Tim Streckert*, Dominik Fromme, Jan Steinbrener, and Jörg Thiem

Distance Measurement in Minimally Invasive Surgery Based on Instrument Segmentation

https://doi.org/10.1515/cdbme-2025-0187

Abstract: The application of instrument segmentation holds significant potential for the future of robotic-assisted minimally invasive surgery, as it has the ability to enhance context awareness of the support system. The objective of this research is the development of a new feature that detects the tips of an instrument and can measures the distance between them. The calculated distances between the tips can then be utilized by the surgeon to assess the geometric structure of organs or lesions, which can play a pivotal role in the decision-making process. The proposed methodology involves the automatic segmentation of instrument parts using Mask R-CNN, which facilitates the selection of a tool tip through a straightforward algorithm. The necessity of connecting jaws to the subsequent instrument part or the boundary is beneficial in the elimination of outliers. The Mask R-CNN achieves an mIoU of 0.6441 for instrument parts segmentation. The tips are detected with a rate of 66.78% compared to the ground truth. The average euclidean distance to the ground truth is 8.37 pixels or 1.02% of the image resolution.

Keywords: Tip Detection, Instrument Segmentation, Deep Learning, Distance Measurement

1 Introduction

Robotic-assisted minimally invasive surgery (RAMIS) minimizes the physical impact of surgery on patients. Systems such as the da Vinci surgical platform enhance RAMIS by using articulated joints to precisely replicate the surgeon's wrist movements, achieving accuracy levels often unattainable with traditional minimally invasive surgery (MIS) [1, 2].

Automatic Instrument Segmentation enhances RAMIS by improving context awareness in surgical support systems through precise instrument positioning data. These enhancements can drive advancements in augmented reality and visualization technologies [3], providing accurate instrument pose information that supports surgeons and improves precision, control, and patient outcomes [1] The field of instrument segmentation remains an active area of research, and has been a focal point of multiple challenges in recent years [4, 5].

This paper introduces a novel feature that measures the distance between instrument tips through parts segmentation using Mask R-CNN [6]. The tool tip is defined as the most distal point opposite the connection between the jaw and the instrument component. This distance metric is crucial for surgical decision-making in size-dependent procedures, such as tumor classification [7]. During RAMIS, surgeons use a console to control robotic arms, and our method provides a simple way to measure organ or lesion size without the need for additional systems. We demonstrate that segmentation and basic tools can create an intuitive distance measurement system that improves surgical precision and decision making.

2 Methods

2.1 Experimental data set

For instance part segmentation the data set from the 2017 Endoscopic Vision Instrument Segmentation Challenge (EndoVis17) [4] is used. The data set consists of ten procedures of 300 frames each, recorded by a da Vinci system. For this paper sequences nine and ten are used as test data, with 600 frames. The eighth sequence serves as the validation data set, while the remaining seven sequences form the training set, with a total of 1,575 images, all resized to 640×512. The present contribution is focused on the process of parts segmentation, wherein the instruments are divided into three segments: shaft, wrist and jaw.

2.2 Mask R-CNN

The Mask R-CNN [6] is a network designed for instance segmentation, which aims to differentiate objects within an image by predicting a mask for each individual instance of a class. Mask R-CNN is a Convolutional Neural Network (CNN) that is based on the well-known object detection model Faster R-CNN and extends it with a segmentation branch [8].

The mask network consists of four main components: an encoder for feature extraction, a region proposal network

***Corresponding author: Tim Streckert,** University of Applied Sciences and Arts Dortmund, Sonnenstr. 96, 44139 Dortmund, Germany, e-mail: tim.streckert@fh-dortmund.de
Dominik Fromme, Jörg Thiem, University of Applied Sciences and Arts Dortmund, Sonnenstr. 96, 44139 Dortmund, Germany
Jan Steinbrener, University of Klagenfurt, Klagenfurt, Austria

Open Access. © 2025 The Author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

(RPN) that generates candidate bounding boxes (anchors) with associated object probabilities, and a region of interest (RoI) pooling layer that refines these proposals to enhance alignment with the feature maps. It employs two parallel branches for object classification and bounding box prediction, along with a fully convolutional network (FCN) that predicts segmentation masks for each class. Losses for the mask, bounding box, and classification are calculated and aggregated to form a comprehensive loss. Finally, non-maximum suppression is utilized to eliminate duplicate detections of the same object.

2.3 Metrics

The mean intersection over union (mIoU) and the mean Average Precision (mAP) are the metrics employed to evaluate the results. These metrics are commonly applied in the context of segmentation tasks. The mIoU is a metric that quantifies the precision with which the model identifies and segments objects, calculated as the ratio of the intersection over union area of the predicted and true objects. The mIoU ranges from zero to one, with zero indicating no overlap and one representing a perfect result and is calculated as follows:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{C}\sum_{c=1}^{C}\frac{A_{i,c}\cap B_{i,c}}{A_{i,c}\cup B_{i,c}}\right) \tag{1}$$

The term A is defined as the ground truth, B as the predictions of the model, C as the number of classes and N the number of total images. Initially, the IoU scores for each class in each image of the test data set are calculated. Subsequently, for each image, the average IoU across all classes is computed. Finally, the mean mIoU for each class is obtained by averaging these average IoU scores across all images in the data set.

The mAP is another commonly used metric, that calculates the area under the precision-recall curve for different IoU thresholds, providing an aggregate measure of the model's ability to correctly identify relevant instances across different recall levels. Precision measures the accuracy of positive predictions, while recall measures the proportion of actual positives identified. Together, they indicate a model's performance and effectiveness in capturing relevant cases.

2.4 Tool tip detection

The tips of instruments are not always clearly definable; for example, forceps have multiple tips, and some may be occluded in the image. To provide a consistent definition, the tip is identified as the most distal point of the instrument's jaw, ensuring it is recognizable and intuitive for users. To detect the tips accurately, several conditions must be met: first, the jaw must be segmented by the Mask R-CNN, as it contains the tip, and second, there must be a connection to the image's boundary or the next instrument part to validate the classification and determine the tip's location. The initial condition is established by using the jaw component as the reference point for detecting the tip. The connection is determined by the intersection area between the jaw mask and another instrument component, excluding other jaw parts or components assigned to different instruments. Segmentation may result in small gaps, leading to a lack of intersection; in such cases, the jaw part is dilated to expand the segmentation mask, repeating this process until a connection is found or the jaw part is deemed incorrectly classified. If a connection is established, the connection point is calculated as the mean of the intersection of the two parts.

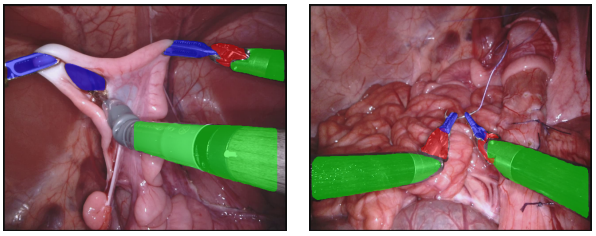
Following this step the contour points of the jaw masks are calculated. A morphological closing operation is applied to bridge disconnected areas between contours, and if multiple contours remain, the part is discarded for lacking distinctiveness at the tip. The tool tip, is expected to be positioned at an angle of $\pm 90^\circ$ from the connection point, on the opposite side of the reference point. From the remaining contour points the tool tip is selected based on the maximum radius from the center.

3 Training the Mask R-CNN

All implementations are executed using *python 3.10* and *Tensorflow 2.15* libraries on an Nvidia H100. The implementation for the Mask R-CNN is adopted from [9] and adapted to the EndoVis17 data set. In this paper geometric and intensity transformations are used, the former consisting of flipping and rotation, and the later consisting of adding Gaussian noise, multiplication, Gaussian blur, hue and saturation. For each image, up to six augmentations are randomly selected and applied.

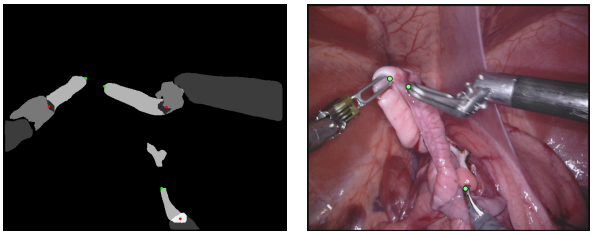
For the Mask R-CNN the ResNet101 is selected as the backbone model, with pretrained COCO weights, which reduces the training time and enhance the robustness of the model. It distinguishes between four classes: shaft, wrist, jaw, and background. The batch size is set to 32.

The training process is consists of two distinct phases. In the initial phase, the RPN, the mask head, and the classification head are trained, the backbone remains frozen. The learning rate is set to 0.001 after exploring several alternative values, with 25 epochs of training, at which point the loss begins to converge. In the second step the model is fine tuned, by incorporating the backbone parameters. The learning rate is reduced to 5×10^{-5}. 40 epochs are trained for fine-tuning, but the validation loss decreases only slightly, indicating the



(a) (b)

Fig. 1: Segmentation results of the trained Mask R-CNN



(a) (b)

Fig. 2: Results of tip detection, left: gray scale image of the parts segmentation; right: image with detected tips; connection points (red), detected tips (green)

end of optimization, and further training does not significantly improve the loss or evaluation metrics.

4 Results

4.1 Results of Parts Segmentation

The trained network demonstrates an overall mIoU of 0.6441 and an overall mAP of 0.4126, with the shaft achieving 0.7754 in mIoU and 0.6162 in mAP, the wrist attaining 0.6037 and 0.3686, and the jaw attaining 0.5544 and 0.2529. Additionally, the network recorded mIoU values of 0.5842 for data set nine and 0.7041 for data set ten.

The results are significantly different for the classes, as shaft has the highest value, followed by wrist and jaw. This can be traced back to the different size of the objects. The shaft is usually, if present, the largest part of the instrument, which makes the impact of small miss predictions of pixel less consequential. Additionally the shaft has a low curvature shape and is therefore more straightforward to predict. A same principle applies to the wrist in an attenuated form. For the jaw, small mask sizes mean that minor misclassifications have a more significant impact on the mIoU. Additionally, the jaws are more likely to be occluded by small tissue, thereby rendering the prediction process more challenging for the model. The lower mAP points out, that the model struggles with higher IoU thresholds, which is underlined by the mIoU values.

Fig. 1 illustrates an examples for the segmentation of the trained Mask R-CNN model on the test data set, including a misclassification, wherein a portion of the tissue has been erroneously designated as a jaw part. Furthermore a jaw part is missing.

4.2 Results Tip detection

The evaluation of tip detection is conducted by calculating the percentage of tips detected from the prediction in relation to the total number of tips detected in the ground truth data. The mean of this percentage is then calculated over all images. Furthermore, the mean distance between the tip of the ground truth and the prediction is calculated. A total of 66.78% of the tips identified in the ground truth could also be detected in the prediction, with a mean euclidean distance of 8.37 pixels or 1.02% of the image resolution. In summary, the system demonstrates a reliable capability to recognize tool tips effectively.

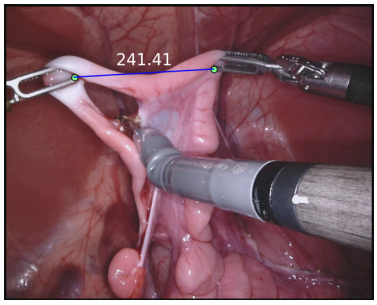


Fig. 3: Distance in pixels between tips (green)

Fig. 2 demonstrates an examples of misclassification. In 2a, a jaw component is identified erroneously. However, as no connection can be identified, no tip is designated for this instrument, illustrated in 2b. It is found that specific jaw parts, in particular forceps, do not have a closed contour. Given their predicted configuration as two distinct segments, the identification of the tip in these regions can potentially result in errors. The results of the pixel distance between the instruments is illustrated in Fig. 3. When the tips of the instruments are successfully detected the pixel distance can be calculated. Using a calibrated 3d endoscope [10] would lead to the metric distance.

4.3 Performance Comparison

The Mask R-CNN is also employed in the study [11] for the EndoVis17 data set, achieving a mIoU of 0.6874. A more

recent network the Segment Anything Model (SAM), has demonstrated encouraging outcomes in the domain of instrument segmentation [12]. The model has achieved an mIoU of 0.8820 for instrument type segmentation. The findings indicate that the model achieves results that are about five percentage points lower than those reported in other studies using the Mask R-CNN. Additionally, recent or revised models show greater potential for improving segmentation outcomes.

To the best of the authors' knowledge, the laparoscopic surgery tip detection approach is the first of its kind in this field, lacking a basis for comparison with other works. In [13], the authors present an automatic tip detection method based on supervised deep learning for surgical instruments in biportal endoscopic spine surgery images. Similarly, [14] describes a method leveraging artificial intelligence for precise localization of needle tips in ultrasound images during robot-assisted interventions, achieving both pixel-accurate representation and metric positioning.

5 Conclusion

The findings indicate that the Mask R-CNN model effectively detects instrument tips and calculates distances between them, though it has limitations in precise segmentation for accurate tip detection. Recent models may enhance this aspect. The Mask R-CNN results are comparable to other studies using the same model for instrument segmentation, and the tool detection implementation is straightforward, incorporating a simple condition to exclude outliers. The results rely heavily on segmentation, which can lead to misclassifications or occlusions affecting tip detection. Although the algorithm may recognize these issues, it can fail if the connection to another part or boundary is not adequately established. The tip is defined based on image data rather than the physical characteristics of the instrument, resulting in variations in detection from different angles and instrument types. A tip may even be identified when not visible in the image, typically as the most distal point, aligning with human intuition and sufficing for this task. This initial approach acknowledges potential improvements for future work, such as using deep learning for direct tip prediction and better instrument segmentation. Future studies will also incorporate stereo images to enhance research and facilitate the calculation of Euclidean distance in metric units.

Overall, this work demonstrates, that a basic approach can lead to a valuable tool to provide surgeons with a new feature for the operating room. Future work should extend beyond the instrument segmentation to include tissue segmentation, which has the potential to directly facilitate the measurement of lesions.

References

[1] Johansson B, Eriksson E,Berglund N., Lindgren I, "Robotic Surgery a Review in minimally invasive techniques," in Fusion of Multidisciplinary Research, An International Journal (FMR), pp. 201–210, 2021.

[2] Boal M. et al., "Evaluation status of current and emerging minimally invasive robotic surgical platforms," in Surgical Endoscopy, vol. 38, pp. 554–585, ISSN: 1432-2218, 2024. Available at: https://link.springer.com/article/10.1007/s00464-023-10554-4

[3] Ahmed F et al., "Deep learning for surgical instrument recognition and segmentation in robotic-assisted surgeries: a systematic review," in Artificial Intelligence Review, vol. 58, 2024.

[4] Allan M. et al., "2017 Robotic Instrument Segmentation Challenge," 2019. Available at: http://arxiv.org/pdf/1902.06426v2

[5] Allan M. et al, "2018 Robotic Scene Segmentation Challenge," 2020. Available at: http://arxiv.org/pdf/2001.11190v3

[6] He K,Gkioxari G, Doll Pár, Girshick R, "Mask R-CNN," 2017. Available at: https://arxiv.org/pdf/1703.06870

[7] Aizza G. et al., "Impact of tumor size on the difficulty of laparoscopic left lateral sectionectomies," in Journal of hepato-biliary-pancreatic sciences, vol. 30, pp. 558–569, 2023.

[8] Ren S, He K, Girshick R, Sun J, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Advances in Neural Information Processing Systems, vol. 28, 2015.

[9] Abdulla W, "Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow," 2017. https://github.com/matterport/Mask_RCNN

[10] Schuldt D, Tanriverdi F, Thiem J, "Performance of Stereo Matching Algorithms in 3D Endoscopy," in Biomedical Engineering / Biomedizinische Technik, vol. 63, pp. 50, ISSN: 0013-5585, 2018.

[11] Kong X. et al., "Accurate instance segmentation of surgical instruments in robotic surgery: model refinement and cross-dataset evaluation," in International journal of computer assisted radiology and surgery, vol. 16, pp. 1607–1614, 2021.

[12] Yu J. et al., "SAM 2 in Robotic Surgery: An Empirical Evaluation for Robustness and Generalization in Surgical Video Segmentation," 2024. Available at: http://arxiv.org/pdf/2408.04593v1

[13] Cho S, Kim Y, Jeong J, Kim I, Lee H, Kim N, "Automatic tip detection of surgical instruments in biportal endoscopic spine surgery," in Computers in biology and medicine, vol. 133, pp. 104384, 2021.

[14] Arapi V, Hardt-Stremayr A, Weiss S, Steinbrener J, "Bridging the simulation-to-real gap for AI-based needle and target detection in robot-assisted ultrasound-guided interventions," in European Radiology Experimental, vol. 7, pp. 30, ISSN: 2509-9280, 2023. Available at: https://eurradiolexp.springeropen.com/articles/10.1186/s41747-023-00344-x