StreoScenNet: surgical stereo robotic scene segmentation

Ahmed Mohammed, Sule Yildirim, Ivar Farup, Marius Pedersen, Øistein Hovde

Ahmed Mohammed, Sule Yildirim, Ivar Farup, Marius Pedersen, Øistein Hovde, "StreoScenNet: surgical stereo robotic scene segmentation," Proc. SPIE 10951, Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling, 109510P (8 March 2019); doi: 10.1117/12.2512518

SPIE.

Event: SPIE Medical Imaging, 2019, San Diego, California, United States

StreoScenNet: Surgical Stereo Robotic Scene segmentation

Ahmed Mohammed[a], Sule Yildirim[b], Ivar Farup[a], Marius Pedersen[a], and Øistein Hovde[c]

[a]Norwegian Colour and Visual Computing Lab, Norwegian University of Science and Technology, Norway.
[b]Norwegian Information Security Lab, Norwegian University of Science and Technology, Norway.
[c]Department of gastroenterology, Innlandet Hospital Trust, Gjøvik and Institute of Clinical Medicine, University of Oslo.

ABSTRACT

Surgical robot technology has revolutionized surgery toward a safer laparoscopic surgery and ideally been suited for surgeries requiring minimal invasiveness. Sematic segmentation from robot-assisted surgery videos is an essential task in many computer-assisted robotic surgical systems. Some of the applications include instrument detection, tracking and pose estimation. Usually, the left and right frames from the stereoscopic surgical instrument are used for semantic segmentation independently from each other. However, this approach is prone to poor segmentation since the stereo frames are not integrated for accurate estimation of the surgical scene. To cope with this problem, we proposed a multi encoder and single decoder convolutional neural network named StreoScenNet which exploits the left and right frames of the stereoscopic surgical system. The proposed architecture consists of multiple ResNet encoder blocks and a stacked convolutional decoder network connected with a novel sum-skip connection. The input to the network is a set of left and right frames and the output is a mask of the segmented regions for the left frame. It is trained end-to-end and the segmentation is achieved without the need of any pre- or post-processing. We compare the proposed architectures against state-of-the-art fully convolutional networks. We validate our methods using existing benchmark datasets that includes robotic instruments as well as anatomical objects and non-robotic surgical instruments. Compared with the previous instrument segmentation methods, our approach achieves a significant improved Dice similarity coefficient.

Keywords: Medical imaging, da Vinci Surgical System, Surgical instruments, Image segmentation, Computer vision, Deep learning

1. INTRODUCTION

The advent of robotics has increased the use of minimally invasive surgery. Advanced laparoscopic surgery is technically more demanding compared to open surgery.[1] The laparoscopic surgeon must view a distant monitor which provides 2-D vision, leading to a change in the normal hand-eye target axis.[2] Moreover, understanding these scenes from 2-D vision involves tracking and pose estimation of surgical instruments and anatomical objects. Therefore, robotic tool detection, segmentation, tracking and pose estimation are bound to become core technologies in a surgical work-flow in improving planning and understanding during the operation. In the context of delicate surgical procedures, such as urology,[3] it is paramount to provide the clinical operator with accurate real-time information about tool-tissue interactions,[4] 3D position and orientation of the instruments,[5] etc., to increase the context-awareness of the operator whilst performing robotic intervention and helping to avoid human errors.

Nowadays, instrument detection and tracking is done through electromagnetic, optical markers, and vision-based techniques. Some of the most commonly used surgical instruments are shown in Fig. (1). External electromagnetic[6] and optical markers based techniques require expensive tracking devices as well as modification to the surgical setup. Therefore, computer vision based approaches are getting more attention as they provide an

Further author information: (Send correspondence to Ahmed Mohammed)
E-mail:mohammed.kedir@ntnu.no,

Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling,
edited by Baowei Fei, Cristian A. Linte, Proc. of SPIE Vol. 10951, 109510P · © 2019 SPIE
CCC code: 1605-7422/19/$18 · doi: 10.1117/12.2512518

(a) Bipolar Forceps (b) Monopolar Curved Scissors (c) Grasping Retractor

(d) Large Needle Driver (e) Prograsp Forceps (f) Vessel Sealer

Figure 1: Commonly used instruments in robotic surgery: Each of these instruments are required for a different task during surgical procedures (Photos courtesy of Intuitive Surgical, Inc.).

alternative that can be realised entirely in software with no modification to the surgical setup. Surgical instrument tracking consists of detecting and identifying objects in video. In some applications, it is also desired to identify different components of the instrument such as shaft, wrist and clasper. There are a number of challenges that need to be addressed for successful tracking of surgical instruments. Endoscopic images typically contain specular reflections, surgical smoke and cluttered background, which causes detection as well as segmentation error.[7]

Vision-based techniques for robotic instrument detection and tracking have been developed for decades.[7–10] Earlier works often relied on using artificial fiducials on tool end effectors. Among the methods that rely on deformable part models (DPM), Kumar et al.[8] proposed a method for tool detection by hypothesizing surgical tool end effector to be the most distinguishable part of a tool and employing cascade object detection with DPM to learn the shape and localize the tool in images. In the last few years, neural networks based on convolutional neural network(CNN) have been producing superior results on various computer vision tasks. This trend has sparked the deep learning based approach for surgical instrument detection. In,[7,11] the authors propose to use an automatic method based on Fully Convolutional Networks (FCN) by replacing the fully connected (FC) layers with convolutions and adding deconvolution layers. Similarly, Shvets et al.[10] proposed using four different modifications of the U-Net[12] deep neural network architecture. However, these approaches are limited to single view (i.e. left or right video frames) tool detection and segmentation without considering the stereo vision system on robotic surgical instruments. Nonetheless, although some datasets are available containing both left and right view with ground truth for one of the views, to the best of our knowledge, there are no previous works that exploit both views of the stereoscopic surgical system tool detection and segmentation.

In this paper, we propose StreoScenNet, a novel encoder-decoder deep neural network that takes both left and right frames as input. Our proposed deep model is capable of detecting and segmenting six surgical instruments shown in Fig. (1) with their corresponding parts such as shaft, wrist and clasper. The StreoScenNet model aims to learn a decoder network from scratch while fine tuning encoder networks. Our method is conceptually simple, relying on the pretrained ResNet network[13] as encoder and a matched decoder network with the novel introduction of sum-skip-concatenation based connections to allow a much deeper network architecture for a more accurate segmentation. The key difference with the existing models is that we introduced an ensemble

of pre-trained encoder networks and a decoder network that uses a discriminative cost function to localize and detect the surgical instruments.

The structure of this paper is organized as follows: Section 2 describes StreoScenNet architecture and design. Section 3 presents the results of the experiments and we show that incorporating both left and right views improves the detection and segmentation accuracy. Finally in Section 4, we present further discussion and conclusions.

2. METHOD

Our proposed StreoScenNet is based on the deep learning model and is inspired by Y-Net,[14] a fully convolutional network. Y-Net uses a pre-trained and untrained VGG19 encoder[15] for a single image polyp detection. However, StreoScenNet explores stereoscopic input frames that uses multiple pre-trained ResNet encoder blocks. The framework (illustrated in Fig. (2)) consists of two 50 layer ResNet[13] encoder networks which are connected to a single decoder network. The main goal for having two encoders network with pre-trained weights is to address the performance loss due to single view and domain-shift from the pre-trained network (natural images) to testing (stereoscopic surgery images), leading to degradation in performance.[16] For example, a pre-trained model trained on natural images do not generalize well when applied to medical images.[16] It is assumed that fine-tuning a pre-trained network works the best when the source and the target tasks have a high degree of similarity. Therefore, our approach focuses on using the pre-trained model features optimally by fine-tuning the pre-trained encoder networks for a better generalization on the test set. In the next sub-sections, we describe each of the network components, and then the loss function used to train the network.

2.1 Network Architecture

The architecture of our model is shown in Fig. (2). It consists of two contracting paths on the left, i.e. encoders, and expanding path to the right, i.e. a decoder, that matches the input dimension. The decoder outputs a binary segmentation mask for each of the classes in a surgical scene.

Encoders: It follows the typical architecture of the ResNet50 network,[13] which has been widely used as the base network in many computer vision applications. These encoders use the pre-trained weights of ResNet50 trained on the ImageNet dataset. The last fully connected layer of the network that was trained on 1000 ImageNet classes is truncated. The usage of a pre-trained model makes training easier and generalizes better in that, the pre-trained model already has learned features that are relevant to our own classification problems such as edges, curves etc. The left and right stereo frames are given as an input for each of the encoders.

Decoder: The decoder network consists of five upsampling blocks and one final convolution block with a filter size of 1×1. Each upsampling block has the structure of upsampling-concatenation followed by three blocks of CONV-BatchNorm(BN)-RELU, except for the final layer which uses a 1×1 convolution with sigmoid for generating the final output mask. Compared with the other encoder-decoder architectures such as U-Net,[12] our decoder is different in: (1) The decoder is not architecturally symmetric with the encoders. (2) The decoder is much deeper than the encoder. This design is due to the fact that with the limited available training data, a deeper decoder network would learn features from each scale of the encoder inputs that are concatenated with the same-scale decoder layer.

Sum-skip concatenation: At each depth of encoder network, the final convolutional output of the left and right encoders network before max-pooling is summed together and skipped to the decoder network. This allows using both the left and right view information since the ground truth is provided for one of the views. Finally, the summed result is skipped and concatenated to the corresponding depth of the decoder network.[14]

2.2 Loss function

The output layer in the decoder consists of an eleven plane for each class of the segmented region. We applied convolution with a sigmoid activation to form the loss. Let p and g be the set of predicted and ground truth binary labels respectively. The weighted binary cross-entropy and dice coefficient loss between the two binary

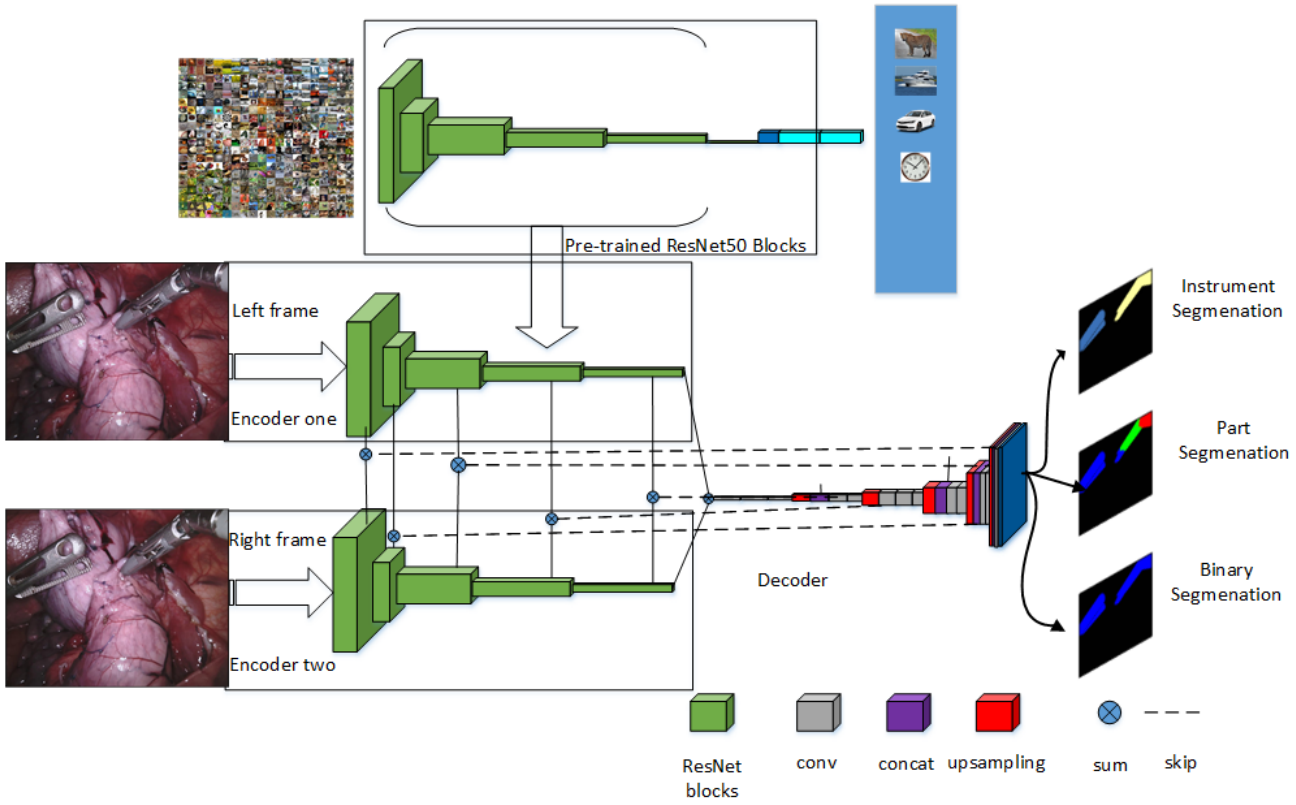


Figure 2: StreoScenNet: the proposed architecture. The top row shows the ResNet50 network pre-trained on the ImageNet dataset with 1000 classes. The weights of the ResNet blocks are transferred to both of the encoders. Given a stereo left and right image with the ground truth mask for the left frame, the network learns fusing the left and right frames for accurate scene segmentation. The output of the decoder is a mask for the instrument, part and binary segmentation tasks.

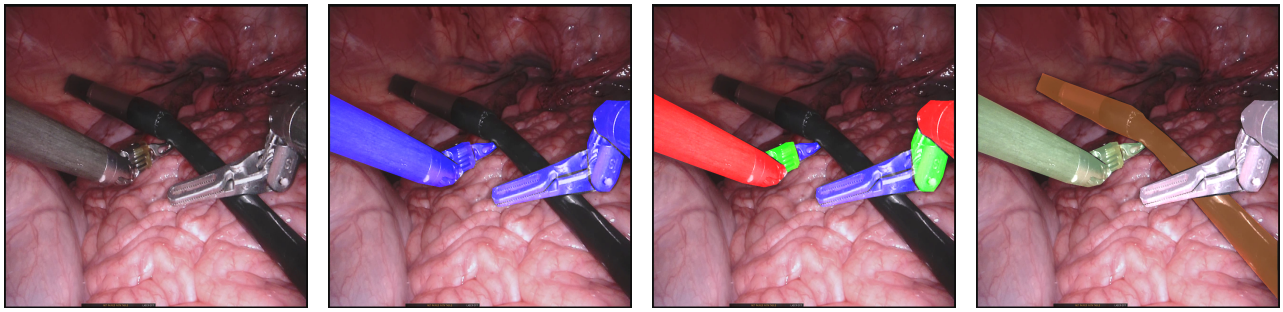
features maps is defined as:

$$\mathcal{L}(p,g) = -\frac{1}{N}\sum_{i=1}^{N}(\frac{\lambda}{2}\cdot g_i \cdot \log p_i) + (1 - \frac{2\sum_{i=1}^{N}(g_i \cdot p_i) + \epsilon}{\sum_{i=1}^{N}(p_i) + \sum_{i=1}^{N}(g_i) + \epsilon}) \tag{1}$$

where λ and ϵ are false negatives (FN) penalty and smoothing factor, respectively. N represents the total number of pixels in the image. In order to penalize FN more than false positives (FP) in training our network for highly imbalanced data, the first term in Eq. (1) penalizes FN and the second term weighs FPs and FNs (precision and recall) equally. In other words, the second term is the same as the negative of F1-score.

The final loss of the network is computed by summing the instrument, part and binary segmentation losses as shown in Eq. (2) :

$$\mathcal{L}(p,g) = \mathcal{L}_{instrument}(p,g) + \mathcal{L}_{part}(p,g) + \mathcal{L}_{binary}(p,g) \tag{2}$$



(a) Input Frame (b) Binary segmentation (c) Multi-label segmentation (d) Instrument recognition

Figure 3: Segmentation and detection problems in surgical procedures. **Best viewed in color**

3. EXPERIMENT AND RESULT

3.1 Dataset

Experiments are conducted on the open dataset from MICCAI 2017 Endoscopic Vision SubChallenge: Robotic Instrument Segmentation.[17] The dataset consists of six different robotic instruments with densely labeled surgical images shown in Fig. (3) and Fig. (1) with three different tasks: binary segmentation, multi-label segmentation and instrument recognition. Binary segmentation involves just separating the image into instruments and background, whereas multi-label segmentation requires the user to also recognize which parts of the instrument body correspond to the different articulated parts of a da Vinci robotic instrument. There are 8×225-frame robotic surgical videos, captured at 2 Hz, to avoid redundancy. The dataset contains left and right stereo views with 1920×1080 pixel resolution with ground truth labels provided for left frames only. The ground truth labels are encoded with numerical values "Background": 0, "Shaft": 10, "Wrist": 20, "Claspers": 30, and "Probe": 40 respectively, for each instrument shown in Fig. (1).

3.2 Implementation Details

Our model is implemented on the Tensorflow and Keras library with a single NVIDIA 12GB Titan X GPU. We first apply data augmentation and resize all images into fixed dimensions with spatial size of 224×224 before feeding to both encoders and finally normalized to [0, 1]. We use RMSProp as the optimizer with a batch size of 10 and the learning rate set to 0.0001. We monitor the dice coefficient and use the early-stop criteria on the validation set error. The network output is an eleven channel mask for seven surgical instruments, three parts and one binary segmentation respectively.

3.3 Baseline Method and Evaluation Metrics

Under the terms of MICCAI 2017 Endoscopic Vision SubChallenge, ground truth data for test dataset is kept with the challenge organizers. Hence, we evaluate the proposed method using a 4-fold cross validation. In such a case, we try to make every fold to contain more or less equal number of instruments. The validation setup is summarized in Table 1.

Table 1: 4-fold validation

Experiments	Training videos	Testing videos
Exp1	(2, 4, 5, 6, 7, 8)	(1, 3)
Exp2	(1, 3, 4, 6, 7, 8)	(2, 5)
Exp3	(1, 2, 3, 5, 6, 7)	(4, 8)
Exp4	(1, 2, 3, 4, 5, 8)	(6, 7)

As a baseline architecture for comparison, we employ different variations of state-of-the-art fully convolutional network, U-Net.[10,12] This network is chosen as a natural baseline for comparison as it represents the state-of-the-art convolutional architecture for robotic surgical tool segmentation.[10]

We evaluate the performance of our network using the mean Intersection over Union (IoU) and Dice score which is similar to F1-Score. IoU is also known as Jaccard index and it is a standard metric commonly used for evaluating segmentation accuracy. The IoU and Dice scores are calculated based on the region overlap between the predicted mask p and the ground truth mask g as follows:

$$IoU = \frac{1}{N}\sum_{i=1}^{N}\frac{p_i q_i}{p_i + q_i - p_i q_i} \tag{3}$$

$$Dice = \frac{2}{N}\frac{\sum_{i=1}^{N} p_i q_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} q_i} \tag{4}$$

where N is the total number of pixels in the image.

3.4 Result

The quantitative and qualitative evaluations examining the performance of StreoScenNet for a binary, instrument detection and segmentation are presented in Table 2 and Fig. (4). Table 2 shows average Dice score and IoU values for a 4-fold validation on test videos as in Table 1. Model1 and Model2 are the variations of the U-net architecture with VGG16 and VGG19 pre-trained encoder.[15] While LinkNet-34[18] model uses an encoder based on a ResNet-type architecture.[6] For instrument segmentation, our model gives an improvement of 10.76 and 10.39 percentage points in mean IoU and mean Dice score respectively (see Table 2). For instrument part recognition, our model gives state-of-the-art result with 1.6 percentage point Dice score improvement with the best reported score in the literature.[10] However, for binary segmentation, the result is slightly less than the state-of-the-art result due to low class performance for ultra-sound probe as it is not included in binary segmentation and looks similar to other parts. The qualitative results in Fig. (4) show how our proposed architecture is able to differentiate different surgical instruments and components.

It is important to note that each of the above models LinkNet, U-net, Model1 and Model2 are trained for each task separately. The final score of each models are computed by taking the ensemble of the three models for each task. However, StreoScenNet is a single model trained for all tasks in an end to end fashion. Hence, it is more efficient for surgical application with the da Vinci platform as it outputs binary, instrument, and part segmentation and detection per prediction.

Table 2: Average IoU and Dice coefficient on 4-fold validation

	Instrument recognition		Multi-label segmentation		Binary segmentation	
Model	IoU	Dice	IoU	DICE	IoU	Dice
U-net[12]	15.80	23.59	48.41	60.75	75.44	84.37
Model1[10]	34.61	45.86	62.23	74.25	81.14	88.07
Model2[10]	33.78	44.95	65.50	75.97	**83.60**	**90.01**
LinkNet-34[18]	22.47	24.71	34.55	41.26	82.36	88.87
StreoScenNet(ours)	**45.37**	**56.25**	**66.23**	**77.57**	80.82	87.86

4. CONCLUSIONS

In this paper, we address surgical stereo robotic scene segmentation problem by proposing a new deep encoder decoder approach. To the best of our knowledge this is the first work that incorporates stereo-information

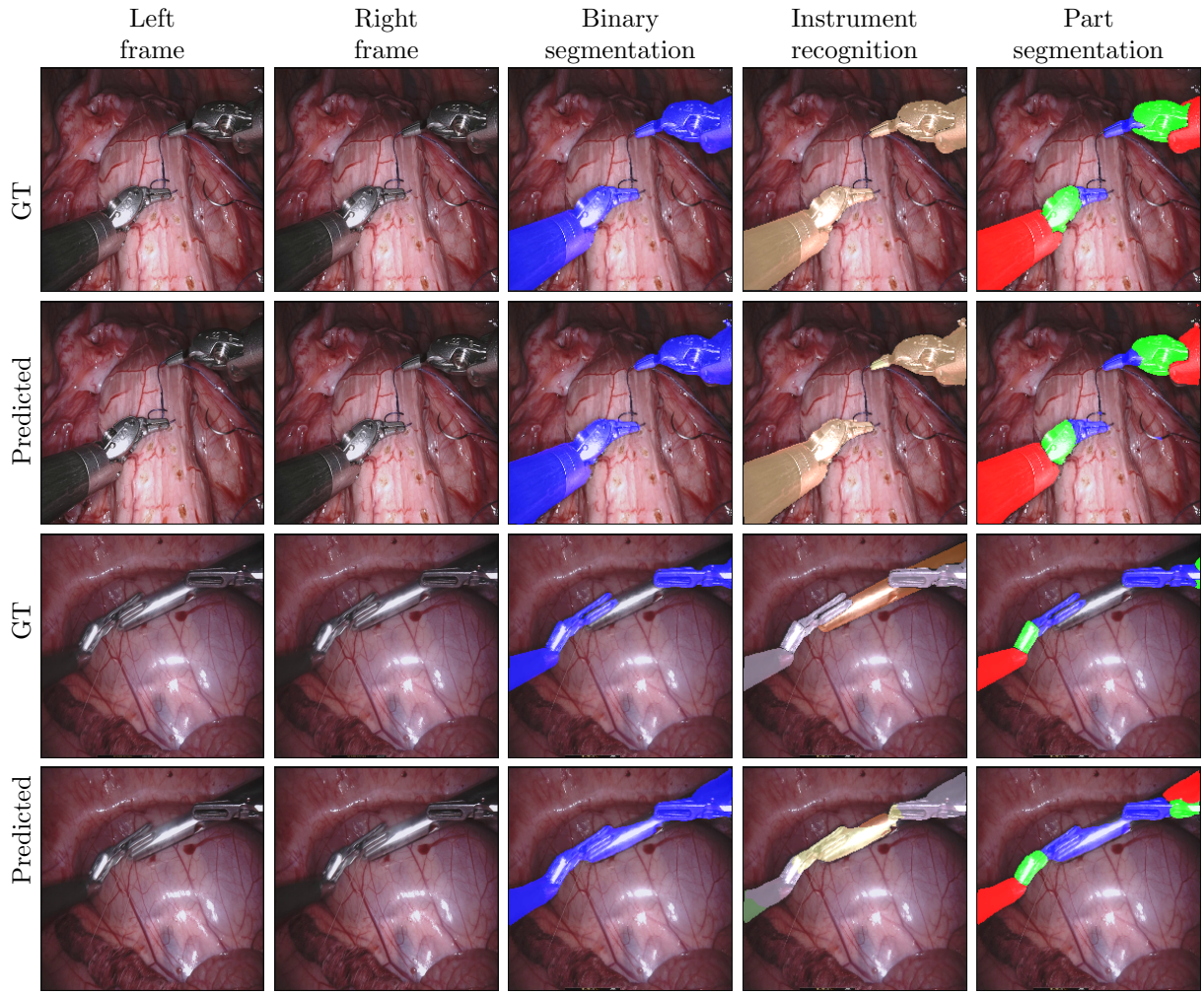


Figure 4: Sample visual result of our proposed method. GT is the ground truth and predicted shows our model output. The top row shows easier sequences and the last row shows more challenging sequence. As it can be seen, our approach gives accurate segmentation for most of the surgical scene. It is also important to note that as mentioned under implementation detail, the output of the network 224 x 224 is up sampled to match the ground truth input dimension of 1280 x 1024 which results in loss of resolution. **Best viewed in color**

from left and right frames for accurate robotic scene segmentation. The proposed architecture relies on the pre-trained ResNet-50 architecture with a novel sum-skip connection. The experimental results show that the proposed StreoScenNet gives promising results for robotic scene segmentation. Further improvements could be achieved in the future by focusing on the following investigations: pre-processing of the stereo laparoscopic images and evaluation of the robustness with an extended test and training dataset.

ACKNOWLEDGMENTS

This research has been supported by the Research Council of Norway through project no. 247689 "IQ-MED: Image Quality enhancement in MEDical diagnosis, monitoring and treatment".

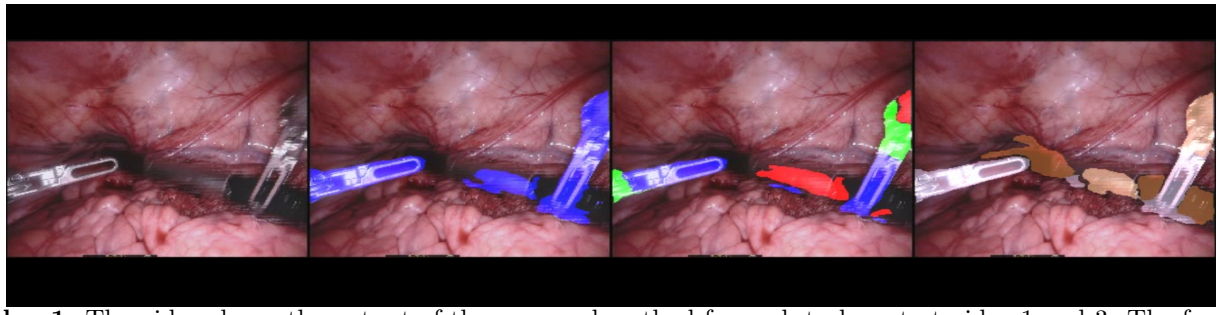
REFERENCES

[1] Palep, J. H., "Robotic assisted minimally invasive surgery," *Journal of Minimal Access Surgery* **5**(1), 1 (2009).

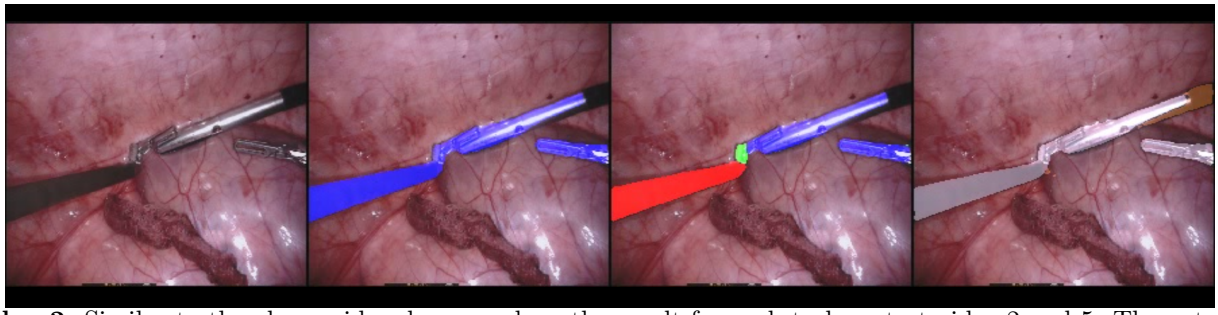
[2] Sackier, J. M. and Wang, Y., "Robotically assisted laparoscopic surgery," *Surgical endoscopy* **8**(1), 63–66 (1994).
[3] Parekattil, S. J. and Moran, M. E., "Robotic instrumentation: evolution and microsurgical applications," *Indian journal of urology: IJU: journal of the Urological Society of India* **26**(3), 395–403 (2010).
[4] Westebring-van der Putten, E. P., Goossens, R. H., Jakimowicz, J. J., and Dankelman, J., "Haptics in minimally invasive surgery–a review," *Minimally Invasive Therapy & Allied Technologies* **17**(1), 3–16 (2008).
[5] Allan, M., Ourselin, S., Hawkes, D. J., Kelly, J. D., and Stoyanov, D., "3-d pose estimation of articulated instruments in robotic minimally invasive surgery," *IEEE transactions on medical imaging* **37**(5), 1204–1213 (2018).
[6] Hu, C., Meng, M. Q.-H., Song, S., and Dai, H., "A six-dimensional magnetic localization algorithm for a rectangular magnet objective based on a particle swarm optimizer," *IEEE Transactions on Magnetics* **45**(8), 3092–3099 (2009).
[7] García-Peraza-Herrera, L. C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., and Ourselin, S., "Real-time segmentation of non-rigid surgical tools based on deep learning and tracking," in [*International Workshop on Computer-Assisted and Robotic Endoscopy*], 84–95, Springer (2016).
[8] Kumar, S., Narayanan, M. S., Singhal, P., Corso, J. J., and Krovi, V., "Product of tracking experts for visual tracking of surgical tools," in [*Automation Science and Engineering (CASE), 2013 IEEE International Conference on*], 480–485, IEEE (2013).
[9] Wang, C., Palomar, R., and Cheikh, F. A., "Stereo video analysis for instrument tracking in image-guided surgery," in [*Visual Information Processing (EUVIP), 2014 5th European Workshop on*], 1–6, IEEE (2014).
[10] Shvets, A., Rakhlin, A., Kalinin, A. A., and Iglovikov, V., "Automatic instrument segmentation in robot-assisted surgery using deep learning," *arXiv preprint arXiv:1803.01207* (2018).
[11] García-Peraza-Herrera, L. C., Li, W., Fidon, L., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., et al., "Toolnet: holistically-nested real-time segmentation of robotic surgical tools," in [*Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*], 5717–5722, IEEE (2017).
[12] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical image computing and computer-assisted intervention*], 234–241, Springer (2015).
[13] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).
[14] Mohammed, A., Yildirim, S., Farup, I., Pedersen, M., and Hovde, Ø., "Y-net: A deep convolutional neural network for polyp detection," *arXiv preprint arXiv:1806.01907* (2018).
[15] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).
[16] Kornblith, S., Shlens, J., and Le, Q. V., "Do better imagenet models transfer better?," *arXiv preprint arXiv:1805.08974* (2018).
[17] "Endovissub2017- robotic instrument segmentation - home." https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/Home/. (Accessed on 01/07/2018).
[18] Chaurasia, A. and Culurciello, E., "Linknet: Exploiting encoder representations for efficient semantic segmentation," in [*International Conference on Visual Communications and Image Processing (VCIP), 2017 IEEE*], 1–4, IEEE (2017).

5. APPENDIX

Additional results Here we show some additional result of our model trained using Exp1 and Exp2 settings.



Video 1: The video shows the output of the proposed method for each task on test video 1 and 3. The frames are resized to 224 x 224. http://dx.doi.org/10.1117/12.2512518.1



Video 2: Similar to the above video, here we show the result for each task on test video 2 and 5. The network is trained by excluding video 2 and 5 from the training set. http://dx.doi.org/10.1117/12.2512518.2