2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
Banff Center, Banff, Canada, October 5-8, 2017

Surgical Tool Segmentation Using A Hybrid Deep CNN-RNN Auto Encoder-Decoder

Mohamed Attia*, Mohammed Hossny*, Saeid Nahavandi* and Hamed Asadi†
*Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University
†School of Medicine, Deakin University

***Abstract*—Surgical tool segmentation is used for detection, tracking and pose estimation of the tools in the vicinity of surgical scenes. It is considered as an essential task in surgical phase recognition and flow identification. Surgical flow identification is an unresolved task in the domain of context-aware surgical systems, which is used extensively on computer assisted intervention (CAI). CAI is used for staff assignment, automated guidance during intervention, surgical alert systems, automatic indexing of surgical video databases and optimisation of the real-time scheduling of operating room. Semantic segmentation is used for accurate delineation of surgical tools from the background. In semantic segmentation, each label is assigned to a class as a tool or a background. In this presented work, we applied a hybrid method utilising both recurrent and convolutional networks to achieve higher accuracy of surgical tools segmentation. The proposed method is trained and tested using a public dataset MICCAI 2016 Endoscopic Vision Challenge Robotic Instruments dataset "EndoVis". We achieved better performance using the proposed method compared to state-of-the-art methods on the same dataset for benchmarking. We achieved a balanced accuracy of 93.3% and Jaccard index of 82.7%.**

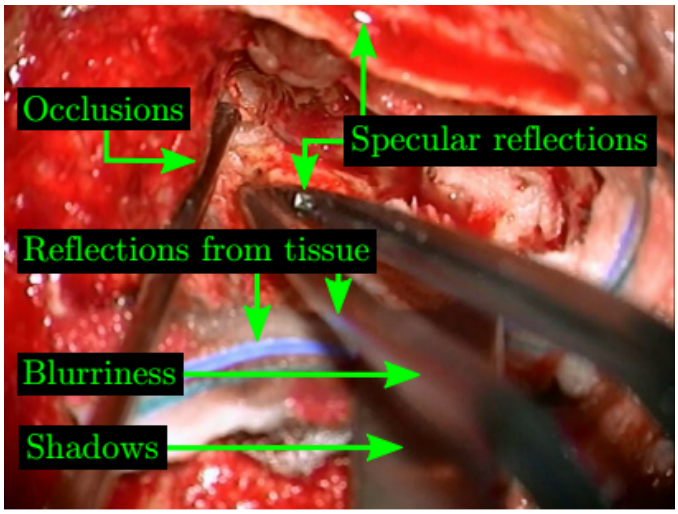

Fig. 1: Surgical tools endoscopic images artifacts, which include occlusion, shadows, reflections and blurriness [9].

I. Introduction

Minimally invasive surgery "MIS" is one of the prominent procedures that enhances the surgery outcomes in terms of operation success and reduction of recovery time. MIS improves the control of the surgeons on articulated instruments over the tissue or anatomy under study. The surgical camera used in MIS is essential for self-localization of the instruments used by surgeons.

However, the surgical camera suffers from reduced field of view. Surgical tool segmentation can help to provide the surgeon with essential information during complex procedures. On the other side, the segmentation task requires accurate identification of spatial relationships between the surgical camera, operating instruments and patient anatomy [22].

Surgical tools segmentation requires accurate delineation of tools in the presence of visual occlusions like smoke and blood. Also, endoscopic images usually suffer from occlusion, shadows, reflections and blurriness, as shown in Fig. 1. These artefacts affect the segmentation process and degrade the quality of predicted masks. Therefore, Surgical tool segmentation task is considered as a challenging task.

Tonet *et al.* proposed one of the earliest attempts for tracking by modifying the visual appearance of the tools to facilitate the tracking task. However, this method had negative impact on sterilization of the tools [23]. Also, changing the outer appearance of the tools requires special setup. Special setups limit the applicability on pre-existing surgical setups or even recordings. Therefore, semantic segmentation provides an accurate and robust solution for the segmentation problem. It is used for pixel-level labelling by assigning a category or class label for every pixel in the image. This can provide accurate identification of the different parts of the surgical tool in addition to delineation of the tool from the surrounding tissue [27].

In this paper, we evaluated our network on publicly accessible annotated surgical tool benchmark dataset. The challenge was hosted by MICCAI 2016 under the name: Endoscopic Vision Challenge Robotic Instruments dataset "EndoVis" [8].

978-1-5386-1645-1/17/$31.00 ©2017 IEEE

This challenge has two tasks: surgical tools segmentation and tracking. We compared our results to other proposed architectures in literature using same reported dataset and same metrics.

The rest of this paper is organized as follows. Section II describes the efforts documented in the literature for state-of-the-art methods. In Section III, a brief description of the proposed methodology is elaborated. In Section IV, results of the conducted experiments are discussed. In Section V, conclusions are drawn from the implementation of the proposed architecture.

II. Related Work

Machine learning approaches were proposed to solve object detection, image classification and semantic segmentation problems. Deep learning is a set of machine learning algorithms that acts on the input data in the form of cascaded layers. Each layer is a non-inear processing unit that employed to extract hierarchical representation for the self-extract data. Very deep convolutional neural network (CNNs) with auto encoder-decoder architectures were utilized for semantic object segmentation and they improved the segmentation task performance [14].

The auto encoder parts are responsible for extraction of deep encoded features using self-learning paradigm. The decoder is required for upsampling of deeply encoded feature maps. The upsampling process is essential for mapping extracted features and segmentation mask, using reconstruction module. Eigen *et al.* utilized a deep auto encoder-decoder architecture that extracts the feature on two scales: coarse and fine predictions. Coarse scale predictions were refined using finer scales to achieve higher accuracy [7]. However, the output prediction maps suffer from lost spatial resolution due to successive unpooling and deconvolutions [14], [15].

More articulated methods addressed error in segmentation due to lost spatial resolution. Ronneberger *et al.* concatenated a copy of encoded feature map during decoding phase to increase spatial accuracy [18]. Conditional random fields were proposed as a post-processing module to refine the coarse predictions. It is based on construction of a model for local and global dependencies. Zheng *et al.* proposed a trainable conditional random fields (CRFs) module to refine segmentation prediction map jointly during training [28]. Chen *et al.* utilized the extracted coarse feature maps and fed them into conditional random fields module as unary potential for image pixel to smooth noisy segmentation masks [4].

Visin *et al.* proposed a recurrent neural network (RNN) as post processing module for the coarse extracted feature maps [24]. Attia *et al.* adopted the utilization of hybrid architecture for joint training of recurrent and convolutional networks for medical image segmentation. This architecture outperformed state-of-the-art architectures in melanoma image segmentation challenge [1].

Traditionally, all deep learning architectures require preprocessing of input images to refrain the saturation of feature maps. Ioffe *et al.* proposed batch normalization techniques to avoid saturation problem [12]. Thus, Badrinarayanan *et al.* modified a fully convolutional networks with batch normalization to avoid preprocessing steps [2].

The semantic segmentation is used for accurate classification of each pixel within the image context into two or more classes [13], [20]. This method successfully solved many computer vision tasks by accurately semantically segment without any prior assumptions about the image from different sources on pixel-wise level [10], [11], [19]

The semantic segmentation task of surgical tools relies on assigning each pixel to a label based on the context of image and surrounding pixels. Complex segmentation is treated as classification problem of binary data, either background or foreground. The discriminative binary classifier is used to assign each pixel into either background or foreground. Binary classifiers can be Random Forests [3], maximum likelihood Gaussian Mixture Models [17] and Naive Bayesian classifiers [21].

However, these aforementioned methods require an extensive feature selection process for accurate classification of the surgical tool pixels. To overcome this problem, deep learning is utilized to solve surgical tool segmentation problem. Deep learning offers end-to-end learning mechanism that exploited great potentials through self-learning paradigm [20]. Garca-Peraza-Herrera *et al.* applied fully convolutional networks in the surgical tools segmentation domain [9]. However, the results lack the required accuracy for this task.

Pakhomov *et al.* proposed an architecture based on deep residual network "ResNet-101" using dilated convolution with stride instead of the average pooling used in the original "ResNet-101" architecture [16]. All these methodologies were based on fully convolutional networks. Fully convolutional networks generate significantly reduced dimension prediction maps (in case of VGG16 and ResNets, the predicted feature maps are reduced by a factor of 32 compared to original input image size). Adding a deconvolutional layer to learn the upsampling restores the reduced dimensions but the segmentation boundaries are course [16].

To solve this problem, two methods were recently adopted. The first approach is fusing features from layers in the encoder part to the decoder part to smooth the predictions [20]. The second approach s mitigating the downsampling of the feature maps by removal of some of the pooling layers in the encoder part of the fully convolutional network [4], [25]. In the

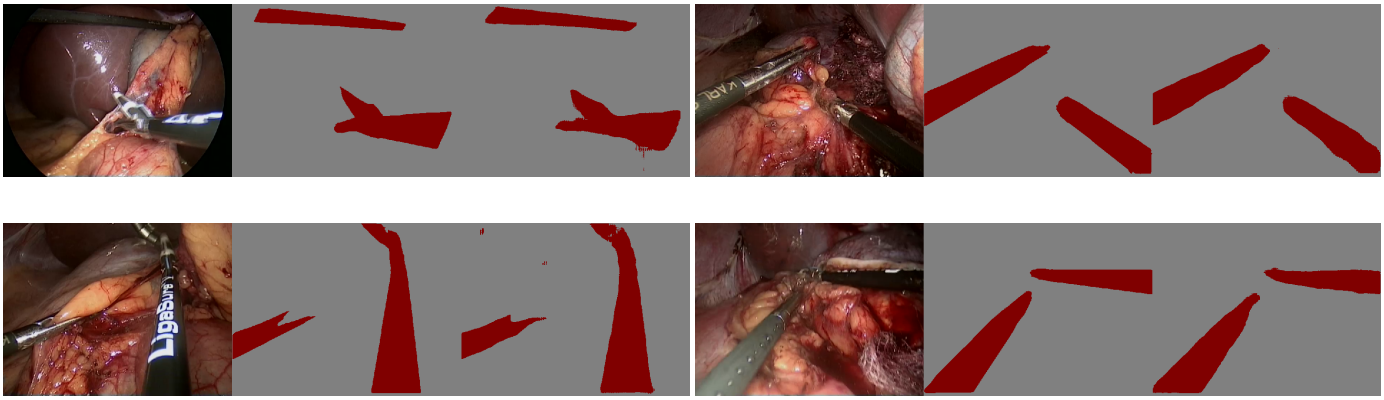

Fig. 2: Example of Surgical tools segmentations. From left to right: image, ground truth and proposed method.

presented work, we proposed using hybrid architecture that exploits the advantages of both recurrent and convolutional networks.

III. Proposed Method

The aim of the presented work is to semantically label each pixel in endoscopic images to one of the semantic classes. The semantic classes are background, surgical tools shaft and surgical tools manipulator. Most of semantic segmentation networks were based on fully convolutional networks such as FCN [14] and SegNet [2]. However, output of these networks were coarse segmentation masks [16]. To overcome this problem, a recurrent neural network (RNN) is trained to model contextual relationships between pixels. RNNs are used to preserve local and global contextual dependencies even over large proximal distances. Long Short Term Memory (LSTM) are the basic building module for RNNs. LSTM is introduced to learn these spatial dependencies between neighbouring pixels. LSTM has four different schemes to scan images in all directions (up to down, down to up, left to right and right to left). Each two coupled directions exhibit parallelization capabilities since they are working independently.

The proposed architecture incorporates a convolutional auto encoder to extract unique learned features of the pixels. The encoder part consists of seven convolutional layers and two max-pooling layers. The convolutional layers are used to encode the extracted feature by using hierarchical representation of the input images. Pooling layers are used for dimension reduction for deeper representation of the input. Four layers of recurrent neural network are employed to find local/global dependencies between pixels in coupled directions, as shown in Fig. 3. The deconvolution network is used to construct the segmentation mask [24].

Recurrent neural layers (RNNs) are utilized to process a sequence of input data to extract feature or produce another sequence. In this presented work, recurrent neural networks are used to process input image data as flattened non-overlapping patches of deep encoded feature maps to model spatial dependencies between them. Let I be the input data such that $I \in \mathbb{R}^{w \times h \times c}$ where w, h and c are width, height and channels respectively. D is splitted into $n \times m$ patches $P_{x,y}$ such that $P_{x,y} \in \mathbb{R}^{w_p \times h_p \times c}$ where $w_p = w/n$ and $h_p = h/m$. Input patches are flattened as 1-D vector to update its hidden state $z^*_{x,y}$ where $*$ is the direction of the sweep direction $\uparrow, \downarrow, \rightarrow$ and $\leftarrow$.

For every patch $P_{x,y}$, the composite activation map feature $O = \{o^*_{x,y}\}^{\{y=1,2,\ldots,m\}}_{\{x=1,2,\ldots,n\}}$ is concatenation of output activation two coupled direction RNN either horizontal (right to left and left to right) or vertical sweep (up to down and down to up) where $o^*_{x,y} \in \mathbb{R}^{2U} \forall * \in \{(\uparrow,\downarrow),(\rightarrow,\leftarrow)\}$ is activation of the recurrent unit at position (x, y) with respect to all patches in the column y in case of coupled vertical direction $\{(\downarrow,\uparrow)\}$ and to all patches in the row i in case of coupled horizontal sweep $\{(\rightarrow,\leftarrow)\}$ and $O^{\updownarrow}$ denotes concatenated output of $o^{\downarrow}$ and $o^{\uparrow}$ and similarly $O^{\leftrightarrow}$ for $O^{\leftarrow}$ and $O^{\rightarrow}$ and U is the number of recurrent units.

$$o^{\downarrow}_{x,y} = f^{\downarrow}(z^{\downarrow}_{x-1,y}, p_{x,y}) \, \forall x = 1, 2, \ldots, n \quad (1)$$

$$o^{\uparrow}_{x,yj} = f^{\uparrow}(z^{\uparrow}_{x+1,y}, p_{x,y}) \, \forall x = n, n-1, \ldots, 1 \quad (2)$$

$$o^{\rightarrow}_{x,y} = f^{\rightarrow}(z^{\rightarrow}_{x,y-1}, p_{x,y}) \, \forall y = 1, 2, \ldots, m \quad (3)$$

$$o^{\leftarrow}_{x,y} = f^{\leftarrow}(z^{\leftarrow}_{x,y+1}, p_{x,y}) \, \forall = ym, m-1, \ldots, 1 \quad (4)$$

where $o^{\downarrow}_{x,y}$for vertical sweep Gated Feedback Recurrent unit

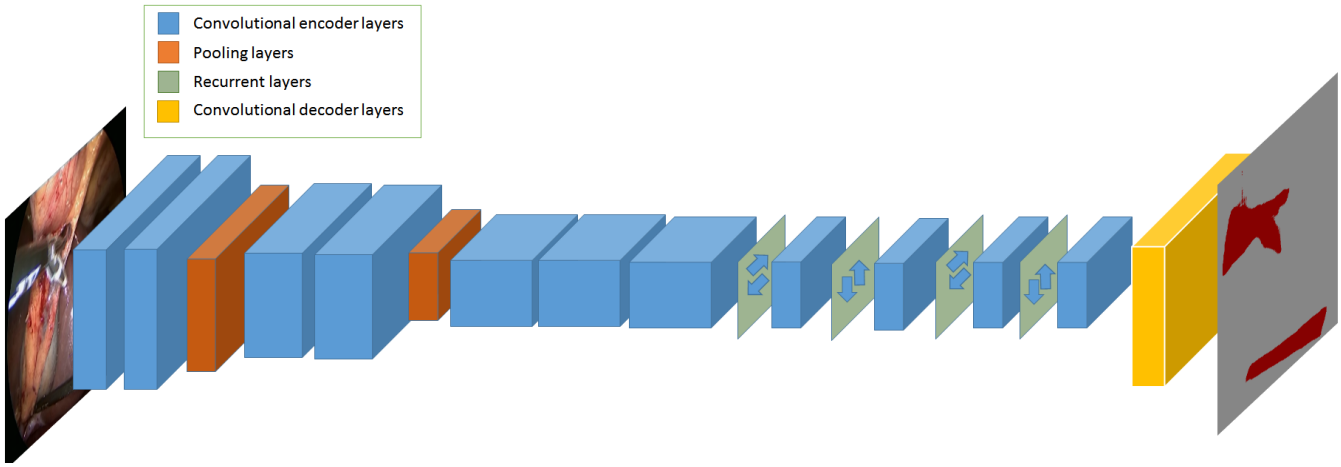

Fig. 3: Proposed architecture for RNN and CNN. Auto encoder network consists of 7-convolutional layers "blue blocks" with 2 max-pooling layers "orange blocks". Then, extracted feature maps are fed into 4 layers of recurrent network "green mask" with 4 decoupled direction. The mask is reconstructed using auto decoder network "yellow blocks".

$f^{\downarrow}(z^{\downarrow}_{x-1,y}, p_{x,y})$ is defined as follows [5]:

$$o^{\downarrow}_{x,y} = \sigma(W_o * p_{x,y} + U_o * z^{\downarrow}_{x-1,y}) \quad (5)$$

$$z^{\downarrow}_{x,y} = o^{\downarrow}_{x,y} * \tanh(C_{x,y}) \quad (6)$$

$$C_{x,y} = f_{x,y} * C_{x-1,y} + A_{x,y} * \hat{C}_{x,y} \quad (7)$$

$$\hat{C}_{x,y} = \tanh(W_c * p_{x,y} + U_c * z^{\downarrow}_{x-1,y}) \quad (8)$$

$$h_{x,y} = \sigma(W_f * p_{x,y} + U_f * z^{\downarrow}_{x-1,y}) \quad (9)$$

$$A_{x,y} = \sigma(W_A * p_{x,y} + U_A * z^{\downarrow}_{x-1,y}) \quad (10)$$

Similarly, $o^{\downarrow}_{x,y}$ and coupled horizontal sweep function can be defined. It is worth noting that both directions are computed independently.

In the presented work, The convolutional neural network layers are used to extract deep encoded feature maps. while, recurrent neural networks are used to model spatial dependencies between these deep encoded feature maps, as shown in Fig. 3. Then, the deep encoded feature maps are used to reconstruct segmentation mask at the same resolution of the input. Fractionally strided convolutions were used in reconstruction of final output. In strided convolutions, predictions are calculated by inner product between the flattened input and a sparse matrix, whose non-zero elements are elements of the convolutional kernel. This method is computationally and memory efficient method to support joint training of convolutional and recurrent neural networks [6].

IV. EXPERIMENTS AND RESULTS

A. Dataset and Evaluation

The proposed architecture was trained using annotated surgical tools videos. These images were provided for the MICCAI 2016 challenge "EndoVis". This dataset is divided into two sub- datasets, robotic and non-robotic. We utilized the same sub-dataset adopted in [16] for evaluation purposes. The robotic sub-dataset for training consists of four 45-seconds videos and the test data is four 15-second and two 60-seconds videos. These videos has a resolution of 720×576 with frame rate 25 fps [8].

The performance of the proposed method is compared to other methods using same dataset and same metrics on the pixel-level, as defined in the challenge: "Balanced" accuracy, as shown in Equ. 14, sensitivity, as shown in Equ. 12 and specificity, as shown in Equ. 13. Also, we reported the Intersection-over-Union metric "IoU", as shown in Equ. 11. IoU is considered one of the best segmentation metric because it penalizes both over and under segmentation.

$$IoU = \frac{TP}{TP + FP + FN} \quad (11)$$

$$SE = \frac{TP}{TP + FN} \quad (12)$$

$$SP = \frac{TN}{TN + FP} \quad (13)$$

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

"TP" is true positive, which represents number of correctly labelled pixels as foreground. "TP" is true negative, which represents correctly labelled pixels as background. "FP" is false positive, which represents mistakenly labelled background pixels as foreground. "FN" is false positive, which represents mistakenly labelled foreground pixels as background.

B. Training

The proposed architecture is trained using annotated surgical tool images from MICCAI 2015 Endoscopic vision "EndoVis" sub-challenge "Instrument segmentation and tracking" [8]. The deep convolutional and recurrent networks were optimized with softmax cross entropy objective for labels of annotations file [14]. The adaDelta optimization algorithm was used for learning process [26].

C. Results

The proposed architecture outperformed other architectures, as shown in Table I. We achieved balanced accuracy of 93.3% in comparison to 92.3% for ResNet-101 based architecture [16]. Also, sensitivity for proposed method was 90.4% compared to 87.8% for FCN-8 [16]. We also had better results for IoU, 82.7% compared to 77.6% for ResNet-101 based architecture [16]. However, precision metric for ResNet-101 based architecture was better due to increased false positive in the hybrid architecture. The proposed method provided high visual accuracy with robust performance, as shown in Fig. 2. In each sub-figure of Fig. 2, the image on the left in is the input endoscopic image, in the middle is the ground truth for the segmentation mask and on the right is the output of the proposed architecture.

It is worth noting that fully convlutional neural networks such as FCN [14] suffer from over-segmentation. Over-segmentation affects the accuracy and jaccard index "IoU". The main cause for over-segmentation is the deficiency of spatial information and dependencies between adjacent image patches fed into the convolutional layers during the deep encoding of the extracted features. Recurrent layers were proposed to address this problem by derivation of the spatial neighbourhood relations between patches before the decoding part of the deep architecture. The Fractionally strided convolutions in the decoder produced smooth masks with better accuracy compared to the predicted masks by simple bilinear interpolation of the feature maps to restore the original size in the presented method in [16].

V. CONCLUSION

We utilized a hybrid architecture that incorporates both deep convolutional and recurrent neural networks for surgical tools segmentation. The evaluation results of the proposed architecture demonstrated superior performance by out performing state-of-the-art methods of surgical tool segmentation. This architecture does not require preprocessing of the input or dimension reduction by grey scale conversion. Also, it is immune, for some extent, to artefacts such as smoke and blood. The utilization of Fractionally strided convolutions in the decoder produced smooth segmentation masks. According to the reported results, the hybrid method that incorporate utilisation of RNN and CNN outperformed methods that rely on CNN and deep residual networks only. Also, it made the segmentation task more robust and consistent.

ACKNOWLEDGMENT

This research was fully supported by the Institute for Intelligent Systems Research and Innovation (IISRI) at Deakin University.

REFERENCES

[1] Mohamed Attia, Mohamed Hossny, Saeid Nahavandi, and Anousha Yazdabadi. Skin melanoma segmentation using recurrent and convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[2] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*, 2015.

[3] David Bouget, Rodrigo Benenson, Mohamed Omran, Laurent Riffaud, Bernt Schiele, and Pierre Jannin. Detecting surgical tools by modelling local appearance and global shape. *IEEE transactions on medical imaging*, 34(12):2603–2617, 2015.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[5] Junyoung Chung, Caglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Gated feedback recurrent neural networks. *CoRR, abs/1502.02367*, 2015.

[6] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[7] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[8] MICCAI 2015 EndoVis. Endovis dataset, April 2017.

[9] Luis Garcia Peraza Herrera, Wenqi Li, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, and Sébastien Ourselin. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In *Lecture Notes in Computer Science*. Springer Verlag (Germany), 2016.

[10] H. Haggag, A. Abobakr, M. Hossny, and S. Nahavandi. Semantic body parts segmentation for quadrupedal animals. *IEEE Conference on Systems, Man, and Cybernetics (SMC)*, 2016.

	Sensitivity	Specificity	"Balanced" Accuracy	IoU
FCN-8 [9]	87.8%	88.7%	88.3%	75.2%
ResNet-101 [16]	85.7%	**98.8%**	92.3%	77.6%
Proposed	**90.4%**	96.1%	**93.3%**	**82.7%**

TABLE I: Surgical tool segmentation. Average Accuracy and Jaccard index comparison results. Higher results are better.

[11] H Haggag, M Hossny, S Nahavandi, S Haggag, and D Creighton. Body parts segmentation with attached props using rgb-d imaging. In *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, pages 1–8. IEEE, 2015.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[13] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proc. NAACL*, 2015.

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[15] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[16] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. Deep residual learning for instrument segmentation in robotic surgery. *arXiv preprint arXiv:1703.08580*, 2017.

[17] Zachary Pezzementi, Sandrine Voros, and Gregory D Hager. Articulated object tracking by rendering consistent appearance parts. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3940–3947. IEEE, 2009.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[19] K. Saleh, M. Hossny, and S. Nahavandi. Kangaroo vehicle collision detection using deep semantic segmentation convolutional neural network. *IEEE Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2016.

[20] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):640–651, 2017.

[21] Stefanie Speidel, Michael Delles, Carsten Gutt, and Rüdiger Dillmann. Tracking of instruments in minimally invasive surgery for surgical skill analysis. In *International Workshop on Medical Imaging and Virtual Reality*, pages 148–155. Springer, 2006.

[22] Russell H Taylor, Arianna Menciassi, Gabor Fichtinger, and Paolo Dario. Medical robotics and computer-integrated surgery. In *Springer handbook of robotics*, pages 1199–1222. Springer, 2008.

[23] Oliver Tonet, TU Ramesh, Giuseppe Megali, and Paolo Dario. Tracking endoscopic instruments without localizer: image analysis-based approach. *Studies in health technology and informatics*, 119:544–549, 2005.

[24] Francesco Visin, Kyle Kastner, Aaron Courville, Yoshua Bengio, Matteo Matteucci, and Kyunghyun Cho. Reseg: A recurrent neural network for object segmentation. *arXiv preprint arXiv:1511.07053*, 2015.

[25] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[26] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[27] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, pages 1–17, 2017.

[28] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.