Anatomy Segmentation in Laparoscopic Surgery: Comparison of Machine Learning and Human Expertise

Fiona R. Kolbinger, MD[1,2,3,✉], Franziska M. Rinner[1], Alexander C. Jenke[4], Matthias Carstens[1], Stefan Leger, PhD[3,4], Marius Distler, MD[1,2], Jürgen Weitz, MD[1,2,3,5], Stefanie Speidel, PhD[3,4,5], Sebastian Bodenstedt, PhD[4,5,✉]

[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
[5] Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI), Technische Universität Dresden, 01062, Dresden, Germany

✉ Corresponding authors:
Dr. Fiona Kolbinger, Department of Visceral, Thoracic and Vascular Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Fetscherstraße 74, 01307 Dresden, Germany
☎ +49 (0) 351 458 19624
✉ fiona.kolbinger@uniklinikum-dresden.de

Dr. Sebastian Bodenstedt, Division of Translational Surgical Oncology, National Center for Tumor Diseases (NCT/UCC), Partner Site Dresden, Fetscherstrasse 74, 01307 Dresden, Germany
☎ +49 (0) 351 5413
✉ sebastian.bodenstedt@nct-dresden.de

Structured Abstract

Background: Lack of anatomy recognition represents a clinically relevant risk in abdominal surgery. Machine learning (ML) methods can help identify visible patterns and risk structures, however, their practical value remains largely unclear.

Materials and Methods: Based on a novel dataset of 13195 laparoscopic images with pixel-wise segmentations of eleven anatomical structures, we developed specialized segmentation models for each structure and a combined model for all anatomical structures, and compared segmentation performance of both algorithms to a cohort of 28 physicians, medical students, and medical laypersons using the example of pancreas segmentation.

Results: Mean Intersection-over-Union for segmentation of intraabdominal structures ranged from 0.28 to 0.83 and from 0.23 to 0.77 for the structure-specific and the combined semantic segmentation model, respectively. At average inference times per model of 28 ms and 71 ms,

respectively, both variants are capable of near-real-time operation. Both models outperformed 26 out of 28 human participants in pancreas segmentation.

Conclusions: These results demonstrate that ML methods have the potential to provide relevant assistance in anatomy recognition in minimally-invasive surgery in near-real-time. Future research should investigate the educational value and subsequent clinical impact of respective assistance systems.

HIGHLIGHTS

- Based on a novel large-scale dataset of 13195 laparoscopic images, two machine learning models were developed for automated identification and delineation of 11 anatomical structures: One model for each structure (abdominal wall, colon, intestinal vessels (inferior mesenteric artery and inferior mesenteric vein with their subsidiary vessels), liver, pancreas, small intestine, spleen, stomach, ureter and vesicular glands) and a combined model for all structures.
- The structure-specific and combined segmentation models demonstrated similar performance in identifying intraabdominal structures and can operate in near-real-time.
- Both models outperformed 26 out of 28 human participants in pancreas segmentation, demonstrating their potential for real-time assistance in recognizing anatomical landmarks during minimally-invasive surgery.

KEYWORDS

Minimally-invasive surgery, laparoscopy, surgical data science, surgical anatomy, surgical innovation, artificial intelligence

INTRODUCTION

Computer vision describes the computerized analysis of digital images aiming at the automation of human visual capabilities, most commonly using machine learning methods, in particular deep learning. This approach has transformed medicine in recent years, with successful applications including computer-aided diagnosis of colonic polyp dignity in endoscopy[1,2], detection of clinically actionable genetic alterations in histopathology[3], and melanoma detection in dermatology[4]. Availability of large amounts of training data is the defining prerequisite for successful application of deep learning methods. With the establishment of laparoscopy as the gold standard for a variety of surgical procedures[5–8] and the increasing availability of computing resources, these concepts have gradually been applied to abdominal surgery. The overwhelming majority of research efforts in the field of Artificial Intelligence (AI)-based analysis of intraoperative surgical imaging data (i.e. video data from laparoscopic or open surgeries) has focused on classifying images with respect to the presence and/or location of previously annotated surgical instruments or anatomical structures[9–13] or on analysis of surgical proficiency[14–16] based on recorded procedures. However, almost all research endeavors in the field of computer vision in laparoscopic surgery have concentrated on preclinical stages and to date, no AI model based on intraoperative surgical imaging data could demonstrate a palpable clinical benefit.[17,18] Among the studies closest to clinical application are recent works on identification of instruments and hepatobiliary anatomy during cholecystectomy for automated assessment of the critical view of safety[13], and on the automated segmentation of safe and unsafe preparation zones during cholecystectomy[19].

In surgery, patient outcome heavily depends on experience and performance of the surgical team.[20,21] In a recent analysis of Human Performance Deficiencies in major cardiothoracic, vascular, abdominal transplant, surgical oncology, acute care, and general surgical operations, more than half of the cases with postoperative complications were associated with identifiable human error. Among these errors, lack of recognition (including misidentified anatomy) accounted for 18.8%, making it the most common Human Performance Deficiency overall.[22] While AI-based systems identifying anatomical risk and target structures would theoretically have the potential to alleviate this risk, limited availability and diversity of (annotated) laparoscopic image data drastically restrict the clinical potential of such applications in practice.

To advance and diversify the applications of computer vision in laparoscopic surgery, we have recently published the Dresden Surgical Anatomy Dataset[23], providing 13195 laparoscopic images with high-quality annotations of the presence and exact location of eleven intraabdominal anatomical structures: abdominal wall, colon, intestinal vessels (inferior mesenteric artery and inferior mesenteric vein with their subsidiary vessels), liver, pancreas, small intestine, spleen, stomach, ureter and vesicular glands. Here, we present the first study evaluating automated

detection and localization of organs and anatomical structures in laparoscopic view based on this dataset, and, using the example of delineation of the pancreas, compare algorithm performance to that of humans.

METHODS

Patient cohort

Video data from 32 robot-assisted anterior rectal resections or rectal extirpations were gathered at the University Hospital Carl Gustav Carus Dresden between February 2019 and February 2021. All included patients had a clinical indication for the surgical procedure, recommended by an interdisciplinary tumor board. The procedures were performed using the da Vinci® Xi system (Intuitive Surgical, Sunnyvale, CA, USA) with a standard Da Vinci® Xi/X Endoscope with Camera (8 mm diameter, 30˚ angle, Intuitive Surgical, Sunnyvale, CA, USA, Item code 470057). 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.

All experiments were performed in accordance with the ethical standards of the Declaration of Helsinki 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). The trial was registered on clinicaltrials.gov (trial registration ID: NCT05268432). Written informed consent to laparoscopic image data acquisition, data annotation, data analysis, and anonymized data publication was obtained from all participants. Before publication, all data was anonymized according to the general data protection regulation of the European Union.

Dataset

Based on the full-length surgery recordings and respective temporal annotations of organ visibility, individual image frames were extracted and annotated as described previously.[23] The resulting Dresden Surgical Anatomy Dataset comprises 13195 distinct images with pixel-wise segmentations of eleven anatomical structures: abdominal wall, colon, intestinal vessels (inferior mesenteric artery and inferior mesenteric vein with their subsidiary vessels), liver, pancreas, small intestine, spleen, stomach, ureter and vesicular glands. Moreover, the dataset comprises binary annotations of the presence of each of these organs for each image. The dataset is publicly available via the following link: https://doi.org/10.6084/m9.figshare.21702600.

For machine learning purposes, the Dresden Surgical Anatomy Dataset was split into training, validation, and test data as follows (Figure 1):

- Training set (at least 12 surgeries per anatomical structure): surgeries 1, 4, 5, 6, 8, 9, 10, 12, 15, 16, 17, 19, 22, 23, 24, 25, 27, 28, 29, 30, 31.
- Validation set (3 surgeries per anatomical structure): surgeries 3, 21, 26.
- Test set (5 surgeries per anatomical structure): surgeries 2, 7, 11, 13, 14, 18, 20, 32.

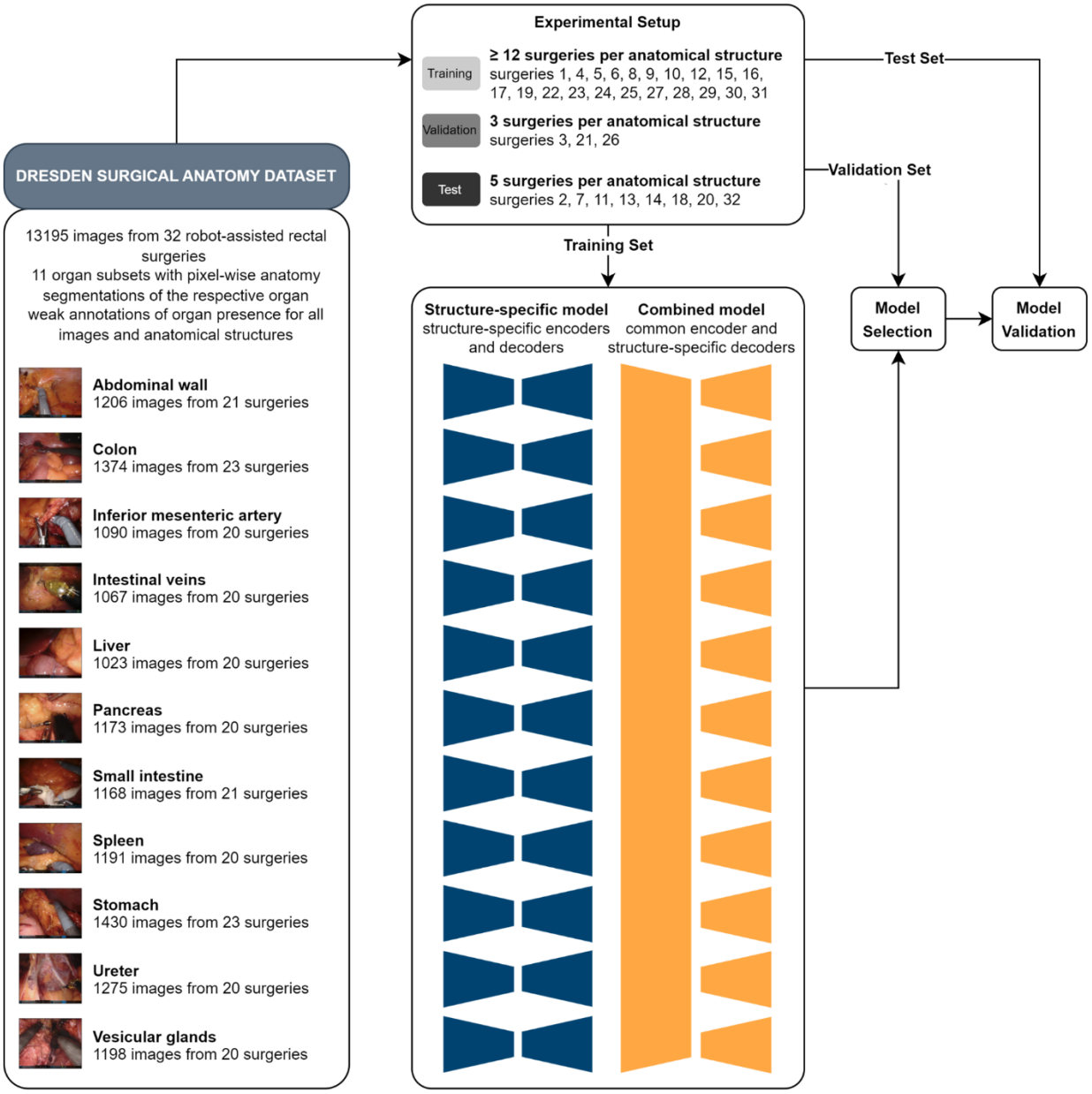

Figure 1: Schematic illustration of the structure-specific and combined machine learning models used for semantic segmentation. The Dresden Surgical Anatomy Dataset was split into a training, a validation, and a test set. For spatial segmentation, two machine learning models were trained: A structure-specific model with individual encoders and decoders, and a combined model with a common encoder and structure-specific decoders.

This split is proposed for future works using the Dresden Surgical Anatomy Dataset to reproduce the variance of the entire dataset within each subset, and to ensure comparability regarding clinical variables between the training, the validation, and the test set. Surgeries for the test set were

selected to minimize variance regarding the number of frames over the segmented classes. Out of the remaining surgeries, the validation set was separated from the training set using the same criterion.

Structure-specific semantic segmentation model

To segment each anatomical structure, a separate convolutional neural network for segmentation a DeeplabV3[24] model with a ResNet50 backbone with default PyTorch pretraining on the COCO dataset[25], was used. The networks were trained using cross-entropy loss and the AdamW optimizer[26] for 100 epochs with a starting learning rate of 10^{-4} and a linear learning rate scheduler decreasing the learning rate by 0.9 every 10 epochs. For data augmentation, we applied random scaling and rotation, as well as brightness adjustments. The final model for each organ was selected via the Intersection-over-Union (IoU) on the validation dataset and evaluated using the Dresden Surgical Anatomy Dataset with the abovementioned training-validation-test split (Figure 1).

Segmentation performance was assessed using F1 score, IoU, precision, recall, and specificity on the test folds. These parameters are commonly used technical measures of prediction exactness, ranging from 0 (least exact prediction) to 1 (entirely correct prediction without any misprediction).

Combined semantic segmentation model

A convolutional neural network with a common encoder and eleven decoders for combined segmentation of the eleven anatomical structures was trained. The used architecture is an extension of DeepLabV3[24]. A shared ResNet50 backbone with default PyTorch pretraining on the COCO dataset[25], was used. For each class, a DeepLabV3 decoder was then run on the features extracted from a given image by the backbone. As the images are only annotated for binary classes, the loss is only calculated for every pixel in images, in which the structure associated with the current decoder is annotated. For images, in which the associated class is not annotated, only the pixels that are annotated as belonging to another class are included in the loss, e.g., pixels that were annotated as belonging to the class "liver" can be used as negative examples for the class "pancreas". The remaining training procedure was identical to the structure-specific model. The model was trained and evaluated using the Dresden Surgical Anatomy Dataset with the abovementioned training-validation-test split (Figure 1).

Segmentation performance was assessed using F1 score, IoU, precision, recall, and specificity on the test folds.

Comparative evaluation of algorithmic and human performance

To determine the clinical potential of automated segmentation of anatomical risk structures, the segmentation performance of 28 humans was compared to that of the structure-specific and the combined semantic segmentation models using the example of the pancreas. The local Institutional Review Board (ethics committee at the Technical University Dresden) reviewed and approved this study (approval number: BO-EK-566122021). All participants provided written informed consent to anonymous study participation, data acquisition and analysis, and publication. In total, 28 participants (physician and non-physician medical staff, medical students, and medical laypersons) marked the pancreas in 35 images from the Dresden Surgical Anatomy Dataset[23] with bounding boxes. These images originated from 26 different surgeries, and the pancreas was visible in 16 of the 35 images. Each of the previously selected 35 images was shown once, the order being arbitrarily chosen but identical for all participants. The open-source annotation software Computer Vision Annotation Tool (CVAT) was used for annotations. In cases where the pancreas was seen in multiple, non-connected locations in the image, participants were asked to create separate bounding boxes for each area.

Based on the structure-specific and the combined semantic segmentation models, axis-aligned bounding boxes marking the pancreas were generated in the 35 images from the pixel-wise segmentation. To guarantee that the respective images were not part of the training data, four-fold cross validation was used, i.e., the origin surgeries were split into four equal-sized batches, and algorithms were trained on three batches that did not contain the respective origin image before being applied to segmentation.

To compare human and algorithm performance, the bounding boxes created by each participant and the structure-specific as well as the combined semantic segmentation models were compared to bounding boxes derived from the Dresden Surgical Anatomy Dataset, which were defined as ground truth. IoU between the manual or automatic bounding box and the ground truth was used to compare segmentation accuracy.

DATA AND CODE AVAILABILITY

Data Availability

The Dresden Surgical Anatomy Dataset is publicly available via the following link: https://doi.org/10.6084/m9.figshare.21702600. All other data generated and analyzed during the

current study are available from the corresponding authors on reasonable request. To gain access, data requestors will need to sign a data access agreement.

Code Availability

The most relevant scripts used for dataset compilation are publicly available via the following link: https://zenodo.org/record/6958337#.YzsBdnZBzOg. The code used for segmentation algorithms is available at https://gitlab.com/nct_tso_public/anatomy-recognition-dsad.

RESULTS

Machine Learning-based anatomical structure segmentation in structure-specific models

Structure-specific multi-layer convolutional neural networks (Figure 1) were trained to segment the abdominal wall, the colon, intestinal vessels (inferior mesenteric artery and inferior mesenteric vein with their subsidiary vessels), the liver, the pancreas, the small intestine, the spleen, the stomach, the ureter, and vesicular glands (Supplementary Table 1). Table 1 displays mean F1 score, IoU, precision, recall, and specificity for individual anatomical structures as predicted by the structure-specific algorithms on the test data.

Out of the analyzed segmentation models, performance was lowest for vesicular glands (mean IoU: 0.28 ± 0.21), the pancreas (mean IoU: 0.28 ± 0.27), and the ureter (mean IoU: 0.36 ± 0.25), while excellent predictions were achieved for the abdominal wall (mean IoU: 0.83 ± 0.14) and the small intestine (mean IoU: 0.80 ± 0.18). In segmentation of the pancreas, the ureter, vesicular glands and intestinal vessel structures, there was a relevant proportion of images with no detection or no overlap between ground truth, while for all remaining anatomical structures, this proportion was minimal (Figure 2 a). While the images, in which the highest IoUs were observed, mostly displayed large organ segments that were clearly visible (Figure 2 b), the images with the lowest IoU were of variable quality with confounding factors such as blood, smoke, soiling of the endoscope lens, or pictures blurred by camera shake (Figure 2 c).

Figure 2: Pixel-wise organ segmentation with structure-specific models trained on the respective organ subsets of the Dresden Surgical Anatomy Dataset. (a) Violin plot illustrations of performance metrics for structure-specific segmentation models on the test dataset. The median and quartiles are illustrated as solid and dashed lines, respectively. **(b)** Example images from the test dataset with the highest IoUs for liver, pancreas, stomach, and ureter segmentation with structure-specific segmentation models. Ground truth is displayed as blue line (upper panel), model segmentations are displayed as white overlay (lower panel). **(c)** Example images from the test dataset with the lowest IoUs for liver, pancreas, stomach, and ureter segmentation with structure-specific segmentation models. Ground truth is displayed as blue line (upper panel), model segmentations are displayed as white overlay (lower panel).

Inference on a single image with a resolution of 640 x 512 pixels required, on average, 28 ms on an Nvidia A5000, resulting in a frame rate of almost 36 frames per second. This runtime includes one decoder, meaning that only the segmentation for one anatomical class is included.

Table 1: Summary of performance metrics for anatomical structure segmentation using structure-specific models based on the DeepLabv3 architecture on the test dataset. For each metric, mean and standard deviation are displayed.

Anatomical structure	F1 score	IoU	Precision	Recall	Specificity
Abdominal wall	0.90 ± 0.10	0.83 ± 0.14	0.89 ± 0.14	0.93 ± 0.07	0.97 ± 0.04
Colon	0.79 ± 0.20	0.69 ± 0.22	0.80 ± 0.21	0.82 ± 0.21	0.97 ± 0.05
Inferior mesenteric artery	0.54 ± 0.26	0.41 ± 0.22	0.55 ± 0.25	0.67 ± 0.33	0.99 ± 0.01
Intestinal veins	0.54 ± 0.33	0.44 ± 0.29	0.70 ± 0.26	0.56 ± 0.36	1.00 ± 0.00
Liver	0.80 ± 0.23	0.71 ± 0.25	0.85 ± 0.21	0.81 ± 0.24	0.98 ± 0.03
Pancreas	0.37 ± 0.32	0.28 ± 0.27	0.59 ± 0.37	0.37 ± 0.36	1.00 ± 0.01
Small intestine	0.87 ± 0.14	0.80 ± 0.18	0.87 ± 0.16	0.91 ± 0.15	0.97 ± 0.04
Spleen	0.79 ± 0.23	0.69 ± 0.24	0.74 ± 0.22	0.90 ± 0.24	0.99 ± 0.01
Stomach	0.71 ± 0.24	0.60 ± 0.25	0.65 ± 0.25	0.89 ± 0.21	0.98 ± 0.02
Ureter	0.47 ± 0.30	0.36 ± 0.25	0.53 ± 0.28	0.57 ± 0.39	1.00 ± 0.00
Vesicular glands	0.40 ± 0.25	0.28 ± 0.21	0.37 ± 0.28	0.62 ± 0.35	0.97 ± 0.03

Machine Learning-based anatomical structure segmentation in a combined model

Using annotated images from the Dresden Surgical Anatomy Dataset across anatomical structure classes, a combined model with a mutual encoder and organ-specific decoders was trained (Figure 1, Supplementary Table 2). Table 2 displays mean F1 score, IoU, precision, recall, and specificity for anatomical structure segmentation in the combined model.

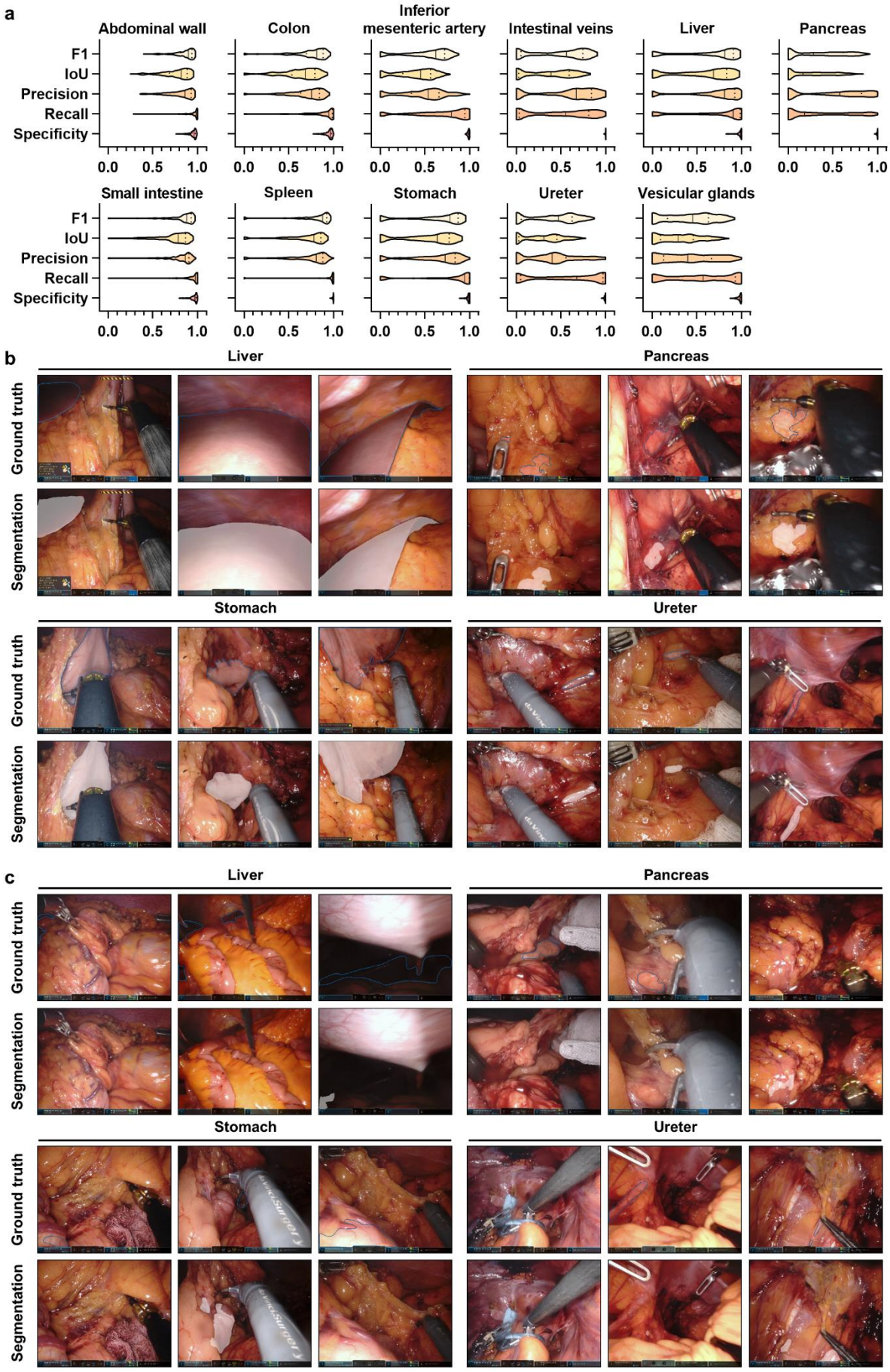

Figure 3: Pixel-wise organ segmentation with the combined model trained on the Dresden Surgical Anatomy Dataset across anatomical structure classes with a common encoder and structure-specific decoders. (a) Violin plot illustrations of performance metrics for the combined segmentation model on the test dataset. The median and quartiles are illustrated as solid and dashed lines, respectively. **(b)** Example images from the test dataset with the highest IoUs for liver, pancreas, stomach, and ureter segmentation with the combined segmentation model. Ground truth is displayed as blue line (upper panel), model segmentations are displayed as white overlay (lower panel). **(c)** Example images from the test dataset with the lowest IoUs for liver, pancreas, stomach, and ureter segmentation with the combined segmentation model. Ground truth is displayed as blue line (upper panel), model segmentations are displayed as white overlay (lower panel).

The performance of the combined model was overall similar to that of structure-specific models (Table 1), with highest segmentation performance for the abdominal wall (mean IoU: 0.77 ± 0.15) and the small intestine (mean IoU: 0.72 ± 0.21), and the lowest performance for the pancreas (mean IoU: 0.23 ± 0.29), the ureter (IoU: 0.29 ± 0.22) and vesicular glands (IoU: 0.30 ± 0.23). In comparison to the respective structure-specific models, the combined model performed notably weaker in liver segmentation, while performance for the other anatomical structures was similar. The proportion of un- or entirely mispredicted images was largest in the ureter, the pancreas, the stomach, the abdominal vessel structures, and the vesicular glands (Figure 3 a). Similar to the structure-specific models, trends towards an impact of segment size, uncommon angles of vision, endoscope lens soiling, blurry images, and presence of blood or smoke were seen when comparing image quality of well-predicted images (Figure 3 b) to images with poor or no prediction (Figure 3 c).

Table 2: Summary of performance metrics for anatomical structure segmentation using the combined model (common encoder with structure-specific decoders) on the test dataset. For each metric, mean and standard deviation are displayed.

Anatomical structure	F1 score	IoU	Precision	Recall	Specificity
Abdominal wall	0.86 ± 0.11	0.77 ± 0.15	0.81 ± 0.15	0.95 ± 0.09	0.95 ± 0.04
Colon	0.75 ± 0.19	0.63 ± 0.21	0.71 ± 0.18	0.84 ± 0.23	0.95 ± 0.04
Inferior mesenteric artery	0.53 ± 0.25	0.40 ± 0.21	0.52 ± 0.22	0.68 ± 0.32	0.99 ± 0.01
Intestinal veins	0.46 ± 0.32	0.35 ± 0.27	0.70 ± 0.23	0.48 ± 0.36	1.00 ± 0.00
Liver	0.65 ± 0.34	0.57 ± 0.33	0.76 ± 0.23	0.69 ± 0.38	0.98 ± 0.03
Pancreas	0.32 ± 0.30	0.23 ± 0.24	0.61 ± 0.33	0.32 ± 0.35	0.99 ± 0.01
Small intestine	0.81 ± 0.19	0.72 ± 0.21	0.81 ± 0.17	0.87 ± 0.23	0.96 ± 0.03
Spleen	0.78 ± 0.24	0.69 ± 0.24	0.76 ± 0.18	0.89 ± 0.26	0.99 ± 0.01
Stomach	0.63 ± 0.32	0.53 ± 0.29	0.68 ± 0.23	0.74 ± 0.37	0.98 ± 0.02
Ureter	0.40 ± 0.28	0.29 ± 0.22	0.44 ± 0.27	0.56 ± 0.40	0.99 ± 0.01
Vesicular glands	0.42 ± 0.27	0.30 ± 0.23	0.41 ± 0.30	0.56 ± 0.36	0.98 ± 0.02

Inference on a single image with a resolution of 640 x 512 pixels required, on average, 71 ms on an Nvidia A5000, resulting in a frame rate of about 14 frames per second. This runtime includes all 11 decoders, meaning that segmentations for all organ classes are included.

Performance of machine learning models in relation to human performance

To approximate the clinical value of the previously described algorithms for anatomical structure segmentation, the performances of the structure-specific and the combined model were compared to that of a cohort of 28 physicians, medical students, and persons with no medical background (Figure 4 a), and different degrees of experience in laparoscopic surgery (Figure 4 b). A vulnerable anatomical structure with – measured by classical metrics of overlap (Tables 1 and 2) – comparably weak segmentation performance of the trained algorithms, the pancreas was selected as an example.

Comparing bounding box segmentations of the pancreas of human annotators, the medical and laparoscopy-specific experience of participants was mirrored by the respective IoUs describing the overlap between the pancreas annotation and the ground truth. The pancreas-specific segmentation model (IoU: 0.29) and the combined segmentation model (IoU: 0.21) outperformed 26 out of the 28 human participants (Figures 4 c and d). Overall, these results demonstrate that the developed models have clinical potential to improve the recognition of vulnerable anatomical structures in laparoscopy.

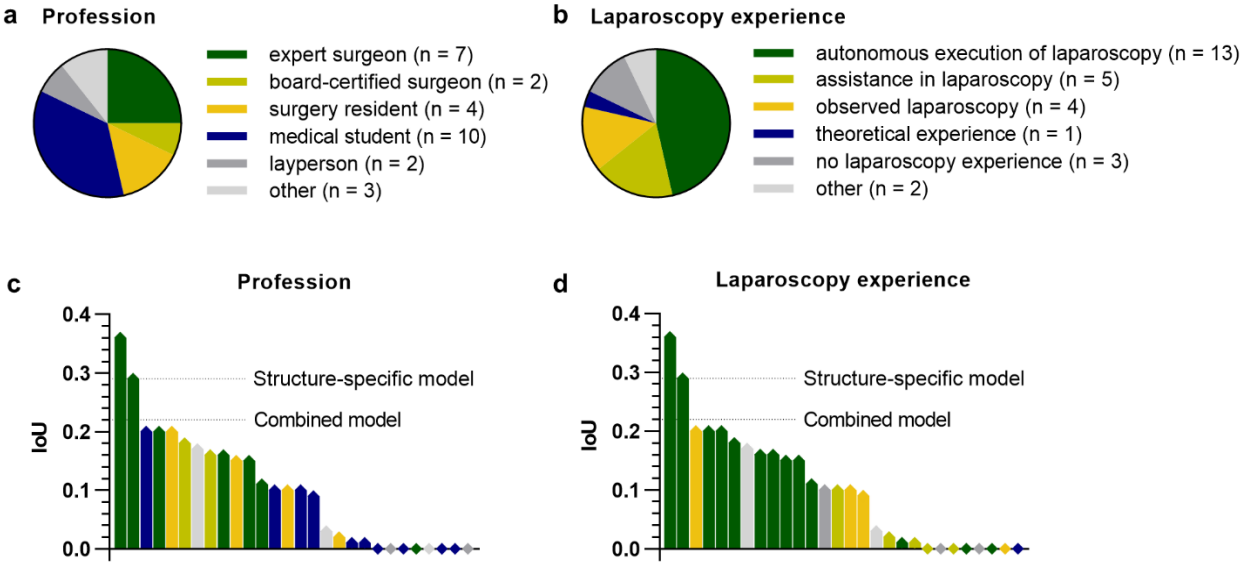

Figure 4: Comparison of pancreas segmentation performance of the structure-specific and the combined semantic segmentation models with a cohort of 28 human participants. (a) Distribution of medical and non-medical professions among human participants. **(b)** Distribution of laparoscopy experience among human participants. **(c)** Waterfall chart displaying the average pancreas segmentation IoUs of participants with different professions as compared to the IoU generated by the structure-specific and the combined semantic segmentation models. **(d)** Waterfall chart displaying the average pancreas segmentation IoUs of participants with varying laparoscopy experience as compared to the IoU generated by the structure-specific and the combined semantic segmentation models.

DISCUSSION

In surgery, misinterpretation of visual cues can result in objectifiable errors with serious consequences.[22] Based on a robust public dataset providing 13195 laparoscopic images with segmentations of eleven intra-abdominal anatomical structures, this study explores the potential of machine learning for automated segmentation of these organs, and compares algorithmic segmentation quality to that of humans with varying experience in minimally-invasive abdominal surgery.

In summary, the presented findings suggest that machine learning-based segmentation of intraabdominal organs and anatomical structures is possible and has the potential to provide clinically valuable information. At an average runtime of 71 ms per image, corresponding to a frame rate of 14 frames per second, the combined model would facilitate near-real-time identification of eleven anatomical structures. These runtimes mirror the performance of a non-optimized version of the model, which can be significantly improved using methods such as TensorRT from Nvidia. Measured by classical metrics of overlap between segmentation and ground truth, predictions were, overall, better for large and similar-appearing organs such as the abdominal wall, the liver, the stomach, and the spleen as compared to smaller and more diverse-appearing organs such as the pancreas, the ureter, or vesicular glands. Furthermore, poor image quality (i.e., images blurred by camera movements, presence of blood or smoke in images) was linked to lower accuracy of machine learning-based segmentations. These findings imply that computer vision studies in laparoscopy should be carefully interpreted taking representativity and potential selection of underlying training and validation data into consideration.

Measured by classical metrics of overlap (e.g., IoU, F1 score, precision, recall, specificity) that are commonly used to evaluate segmentation performance, the structure-specific models and the combined model performed similarly for most anatomical structures with average IoUs ranging from 0.28 to 0.83 and from 0.23 to 0.77, respectively. The structure-specific model was superior to the combined segmentation model with respect to liver segmentation. Interpretation of such metrics of overlap, however, represents a major challenge in computer vision applications in medical domains such as dermatology and endoscopy[27–29] as well as non-medical domains such as autonomous driving[30]. In the specific use case of laparoscopic surgery, evidence suggests that such technical metrics alone are not sufficient to characterize the clinical potential and utility of segmentation algorithms.[31,32] In this context, the 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, demonstrated by Tokuyasu *et al.*, supports this hypothesis.[12] In the presented analysis, the trained structure-specific and combined machine learning algorithms outperformed all human participants in the specific task of bounding box segmentation of the

pancreas except for two surgical specialists with over 10 years of experience. This suggests that even for structures such as the pancreas with seemingly poor segmentation quality (segmentation IoU of the best-performing model: 0.28 ± 0.27 in the test set) have the potential to provide clinically valuable help in anatomy recognition. Notably, the best average IoUs achieved in this comparative study were 0.29 (for the structure-specific model) and 0.36 (for the best human participant), which would both be considered less reliable segmentation quality measures on paper. This encourages further discussion about metrics for segmentation quality assessment in clinical AI. In the future, the potential of the described dataset[23] and organ segmentation algorithms could be exploited for educational purposes[33,34], for guidance systems facilitating real-time detection of risk and target structures[19,32,35], or as an auxiliary function integrated in more complex surgical assistance systems, such as guidance systems relying on automated liver registration[36].

The limitations of this work are mostly related to the dataset and general limitations of machine learning-based segmentation: First, the Dresden Surgical Anatomy Dataset is a monocentric dataset based on 32 robot-assisted rectal surgeries. Therefore, the images used for algorithm training and validation originate from one set of hardware and display organs from specific angles, which could limit generalizability and transferability of the presented findings to other centers and other minimally-invasive abdominal surgeries, particularly non-robotic procedures. Second, annotations were required for training of machine learning algorithms, potentially inducing some bias towards the way that organs were annotated in the algorithms, which may differ from individual healthcare professionals' way of recognizing an organ. This is particularly relevant for organs such as the ureters or the pancreas, which often appear covered by layers of tissue. Here, computer vision-based algorithms that solely consider the laparoscopic images provided by the Dresden Surgical Anatomy Dataset for identification of risk structures will only be able to identify an organ once it is visible. For an earlier recognition of such hidden risk structures, more training data with meaningful annotations would be necessary. Importantly, the presented comparison to human performance focused on segmentation of visible anatomy as well, neglecting that humans (and possibly computers, too) could already identify a risk structure hidden underneath layers of tissue. Third, the dataset only includes individual annotated images. In some structures such as the ureter, video data offers considerably more information than still image data. In this context, it is conceivable that an incorporation of temporal aspects could result in major improvements of both human and algorithm recognition performance. The existing limitations notwithstanding, the presented study represents an important addition to the growing body of research on medical image analysis in laparoscopic surgery, particularly by linking technical metrics to human performance.

In conclusion, this study demonstrates that machine learning methods have the potential to provide clinically relevant near-real-time assistance in anatomy recognition in minimally-invasive surgery. This study is the first to use the recently published Dresden Surgical Anatomy Dataset, providing baseline algorithms for organ segmentation and evaluating the clinical relevance of such algorithms by introducing more clinically meaningful comparators beyond classical computer vision metrics. Future research should investigate other segmentation methods, the transferability of these results to other surgical procedures, and the clinical impact of real-time surgical assistance systems and didactic applications based on automated segmentation algorithms.

ACKNOWLEDGMENTS

Assistance with the study

The authors gratefully acknowledge excellent project coordination by Dr. Elisabeth Fischermeier and Dr. Grit Krause-Jüttler.

Financial support and sponsorship

This work has been funded by the Else Kröner Fresenius Center for Digital Health (EKFZ), Dresden, Germany (project "CoBot"), by the German Research Foundation DFG within the Cluster of Excellence EXC 2050: "Center for Tactile Internet with Human-in-the-Loop (CeTI)" (project number 390696704) and by the German Federal Ministry of Health (BMG) within the "Surgomics" project (grant number BMG 2520DAT82). Furthermore, FRK received funding from the Medical Faculty of the Technical University Dresden within the MedDrive Start program (grant number 60487) and from the Joachim Herz Foundation (Add-On Fellowship for Interdisciplinary Life Science). FMR received a doctoral student scholarship from the *Carus Promotionskolleg* Dresden.

Conflicts of interest

The authors declare no conflicts of interest.

Presentation

None.

REFERENCES

1. Wang P, Liu X, Berzin TM, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol Hepatol*. 2020;5(4):343-351. doi:10.1016/S2468-1253(19)30411-X
2. Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*. 2019;68(10):1813-1819. doi:10.1136/GUTJNL-2018-317500
3. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer 2020 18*. 2020;1(8):789-799. doi:10.1038/s43018-020-0087-6
4. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056
5. Simillis C, Lal N, Thoukididou SN, et al. Open Versus Laparoscopic Versus Robotic Versus Transanal Mesorectal Excision for Rectal Cancer: A Systematic Review and Network Meta-analysis. *Ann Surg*. 2019;270(1):59-68. doi:10.1097/SLA.0000000000003227
6. Zhao JJ, Syn NL, Chong C, et al. Comparative outcomes of needlescopic, single-incision laparoscopic, standard laparoscopic, mini-laparotomy, and open cholecystectomy: A systematic review and network meta-analysis of 96 randomized controlled trials with 11,083 patients. *Surgery*. 2021;170(4):994-1003. doi:10.1016/J.SURG.2021.04.004
7. Luketich JD, Pennathur A, Awais O, et al. Outcomes after minimally invasive esophagectomy: review of over 1000 patients. *Ann Surg*. 2012;256(1):95-103. doi:10.1097/SLA.0B013E3182590603
8. Thomson JE, Kruger D, Jann-Kruger C, et al. Laparoscopic versus open surgery for complicated appendicitis: a randomized controlled trial to prove safety. *Surg Endosc*. 2015;29(7):2027-2032. doi:10.1007/S00464-014-3906-Y
9. Islam M, Atputharuban DA, Ramesh R, Ren H. Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. *IEEE Robot Autom Lett*. 2019;4(2):2188-2195. doi:10.1109/LRA.2019.2900854
10. Roß T, Reinke A, Full PM, et al. Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge. *Med Image Anal*. 2020;70:101920. doi:10.1016/j.media.2020.101920
11. Shvets AA, Rakhlin A, Kalinin AA, Iglovikov VI. Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning. In: *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*. Institute of Electrical and

Electronics Engineers Inc.; 2019:624-628. doi:10.1109/ICMLA.2018.00100

12. Tokuyasu T, Iwashita Y, Matsunobu Y, et al. Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surg Endosc 2020 354*. 2020;35(4):1651-1658. doi:10.1007/S00464-020-07548-X

13. Mascagni P, Vardazaryan A, Alapatt D, et al. Artificial Intelligence for Surgical Safety: Automatic Assessment of the Critical View of Safety in Laparoscopic Cholecystectomy Using Deep Learning. *Ann Surg*. Published online 2020. doi:10.1097/SLA.0000000000004351

14. Jin A, Yeung S, Jopling J, 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;2018-January:691-699. Accessed July 14, 2021. https://arxiv.org/abs/1802.08774v2

15. Funke I, Bodenstedt S, Oehme F, von Bechtolsheim F, Weitz J, Speidel S. Using 3D Convolutional Neural Networks to Learn Spatiotemporal Features for Automatic Surgical Gesture Recognition in Video. *Med Image Comput Comput Assist Interv – MICCAI 2019 Lect Notes Comput Sci*. 2019;11768:467-475. doi:10.1007/978-3-030-32254-0_52/TABLES/2

16. Lavanchy JL, Zindel J, Kirtac K, et al. Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci Reports 2021 111*. 2021;11(1):1-9. doi:10.1038/s41598-021-84295-6

17. Maier-Hein L, Eisenmann M, Sarikaya D, et al. Surgical data science – from concepts toward clinical translation. *Med Image Anal*. 2022;76:102306. doi:10.1016/J.MEDIA.2021.102306

18. Kolbinger FR, Leger S, Carstens M, et al. Artificial Intelligence for context-aware surgical guidance in complex robot-assisted oncological procedures: an exploratory feasibility study. *medRxiv*. Published online May 3, 2022:2022.05.02.22274561. doi:10.1101/2022.05.02.22274561

19. Madani A, Namazi B, Altieri MS, et al. Artificial Intelligence for Intraoperative Guidance. *Ann Surg*. Published online 2020. doi:10.1097/sla.0000000000004594

20. Fecso AB, Szasz P, Kerezov G, Grantcharov TP. The effect of technical performance on patient outcomes in surgery. *Ann Surg*. 2017;265(3):492-501. doi:10.1097/SLA.0000000000001959

21. Mazzocco K, Petitti DB, Fong KT, et al. Surgical team behaviors and patient outcomes. *Am J Surg*. 2009;197(5):678-685. doi:10.1016/J.AMJSURG.2008.03.002

22. Suliburk JW, Buck QM, Pirko CJ, et al. Analysis of Human Performance Deficiencies Associated With Surgical Adverse Events. *JAMA Netw Open*. 2019;2(7):e198067-e198067. doi:10.1001/JAMANETWORKOPEN.2019.8067

23. Carstens M, Rinner FM, Bodenstedt S, et al. The Dresden Surgical Anatomy Dataset for Abdominal Organ Segmentation in Surgical Data Science. *Sci Data*. 2023;10(1):1-8. doi:10.1038/s41597-022-01719-2

24. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv*. Published online June 17, 2017. doi:10.48550/arxiv.1706.05587

25. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2014;8693 LNCS(PART 5):740-755. doi:10.48550/arxiv.1405.0312

26. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. *7th Int Conf Learn Represent ICLR 2019*. Published online November 14, 2017. doi:10.48550/arxiv.1711.05101

27. Renard F, Guedria S, Palma N De, Vuillerme N. Variability and reproducibility in deep learning for medical image segmentation. *Sci Rep*. 2020;10(1):1-16. doi:10.1038/s41598-020-69920-0

28. Powers DMW, Ailab. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*. Published online October 11, 2020. doi:10.48550/arxiv.2010.16061

29. Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. *JAMA*. 2019;322(24):2377-2378. doi:10.1001/JAMA.2019.18058

30. Zhang Y, Mehta S, Caspi A. Rethinking Semantic Segmentation Evaluation for Explainability and Model Selection. Published online January 21, 2021. Accessed July 27, 2021. https://arxiv.org/abs/2101.08418v1

31. Reinke A, Tizabi MD, Sudre CH, et al. Common Limitations of Image Processing Metrics: A Picture Story. *arXiv*. Published online April 12, 2021. doi:10.48550/arxiv.2104.05642

32. Hashimoto DA, Rosman G, Witkowski ER, et al. Computer Vision Analysis of Intraoperative Video: Automated Recognition of Operative Steps in Laparoscopic Sleeve Gastrectomy. *Ann Surg*. 2019;270(3):414-421. doi:10.1097/SLA.0000000000003460

33. Hu YY, Mazer LM, Yule SJ, et al. Complementing Operating Room Teaching With Video-Based Coaching. *JAMA Surg*. 2017;152(4):318-325. doi:10.1001/JAMASURG.2016.4619

34. Mizota T, Anton NE, Stefanidis D. Surgeons see anatomical structures faster and more accurately compared to novices: Development of a pattern recognition skill assessment

platform. *Am J Surg*. 2019;217(2):222-227. doi:10.1016/J.AMJSURG.2018.10.011

35. Ward TM, Mascagni P, Ban Y, et al. Computer vision in surgery. *Surgery*. 2021;169(5):1253-1256. doi:10.1016/J.SURG.2020.10.039

36. Docea R, Pfeiffer M, Bodenstedt S, et al. Simultaneous localisation and mapping for laparoscopic liver navigation : a comparative evaluation study. In: Linte CA, Siewerdsen JH, eds. *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol 11598. SPIE; 2021:8. doi:10.1117/12.2582121

ABBREVIATIONS

AI	Artificial Intelligence
IoU	Intersection-over-Union
SD	Standard deviation

AUTHOR CONTRIBUTIONS

FRK, JW, MD, SS, and SB conceptualized the study. FRK, FMR, and MC collected and annotated clinical and video data and contributed to data analysis. ACJ, SL, and SB implemented and trained the neural networks and contributed to data analysis. JW, MD, 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.