Artificial Intelligence for context-aware surgical guidance in complex robot-assisted oncological procedures: an exploratory feasibility study

Fiona R. Kolbinger[1,2,3,*], Stefan Leger[3,4,*], Matthias Carstens[1], Franziska M. Rinner[1], Stefanie Krell[4], Alexander Chernykh[4], Thomas P. Nielen[1], Sebastian Bodenstedt[4], Thilo Welsch[1,2], Johanna Kirchberg[1,2], Johannes Fritzmann[1,2], Jürgen Weitz[1,2,3], Marius Distler[1,2,§], and Stefanie Speidel[3,4,§,✉]

[1] Department of Visceral, Thoracic and Vascular Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Fetscherstraße 74, 01307 Dresden, Germany

[2] National Center for Tumor Diseases (NCT/UCC), Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany; Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Fetscherstraße 74, 01307 Dresden, Germany

[3] Else Kröner Fresenius Center for Digital Health (EKFZ), Technische Universität Dresden, Fetscherstraße 74, 01307 Dresden, Germany

[4] Division of Translational Surgical Oncology, National Center for Tumor Diseases (NCT), Partner Site Dresden, Fetscherstraße 74, 01307 Dresden, Germany

* These authors contributed equally

§ These authors jointly supervised this work

✉ Corresponding author: Prof. Dr. Stefanie Speidel, Division of Translational Surgical Oncology, National Center for Tumor Diseases (NCT/UCC), Partner Site Dresden, Fetscherstrasse 74, 01307 Dresden, Germany

☎ +49 (0) 351 5413

✉ stefanie.speidel@nct-dresden.de

ABSTRACT

Background: Complex oncological procedures pose various surgical challenges including dissection in distinct tissue planes and preservation of vulnerable anatomical structures throughout different surgical phases. In rectal surgery, a violation of dissection planes increases the risk of local recurrence and autonomous nerve damage resulting in incontinence and sexual dysfunction. While deep learning-based identification of target structures has been described in basic laparoscopic procedures, feasibility of artificial intelligence-based guidance has not yet been investigated in complex abdominal surgery.

Methods: A dataset of 57 robot-assisted rectal resection (RARR) videos was split into a pre-training dataset of 24 temporally non-annotated videos and a training dataset of 33 temporally annotated videos. Based on phase annotations and pixel-wise annotations of randomly selected image frames, convolutional neural networks were trained to distinguish surgical phases and phase-specifically segment anatomical structures and tissue planes. To evaluate model performance, F1 score, Intersection-over-Union (IoU), precision, recall, and specificity were determined.

Results: We demonstrate that both temporal (average F1 score for surgical phase recognition: 0.78) and spatial features of complex surgeries can be identified using machine learning-based image analysis. Based on analysis of a total of 8797 images with pixel-wise target structure segmentations, mean IoUs for segmentation of anatomical target structures range from 0.09 to 0.82 and from 0.05 to 0.32 for dissection planes and dissection lines throughout different phases of RARR in our analysis.

Conclusions: Image-based recognition is a promising technique for surgical guidance in complex surgical procedures. Future research should investigate clinical applicability, usability, and therapeutic impact of a respective guidance system.

INTRODUCTION

Despite advances in nonsurgical management of rectal cancer, the vast majority of curative treatment approaches require oncologically radical surgical removal of the primary tumor and locoregional lymph nodes within the mesorectal envelope (total mesorectal excision, TME). This surgical step requires delicate preparation along the mesorectal fascia to ensure complete resection and minimize the risk of local tumor recurrence.[1] In parallel, autonomous nerves running in close proximity to the resection area must be protected to avoid postoperative incontinence and sexual dysfunction. With highly varying numbers depending on surgeon expertise and center volume, these functional sequelae affect up to 85% of patients after oncological rectal resection and can be a direct result of suboptimal surgical preparation outside of anatomical planes, leading to nerve injury.[2,3]

Surgical robots offer mechanical enhancements such as improved instrument dexterity, hand-eye coordination, and a three-dimensional view, however, no robust clinical benefit over conventional laparoscopic surgery could be demonstrated to date.[4] Based on the hypothesis that an integration of context-dependent surgical guidance into minimally-invasive surgical systems could markedly improve surgical quality and outcome and diminish quality variation between surgeons, extensive research on the development of image recognition algorithms has been conducted in recent years using laparoscopic imaging data. Most of these endeavors, however, have focused on tasks with indirect surgical benefit, such as automated instrument detection.[5,6] Thus far, translational Artificial Intelligence (AI)-based success stories in the field of surgery are lacking and clinical applications are mostly limited to orthopedic, neurosurgical, and hepatic surgical procedures.[7,8] With regard to approaches with high translational potential in laparoscopic surgery, deep learning-based algorithms have recently been shown to identify relevant anatomical areas during cholecystectomy.[9,10] Of note, such methods have only been established for less complex surgical procedures, for which – in part – open-access datasets exist for the purpose of algorithm development and validation.[11,12] In these operations, however, the benefit of AI-based assistance is marginal.

In this project, we aim to explore the potential of deep learning-based detection algorithms based on convolutional neural networks (CNNs) to provide context-dependent guidance in highly complex oncological procedures. Considering the example of robot-assisted rectal resection (RARR), we created a dataset of spatio-temporal data in the form of temporally annotated surgery videos and semantically segmented video frames. Based on this data, we trained different CNNs to automatically detect surgical phases and identify a variety of visually delineable anatomical structures and dissection planes to enable context-aware surgical guidance.

Our findings suggest that surgical phase differentiation and recognition of selected anatomical structures including the mesocolon, the mesorectum, Gerota's fascia, or the abdominal wall, can satisfactorily be implemented in complex minimally-invasive surgical procedures. We thereby establish essential components of a clinically applicable AI-based system for context-aware guidance in complex minimally-invasive oncological surgery. On a general level, this work highlights challenges in the implementation of clinically relevant surgical guidance functions, in particular limited data availability, and encourages inter-institutional collaboration and improved data accessibility to facilitate data science-based approaches in a diverse range of surgical procedures with varying complexity.

METHODS

Patient population and dataset

Between February 2019 and January 2021, video data from a total of 57 robot-assisted anterior rectal resections, intersphincteric resections, or abdominoperineal resections with either partial (PME) or TME performed at the University Hospital Carl Gustav Carus Dresden, Germany were collected. All included patients had a clinical indication for the surgical procedure, recommended by an interdisciplinary tumor board. All procedures were performed using the da Vinci® Xi system (Intuitive Surgical, Sunnyvale, CA, USA). Surgeries were recorded using the CAST system (Orpheus Medical GmBH, Frankfurt a.M., Germany). Each record was saved at a resolution of 1920 x 1080 pixels in MPEG-4 format.

This study was performed in accord with the ethical standards of the Helsinki declaration and its later amendments. The local Institutional Review Board (ethics committee at the Technical University Dresden) reviewed and approved this study (approval number: BO-EK-137042018). Written informed consent was obtained from all participants. The trial was registered on clinicaltrials.gov (trial registration ID: NCT05268432).

Annotations

The total dataset of 57 RARR recordings was randomly split into a pre-training dataset (24 temporally non-annotated videos) and an annotated training dataset (33 videos). Annotations were performed by two members of the research team (MC, FMR) after thorough education and several rounds of training annotations with two experts in robot-assisted rectal surgery (JW, MD), and reviewed by a surgeon in training with considerable experience in robot-assisted rectal surgery (FRK). According to a previously created annotation protocol (Supplementary Material 1), the surgical process was temporally annotated using b<>com *Surgery Workflow Toolbox* [Annotate] version 2.2.0 (b<>com, Cesson-Sévigné, France) with regard to the sequence of surgical phases (Figure 1). The definition and order of the five phases – preparation and intraabdominal orientation, medial mobilization of descending colon, lateral mobilization of descending colon, mesorectal excision, and extraabdominal preparation – was based on institutional standards and international recommendations for RARR.[13,14] Of note, splenic flexure and descending colon mobilization was carried out from medial to lateral as described previously by Ahmed *et al.*[13].

For training and validation of phase-specific semantic segmentation algorithms, equidistant frames from at least ten selected video recordings were annotated per phase with regard to

visibility of anatomical structures and dissection planes (Figure 1, Supplementary Table 1) using the open-source software 3D Slicer with the SlicerRT program extension.[15] The videos for each phase were selected based on criteria of target structure visibility and exposition (Supplementary Table 1).

Model development and evaluation: Phase detection

For surgical phase recognition a long short-term memory network (LSTM)[16] that learns the temporal relationships combined with a residual neural network (ResNet)[17] as feature extractor was trained based on the 33 annotated RARR videos to detect the five surgical phases (preparation, medial mobilization of descending colon, lateral mobilization of descending colon, (total) mesorectal excision, and extraabdominal preparation) (Figure 2 A).

The pre-training of the feature extractor was performed in two different configurations: (a) the ResNet was pre-trained on the ImageNet database[18] only and (b) the pre-trained feature extractor was fine-tuned on 24 temporally non-annotated RARR videos using self-supervised training. Subsequently, the two different models were trained and evaluated using a 4-fold cross validation scheme. For this purpose, the training folds were again divided into a training and validation (four randomly chosen surgeries) set to select the final model based on the network accuracy and to improve the generalizability of the network. Afterwards, the final model was assessed on the test folds and the performance was measured by the F1 score and the accuracy.

Model development and evaluation: Semantic segmentation

For each surgery phase, a separate convolutional neural network was trained for the phase-specific segmentation of anatomical structures and tissue planes (Figure 2 B). For this purpose, the Detectron2[19] framework was used in combination with a ResNet and a feature pyramid network as backbone, which was pre-trained on the ImageNet dataset[18]. The model was trained and evaluated within a leave-one-out cross validation scheme. To improve the generalization and the stability of the segmentation network the training fold was split randomly into a training and a validation fold at a ratio of 80:20. Subsequently, the trained model was assessed on the validation fold every 5000 training iterations using the average Intersection-over-Union (IoU) to select the final model. Finally, the average segmentation performance was assessed using F1 score, IoU, precision, recall, and specificity on the test folds.[20] These parameters are commonly used technical measures of prediction exactness, ranging from 0 (least exact prediction) to 1 (entirely correct prediction without any misprediction).

Statistics and Reproducibility

In total, recorded surgery videos from 57 surgeries were considered in this analysis. For surgical phase recognition, these data were randomly split into a non-annotated pre-training dataset (24 videos) and a temporally annotated training and test dataset (33 videos). For training of the segmentation algorithms for each individual phase, at least ten video recordings were selected per phase with regard to visibility of anatomical structures and dissection planes and annotated as described above.

Training of phase detection and segmentation networks were carried out as follows: The ResNet50 described by He et al.[17] was selected as image feature extractor, having demonstrated a high performance on the Cholec80 dataset[21]. For the specific detection task, the last layer was replaced by a layer representing the five surgical phases. Afterwards, the network was trained frame-wise with the label information of the current phase. To enhance the robustness of the model, data augmentation consisting of image rotation, image flipping and image noise was used. Furthermore, the network was trained for 30 epochs, where each epoch consists of 3000 random chosen batches with a size of 64. As optimizer, Adam with an initial learning rate of 0.0001 was used and the weighted cross entropy loss was applied to handle the class-imbalanced problem. The self-supervised learning approach described by Funke et al.[22] was adapted, so that a ResNet50 was trained based on temporally coherent video frame embedding using the 1st and 2nd order contrastive loss.[23,24]

The segmentation network was trained for 200 epochs and a batch size of 2. The data augmentation consisting of image resizing and random flipping of the image was used to enhance the training process. For the optimization of the network the gradient descent optimization algorithm was used with a learning rate of 0.0002.

RESULTS

Dataset

A total of 57 RARR with an average duration of 414 ± 112 minutes performed by nine experts in robot-assisted oncological surgery were recorded and analyzed within this study. The majority of patients (n = 43, 75.4%) were male and the predominant indication for the surgical procedure was colorectal cancer (n = 54, 94.7%). Twenty-five patients (43.9%) had undergone neoadjuvant irradiation to the surgical field. Overall, the sample represents a wide variety of rectal cancer manifestations, including a considerable number of patients (n = 27, 47.4%) with locally advanced tumors (Table 1).

Thirty-three videos were temporally annotated with regard to the sequence of five standardized surgical phases (Figure 1). For semantic segmentation, individual models were trained for each surgical phase to phase-specifically identify distinct target structures. Based on the visibility of these target structures (Supplementary Table 1), at least ten videos were selected from the available video dataset for training of the algorithm for the respective surgical phase. In total, 8797 annotated frames were used for training of the semantic segmentation algorithms (Figure 3, Supplementary Table 2).

Model performance: Surgical phase recognition

The sequence of the five surgical phases (Figure 1) was recognized using a ResNet50 architecture for recognition of visual features with or without a long-short-term memory (LSTM) network for extraction of temporal features (Figure 2 A). Using the ResNet50 alone, surgical phases were predicted at an average accuracy of 0.74 ± 0.02 and an F1 score of 0.68 ± 0.01 both with and without pre-training with non-annotated RARR videos. Paired with an LSTM, the accuracy for surgical phase prediction was 0.82 ± 0.01. As for the ResNet50 alone, self-supervised pre-training of the network did not result in any improvement of the prediction (accuracy: 0.83 ± 0.02) by the combined ResNet50 and the LSTM (Table 2).

Model performance: Segmentation of anatomical structures and dissection planes

Anatomical structures and dissection planes were localized using a multi-layer convolutional neural network (Figure 2 B). Table 3 displays mean F1 score, IoU, precision, recall, and specificity for individual segments predicted by the trained phase-specific algorithms.

Overall, the most reliable detections were achieved for Gerota's fascia (F1 score: 0.78 ± 0.05) and the mesocolon (F1 score: 0.71 ± 0.07) during medial mobilization of the descending colon and for the abdominal wall (F1 score: 0.82 ± 0.06) and fatty tissue (F1 score: 0.68 ± 0.11) during lateral mobilization of the descending colon. The dissection plane ("angel's hair") in the Mesorectal Excision phase could be predicted with an F1 score of 0.32 ± 0.09, whereas prediction of the exact dissection line was largely unsuccessful at an F1 score of 0.05 ± 0.02 both during medial mobilization of the descending colon and during Mesorectal Excision (Table 3). Figure 3 illustrates examples of semantic segmentations by the developed CNNs, which were trained for each phase of the procedure.

The target structures varied considerably with regard to the proportion of images without any prediction, resulting in a value of "0" for the analyzed metrics (Figure 4). For instance, Gerota's fascia and the mesocolon were correctly identified in most of the respective images, while dissection lines, vessel structures, and seminal vesicles were segmented in very few images displaying the structures. Other structures, such as the colon, the mesorectum, the small intestine, plastic clips, and the dissection plane during mesorectal excision, showed two significant fractions of images: one fraction without any prediction and one fraction with good to excellent performance metrics.

DISCUSSION

Correct recognition of visual cues is a key challenge in minimally-invasive surgery, as the majority of currently available laparoscopic and robotic instruments and devices do not provide any haptic feedback or additional imaging to visualize hidden structures. Consequently, the concept of enhancing surgeons' visual perception through computational techniques has attracted increasing scientific interest in the surgical community, mirrored by a growing body of research in the field of surgical data science.[25] The overwhelming majority of these studies, however, has investigated potential assistance tasks with limited clinical benefit, such as automated instrument classification.[5,6] Projects investigating the recognition of surgically meaningful structures in abdominal surgery are rare and mostly limited to laparoscopic cholecystectomy[9,10], a far less complex surgical procedure performed in large numbers all over the world. The availability of a range of open-access datasets of cholecystectomy recordings[6,11,26] make this procedure an attractive example for proof-of-principle studies. Considering the example of RARR, we aimed to explore the possibility of an integration of image-based recognition for surgical guidance in complex oncological surgical procedures, for which no open-access video datasets are available. In summary, our findings suggest that even on the basis of limited annotated data, essential CNN-based temporal and spatial recognition tasks can be implemented in complex robot-assisted oncological procedures, providing the basis for context-dependent surgical guidance. State-of-the-art approaches using a ResNet50 paired with an LSTM satisfactorily differentiated five surgical phases at an average accuracy of 0.83 and F1 score of 0.78. Our results on image-based identification of anatomical structures and dissection planes are promising for selected target structures. Overall, however, they indicate that further methodological improvements and intensified interdisciplinary efforts are necessary for entirely convincing implementation of specific image recognition functions, in particular for tasks that require considerable specialized knowledge for annotation.

In general, the presented observations are in line with previous studies on visual analysis of laparoscopic images.[27] With regard to operative phase recognition, similar accuracies around 0.8 have been reported for other laparoscopic procedures including cholecystectomy and sigmoidectomy.[28] Previous studies investigating the identification of anatomical structures have mostly focused on the classification of the presence of certain organs[12] or rough localization of the detected organs using bounding boxes[29]. We performed semantic segmentation of the exact boundaries of previously annotated structures and observed substantial variation of recognition quality, ranging from a mean IoU of < 0.1 for dissection lines and seminal vesicles to 0.82 for the abdominal wall during the lateral mobilization of the descending colon. It is conceivable that

unsuccessful recognition of some structures was a consequence of the limited amount of training data. For instance, seminal vesicles during mesorectal excision (mean IoU: 0.09) could, overall, not be segmented successfully, and input data was limited at 336 annotated images (Supplementary Table 2). In turn, four out of the five structures with the largest number of annotated images – the mesocolon (mean IoU: 0.71), Gerota's fascia (mean IoU: 0.78), the mesorectum (mean IoU: 0.48), and the dissection plane during mesorectal excision (mean IoU: 0.32) – could be recognized in a significant proportion of images (Figure 4). Notably, several structures (including the colon, the small intestine, the mesorectum, and the dissection plane during mesorectal excision) were identified either at very high accuracy or not identified in the image at all, resulting in two large subsets of images at either end of the scale. Considering a video stream as a sequence of images, the successful recognition of a structure in a significant proportion of images may imply that even at this stage, the recognition algorithms may be clinically useful. The reported subjective clinical utility of a bounding box-based detection system recognizing the common bile duct and the cystic duct at average precisions of 0.32 and 0.07, respectively, supports this hypothesis[29] and encourages discussion about the clinical value of classical metrics quantifying segment overlap. Notably, explainability and real-world applicability of the classical metrics describing segment overlap have recently been explored in the context of autonomous driving, with similar findings on the limitations of these metrics.[30]

Of note, the exact location of the dissection line could neither be successfully identified during medial mobilization of the descending colon nor during mesorectal excision. Two explanations are conceivable: First, thin and small structures displayed a trend towards being more challenging to predict. Another example for this observation is the largely unsuccessful segmentation of metal clips during vascular dissection. Second, the dissection lines show very diverse visual appearance. Especially during mesorectal excision, the appearance of embryonal tissue planes and the closely related dissection line at the mesorectal fascia can vary with regard to neoadjuvant (radio-)therapy and individual factors such as body composition, making identification particularly challenging.[31] Of note, this study included a variety of patients with rectal cancer, and no special patient factors were considered during selection of the video subsets for semantic segmentation for each individual phase (Table 1). An integration of more and diverse video data from multiple centers would likely improve recognition of the dissection lines and explore the limits of the applied state-of-the-art neural networks. In general, diversification of input data would be massively facilitated through publication of anonymized video data documenting complex surgical procedures.

For improved applicability, ongoing and future work will focus on two technical improvements: On the algorithmical level, recognition uncertainty will be implemented in display of the target structure segmentations, for instance through Bayesian computation methods.[32] Such methods facilitate a heatmap-like display of the image-based segmentations, which would visually approximate the ambiguity inherent to the distinction of adjacent and, in part, similar appearing anatomical structures. The integration of uncertainty estimation could also contribute to an increased acceptance of the system among clinicians. On the level of clinical implementation, the system needs to meet real-time requirements for clinical usability. This comprises the exact determination and subsequent minimization of the computation-associated delay to the range of tenths of a second. During the computation process for the presented study, the entire time required for image loading, computation of all outcome metrics and saving of the image ranged around four seconds. Moreover, a graphical user interface that allows for interaction of the surgeon with the guidance system (i.e. selecting target structures for display) will need to be integrated.

The limitations of this study are mostly related to limited data availability, which is generally one of the major bottlenecks for successful translation of AI-based methods in surgery. One of the key reasons for this is the high annotation effort required for an integration of clinically meaningful knowledge into a dataset.[25] The monocentric study design is another major limitation of this study, considerably restricting generalizability and transferability of the implemented functions. Future research will therefore aim at an integration of multicentric video data. In this context, standardization of (video) data recording and adaptability of the annotation process to differences in surgical techniques and approaches pose important challenges to consider during study design. Of note, the Society of American Gastrointestinal and Endoscopic Surgeons recently published recommendations for surgical video annotation to increase the amount of publicly available datasets and enable the creation of benchmark AI algorithms with a clinical impact.[33] The existing limitations notwithstanding, the presented dataset and study represent an important addition to the growing body of research on context-aware guidance in abdominal surgery, as it significantly expands the horizon of data science-assisted surgery to procedures of higher complexity and technical difficulty.

In conclusion, this study has established the principal components of a clinically applicable AI-based system for context-aware guidance in RARR. To our knowledge, this is the first approach to an implementation of machine learning-based guidance into a complex robot-assisted oncological procedure. A subsequent pilot study will merge the established functionalities into a basic guidance system and evaluate its clinical applicability, usability, and therapeutic impact on oncological and functional outcomes of RARR. Importantly, the basic concept of this system is

transferable to other complex minimally invasive surgical procedures with several critical phases such as minimally-invasive pancreatectomy or esophagectomy. With regard to the identified challenges in implementation of clinically relevant guidance functions, in particular limited data availability, our findings moreover encourage inter-institutional collaboration and improved data accessibility to stimulate research on a more diverse range of surgical procedures. The creation of more potent and impactful networks could ultimately result in a true clinical benefit for patients and healthcare professionals.

REFERENCES

1. Heald, R. J., Husband, E. M. & Ryall, R. D. H. The mesorectum in rectal cancer surgery—the clue to pelvic recurrence? *Br. J. Surg.* (1982) doi:10.1002/bjs.1800691019.
2. Chew, M. H., Yeh, Y. T., Lim, E. & Seow-Choen, F. Pelvic autonomic nerve preservation in radical rectal cancer surgery: Changes in the past 3 decades. *Gastroenterology Report* (2016) doi:10.1093/gastro/gow023.
3. Sturiale, A. *et al.* Long-term functional follow-up after anterior rectal resection for cancer. *Int. J. Colorectal Dis.* (2017) doi:10.1007/s00384-016-2659-6.
4. Jayne, D. *et al.* Effect of robotic-assisted vs conventional laparoscopic surgery on risk of conversion to open laparotomy among patients undergoing resection for rectal cancer the rolarr randomized clinical trial. *JAMA - J. Am. Med. Assoc.* **318**, 1569–1580 (2017).
5. Alsheakhali, M., Eslami, A., Roodaki, H. & Navab, N. CRF-Based Model for Instrument Detection and Pose Estimation in Retinal Microsurgery. *Comput. Math. Methods Med.* **2016**, (2016).
6. Jin, A. *et al.* Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. *Proc. - 2018 IEEE Winter Conf. Appl. Comput. Vision, WACV 2018* **2018-January**, 691–699 (2018).
7. Burström, G. *et al.* Feasibility and accuracy of a robotic guidance system for navigated spine surgery in a hybrid operating room: a cadaver study. *Sci. Rep.* **10**, (2020).
8. Hu, Z. *et al.* First-in-human liver-tumour surgery guided by multispectral fluorescence imaging in the visible and near-infrared-I/II windows. *Nat. Biomed. Eng.* **4**, 259–271 (2020).
9. Madani, A. *et al.* Artificial Intelligence for Intraoperative Guidance. *Ann. Surg.* (2020) doi:10.1097/sla.0000000000004594.
10. Mascagni, P. *et al.* Artificial Intelligence for Surgical Safety: Automatic Assessment of the Critical View of Safety in Laparoscopic Cholecystectomy Using Deep Learning. *Ann. Surg.* (2020) doi:10.1097/SLA.0000000000004351.
11. Twinanda, A. P. *et al.* EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Trans. Med. Imaging* **36**, 86–97 (2016).
12. Leibetseder, A. *et al.* LapGyn4: A Dataset for 4 Automatic Content Analysis Problems in the Domain of Laparoscopic Gynecology. *Proc. 9th ACM Multimed. Syst. Conf.* **18**, (2018).
13. Ahmed, J., Kuzu, M. A., Figueiredo, N., Khan, J. & Parvaiz, A. Three-step standardized approach for complete mobilization of the splenic flexure during robotic rectal cancer surgery. *Color. Dis.* **18**, O171–O174 (2016).
14. Panteleimonitis, S. *et al.* Precision in robotic rectal surgery using the da Vinci Xi system

and integrated table motion, a technical note. *J. Robot. Surg.* **12**, 433–436 (2018).

15. Kikinis, R., Pieper, S. D. & Vosburgh, K. G. 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. in *Intraoperative Imaging and Image-Guided Therapy* 277–289 (Springer New York, 2014). doi:10.1007/978-1-4614-7657-3_19.
16. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).
17. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2016-December**, 770–778 (2015).
18. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. 248–255 (2010) doi:10.1109/CVPR.2009.5206848.
19. Wu, Y., Kirillov, A., Lo, F. W.-Y. & Girshick, R. Detectron2. https://github.com/facebookresearch/detectron2 (2019).
20. Leger, S. *et al.* A comparative study of machine learning methods for time-To-event survival data for radiomics risk modelling. *Sci. Rep.* **7**, 11 (2017).
21. Czempiel, T. *et al.* TeCNO: Surgical Phase Recognition with Multi-Stage Temporal Convolutional Networks.
22. Funke, I. *et al.* Temporal coherence-based self-supervised learning for laparoscopic workflow analysis. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **11041 LNCS**, 85–93 (2018).
23. Goroshin, R., Bruna, J., Tompson, J., Eigen, D. & LeCun, Y. Unsupervised Learning of Spatiotemporally Coherent Metrics. *Proc. IEEE Int. Conf. Comput. Vis.* **2015 Inter**, 4086–4093 (2015).
24. Jayaraman, D. & Grauman, K. Slow and steady feature analysis: higher order temporal coherence in video. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2016-December**, 3852–3861 (2015).
25. Maier-Hein, L. *et al.* Surgical data science for next-generation interventions. *Nature Biomedical Engineering* (2017) doi:10.1038/s41551-017-0132-7.
26. Hong, W.-Y. *et al.* CholecSeg8k: A Semantic Segmentation Dataset for Laparoscopic Cholecystectomy Based on Cholec80. (2020).
27. Anteby, R. *et al.* Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. *Surg. Endosc.* **35**, 1521–1533 (2021).
28. Garrow, C. R. *et al.* Machine Learning for Surgical Phase Recognition: A Systematic Review. *Ann. Surg.* **273**, 684–693 (2021).

29. Tokuyasu, T. *et al.* Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surg. Endosc. 2020 354* **35**, 1651–1658 (2020).
30. Zhang, Y., Mehta, S. & Caspi, A. Rethinking Semantic Segmentation Evaluation for Explainability and Model Selection. (2021).
31. Ward, T. M. *et al.* Challenges in surgical video annotation. *Comput. Assist. Surg.* **26**, 58–68 (2021).
32. Kwon, Y., Won, J. H., Kim, B. J. & Paik, M. C. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Comput. Stat. Data Anal.* **142**, 106816 (2020).
33. Meireles, O. R. *et al.* SAGES consensus recommendations on an annotation framework for surgical video. *Surg. Endosc. 2021* **1**, 1–12 (2021).
34. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. (2019).

TABLES

Table 1

Tab. 1 Patient characteristics. For age, BMI and surgery duration, mean ± SD are displayed. All other data are presented as total numbers and percentages of the (sub-)cohort. Abbreviations: Abdominoperineal resection (APR), Anterior Resection (AR), Body Mass Index (BMI), Colorectal cancer (CRC), Low anterior resection (LAR), Intersphincteric resection (ISR), Partial mesorectal excision (PME), Total mesorectal excision (TME).

	Total (n = 57)	**Temporal annotation** (n = 33)	**Vascular Dissection** (n = 10)	**Medial Mobilization** (n = 11)	**Lateral Mobilization** (n = 10)	**Mesorectal Excision** (n = 10)
Age [years]	62.4 ± 11.2	62.2 ± 10.2	62.1 ± 5.2	65.6 ± 7.7	63.9 ± 5.0	62.8 ± 8.1
BMI [kg/m²]	26.5 ± 3.3	26.7 ± 2.9	26.1 ± 2.1	26.7 ± 2.9	26.6 ± 2.8	26.9 ± 2.0
Surgery duration [min]	414 ± 112	415 ± 117	444 ± 148	410 ± 98	466 ± 141	377 ± 130
Sex						
Female	14 (24.6)	5 (15.2)	1 (10.0)	2 (18.2)	1 (10.0)	1 (10.0)
Male	43 (75.4)	28 (84.8)	9 (90.0)	9 (81.8)	9 (90.0)	9 (90.0)
Indication						
CRC	54 (94.7)	31 (93.9)	10 (100.0)	11 (100.0)	10 (100.0)	10 (100.0)
Other	3 (5.3)	2 (6.1)	0	0	0	0
Distance from anocutaneous line						
<6 cm	19 (33.3)	14 (42.4)	3 (30.0)	3 (27.3)	3 (30.0)	5 (50.0)
6 - <12 cm	26 (45.6)	14 (42.4)	4 (40.0)	5 (45.5)	5 (50.0)	3 (30.0)
≥12 cm	12 (21.1)	5 (15.2)	3 (30.0)	3 (27.3)	2 (20.0)	2 (20.0)
Tumor stenosis						
Yes	23 (40.4)	11 (33.3)	4 (40.0)	5 (45.5)	3 (30.0)	4 (40.0)
No	34 (59.6)	22 (66.7)	6 (60.0)	6 (54.5)	7 (70.0)	6 (60.0)
Neoadjuvant irradiation						
Yes	25 (43.9)	16 (48.5)	4 (40.0)	4 (36.4)	4 (40.0)	4 (40.0)
No	32 (56.1)	17 (51.5)	6 (60.0)	7 (63.6)	6 (60.0)	6 (60.0)
Previous intraabdominal surgery						
Yes	20 (35.1)	10 (30.3)	1 (10.0)	0	1 (10.0)	3 (30.0)
No	37 (64.9)	23 (69.7)	9 (90.0)	11 (100.0)	9 (90.0)	7 (70.0)
Surgical resection technique						
LAR, TME	33 (57.9)	18 (54.5)	4 (40.0)	7 (63.6)	6 (60.0)	4 (40.0)
ISR, TME	9 (15.8)	5 (15.2)	1 (10.0)	0	1 (10.0)	2 (20.0)
APR, TME	6 (10.5)	4 (12.1)	1 (10.0)	0	0	2 (20.0)
AR, PME	9 (15.8)	6 (18.2)	4 (40.0)	4 (36.4)	3 (30.0)	2 (20.0)
T status (for rectal cancers, n = 54)						
pT0	4 (7.4)	2 (6.5)	0	1 (9.1)	0	2 (20.0)
pTis	1 (1.9)	0	0	0	0	0
pT1	8 (14.8)	4 (12.9)	3 (30.0)	2 (18.2)	3 (30.0)	3 (30.0)
pT2	14 (25.9)	9 (29.0)	2 (20.0)	1 (9.1)	2 (20.0)	3 (30.0)
pT3a	15 (27.8)	10 (32.3)	3 (30.0)	6 (54.5)	3 (30.0)	1 (10.0)
pT3b	12 (22.2)	6 (19.4)	2 (20.0)	1 (9.1)	2 (20.0)	1 (10.0)
N status (for rectal cancers, n = 54)						
pN0	32 (59.3)	14 (45.2)	4 (40.0)	4 (36.4)	4 (40.0)	5 (50.0)
pN1	14 (25.9)	10 (32.3)	5 (50.0)	3 (27.3)	5 (50.0)	3 (30.0)
pN2a	5 (9.3)	4 (12.9)	1 (10.0)	2 (18.2)	0	1 (10.0)
pN2b	3 (5.6)	3 (9.7)	0	2 (18.2)	1 (10.0)	1 (10.0)

Table 2

Tab. 2 Summary of performance metrics for recognition of surgical phases using a ResNet50 alone and in combination with an LSTM network. For each metric, mean and standard deviation are displayed.

Model	No pre-training		Self-supervised pre-training	
	F1	Accuracy	F1	Accuracy
ResNet50	0.68 ± 0.01	0.74 ± 0.02	0.68 ± 0.01	0.74 ± 0.02
ResNet50 + LSTM	0.79 ± 0.02	0.82 ± 0.01	0.78 ± 0.01	0.83 ± 0.02

Table 3

Tab. 3 Summary of performance metrics for phase-specific semantic segmentation of (anatomical) structures and dissection planes. For each metric, mean and standard deviation are displayed. Abbreviations: Intersection over union (IoU), Lateral mobilization (LM), Medial mobilization (MM), Mesorectal excision (ME), Vascular dissection (VD).

Phase	Target structure	F1	IoU	Precision	Recall	Specificity
VD	Inferior mesenteric artery	0.18 ± 0.11	0.16 ± 0.09	0.20 ± 0.12	0.20 ± 0.12	0.23 ± 0.14
	Inferior mesenteric vein	0.27 ± 0.19	0.25 ± 0.17	0.29 ± 0.19	0.29 ± 0.19	0.32 ± 0.21
	Plastic clip	0.59 ± 0.22	0.55 ± 0.21	0.60 ± 0.22	0.61 ± 0.22	0.64 ± 0.24
	Metal clip	0.20 ± 0.10	0.17 ± 0.08	0.23 ± 0.10	0.23 ± 0.10	0.26 ± 0.14
MM	Gerota's fascia	0.78 ± 0.05	0.74 ± 0.03	0.80 ± 0.05	0.81 ± 0.05	0.83 ± 0.06
	Mesocolon	0.71 ± 0.07	0.65 ± 0.05	0.73 ± 0.07	0.74 ± 0.07	0.77 ± 0.09
	Dissection line (MM)	0.05 ± 0.02	0.04 ± 0.01	0.06 ± 0.02	0.06 ± 0.03	0.10 ± 0.08
	Exploration area	0.16 ± 0.07	0.14 ± 0.05	0.18 ± 0.07	0.19 ± 0.08	0.23 ± 0.12
LM	Abdominal wall	0.82 ± 0.06	0.78 ± 0.04	0.84 ± 0.06	0.85 ± 0.05	0.87 ± 0.06
	Adhesion	0.09 ± 0.07	0.08 ± 0.06	0.11 ± 0.09	0.11 ± 0.08	0.13 ± 0.11
	Colon	0.49 ± 0.10	0.46 ± 0.09	0.51 ± 0.11	0.51 ± 0.11	0.54 ± 0.12
	Fat	0.68 ± 0.11	0.64 ± 0.10	0.71 ± 0.12	0.71 ± 0.12	0.73 ± 0.12
	Small intestine	0.35 ± 0.25	0.33 ± 0.24	0.36 ± 0.25	0.36 ± 0.25	0.37 ± 0.26
ME	Mesorectum	0.48 ± 0.10	0.45 ±0.08	0.50 ± 0.10	0.50 ± 0.10	0.53 ± 0.12
	Dissection plane (ME)	0.32 ± 0.09	0.28 ±0.08	0.34 ± 0.10	0.35 ± 0.10	0.39 ± 0.12
	Dissection line (ME)	0.05 ± 0.02	0.04 ±0.02	0.08 ± 0.03	0.06 ± 0.03	0.12 ± 0.12
	Seminal vesicles	0.09 ± 0.10	0.08 ±0.09	0.10 ± 0.10	0.10 ± 0.10	0.11 ± 0.12

FIGURE CAPTIONS

Figure 1

Fig. 1 Annotated surgical phases and phase-specifically segmented target structures.

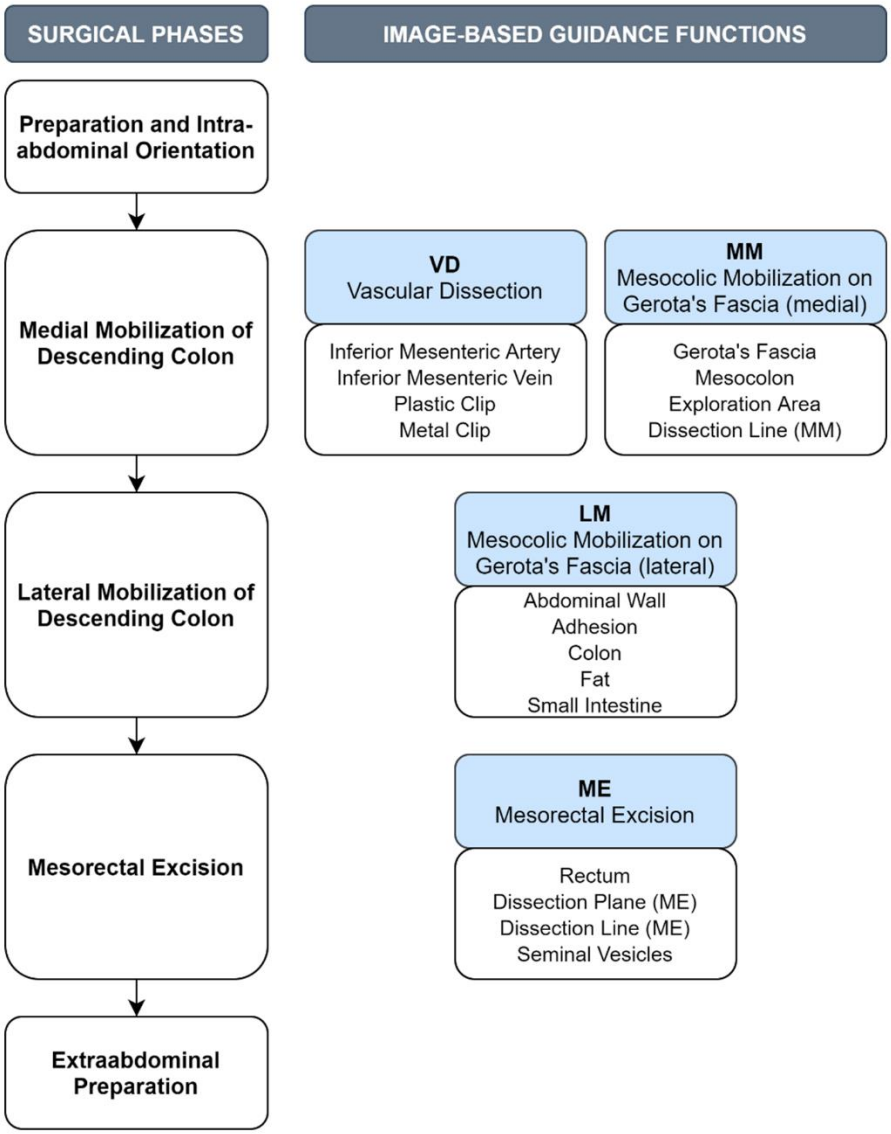


Figure 2

Fig. 2 Schematic illustration of the networks used for phase identification (A) and semantic segmentation (B). (A) A neural network consisting of a ResNet50 backbone with or without an LSTM was used for temporal prediction of surgical phases. Non-annotated videos were used for self-supervised pre-training of the algorithms. The model was validated using 4-fold cross

validation. **(B)** For spatial segmentation, a neural network based on a ResNet and a feature pyramid network as backbone was used. The network was pre-trained on the ImageNet dataset and validated using leave-one-out cross validation.

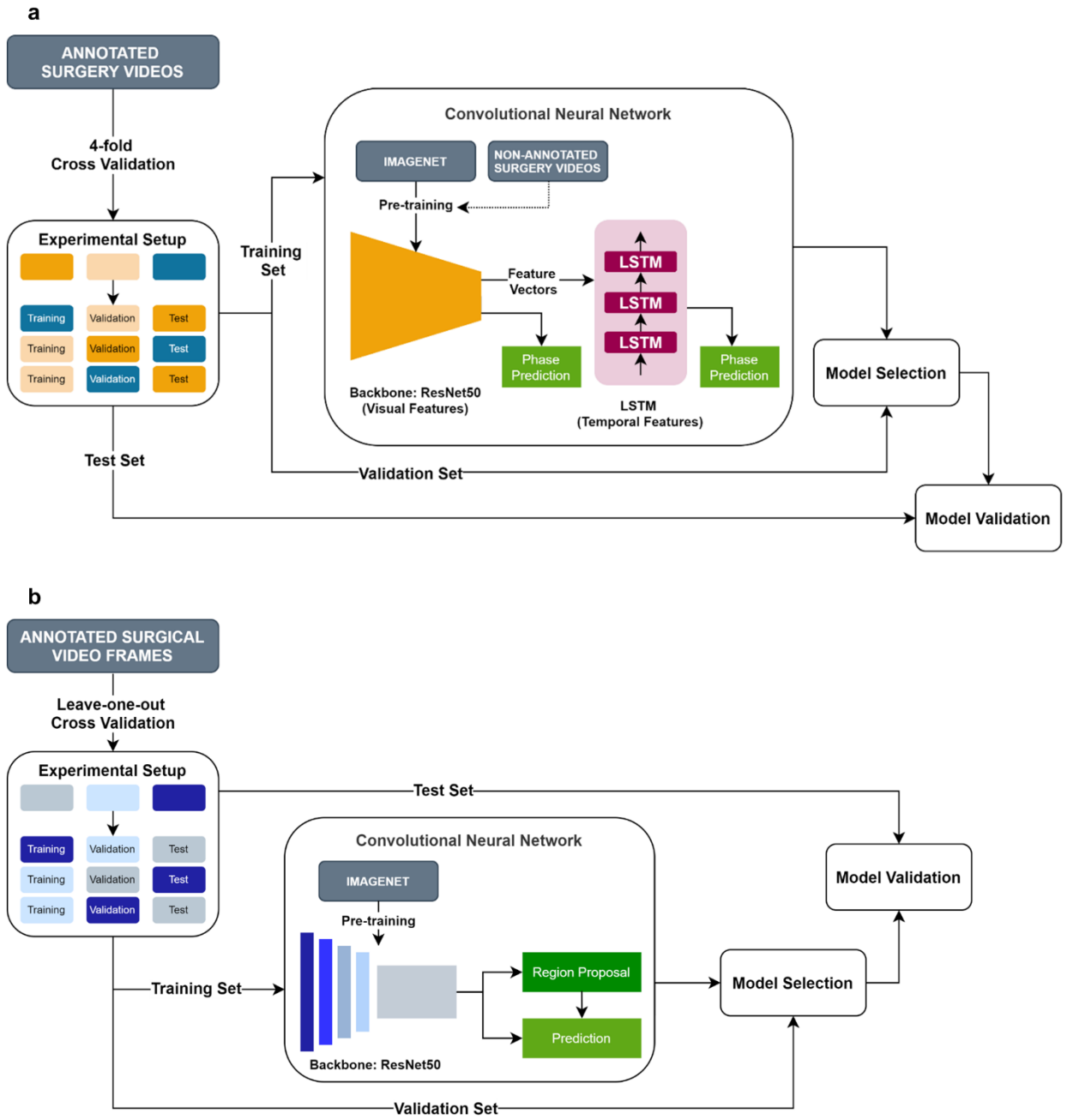


Figure 3

Fig. 3 Phase-specific annotation and model outputs for recognition of anatomical structures and dissection planes. Vascular Dissection (VD): Inferior mesenteric artery (red), plastic clip (green), and metal clip (orange) are displayed. **Medial mobilization (MM)**: Mesocolon

(light green), Gerota's fascia (light pink), dissection line (yellow), and exploration area (pink) are displayed. **Lateral mobilization (LM)**: Abdominal wall (red), colon (light pink), fat (yellow), and adhesions (blue) are displayed. **Mesorectal excision (ME)**: Mesorectum (light brown), dissection plane (green), and dissection line (red) are displayed.

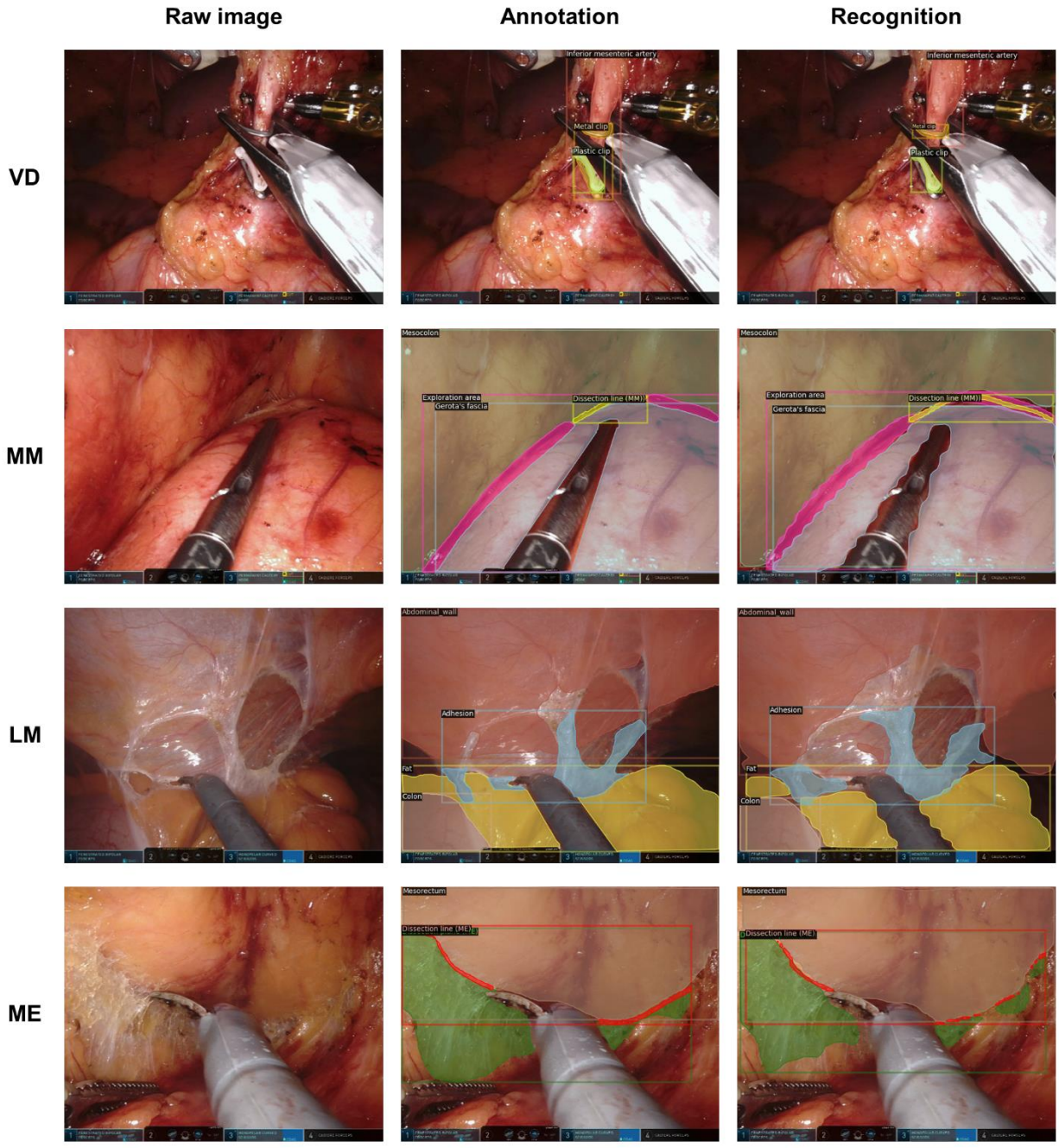
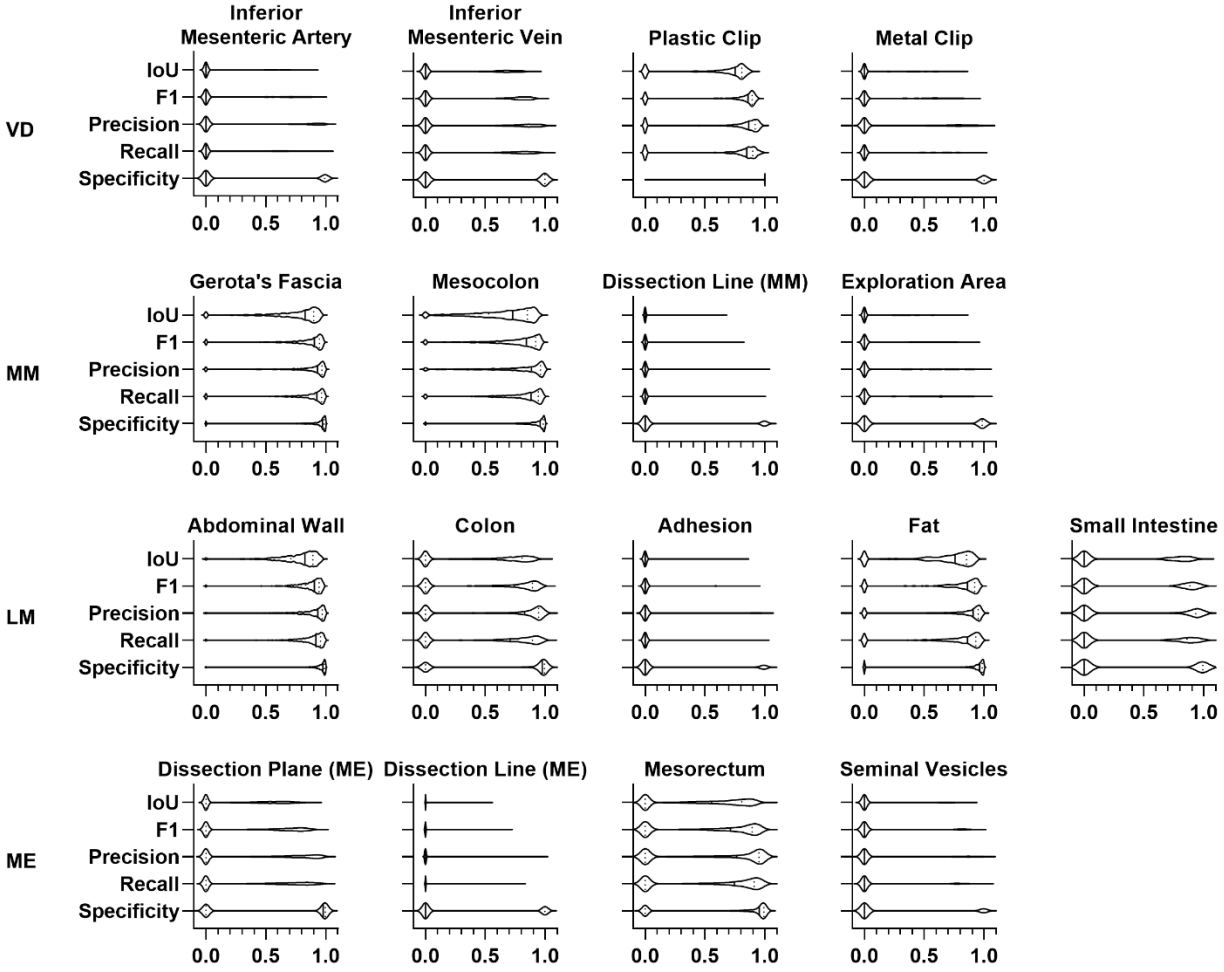


Figure 4

Fig. 4 Violin plot illustrations of the performance metrics for model outputs for phase-specific segmentation of anatomical structures and dissection planes. The median and quartiles are illustrated as solid and dashed lines, respectively. Abbreviations: Intersection over union (IoU), Lateral mobilization (LM), Medial mobilization (MM), Mesorectal excision (ME), Vascular dissection (VD).



ABBREVIATIONS

AI	Artificial Intelligence
AUC	Area under the curve
CNN	Convolutional neural network
F1	F1 score
IoU	Intersection over union
LM	Lateral mobilization (phase)
ME	Mesorectal excision (phase)
MM	Medial mobilization (phase)
PME	Partial mesorectal excision
SD	Standard deviation
TME	Total mesorectal excision
VD	Vascular dissection (phase)

ACKNOWLEDGEMENTS AND FUNDING

FRK, SL, JW, and SS were supported through project funding within the Else Kröner Fresenius Center for Digital Health (EKFZ), Dresden, Germany (project "CoBot"). FRK received funding from the Medical Faculty of the Technical University Dresden within the MedDrive Start program (grant number 60487). FMR received a doctoral student scholarship from the *Carus Promotionskolleg* Dresden.

AUTHOR CONTRIBUTIONS

FRK, SL, JW, MD, and SS conceptualized the study. FRK, MC, FMR, and TN collected and annotated clinical and video data and contributed to data analysis. SL, SK, AC, and SB implemented and trained the neural networks and contributed to data analysis. TW and JK gave substantial clinical input. JW, MD, JF, and SS supervised the project, provided infrastructure and gave important scientific input. FRK drafted the initial manuscript text. All authors reviewed, edited, and approved the final manuscript.

COMPETING INTERESTS

The authors declare no conflicts of interest.

DATA AVAILABILITY

The datasets generated and analysed during the current study are available from the corresponding author (Prof. Stefanie Speidel, stefanie.speidel@nct-dresden.de) on reasonable request. To gain access, data requestors will need to sign a data access agreement.

CODE AVAILABILITY

The developed framework in this current study based on the Detectron2 model[19] using the PyTorch library[34]. The source code for surgical phase recognition and the source code for image segmentation are publicly available: https://gitlab.com/nct_tso_public/cobot_phase_recognition (phase recognition) and https://gitlab.com/nct_tso_public/cobot_segmentation (segmentation).