Segmentation of surgical instruments in laparoscopic videos: training dataset generation and deep-learning-based framework

Lee, Eung-Joo, Plishker, William, Liu, Xinyang, Kane, Timothy, Bhattacharyya, Shuvra, et al.

Eung-Joo Lee, William Plishker, Xinyang Liu, Timothy Kane, Shuvra S. Bhattacharyya, Raj Shekhar, "Segmentation of surgical instruments in laparoscopic videos: training dataset generation and deep-learning-based framework," Proc. SPIE 10951, Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling, 109511T (8 March 2019); doi: 10.1117/12.2512994

SPIE.

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

Segmentation of Surgical Instruments in Laparoscopic Videos: Training Dataset Generation and Deep Learning-based Framework

Eung-Joo Lee[a,b], William Plishker[b],Xinyang Liu[c], Timothy Kane[c], Shuvra S. Bhattacharyya[a], Raj Shekhar[b,c]

[a]The Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA; [b]IGI Technologies, Inc., College Park, MD, USA; [c]Sheikh Zayed Institute for Pediatric Surgical Innovation, the Children's National Medical Center, Washington DC, USA

ABSTRACT

Surgical instrument segmentation in laparoscopic image sequences can be utilized for a variety of applications during surgical procedures. Recent studies have shown that deep learning-based methods produce competitive results in surgical instrument segmentation. Difficulties, however, lie in the limited number of training datasets involving surgical instruments in laparoscopic image frames. Even though there are publicly available pixel-wise training datasets along with trained models from the Robotic Instrument Segmentation challenge, we are not able to relate them to laparoscopic image frames from different surgical scenarios without any pre- or post-processing. This is because they contain different instrument shapes, image backgrounds, and specular reflections, which implies laborious manual segmentation for training dataset generation. In this work, we propose a novel framework for semi-automated training dataset generation for the purpose of robust segmentation using deep learning. To generate training datasets in various surgical scenarios faster and more accurately, we utilize the publicly available trained model from the Robotic Instrument Segmentation challenge and then use the Watershed Segmentation-based method. For robust segmentation, we use a two-step approach: first, we obtain a coarse segmentation obtained from a deep convolutional neural network architecture, and then we refine the segmentation result via the GrabCut algorithm. Through experiments using four different laparoscopic image sequences, we demonstrate the ability of our proposed framework to provide robust segmentation quality.

Keywords: Medical Imaging, Laparoscopic Surgery, Image Segmentation, Deep Learning, Training Dataset, GrabCut

1. INTRODUCTION AND RELATED WORK

Accurate and robust segmentation of surgical instruments in laparoscopic camera images becomes an important capability as the interest in computer-assisted surgical systems increases. It is also a prerequisite to developing more advanced capabilities such as instrument tracking, which enables diverse laparoscopic applications. By using precise segmentation mask, efficient tracking and pose estimation of the surgical instrument of interest is possible during laparoscopic surgery.[1] For this application, it is essential to obtain fine segmentation with boundary delineation of the surgical instrument. However, the dynamic setting of laparoscopic surgery causes difficulties in segmentation tasks using laparoscopic videos. For instance, image blur, smoke, complex tissues, and light reflectance from surgical scenes act as challenging factors for acquiring precise segmentation.

To overcome these challenges, diverse approaches have been proposed for surgical instrument segmentation tasks. Feature extraction in laparoscopic videos has been performed using handcrafted features, which include gradient, textual, and color information.[2] This approach, however, has limitations due to the motion of surgical instruments and lighting changes. The method of using a machine learning technique was also proposed by employing Random Forests with color and structural information.[3]

As deep learning methods have shown to be successful in segmentation on medical images,[4] different deep learning-based approaches have been presented for surgical instrument segmentation. EndoNet[5] presented a segmentation method for instrument detection and surgical phase recognition using convolutional neural networks. ToolNet[6] introduced a lightweight deep learning architecture for real-time surgical instrument segmentation.

Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling, edited by Baowei Fei, Cristian A. Linte, Proc. of SPIE Vol. 10951, 109511T · © 2019 SPIE
CCC code: 1605-7422/19/$18 · doi: 10.1117/12.2512994

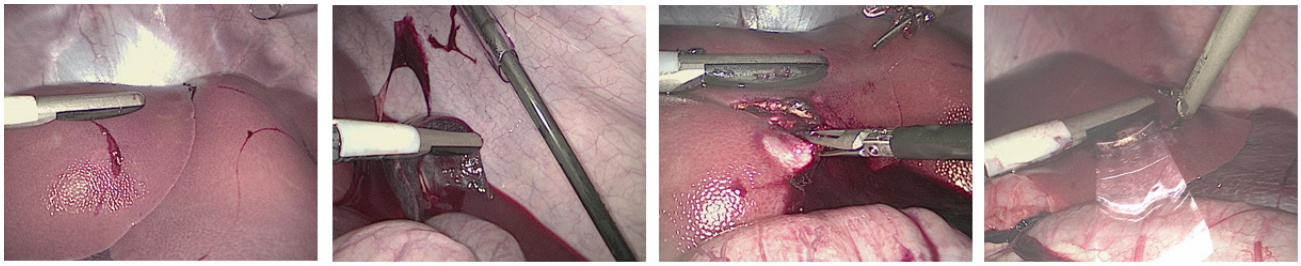

Figure 1. Different characteristics of laparoscopic images in diverse surgical scenarios: (a) specular reflection; (b), (c) blood with miscellaneous objects; (d) instruments for an advanced laparoscopic visualization application.[9]

Attia et al.[7] presented a hybrid deep architecture, which combines convolutional and recurrent neural networks for instrument segmentation. They used the recurrent neural network to model contextual relationships between pixels and enhance the accuracy of segmentation results. Alexey et al.[8] proposed several deep encoder-decoder architectures using transfer learning. These architectures were shown to improve learning in surgical instrument segmentation tasks by leveraging knowledge from a natural image model.

The previous deep learning-based frameworks used publicly available datasets released for the Endoscopic Vision Challenges by the Medical Image Computing and Computer Assisted Intervention (MICCAI) society. These datasets provide a limited set of pixel-wise labeled images. However, there is a limitation on employing deep segmentation networks with limited amounts of laparoscopic training data. This is because to achieve high accuracy, deep learning-based approaches generally require large amounts of labeled training data.s

Figure 1 illustrates different surgical scenes in laparoscopic videos where various surgical instruments such as laparoscopic instruments and an ultrasound probe are utilized. As can be seen, specular reflection, miscellaneous objects, blood or tissues, and instruments for advanced laparoscopic visualization introduce complex visual features, which restricts deployment of existing segmentation models that have been trained using public datasets. To tackle this issue, more training datasets in the dynamic setting of laparoscopic surgery need to be generated. Furthermore, a robust segmentation method is required for such imaging conditions.

In this work, we present a deep learning-based framework for training dataset generation and segmentation of surgical instruments of interest in diverse laparoscopic surgery scenarios. The main contributions of this paper are as follows:

- We present a semi-automated annotation method that has the ability to label surgical instruments of interest using a limited amount of publicly available training data.
- We propose a deep learning-based segmentation method using GrabCut refinement.[10]

To evaluate the proposed method, we obtained retrospectively collected laparoscopic image sequences in different surgical scenarios and applied our proposed framework to the set of collected sequences. The findings from our experiments demonstrate the capability of our framework to consistently provide robust segmentation results.

2. METHOD

The overall structure of the proposed framework is presented in Figure 2. The framework consists of two parts: (1) a semi-automated method for generating a training dataset for laparoscopic videos in diverse surgical scenarios, and (2) a two-step deep learning-based segmentation method using the generated labeled dataset. As can be seen, the inputs of each part are laparoscopic image frames, and the outputs are labeled image frames and semantic binary segmentations, respectively. For the deep segmentation network, we used the LinkNet-34 model[8] in the convolutional encoder-decoder architecture, where a pre-trained ResNet-34 network was used in the

encoder. We then refined the network by retraining refinements in the decoder parts. For binary segmentation of the instrument, the intermediate results are created using the deep learning framework and then refined using morphological operations with the GrabCut algorithm.

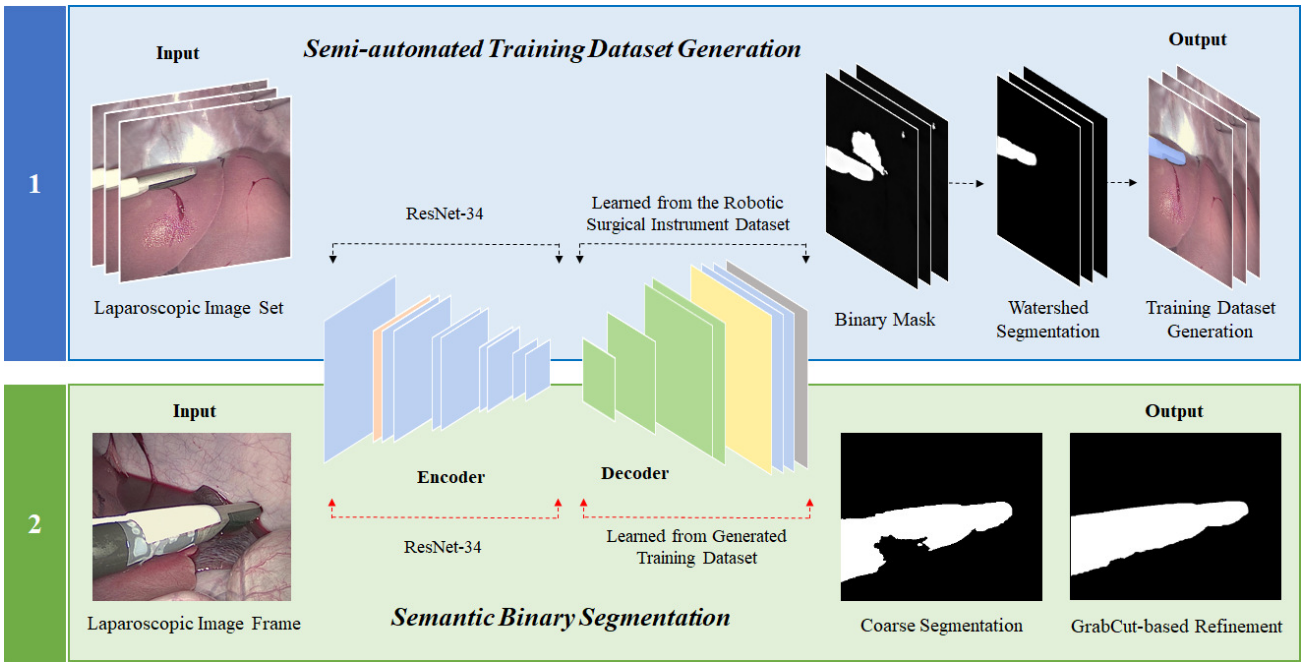

Figure 2. Overall structure of our proposed framework.

2.1 Semi-automated Training Dataset Generation

The framework that we have developed for generating training data is illustrated in Figure 3(a). In this framework, we propose a semi-automated annotation method to generate training samples using publicly available training data from the MICCAI 2017 Robotic Instrument Segmentation challenge *. We refer to this dataset as the MICCAI Robotic Instrument Segmentation (MRIS) dataset. The MRIS dataset is part of the 2017 Endoscopic Vision Challenge. The dataset provides a pixel-wise labeled training dataset of laparoscopic image frames acquired from a da Vinci Xi surgical system.

In this work, we have developed a framework that utilizes the MRIS training samples as a starting point for generating a much larger training dataset that is applicable across a large variety of surgical scenarios. Due to the comparable shape and texture of surgical instruments, we can correlate features of different surgical instruments with those of robotic instruments represented in the MRIS dataset. However, we cannot directly relate them to one other since the image characteristics in different laparoscopic surgical scenarios vary, as illustrated in Figure 1.

In this study, we propose a framework that utilizes segmentation results acquired from the MRIS dataset. These results are then refined using watershed segmentation. We use the segmentation model of Shvets et al.[8] in the convolutional encoder-decoder neural network. This segmentation model was applied in the winning solution in a MICCAI Society competition pertaining to the best instrument segmentation using the MRIS dataset.

As described previously, our framework enables transfer learning to be applied to extend limited training data so that sufficient training data is available for high accuracy deep learning. The framework produces a probability map having the same dimensions as the input image frame. Each pixel value in the probability map indicates the probability of the area or class of interest. From this probability map, we generate a binary mask by setting

*MICCAI 2017 Robotic Instrument Segmentation Sub-challenge
https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/

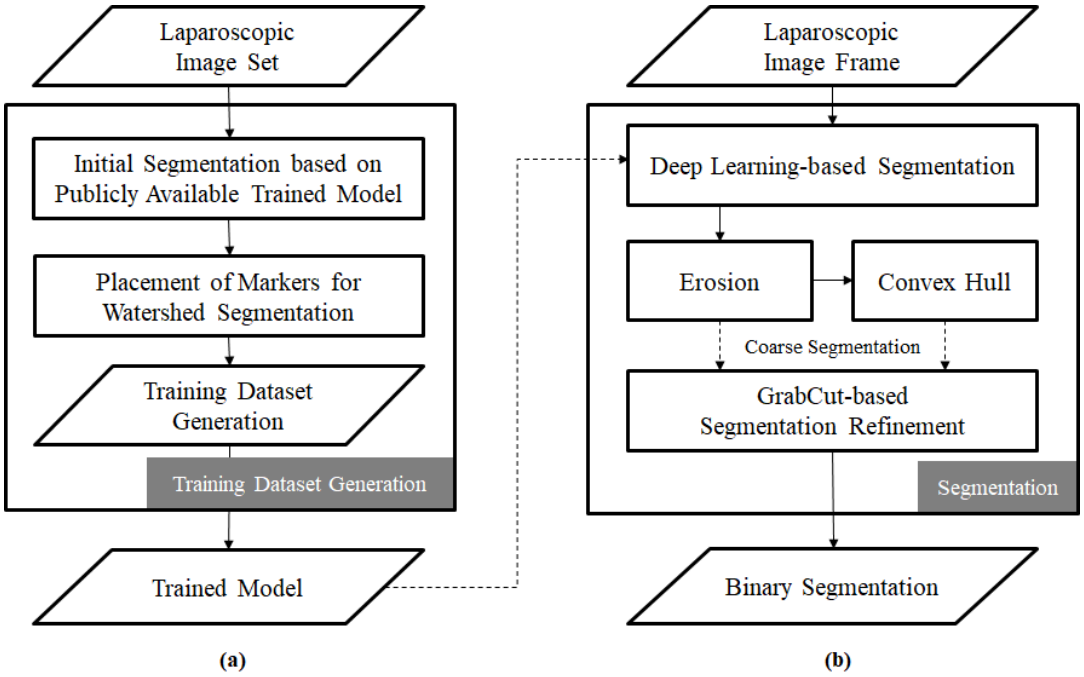

Figure 3. Flowcharts of (a) semi-automated training dataset generation, and (b) two-step deep learning-based semantic binary segmentation.

each pixel to 1 when the probabilistic pixel output is nonzero, and 0 otherwise. Such a binary mask conforms to the expected input of watershed segmentation. Watershed segmentation is a mathematical morphology that has been shown to be effective in separating objects into different regions. We used this type of segmentation to partition the detected pixels into two classes — the surgical instrument of interest and background — and then generate a labeled dataset from the partitioned pixels.

For faster and more accurate generation of training datasets, we utilize the marker-controlled watershed segmentation tool[†] and place two sets of markers at the surgical instrument of interest and image background. This allows us to separate a given surgical instrument from incorrectly segmented background pixels produced by different surgical scenarios and image characteristics. Using this approach, we are able to remove exterior artifacts from the generated dataset with only a small amount of manual marking required.

In summary, our framework generates large amounts of useful new training data with our laparoscopic surgical setup by starting with segmentation results from a publicly available trained model and then applying a semi-automated annotation method using watershed segmentation. The non-automated aspect of the method involves placing single-pixel markers on the surgical instrument and background, which is much less laborious and time consuming compared to performing a full manual segmentation and labeling process.

2.2 Two-phase Deep Learning-based Segmentation

Our approach to deep learning-based segmentation is illustrated in Figure 3(b). We refer to this approach as Two-phase Deep learning Segmentation for Laparoscopic Images (TDSLI) In TDSLI, we perform binary segmentation using the retrained segmentation network constructed from generated label datasets. The binary segmentation results derived from the retrained model are improved with fewer background artifacts. However, they still in general have inaccurate instrument boundaries. These boundaries need refinement to delineate the surgical instrument of interest. The design of the two-phase approach of TDSLI, illustrated in Figure 3(b), is motivated in part by this need for boundary refinement from the initial segmentation results.

[†]Pixel Annotation Tool. https://github.com/abreheret/PixelAnnotationTool

To perform the refinement process (second phase of TDSLI) described above, we apply the GrabCut algorithm.[10] GrabCut is an iterative energy minimization scheme that applies a Gaussian Mixture Model to extract foreground pixels from a complex background. This algorithm can be utilized for boundary refinement with user interaction by placing seed points on the target object. However, in TDSLI we develop an automatic segmentation tool based on the GrabCut algorithm that does not rely on user interaction.

The two-phase approach of TDSLI involves first creating an initial (coarse) segmentation result by applying deep learning followed by erosion and convex hull operations. Then in the second phase, the coarse segmentation result is refined using GrabCut, where, as described above, GrabCut is configured to operate in a fully automated manner. After GrabCut is applied to identify the boundaries of the surgical instrument, TDSLI produces the final binary segmentation. The end-to-end process of TDSLI — from the input to the deep learning block to the generated binary segmentation output — operates in a fully automated manner.

Suppose that n is the number of pixels in a laparoscopic input image to the TDSLI system (the system is parameterized to allow different settings for the image size), and let $Z = \{z_1, z_2, \ldots, z_n\}$ denote an ordered labeling of the pixels in an image. From the coarse segmentation result of TDSLI, we can identify three sets of pixel labels in the input image: *sure foreground* T_f, *probable foreground* T_p, and *sure background*. The set T_f consists of the pixels that are identified as foreground from the Erosion block (see Figure 3(b)), while T_p consists of the foreground pixels as determined by the Convex Hull block. The set T_b is simply the complement of the probable foreground set: $T_u = \{z \mid z \notin T_p\}$.

Figure 4 illustrates operation of the GrabCut-based refinement process in terms of the three sets T_f, T_p, T_u for a sample input image.

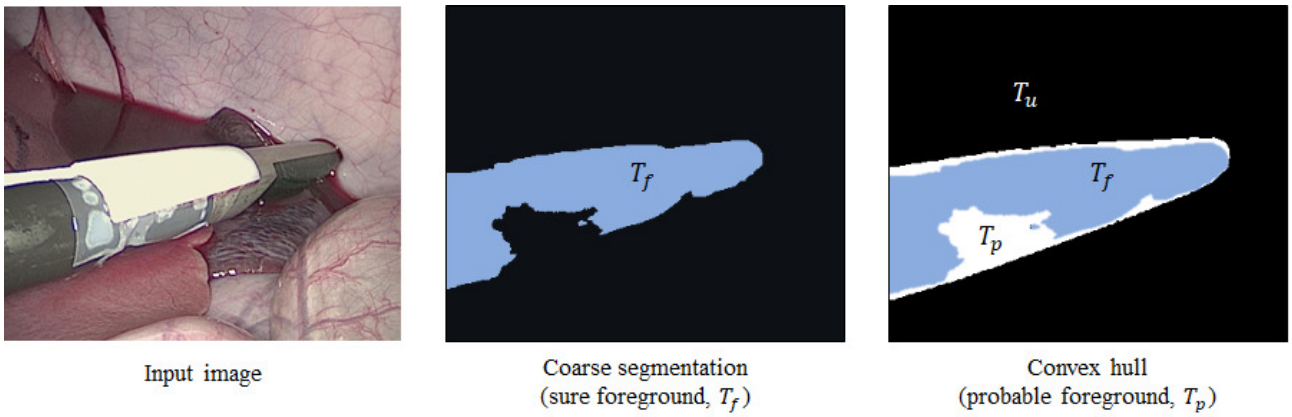

Figure 4. An illustration of the GrabCut-based refinement process for coarse segmentation results.

2.3 Experimental Setup

As described in Section 2.1 we used the MRIS dataset. This dataset consists of 8 frame-sequences, with each sequence consisting of 225 frames. We also acquired four retrospectively collected laparoscopic image sequences in different surgical scenarios from an IACUC-approved animal study. Each of these four sequences contains 3,600 RGB frames with a frame rate of 30 fps, and a resolution of 1280×1024 pixels. We refer to this four-sequence dataset as the *laparoscopic dataset* in the remainder of this paper. The laparoscopic dataset includes different surgical instruments with different poses, and with multiple instruments present in some of the frames.

Among the approximately 14,000 images contained in the laparoscopic dataset, a significant proportion is not suitable for training due to characteristics such as motion blur, image blur or not having any instruments. Additionally, there is some redundancy among the images within the set. We therefore extracted, through manual inspection, a subset of diverse (non-redundant) images that contain instruments and are of suitable quality to be useful for training. The resulting subset of laparoscopic images that we selected for training contains 3,200 images.

For the Deep Learning-based Segmentation block in Figure 3(b), we used the LinkNet-34 network. We selected this network architecture because it is designed not to lose spatial information and incorporates a lightweight encoder to promote fast runtime. However, we would like to emphasize that the overall TDSLI system is not dependent on any particular neural network structure; different network architectures can be applied — for example, based on different image characteristics or different requirements in terms of runtime performance.

In our experiments, we used a batch size of 16, learning rate of 0.0001, and 20 epochs to train the neural network. As a loss function, we used binary cross entropy, which is commonly used for binary pixel classification. The Jaccard index was used for the evaluation metric, which is a similarity measure of overlap that is frequently used to assess the accuracy of segmentation results. This index is defined as:

$$\text{Jaccard Index}(A, B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}, \tag{1}$$

where A is the set of ground truth foreground pixels, B is the set of foreground pixels that is to be evaluated, and $\mid S \mid$ represents the cardinality of (number of elements in) the set S.

As another metric for quantitative evaluation, we used the DICE similarity coefficient (DSC), which is another common metric for assessing segmentation accuracy. This metric can be expressed as:

$$\text{DSC}(A, B) = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid}. \tag{2}$$

3. RESULTS

The proposed segmentation framework for laparoscopic images provides robust binary segmentation to identify surgical instruments of interest. For evaluation (testing), we selected 12 laparoscopic image frames that each contain a single instance of the surgical instrument (target instrument) that were are trying to detect, and across which the target instrument has a diverse variety of poses, and image backgrounds. In both the training and testing datasets, a given image may contain multiple instruments, but only one instance of the target instrument. The 12 images in the testing dataset were taken from the laparoscopic dataset introduced in Section 2.3. The testing dataset was selected to represent diverse surgical scenarios, including different combinations of instruments that are present in addition to the target instrument. We manually segmented the 12 selected testing images to generate ground truth associated with the testing dataset.

Figure 5 illustrates the process of semi-automated training data generation using our framework on four sample laparoscopic images. The input images are shown in Column (a) of the figure. Each of the four rows of this figure corresponds to the processing of one of these four images, and shows (in Columns (b) and (c)) intermediate images that are generated during the process. Column (d) illustrates the labeled result generated from each image.

The derivation of binary segmentation results using deep learning together with boundary refinement is illustrated in Figure 6 for four sample laparoscopic image frames. The sample input frames shown here are taken from the testing dataset described above. The top three rows illustrate typical GrabCut refinements that were derived from among the 12 testing frames, while the bottom row shows the most significant improvement that was observed from GrabCut among the testing frames.

Table 1 presents quantitative results of instrument segmentation that are obtained from the testing dataset. The first two columns of results give statistics using the DSC metric, while the last two columns are based on the Jaccard index. The first row of results is derived using deep neural network processing. The second row of results represents the complete the TDSLI approach, and therefore demonstrates the improvement obtained by applying erosion, and convex hull computation, followed by GrabCut refinement to the result of deep learning-based segmentation.

The first row of results in Table 1 helps to demonstrate the utility of our approach for semi-automated training data generation. In particular, the results demonstrate that the approach can generate training data that is effective in training a deep neural network so that the network produces results of high accuracy.

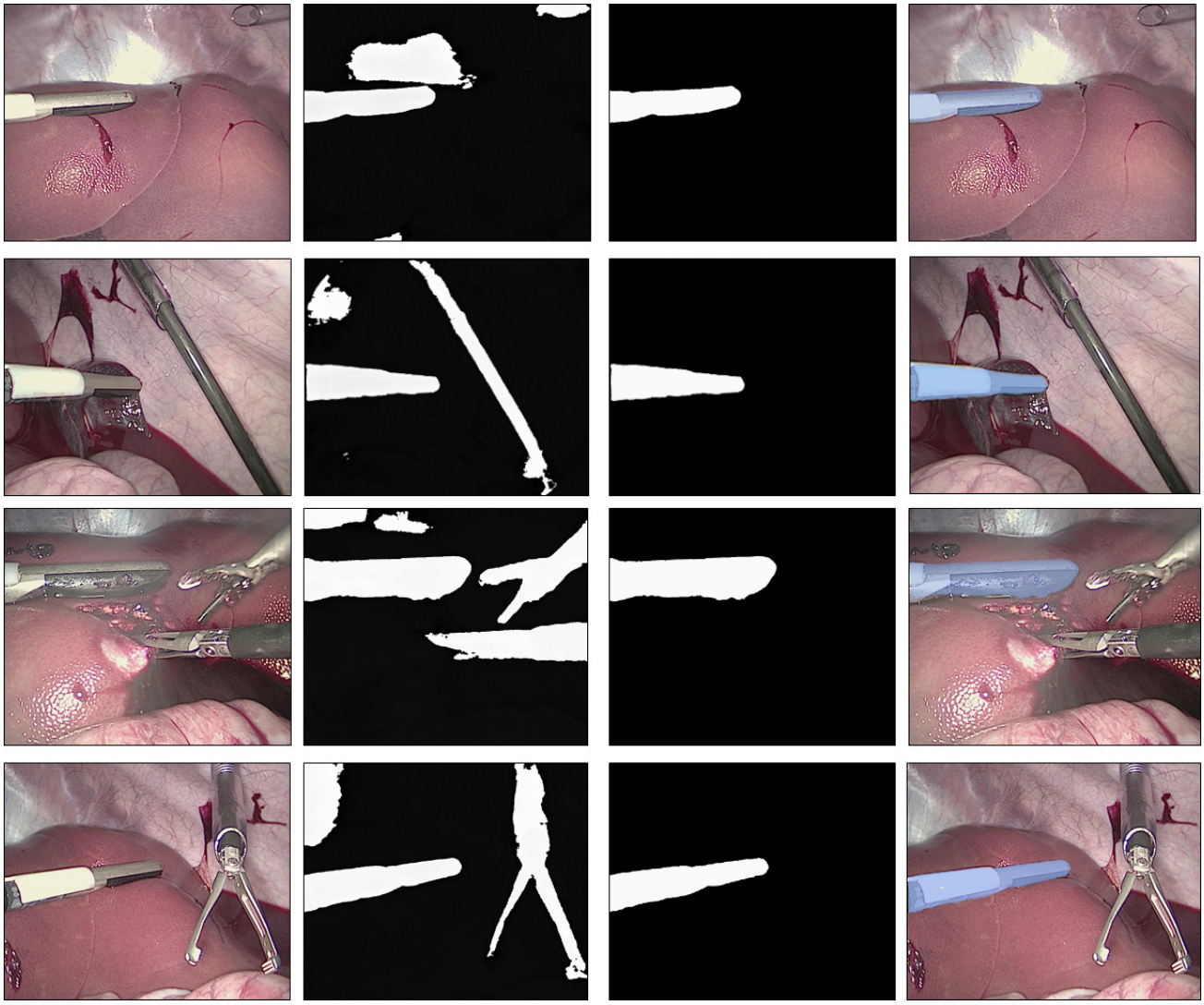

Figure 5. An illustration of semi-automated training dataset generation for a surgical instrument of interest (target instrument) using the proposed framework: (a) laparoscopic image input, (b) initial feature map obtained by applying the model of Shvets et al.[8] that is trained using the MRIS dataset, (c) semi-automated watershed segmentation, and (d) the generated (output) labeled image frames.

Similarly, the second row of results in Table 1 demonstrates the utility of our proposed use of GrabCut refinement in TDSLI together with the erosion and convex hull operations that pre-process the input to GrabCut. From the second row of results, we see that the application of GrabCut helps to further increase segmentation accuracy, beyond the already high accuracy level produced by the deep neural network alone.

4. CONCLUSION

In this paper, we develop new methods to improve the accuracy of surgical instrument segmentation from laparoscopic video streams. First, we present an efficient semi-automated annotation method to create labeled training datasets. This method is motivated by the lack of labeled training data that is available in typical

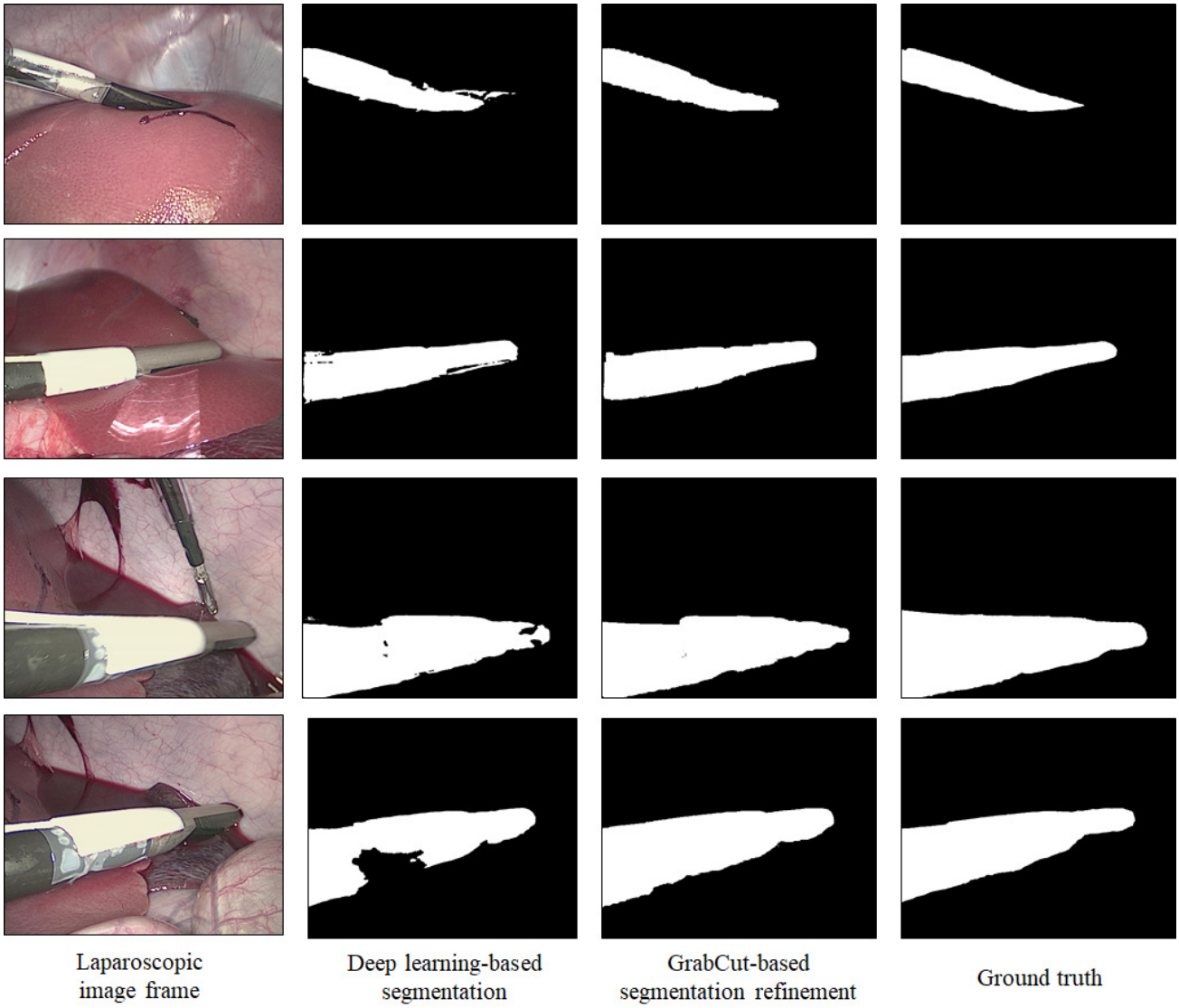

Figure 6. An illustration of deep learning-based segmentation with GrabCut refinement.

scenarios for laparoscopic image segmentation, and the critical need for training data when applying semi-automated segmentations tools, such as those based on deep neural networks (DNNs). Our tool for training dataset generation uses a novel workflow involving (limited) publicly available training data, DNN-based initial segmentation, and marker-controlled watershed segmentation. Second, we present a DNN-based tool, called Two-phase Deep learning Segmentation for Laparoscopic Images (TDSLI), for efficient, automated, laparoscopic image segmentation. The DNN in TDSLI is trained using image-level annotation datasets that is generated from our tool for training data generation. Like our tool for training data generation, TDSLI employs its DNN subsystem for first-phase image processing. The results produced by the DNN are then post-processed using erosion, convex hull computation, and GrabCut refinement to improve segmentation accuracy.

Through experiments using manually-labeled ground truth testing data, we have demonstrated the effectiveness of both of the tools presented in this paper — the DNN-based training data generation tool and TDSLI. The proposed DNN-based framework, which encompasses both tools, can be applied to generate training datasets for diverse surgical instruments and extract the target instruments in various surgical scenarios. The framework

Table 1. Quantitative results of laparoscopic instrument segmentation that are obtained from the testing dataset.

	DSC		Jaccard Index	
	Average (%)	Standard Deviation	Average (%)	Standard Deviation
Deep Neural Network	93.83	2.21	88.46	3.91
Segmentation with GrabCut-based Refinement	95.19	1.41	90.85	2.54

therefore provides practical solutions that are relevant to a variety of applications involving laparoscopic videos.

ACKNOWLEDGEMENT

This work was supported by the National Institutes of Health research grant R42CA192504.

REFERENCES

[1] Bouget, D., Allan, M., Stoyanov, D., and Jannin, P., "Vision-based and marker-less surgical tool detection and tracking: a review of the literature," *Medical image analysis* **35**, 633–654 (2017).

[2] Speidel, S., Benzko, J., Krappe, S., Sudra, G., Azad, P., Müller-Stich, B. P., Gutt, C., and Dillmann, R., "Automatic classification of minimally invasive instruments based on endoscopic image sequences," in [*Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling*], **7261**, 72610A, International Society for Optics and Photonics (2009).

[3] Allan, M., Ourselin, S., Thompson, S., Hawkes, D. J., Kelly, J., and Stoyanov, D., "Toward detection and localization of instruments in minimally invasive surgery," *IEEE Transactions on Biomedical Engineering* **60**(4), 1050–1058 (2013).

[4] 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).

[5] Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., and Padoy, N., "Endonet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging* **36**(1), 86–97 (2017).

[6] 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).

[7] Attia, M., Hossny, M., Nahavandi, S., and Asadi, H., "Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder," in [*Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*], 3373–3378, IEEE (2017).

[8] 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).

[9] Liu, X., Kang, S., Plishker, W., Zaki, G., Kane, T. D., and Shekhar, R., "Laparoscopic stereoscopic augmented reality: toward a clinically viable electromagnetic tracking solution," *Journal of Medical Imaging* **3**(4), 045001 (2016).

[10] Rother, C., Kolmogorov, V., and Blake, A., ""GrabCut": interactive foreground extraction using iterated graph cuts," in [*ACM transactions on graphics (TOG)*], **23**(3), 309–314, ACM (2004).