey H. Siewerdsen, Proc. of SPIE Vol. 11598, 115980A

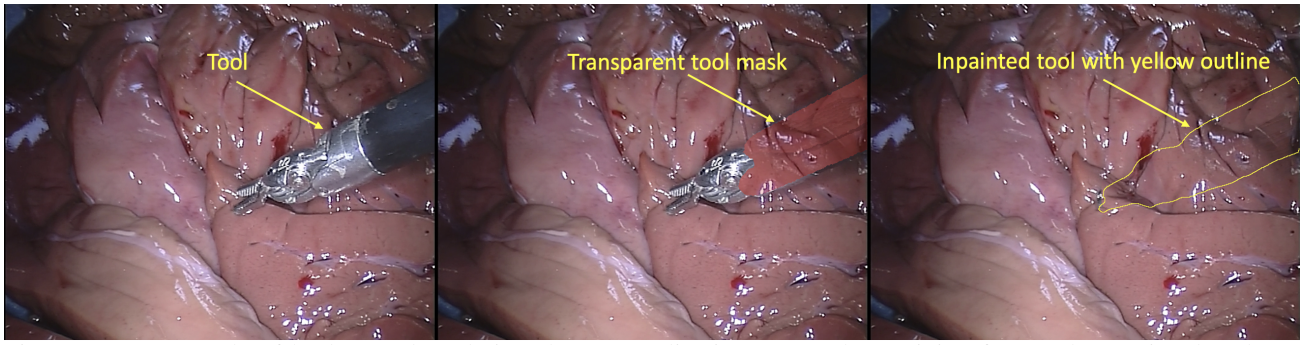

Figure 1: An example of background renderings by our application: (a) tool containing frame; (b) Inpainted tool; (c) Inpainted tool with yellow outline.

a hardware/software-based solution to visualize areas obscured by surgical instruments. Nevertheless, their method poses some limitations related to the need for multiple endoscopes present, which may increase patient invasiveness. In this paper, we have developed two image-driven approaches for surgical tool removal; both approaches rely on the use of information from the images captured by the laparoscope / endoscope to "paint over" the surgical tool mask identified by our automated surgical tool segmentor. We show two example renderings of the background otherwise hidden behind the surgical tool "removed" using our proposed application in Figure 1.

2. METHODS

2.1 SURGICAL TOOL SEGEMENTATION METHOD

To segment the surgical instruments, we use a modified U-Net architecture (U-NetPlus)[1] composed of a pre-trained encoder (VGG-11/VGG-16) and a decoder redesigned by replacing the transposed convolution operation with an up-sampling operation based on nearest-neighbor (NN) interpolation.[3] The network was trained on a dataset obtained from the MICCAI 2015 EndoVis Challenge focused on surgical instrument segmentation and tracking. The dataset set is composed of video sequences collected with the da Vinci surgical system during laparoscopic procedures. In each frame, the articulated parts of the instrument consist of the shaft and claspers, and are accompanied by labeled ground truth masks automatically generated by the da Vinci Research Kit using joint encoder information and forward kinematics.

2.2 SURGICAL TOOL REMOVAL METHOD A: OPTICAL FLOW-BASED VIDEO OBJECT REMOVAL ALGORITHMS

The first approach is based on video object removal algorithms[4,5] that employ data from previous frames that contain the surgical tool to replace the pixels of the segmented tools in the current frame. The method works by establishing correspondences between pixels (regions) occluded by the surgical tool Ω_t in the current frame $I_t(x, y)$ to the pixels (i.e. regions) observed in the background region of a previous frame $I_{(t-1)}(x, y))$. The background region corresponds to pixels (i.e. regions) not occluded by the foreground surgical tool. The correspondences between the frames can be identified by using a parametric warp model,[6] such as an affine warp described in Equation 1:

$$\mathbf{p}^* = \min_{p} \sum_{\substack{x,y\epsilon\Omega, \\ (x,y)\neq\Omega_t, \\ (x+u,y+v)\neq\Omega_{t-1}}} [I_t(x,y) - I_{t-1}(x+u(x,y;\mathbf{p}), y+v(x,y;\mathbf{p}))]^2 \tag{1}$$

where

$$\begin{bmatrix} u(x,y;\mathbf{p}) \\ v(x,y;\mathbf{p}) \end{bmatrix} = \begin{bmatrix} p_1 & p_3 & p_5 \\ p_2 & p_4 & p_6 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

represents the displacement vector at pixel (x, y) from I_t to $I_{(t-1)}$and Ω represent the region used to determine the affine parameters **p**. The region Ω can represent the whole frame or the union of the dilated tool mask from the current and previous frame. The displacement field in the missing tool region Ω_t is determined by evaluating Equation 1 within the region Ω_t using the determined affine parameters **p***.

Alternatively, the correspondences can be determined by a non-parametric optical flow-based model[7] such as

$$\mathbf{u^*, v^*} = \min_{u,v} \sum_{x,y \epsilon \Omega} [I_t(x,y) - I_{t-1}(x + u(x,y), y + v(x,y))]^2 + \alpha(|\nabla u(x,y)|^2 + |\nabla v(x,y)|^2) \tag{2}$$

where α is the weight between the data (first) and smoothness (second) term. The data term represents the similarity between the pixel values of adjacent frames, while the smoothness term enforces the smoothness of the flow fields. The data term is undefined inside the tool regions Ω_t and Ω_{t-1}, so the smoothness term becomes the only constraint resulting in the optical flow field being smoothly interpolated into the missing tool region.

The correspondences (u, v) are used to trace the backward displacement at each pixel of the tool region Ω_t to find its corresponding location in a previous frame where the background region is visible. The occluded pixel is then replaced by the pixel value of the un-occluded pixel using bilinear interpolation. If an un-occluded pixel doesn't exist in the previous frame, then the pixel value of the tool region is left unchanged.

2.3 SURGICAL TOOL REMOVAL METHOD B: REFERENCE IMAGE FRAME INPAINTING

This approach relies on the collection of a number of reference image frames before the surgical instruments are introduced into the surgical environment and appear in the field-of-view of the laparoscope / endoscope. These reference images $R_i(x, y)$ are then used by the inpainting algorithm to replace the segmented surgical tools. The method works by establishing correspondences between regions occluded by the surgical tool Ω_t in the current frame $I_t(x, y)$ to the regions observed in a reference frame. From the set of frame reference frames captured before the tools were introduced, we determine the closest matching reference frame and then further spatially transform the reference image to match current image and fill the tool mask region with the pixels from the warped reference images. For the current frame, we first find the reference image having the smallest sum of the square differences (SSD) between the reference and the current image within a region of interest surrounding the tool mask Ω in the current image.

$$\sum_{x \epsilon \Omega} [R_i(x,y) - I_t(x,y)]^2 \tag{3}$$

where i is the index of the reference frame. This term enforces spatial continuity between the selected reference and the region surrounding the tool mask. The chosen reference frame is then spatially transformed to improve its registration to the current frame and to determine the displacement field in the missing tool region. Similar to the previous method A, the spatial transforms can be defined by an affine parametric motion model defined via Equation 1 or by non-parametric optical flow-based model Equation 2 .

2.4 ILLUMINATION/APPEARANCE ADJUSTMENT

Nonuniform illumination of the operating environment results in variations in the appearance of tissue in different frames. As a result, copying pixels from the reference images or previous frames into the tool mask region can result in noticeable boundaries (seams) between the inpainted and existing regions. To mitigate these seaming artifacts, we use modified Poisson blending[8] to blend the current frame background I_B with the inpainted tool region, whereby instead of combining pixels from the two regions, their gradient fields are combined, using the model described in Eqution 3:

$$I^* = \min_I \sum_{x,y \epsilon \Omega_t} |\nabla I(x,y) - \mathbf{v}(x,y)|^2 \tag{4}$$

$$I_B|_{\partial \Omega} = I^*|_{\partial \Omega} \tag{5}$$

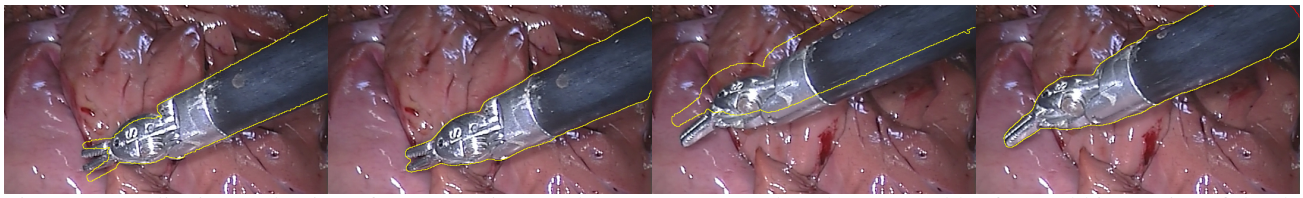

Figure 2: Qualitative evaluation of segmentation results: (a)&(c) ground truth generated by forward kinematics of the da Vinci Research Kit; (b)&(d) segmentation results from our U-NetPlus segementor.

where I^* is the Poisson blended inpainted tool image, $\mathbf{v}$ is the gradient of inpainted tool image determined by the tool removal algorithms, $\partial\Omega$ is the boundary between the inpainted region and the background, and Ω_t is the tool mask region. The current image provides Dirichlet boundary conditions for the equation around the inpainted region. If the inpainted region does not span the entire tool mask region, pixels bordering the remaining unfilled region take on Neumann boundary conditions.

3. RESULTS

Overall, our tool segmentation architecture shows sufficient accuracy for reliable binary segmentation of the surgical tools. For the training set the DICE score was $90.84 \pm 0.046\%$ and for the test set the DICE score was $89.56 \pm 0.103\%$.

It should be noted that the da Vinci labeled ground truth data does not always represent an accurate segmentation of the surgical tool (see Figure 2 (b) & (d)). There are significant limitations that essentially discredit the reliability of the ground truth data due to the misalignments associated with the tool outline reconstructed from the forward kinematics of the da Vinci Research Kit and the actual tool appearance in the image frame. Nevertheless, our segmentation technique learns how to compensate for these limitations and yields more accurate tool outlines than those generated from the ground truth forward kinematics (see Figure 2 (a) & (c)).

The first surgical video demonstrates that our tool segmentor can successfully segment and generate a mask that can be used to remove the tool from the video images. In this video, the camera is stationary, while viewing *in vivo* anatomy with minimal surface deformation. In Figure 3, we show the results of the tool segmentor (top row (a), red outline) and tool removal method that uses an affine parametric motion model to inpaint the segmentation mask region (bottom row (b)). The majority of frames show tool segmentation results that are comparable to the results shown in columns 1 and 3. Occasionally the tool segmentor misses parts of the tool calipers as shown in column 2. To compensate for under segmentation and to ensure complete inpainting of the tool, the segmentation mask was dilated by 20 pixels.

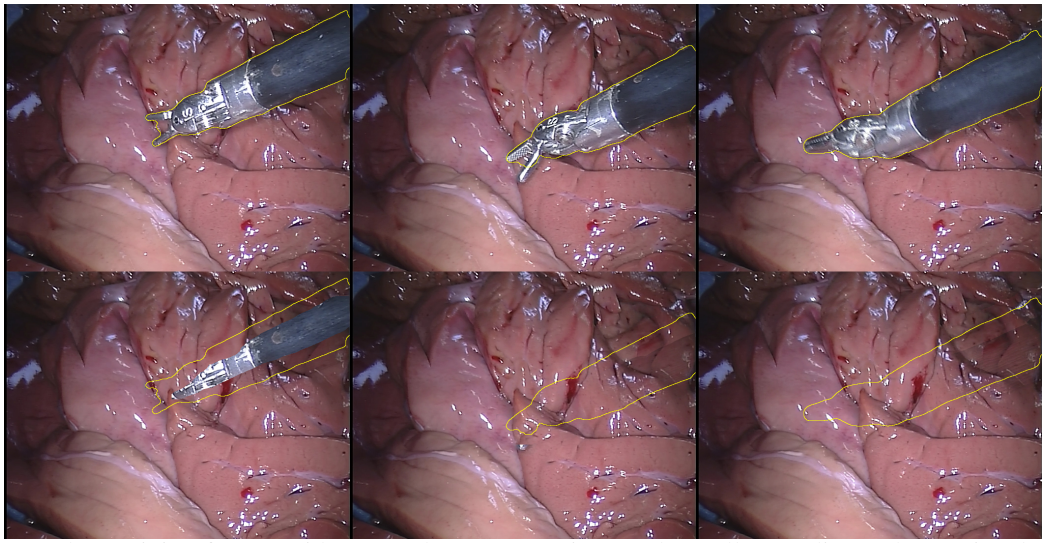

Figure 3: (a) Tool containing frames with U-NetPlus segmentation results (yellow outline). (b) Inpainted results using Method A; yellow arrow in mid-column shows residual tool caliper.

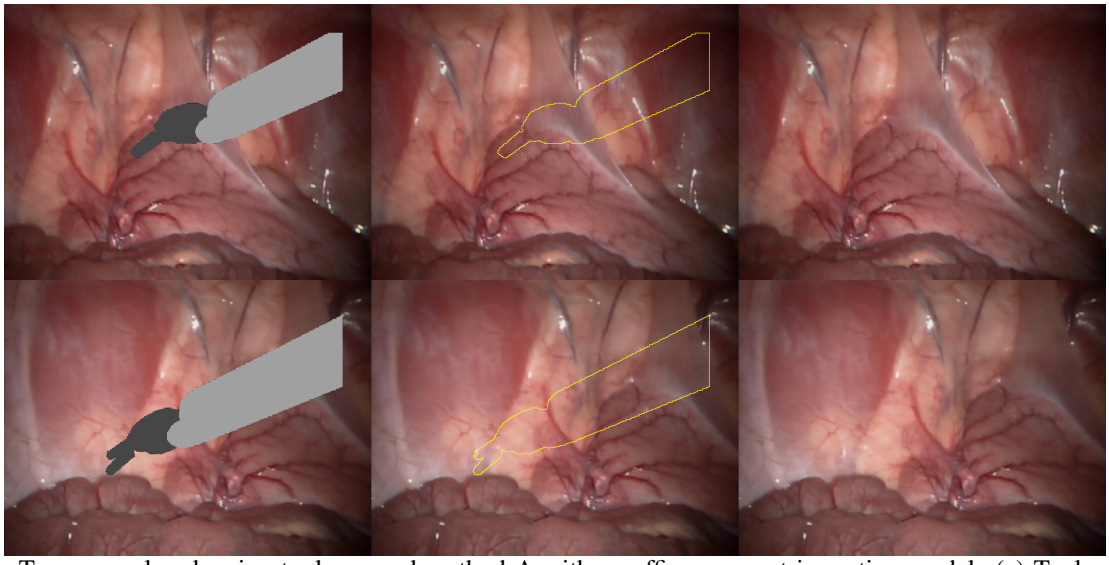

Figure 4: Two examples showing tool removal method A with an affine parametric motion model: (a) Tool containing frames; (b) Poisson blended inpainted results using Method A; (c) ground truth frames.

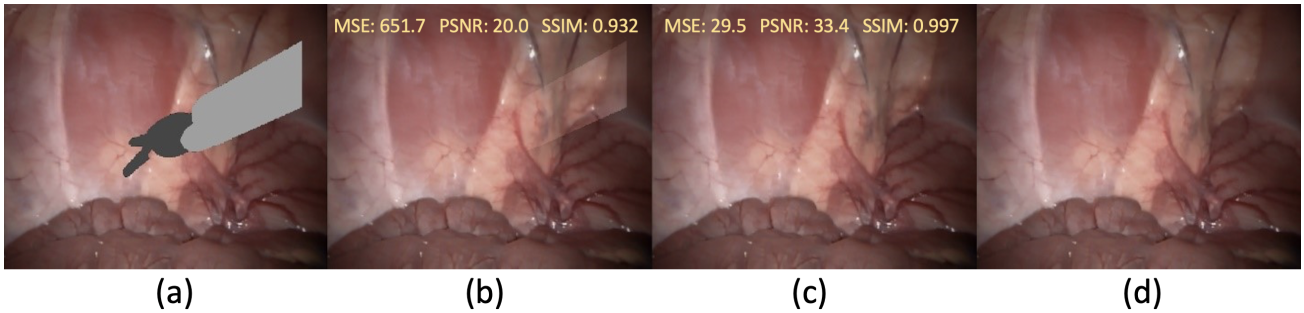

Figure 5: Example showing the effect of Poisson blending: (a) tool containing frame; (b) inpainted results; (c) Poisson blended inpainted results; (d) ground truth.

To test our tool removal algorithms on more difficult cases where the camera and/or anatomy are in motion, we generated videos containing surgical tools from surgical tool-free videos by embedding a moving surgical tool into the surgical tool-free video. The surgical tool-free videos were obtained from the Hamlyn Centre Laparoscopic / Endoscopic Video Datasets and the surgical tool was the ground truth mask obtained from the MICCAI 2015 dataset. In these cases, the tool segmentation mask was obtained from the ground truth mask and was dilated by 1 pixel.

In Figure 4, we show representative examples of using tool removal method A with an affine parametric motion model to remove the tool from a video where the camera is in motion while viewing a porcine abdomen with minimal deformation of the abdomen. Column (a) shows the tool containing frame, column (b) shows the Poisson blended inpainting results and column (c) shows the ground truth.

In Figure 5, we show the efficacy of using Poisson blending to mitigate illumination seams. Column (a) shows the tool containing frame, column (b) shows the inpainting results, column (c) shows the Poisson blended inpainting results, and column (d) shows the ground truth.

In Figure 6, we show the results of using tool removal method B using a nonparametric optical flow-based model to remove the tool from a video where the camera is stationary while viewing a cardiac surface deforming due to both respiration and cardiac motion. The reference frames are captured before introducing the surgical tools and consist of 150 consecutive frames that encompass multiple cycles of the deforming cardiac surface. Column (a) shows the tool containing frame, column (b) shows the inpainted results and column (c) shows the ground truth.

In Table I, we report the quantitative evaluation of the inpainted videos using mean squared error (MSE), peak signal to

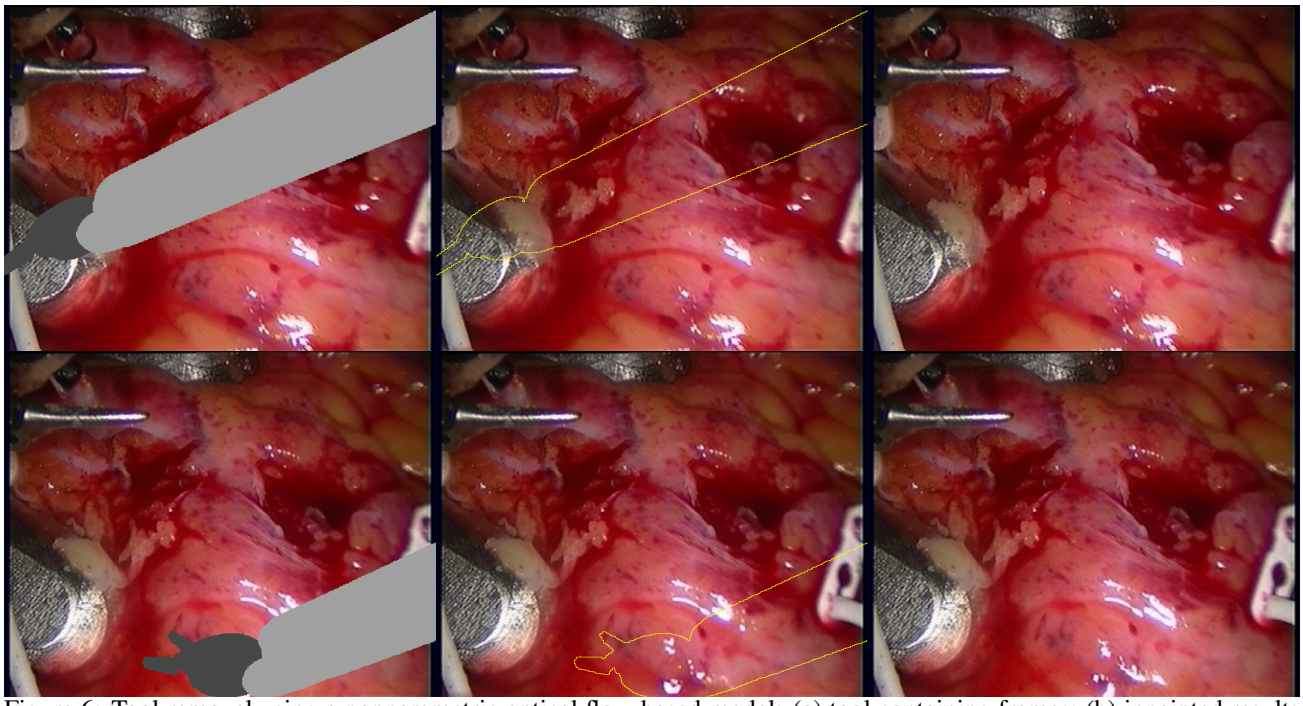

Figure 6: Tool removal using a nonparametric optical flow-based model: (a) tool containing frames; (b) inpainted results using Method B; (c) ground truth frames.

noise ratio (PSNR), and structural similarity index (SSIM)[9] image quality metrics. It can be noted that MSE and PSNR are not always well-correlated with perceived/subjective visual quality, whereas SSIM can show better correlations.

Table 1: Quantitative evaluation of the tool removal methods for synthetic tools in terms of mean squared error (MSE), peak signal to noise ratio (PSNR), and structural similarity index (SSIM).

Method	Metric		
	MSE (avg / min / max) (smaller better)	PSNR (avg / min / max) (larger better)	SSIM (avg / min / max) (larger better)
Method A: Affine Transformation (640 x 480 x 135)	690.9 / 58.0 / 2111.6	22.5 / 14.9 / 30.5	0.932 / 0.797 / 0.993
Method A: Affine Transformation with Poisson Blending	41.5 / 6.5 / 163.9	33.3 / 26.0 / 40.0	0.993 / 0.958 / 0.999
Method B: Copy and Paste (720 x 576 x 500)	223.7 / 40.8 / 1183.5	25.4 / 17.4 / 32.0	0.971 / 0.937 / 0.994
Method B: Optical Flow Warping	125.0 / 16.7 / 641.7	28.1 / 20.1 / 35.9	0.980 / 0.948 / 0.994

For the method A example, we show the comparison between the inpainted and Poisson blended inpainted results. For this example, the algorithm performs well in finding the appropriate pixels from previous frames to fill in the occluded region. But these image pixels originate from frames where the illumination of the occlude anatomy was not the same as the current frame (see Fig. 5b). Therefore, the errors for this example are mostly nonstructural errors and can be reduced by using the Poisson blending algorithm to help to minimize illumination mismatches.

For the method B example, we show a comparison between copying and pasting the pixels of the closest reference frame before and after applying the optical flow transformation. For this case, since the camera is stationary, the illumination is fairly constant albeit there are variations in the specular highlights due the variations in the surface in the beating heart. The errors in this example are mostly structural errors due to the potential lack of an appropriate match between the reference frames and current frame. The lack of a matching frame is most likely due to an insufficient frame rate of the video capture. Although it is also known that the beating heart has an underlying stochastic component partly due to the stochastic properties of the ion channels.[10] The spatial transformation helps to minimize these errors, but can never fully alleviate the structural errors. The complete videos for our inpainting results can be seen at https://smkamrulhasan.github.io/.

4. CONCLUSION AND FUTURE WORK

This paper demonstrates a novel application of segmenting and digitally removing the surgical instruments from laparoscopic / endoscopic video using digital inpainting to allow the visualization of the anatomy being obscured by the tool. To segment the surgical instruments, we use a modified U-Net architecture (U-NetPlus) composed of a pre-trained encoder and re-designed decoder. The tool removal algorithms use tool segmentation mask and either instrument-free reference frames or previous instrument containing frames to fill in (inpaint) the instrument segmentation mask. We have demonstrated the performance of the proposed surgical tool segmentation/removal algorithms on a robotic instruments dataset from the MICCAI 2015 EndoVis Challenge. We also showed successful performance of the tool removal algorithm from synthetically generated surgical instruments containing videos obtained by embedding a moving surgical tool into surgical tool-free videos. Our application successfully segments and removes the surgical tool producing visually comparable results to the ground truth.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award No. R35GM128877 and by the Office of Advanced Cyber infrastructure of the National Science Foundation under Award No. 1808530.

REFERENCES

[1] Hasan, S. M. K. and Linte, C. A., "U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images," in [*2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*], 7205–7211, IEEE (2019).

[2] Koreeda, Y., Kobayashi, Y., Ieiri, S., Nishio, Y., Kawamura, K., Obata, S., Souzaki, R., Hashizume, M., and Fujie, M. G., "Virtually transparent surgical instruments in endoscopic surgery with augmentation of obscured regions," *International Journal of Computer Assisted Radiology and Surgery* **11**(10), 1927–1936 (2016).

[3] Hasan, S. M. K. and Linte, C. A., "A modified U-Net convolutional network featuring a nearest-neighbor re-sampling-based elastic-transformation for brain tissue characterization and segmentation," in [*2018 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*], 1–5, IEEE (2018).

[4] Bokov, A. and Vatolin, D., "100+ times faster video completion by optical-flow-guided variational refinement," in [*2018 25th IEEE International Conference on Image Processing (ICIP)*], 2122–2126, IEEE (2018).

[5] Bokov, A. and Vatolin, D., "Toward efficient background reconstruction for 3D-view synthesis in dynamic scenes," in [*2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*], 37–42, IEEE (2017).

[6] Baker, S. and Matthews, I., "Lucas-Kanade 20 years on: A unifying framework," *International Journal of Computer Vision* **56**(3), 221–255 (2004).

[7] Meinhardt-Llopis, E., Pérez, J. S., and Kondermann, D., "Horn-Schunck optical flow with a multi-scale strategy," *Image Processing On Line* **3**, 151–172 (2013).

[8] Pérez, P., Gangnet, M., and Blake, A., "Poisson image editing," in [*ACM SIGGRAPH 2003 Papers*], 313–318 (2003).

[9] Samajdar, T. and Quraishi, M. I., "Analysis and evaluation of image quality metrics," in [*Information Systems Design and Intelligent Applications*], 369–378, Springer (2015).

[10] Qu, Z., Hu, G., Garfinkel, A., and Weiss, J. N., "Nonlinear and stochastic dynamics in the heart," *Physics Reports* **543**(2), 61–162 (2014).