Better than humans? Machine learning-based anatomy recognition in minimally-invasive abdominal surgery

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

✉ 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

Abstract

Background: Lack of anatomy recognition represents a clinically relevant risk factor in abdominal surgery. While machine learning methods have the potential to aid in recognition of visible patterns and structures, limited availability and diversity of (annotated) laparoscopic image data restrict the clinical potential of such applications in practice. This study explores the potential of machine learning algorithms to identify and delineate abdominal organs and anatomical structures using a robust and comprehensive dataset, and compares algorithm performance to that of humans.

Methods: Based on the Dresden Surgical Anatomy Dataset providing 13195 laparoscopic images with pixel-wise segmentations of eleven anatomical structures, two machine learning algorithms were developed: individual segmentation algorithms for each structure, and a combined algorithm with a common encoder and structure-specific decoders. Performance was assessed using F1 score, Intersection-over-Union (IoU), precision, recall, and specificity. Using the example of pancreas segmentation on a sample dataset of 35 images, algorithm performance was compared to that of a cohort of 28 physicians, medical students, and medical laypersons.

Results: Mean IoU for segmentation of intraabdominal structures ranged from 0.28 to 0.83 and from 0.32 to 0.81 for the structure-specific and the combined semantic segmentation model, respectively. Average inference for the structure-specific (one anatomical structure) and the combined model (eleven anatomical structures) took 20 ms and 54 ms, respectively. The structure-specific model performed equal to or better than 27 out of 28 human participants in pancreas segmentation.

Conclusions: Machine learning 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.

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] 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[18].

In surgery, patient outcome heavily depends on experience and performance of the surgical team.[19,20] 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.[21] 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[22], 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. 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://figshare.com/s/d7a60b74989a9cab2f7f.

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.

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[23] model with a ResNet50 backbone with default PyTorch pretraining on the COCO dataset[24], was used. The networks were trained using cross-entropy loss and the AdamW optimizer[25] 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[23]. A shared ResNet50 backbone with default PyTorch pretraining on the COCO dataset[24], 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 the decoder associated with the class annotated in a given image. 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.[26]

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 semantic segmentation model 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 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 semantic segmentation model, 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 semantic segmentation model were compared to bounding boxes derived from the Dresden Surgical Anatomy Dataset, which were defined as ground truth. IoU between the manual or algorithmical bounding box and the ground truth was used to compare segmentation accuracy.

Data Availability

The Dresden Surgical Anatomy Dataset is publicly available via the following link: https://figshare.com/s/d7a60b74989a9cab2f7f. 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.

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. Table 1 displays mean F1 score, IoU, precision, recall, and specificity for individual anatomical structures as predicted by the structure-specific algorithms. Out of the analyzed segmentation models, performance was lowest for the vesicular glands (mean IoU: 0.28 ± 0.21) and the pancreas (mean IoU: 0.28 ± 0.27), 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, the vesicular glands and the intestinal veins, there was a 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, extreme zoom, or pictures blurred by camera shake (Figure 2 c).

Inference on a single image with a resolution of 640 x 512 pixels required, on average, 20 ms on an Nvidia A5000, resulting in a frame rate of 50 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. 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 all annotated images from the Dresden Surgical Anatomy Dataset, a combined model with a mutual encoder and organ-specific decoders was trained (Figure 1). Table 2 displays mean F1 score, IoU, precision, recall, and specificity for anatomical structure segmentation in the combined model. 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 (IoU: 0.81 ± 0.16) and the small intestine (IoU: 0.77 ± 0.19), and the lowest performance for the vesicular glands (IoU: 0.32 ± 0.24) and the pancreas (IoU: 0.33 ± 0.25). As for the structure-specific models, few images were not or entirely mispredicted; the proportion of such images was largest in the pancreas and the vesicular glands (Figure 3 a). Similar trends towards an impact of segment size, zoom, endoscope lens soiling, blurry images, and presence of blood or smoke were seen as for structure-specific models when comparing image quality of well-predicted images (Figure 3 b) and images with poor or no prediction (Figure 3 c).

Inference on a single image with a resolution of 640 x 512 pixels required, on average, 54 ms on an Nvidia A5000, resulting in a frame rate of about 18.5 frames per second. This runtime includes all 11 decoders, meaning that segmentations for all organ classes are included.

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

Anatomical structure	F1 score	IoU	Precision	Recall	Specificity
Abdominal wall	0.89 ± 0.12	0.81 ± 0.16	0.85 ± 0.16	0.95 ± 0.07	0.96 ± 0.05
Colon	0.79 ± 0.17	0.69 ± 0.20	0.77 ± 0.18	0.86 ± 0.19	0.96 ± 0.05
Inferior mesenteric artery	0.63 ± 0.21	0.49 ± 0.20	0.57 ± 0.22	0.78 ± 0.23	0.99 ± 0.01
Intestinal veins	0.60 ± 0.25	0.47 ± 0.22	0.60 ± 0.24	0.72 ± 0.31	0.99 ± 0.01
Liver	0.81 ± 0.19	0.72 ± 0.22	0.80 ± 0.22	0.89 ± 0.15	0.98 ± 0.05
Pancreas	0.44 ± 0.30	0.33 ± 0.25	0.54 ± 0.34	0.49 ± 0.36	0.98 ± 0.10
Small intestine	0.85 ± 0.15	0.77 ± 0.19	0.80 ± 0.19	0.94 ± 0.12	0.95 ± 0.06
Spleen	0.80 ± 0.19	0.70 ± 0.22	0.72 ± 0.20	0.95 ± 0.15	0.99 ± 0.01
Stomach	0.70 ± 0.27	0.60 ± 0.28	0.70 ± 0.26	0.81 ± 0.29	0.99 ± 0.01
Ureter	0.52 ± 0.26	0.39 ± 0.23	0.45 ± 0.25	0.71 ± 0.34	0.99 ± 0.01
Vesicular glands	0.44 ± 0.29	0.32 ± 0.24	0.47 ± 0.31	0.53 ± 0.35	0.98 ± 0.02

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 performance of the structure-specific model was compared to that of 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 participants and the machine learning model, the medical and laparoscopy-specific experience of human 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) performed equal to or better than 27 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.

Discussion

In surgery, misinterpretation of visual cues can result in objectifiable errors with serious consequences.[21] 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 54 ms per image, corresponding to a frame rate of 18.5 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 with average IoUs ranging from 0.28 to 0.83 and from 0.32 to 0.81, respectively. Interpretation of these metrics, 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 machine learning algorithm performed equal to or better than all human participants in the specific task of bounding box segmentation of the pancreas except for one expert 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: 0.33 ± 0.25 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[22] and organ segmentation algorithms could be exploited for educational purposes[33,34], for guidance systems facilitating real-time detection of risk and target structures[18,32,35], 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 display organs from specific angles, which could limit generalizability and transferability of the presented findings to 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 tissue layers. 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. 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.

References

1. Wang, P. *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.* **5**, 343–351 (2020).
2. Wang, P. *et al.* Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* **68**, 1813–1819 (2019).
3. Kather, J. N. *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer 2020 18* **1**, 789–799 (2020).
4. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
5. Simillis, C. *et al.* Open Versus Laparoscopic Versus Robotic Versus Transanal Mesorectal Excision for Rectal Cancer: A Systematic Review and Network Meta-analysis. *Ann. Surg.* **270**, 59–68 (2019).
6. Zhao, J. J. *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* **170**, 994–1003 (2021).
7. Luketich, J. D. *et al.* Outcomes after minimally invasive esophagectomy: review of over 1000 patients. *Ann. Surg.* **256**, 95–103 (2012).
8. Thomson, J. E. *et al.* Laparoscopic versus open surgery for complicated appendicitis: a randomized controlled trial to prove safety. *Surg. Endosc.* **29**, 2027–2032 (2015).
9. Islam, M., Atputharuban, D. A., Ramesh, R. & Ren, H. Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. *IEEE Robot. Autom. Lett.* **4**, 2188–2195 (2019).
10. Roß, T. *et al.* Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge. *Med. Image Anal.* **70**, 101920 (2020).
11. Shvets, A. A., Rakhlin, A., Kalinin, A. A. & Iglovikov, V. I. Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning. in *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018* 624–628 (Institute of Electrical and Electronics Engineers Inc., 2019). doi:10.1109/ICMLA.2018.00100.
12. 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).

13. 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.
14. 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).
15. Funke, I. *et al.* 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.* **11768**, 467–475 (2019).
16. Lavanchy, J. L. *et al.* Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci. Reports 2021 111* **11**, 1–9 (2021).
17. Maier-Hein, L. *et al.* Surgical data science – from concepts toward clinical translation. *Med. Image Anal.* **76**, 102306 (2022).
18. Madani, A. *et al.* Artificial Intelligence for Intraoperative Guidance. *Ann. Surg.* (2020) doi:10.1097/sla.0000000000004594.
19. Fecso, A. B., Szasz, P., Kerezov, G. & Grantcharov, T. P. The effect of technical performance on patient outcomes in surgery. *Ann. Surg.* **265**, 492–501 (2017).
20. Mazzocco, K. *et al.* Surgical team behaviors and patient outcomes. *Am. J. Surg.* **197**, 678–685 (2009).
21. Suliburk, J. W. *et al.* Analysis of Human Performance Deficiencies Associated With Surgical Adverse Events. *JAMA Netw. Open* **2**, e198067–e198067 (2019).
22. Carstens, M. *et al.* The Dresden Surgical Anatomy Dataset for abdominal organ segmentation in surgical data science. *Figshare* (2022).
23. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* (2017) doi:10.48550/arxiv.1706.05587.
24. Lin, T. Y. *et al.* Microsoft COCO: Common Objects in Context. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **8693 LNCS**, 740–755 (2014).
25. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *7th Int. Conf. Learn. Represent. ICLR 2019* (2017) doi:10.48550/arxiv.1711.05101.
26. 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).

27. Renard, F., Guedria, S., Palma, N. De & Vuillerme, N. Variability and reproducibility in deep learning for medical image segmentation. *Sci. Rep.* **10**, 1–16 (2020).
28. Powers, D. M. W. & Ailab. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* (2020) doi:10.48550/arxiv.2010.16061.
29. Parikh, R. B., Teeple, S. & Navathe, A. S. Addressing Bias in Artificial Intelligence in Health Care. *JAMA* **322**, 2377–2378 (2019).
30. Zhang, Y., Mehta, S. & Caspi, A. Rethinking Semantic Segmentation Evaluation for Explainability and Model Selection. (2021).
31. Reinke, A. *et al.* Common Limitations of Image Processing Metrics: A Picture Story. *arXiv* (2021) doi:10.48550/arxiv.2104.05642.
32. Hashimoto, D. A. *et al.* Computer Vision Analysis of Intraoperative Video: Automated Recognition of Operative Steps in Laparoscopic Sleeve Gastrectomy. *Ann. Surg.* **270**, 414–421 (2019).
33. Hu, Y. Y. *et al.* Complementing Operating Room Teaching With Video-Based Coaching. *JAMA Surg.* **152**, 318–325 (2017).
34. Mizota, T., Anton, N. E. & Stefanidis, D. Surgeons see anatomical structures faster and more accurately compared to novices: Development of a pattern recognition skill assessment platform. *Am. J. Surg.* **217**, 222–227 (2019).
35. Ward, T. M. *et al.* Computer vision in surgery. *Surgery* **169**, 1253–1256 (2021).
36. Docea, R. *et al.* Simultaneous localisation and mapping for laparoscopic liver navigation : a comparative evaluation study. in *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling* (eds. Linte, C. A. & Siewerdsen, J. H.) vol. 11598 8 (SPIE, 2021).

Figures and Figure Captions

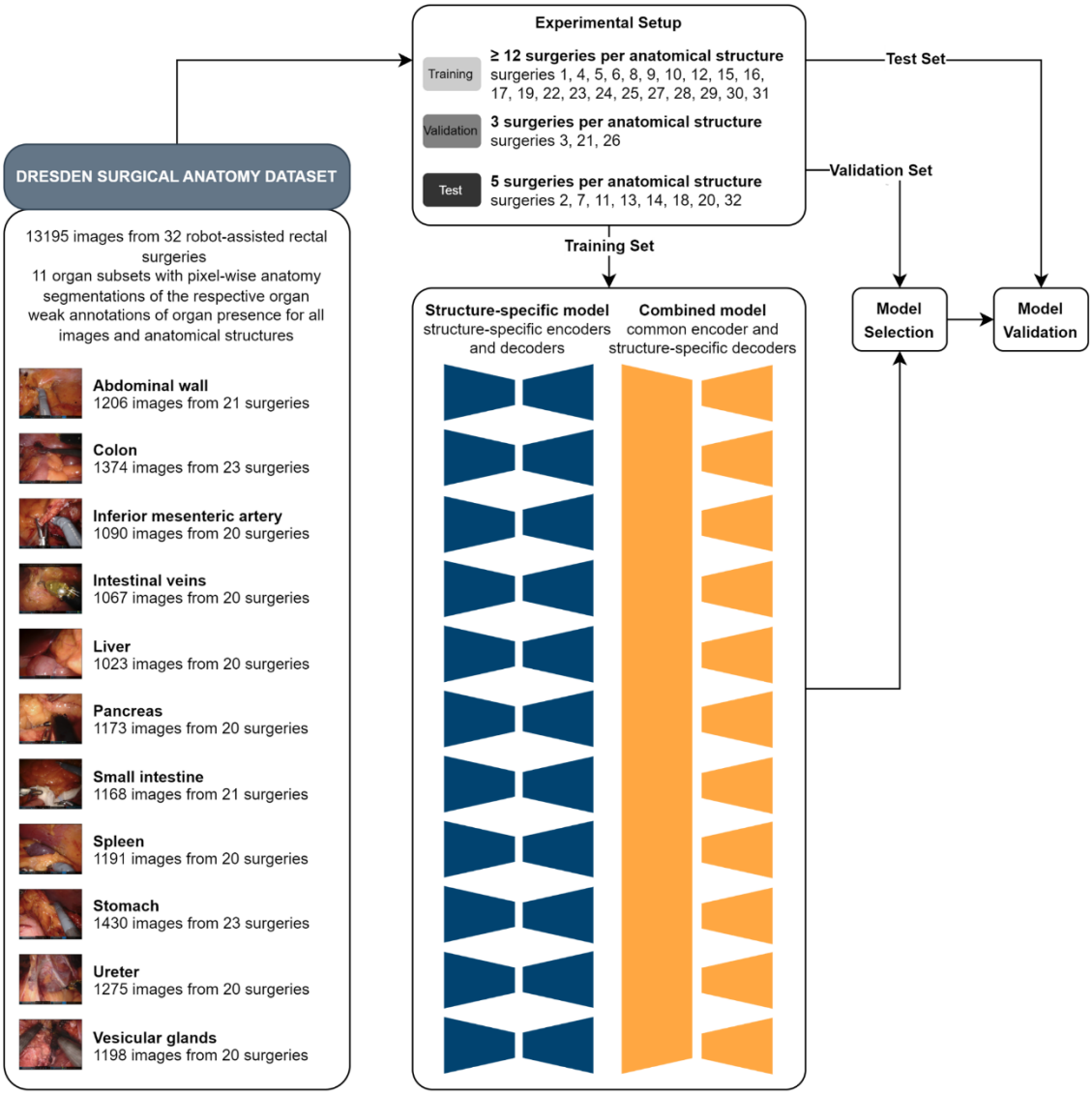

Figure 1

Fig. 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.

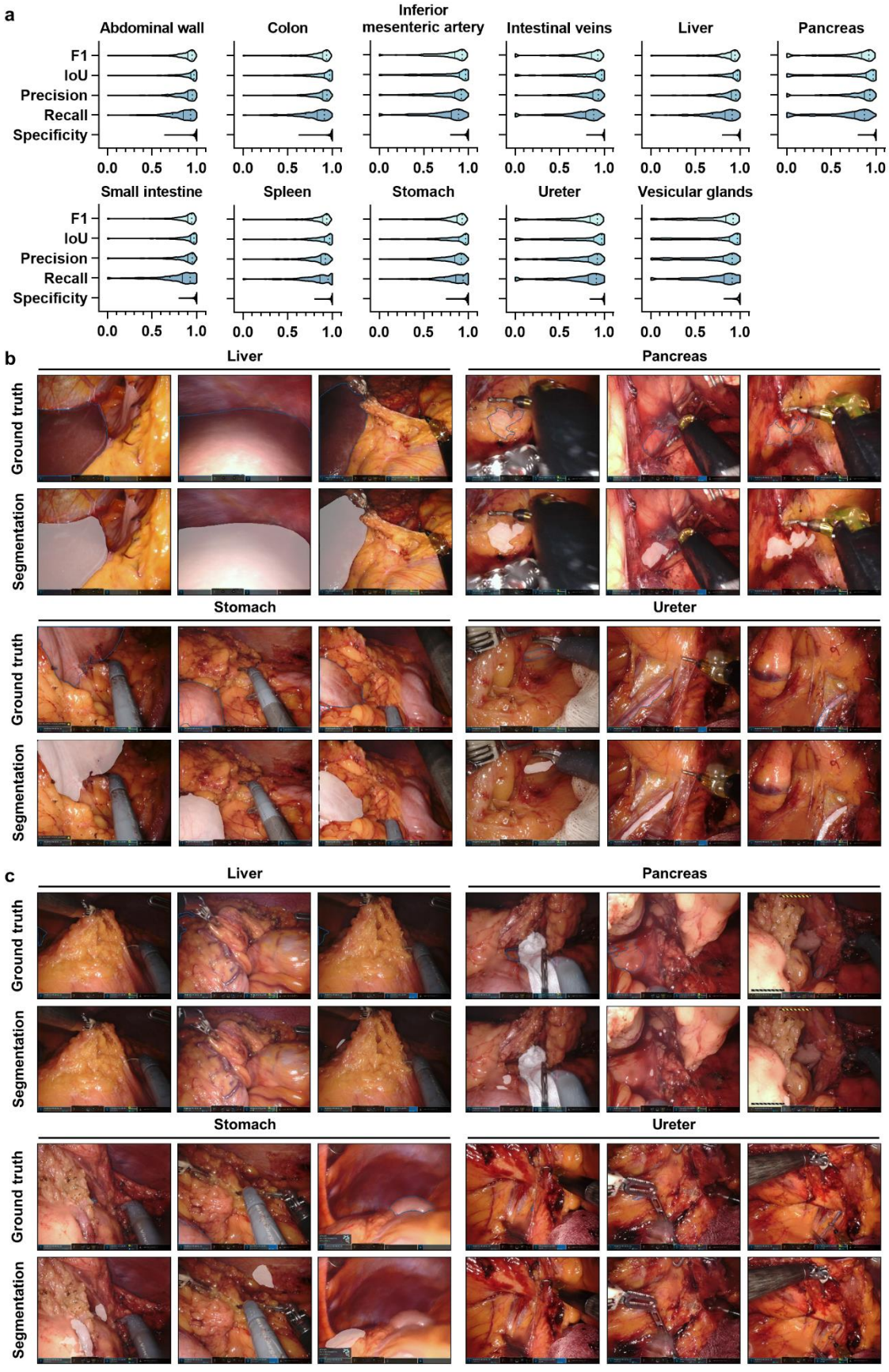

Figure 2

Fig. 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. The median and quartiles are illustrated as solid and dashed lines, respectively. **(b)** Example images with the highest IoUs for liver, pancreas, stomach, and ureter segmentation with structure-specific segmentation models. Ground truth is displayed in blue, and proposed segmentations are displayed as white overlay. **(c)** Example images with the lowest IoUs for liver, pancreas, stomach, and ureter segmentation with structure-specific segmentation models. Ground truth is displayed in blue, and proposed segmentations are displayed as white overlay.

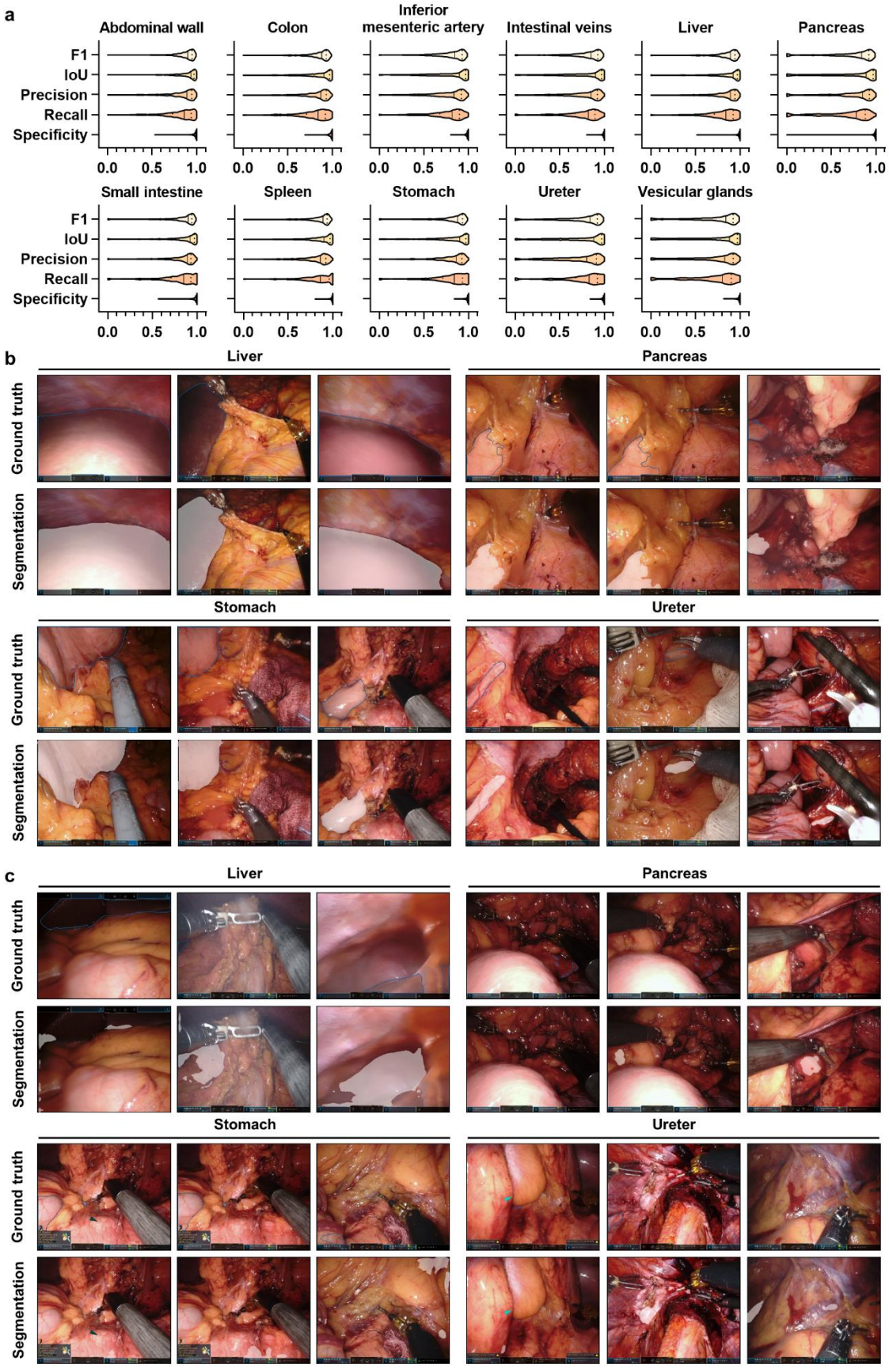

Figure 3

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

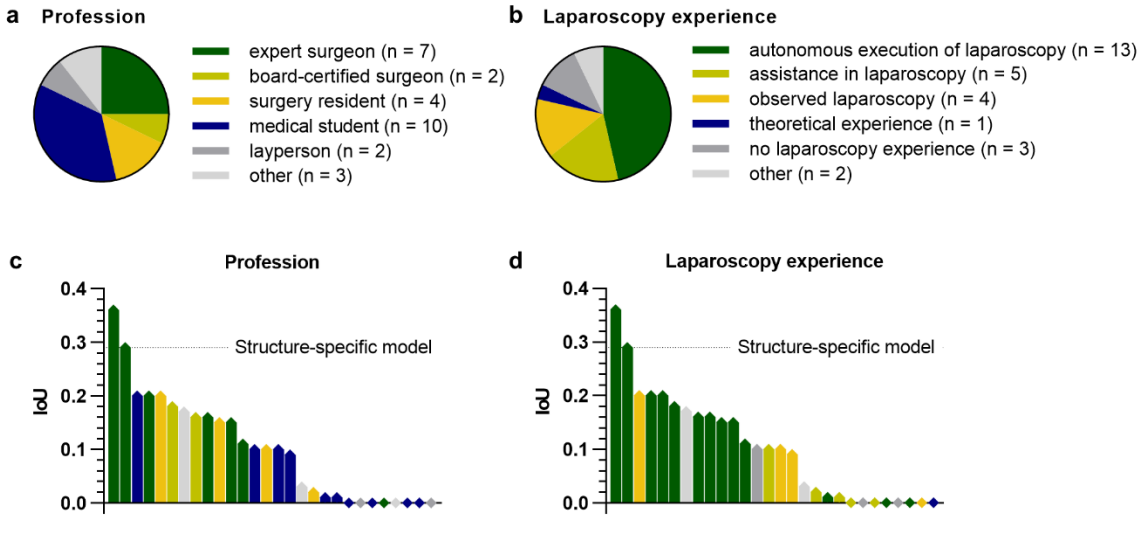

Figure 4

Fig. 4 Comparison of pancreas segmentation performance of the structure-specific model 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 model. **(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 model.

Abbreviations

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

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) and from the Joachim Herz Foundation (Add-On Fellowship for Interdisciplinary Life Science). FMR received a doctoral student scholarship from the *Carus Promotionskolleg* Dresden. The authors gratefully acknowledge excellent project coordination by Dr. Elisabeth Fischermeier and Dr. Grit Krause-Jüttler.

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.

Competing interests

The authors declare no conflicts of interest.