Medical Image Analysis 86 (2023) 102770

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark

Martin Wagner [a,b,*], Beat-Peter Müller-Stich [a,b], Anna Kisilenko [a,b], Duc Tran [a,b], Patrick Heger [a], Lars Mündermann [c], David M Lubotsky [a,b], Benjamin Müller [a,b], Tornike Davitashvili [a,b], Manuela Capek [a,b], Annika Reinke [d,e,f], Carissa Reid [g], Tong Yu [h,i], Armine Vardazaryan [h,i], Chinedu Innocent Nwoye [h,i], Nicolas Padoy [h,i], Xinyang Liu [j], Eung-Joo Lee [k], Constantin Disch [l], Hans Meine [l,m], Tong Xia [n], Fucang Jia [n], Satoshi Kondo [o,2], Wolfgang Reiter [p], Yueming Jin [q], Yonghao Long [q], Meirui Jiang [q], Qi Dou [q], Pheng Ann Heng [q], Isabell Twick [r], Kadir Kirtac [r], Enes Hosgor [r], Jon Lindström Bolmgren [r], Michael Stenzel [r], Björn von Siemens [r], Long Zhao [s], Zhenxiao Ge [s], Haiming Sun [s], Di Xie [s], Mengqi Guo [t], Daochang Liu [u], Hannes G. Kenngott [a], Felix Nickel [a], Moritz von Frankenberg [v], Franziska Mathis-Ullrich [w], Annette Kopp-Schneider [g], Lena Maier-Hein [d,e,f,x], Stefanie Speidel [y,z,1], Sebastian Bodenstedt [y,z,1]

[a] Department for General, Visceral and Transplantation Surgery, Heidelberg University Hospital, Im Neuenheimer Feld 420, 69120 Heidelberg, Germany
[b] National Center for Tumor Diseases (NCT) Heidelberg, Im Neuenheimer Feld 460, 69120 Heidelberg, Germany
[c] Data Assisted Solutions, Corporate Research & Technology, KARL STORZ SE & Co. KG, Dr. Karl-Storz-Str. 34, 78332 Tuttlingen
[d] Div. Computer Assisted Medical Interventions, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 223, 69120 Heidelberg Germany
[e] HIP Helmholtz Imaging Platform, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 223, 69120 Heidelberg Germany
[f] Faculty of Mathematics and Computer Science, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg
[g] Division of Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 280, Heidelberg, Germany
[h] ICube, University of Strasbourg, CNRS, France. 300 bd Sébastien Brant - CS 10413, F-67412 Illkirch Cedex, France
[i] IHU Strasbourg, France. 1 Place de l'hôpital, 67000 Strasbourg, France
[j] Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, 111 Michigan Ave NW, Washington, DC 20010, USA
[k] University of Maryland, College Park, 2405 A V Williams Building, College Park, MD 20742, USA
[l] Fraunhofer Institute for Digital Medicine MEVIS, Max-von-Laue-Str. 2, 28359 Bremen, Germany
[m] University of Bremen, FB3, Medical Image Computing Group, % Fraunhofer MEVIS, Am Fallturm 1, 28359 Bremen, Germany
[n] Lab for Medical Imaging and Digital Surgery, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[o] Konika Minolta, Inc., 1-2, Sakura-machi, Takatsuki, Oasak 569-8503, Japan
[p] Wintegral GmbH, Ehrenbreitsteiner Str. 36, 80993 München, Germany
[q] Department of Computer Science and Engineering, Ho Sin-Hang Engineering Building, The Chinese University of Hong Kong, Sha Tin, NT, Hong Kong
[r] Caresyntax GmbH, Komturstr. 18A, 12099 Berlin, Germany
[s] Hikvision Research Institute, Hangzhou, China
[t] School of Computing, National University of Singapore, Computing 1, No.13 Computing Drive, 117417, Singapore
[u] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, China
[v] Department of Surgery, Salem Hospital of the Evangelische Stadtmission Heidelberg, Zeppelinstrasse 11-33, 69121 Heidelberg, Germany
[w] Health Robotics and Automation Laboratory, Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Geb. 40.28, KIT Campus Süd, Engler-Bunte-Ring 8, 76131 Karlsruhe, Germany
[x] Medical Faculty, Heidelberg University, Im Neuenheimer Feld 672, 69120 Heidelberg
[y] Div. Translational Surgical Oncology, National Center for Tumor Diseases Dresden, Fetscherstraße 74, 01307 Dresden, Germany
[z] Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden, 01062 Dresden, Germany

* Corresponding author at: Universitätsklinikum Heidelberg, Chirurgische Klinik, Im Neuenheimer Feld 420, 69120 Heidelberg, Germany.
E-mail address: martin.wagner@med.uni-heidelberg.de (M. Wagner).

[1] Stefanie Speidel and Sebastian Bodenstedt contributed equally to this work.

[2] Present address: Muroran Institute of Technology, Muroran, Hokkaido, 050-8585, Japan.

https://doi.org/10.1016/j.media.2023.102770
Received 25 September 2021; Received in revised form 3 February 2023; Accepted 8 February 2023
Available online 21 February 2023

ARTICLE INFO

Keywords:
Surgical workflow analysis
Endoscopic vision
Surgical data science
Laparoscopic cholecystectomy

ABSTRACT

Purpose: Surgical workflow and skill analysis are key technologies for the next generation of cognitive surgical assistance systems. These systems could increase the safety of the operation through context-sensitive warnings and semi-autonomous robotic assistance or improve training of surgeons via data-driven feedback. In surgical workflow analysis up to 91% average precision has been reported for phase recognition on an open data single-center video dataset. In this work we investigated the generalizability of phase recognition algorithms in a multicenter setting including more difficult recognition tasks such as surgical action and surgical skill.

Methods: To achieve this goal, a dataset with 33 laparoscopic cholecystectomy videos from three surgical centers with a total operation time of 22 h was created. Labels included framewise annotation of seven surgical phases with 250 phase transitions, 5514 occurences of four surgical actions, 6980 occurences of 21 surgical instruments from seven instrument categories and 495 skill classifications in five skill dimensions. The dataset was used in the 2019 international Endoscopic Vision challenge, sub-challenge for surgical workflow and skill analysis. Here, 12 research teams trained and submitted their machine learning algorithms for recognition of phase, action, instrument and/or skill assessment.

Results: F1-scores were achieved for phase recognition between 23.9% and 67.7% ($n = 9$ teams), for instrument presence detection between 38.5% and 63.8% ($n = 8$ teams), but for action recognition only between 21.8% and 23.3% ($n = 5$ teams). The average absolute error for skill assessment was 0.78 ($n = 1$ team).

Conclusion: Surgical workflow and skill analysis are promising technologies to support the surgical team, but there is still room for improvement, as shown by our comparison of machine learning algorithms. This novel HeiChole benchmark can be used for comparable evaluation and validation of future work. In future studies, it is of utmost importance to create more open, high-quality datasets in order to allow the development of artificial intelligence and cognitive robotics in surgery.

1. Introduction

Surgical workflow and skill analysis are key technologies for the development and seemingless integration of artificial intelligence systems (AI) in the operating room (OR) and thus a main building block of surgical data science (Maier-Hein et al., 2022). Surgical AI systems may increase the safety and efficiency of the operation through early context-sensitive warnings (Katić et al., 2013), OR management (Tanzi et al., 2020) and procedure time prediction (Aksamentov et al., 2017; Bodenstedt et al., 2019), continuing surgical education (Maier-Hein et al., 2017) and professional development (Greenberg et al., 2018) by objective assessment of surgical skill and competency (Funke et al., 2019; Hashimoto et al., 2019; Vedula et al., 2017), as well as semi-autonomous assistance (Lalys and Jannin, 2014; Vercauteren et al., 2020). Surgical AI may also help to locate critical events in surgical videos and document safety measures (Korndorffer et al., 2020; Mascagni et al., 2021). Furthermore, if today's robotic telemanipulators are to become cognitive assistance systems, which perceive their environment, interpret it according to previous experience and perform a context-aware action in a semi-autonomous way, they will have to understand the surgical workflow and learn from the skilled surgeons (Maier-Hein et al., 2022; Wagner et al., 2021). Thus, surgical workflow and skill analysis are a prerequisite for the next generation of surgical robotics.

A great obstacle to train the underlying machine learning (ML) algorithms to created AI applications in surgery is the lack of high quality annotated datasets (Maier-Hein et al., 2021). The use of ML has been successfully researched on the basis of annotated non-surgical data (Topol, 2019). At the same time, research on surgical data, especially laparoscopic videos is comparatively underrepresented, even though the availability of surgical video data is increasing thanks to time-efficient and cost-effective recording and storage of those videos. A reason may be that the process of video annotation with meaningful information for ML is time-consuming and laborious. As a result, only few surgical video datasets are openly available for research, even though the publication rate for the analysis of surgical procedures has increased in recent years (Loukas, 2018). For instance, the Cholec80 dataset contains videos of 80 laparoscopic cholecystectomies from a single center and annotation of surgical phase and instrument presence with a phase recognition average precision of up to 91% (Twinanda et al., 2017). Similar results have been reported on a larger multicenter dataset of 1243 laparoscopic cholecystectomies, but this data is not openly available (Bar et al., 2020). Another open dataset is the Heidelberg colorectal dataset. It comprises 30 videos of three different types of laparoscopic colorectal surgeries, corresponding sensor data from medical devices in the OR and pixel-wise semantic segmentation of surgical instruments (Maier-Hein et al., 2021). The successful use of the Heidelberg Colorectal dataset during the Endoscopic Vision (EndoVis) challenges 2017 (https://endovissub2017-workflow.grand-challenge.org/) and 2019 (https://robustmis2019.grand-challenge.org/) is an example for the comparative validation of ML algorithms to explore an optimal solution for surgical problems. Apart from laparoscopy, surgical workflow has been investigated in ophthalmology. The CATARACTS Challenge presented successful results on instrument presence detection during cataract surgery using computer vision algorithms (Al Hajj et al., 2019). Though all three datasets used clinical patient videos as the basis for their annotations, they are limited in their transferability, because they do not sufficiently reflect the diversity of clinical data in a multicenter setting. Furthermore, they focus on a limited variety of annotated features.

In contrast, the preclinical JIGSAWS dataset for gesture and skill analysis contains kinematic and video data of surgical tasks with detailed action and skill annotation (Ahmidi et al., 2017). However, JIGSAWS does not contain real patient data. Due to their achievements, EndoVis (https://endovis.grand-challenge.org/) (28%), Cholec80 (21%), and JIGSAWS (17%) were mentioned as the most useful publicly available datasets for surgical data science (Maier-Hein et al., 2022), but there is still an urgent demand in the scientific community and medical device industry for high-quality datasets from laparoscopic surgery that allow a comparison of ML algorithms (Maier-Hein et al., 2022).

Apart from this, it is generally important to view medical recognition challenges with caution due to the lack of standardized and evidence-based quality control. For example, it is possible for later participating teams to boost their performance if the test data has been published (Maier-Hein et al., 2018). Moreover, standardized phase definitions are missing in the existing datasets, especially for laparoscopic cholecystectomy (Garrow et al., 2020; Meireles et al., 2021). Since the reproducibility of research results is an important element of science, a standardized benchmark for comparing such results is of great importance.

In this study we aim to counteract this deficiency and propose an open benchmark for surgical workflow and skill analysis by providing a state of the art comparison of ML algorithms on a novel and publicly

accessible clinical multicenter dataset. Specifically, our study aimed at answering the following research questions:

(1) Can the previously reported performance in recognition of surgical phase and surgical instrument be reproduced on this dataset by independent researchers?
(2) What performance can be achieved for recognition tasks more difficult than phase recognition such as surgical action (often brief and subtle) and surgical skill (holistic assessment of the whole video)?

2. Dataset

The structure of this paper follows the BIAS statement for transparent reporting of biomedical image analysis challenges (Maier-Hein et al., 2020) and includes the structured challenge design in Appendix B. The creation of the challenge dataset is described including annotations for surgical phase, action, instrument and skill. Then, the challenge design and a description of the competing ML algorithms are described.

2.1. Data collection

The dataset contains $n = 33$ videos of laparoscopic cholecystectomies from three surgical centers in Germany with a total video length of 22 h. The total number of cases was chosen based on annotation capacity. The operation videos at Heidelberg University Hospital ($n = 15$) were recorded with a laparoscopic 2D camera (Karl Storz SE & Co KG, Tuttlingen Germany) with 30° optics, a resolution of 960 × 540 pixels and 25 frames per second. The operation videos at Salem Hospital ($n = 15$) and the GRN-hospital Sinsheim ($n = 3$) were recorded with the laparoscopic 2D camera ENDOCAM Logic HD (Richard Wolf GmbH, Knittlingen, Germany) with 30° optics, a resolution of 1920 × 1080 pixels and for the greater part 50 frames per second. Three operations at Salem Hospital were recorded with a resolution of 720 × 576 pixels and 25 frames per second. Every video starts at the first insertion of the laparoscopic camera into the patient's abdomen and ends with the last removal of the laparoscopic camera.

The videos were split into the training ($n = 24$) and test ($n = 9$) dataset. In the training dataset, videos from Heidelberg University Hospital and Salem Hospital are equally represented ($n = 12$ each). In the test dataset, all three centers are equally represented ($n = 3$ each). Assignment to training or test dataset was performed randomly with stratification by center. The total number of test cases was chosen to maximize the ability to generalize and evaluate while maintaining a large enough training set.

To comply with ethical standards and the general data protection regulation of the European Union, routinely collected data was used and anonymized. To this end, scenes outside the abdominal cavity, for example when the camera was pulled out for cleaning purposes, were manually censored (frames were replaced with white frames) and files were renamed anonymously (HeiChole-1, HeiChole-2 etc.).

2.2. Data annotation

The anonymized video data were annotated with surgical knowledge by specifically instructed medical students following annotation rules in Appendix A. Annotation, i.e. labeling of each video frame with information about what is depicted in this frame, was performed using the video annotation research tool Anvil (Kipp, 2014). The annotation included framewise annotation of surgical phases, actions and instruments as well as skill and difficulty classification for procedures and selected phases. Thus, different perspectives of the surgical activity were annotated. According to Neumuth et al. a surgical activity consists of five components (Neumuth et al., 2009), which are the functional, organizational, operational, spatial and behavioral perspectives. In this study, all perspectives ecxemt the spatial were annotated. The performed action describes what is done (functional, e.g. "grasp", see paragraph "Action") and is defined as a sequence of related gestures. The performer of the action (organizational, e.g. "left hand of the surgeon") and the surgical instrument used (operational, e.g. "atraumatic grasper", see paragraph "Instrument") were annotated in relation to the exact time (behavioral, framewise annotation of the video). An example of a comprehensive annotation would be "the left hand of the surgeon performs the grasping and holding action with the atraumatic grasper at 10 min and 15 s after start of the operation". Whereas the hand is not visible in the image, in the standardized procedure of laparoscopic cholecystectomy the performing hand can be deducted from the position of the instrument.

To ensure standardization and reproducibility of annotation as well as to minimize sources of error, explicit rules were formulated for phase, action and instrument annotation. An identical procedure was followed for both training and test cases. The annotation rules are enclosed in Appendix A. Surgical phase (see Section 2.2.1. Phase) was annotated analogous to the Cholec80 dataset (Twinanda et al., 2017), surgical skill and difficulty (see Section 2.2.4. Skill) were annotated using modified Global Operative Assessment of Laparoscopic Skills (GOALS) score (Vassiliou et al., 2005) as extended by Chang et al. (Chang et al., 2007). In order to increase the reliability of the annotation, the phases were annotated independently by three specifically instructed medical students and the surgical skill and difficulty by two specifically instructed medical students. Possible error sources occurred with disagreement on the beginning or end of a phase or the skill level. Deviations were discussed and resolved by consensus between the same students. Final consensus annotations as well as raw annotations for phases before consensus can be downloaded from the challenge website on Synapse (see Section 4.3. HeiChole benchmark & online leaderboard).

According to the BIAS statement, a case in our dataset encompassed all data for which the algorithm(s) participating in a specific challenge task produced one result. One case comprises three videos, a full laparoscopic cholecystectomy, all frames of the phase calot triangle dissection (P1) and all frames of the phase gallbladder dissection (P3), respectively. Annotations for surgical phase were one value per frame. Annotations for action were a 4D binary vector per frame indicating if the corresponding action is being performed (1) or not (0)). Annotations for instrument category were a 21D binary vector per frame, consisting of 7 instrument categories used during the EndoVis challenge, one undefined instrument shaft plus 14 unused categories reserved for future additions like further grasping instruments, with each entry indicating if the corresponding instrument category is visible (1) or not (0)). Annotations for instruments were a 31D binary vector per frame, consisting of 21 instruments used during the EndoVis challenge, one undefined instrument shaft plus 9 unused instruments reserved for future additions, with each entry indicating if the corresponding instrument category is visible (1) or not (0)). Surgical skill was annotated in five different dimensions, each ranked with integer values between 1 and 5, for each of the three videos full operation, P1 and P3.

2.2.1. Phase

For the surgical phases, one of seven surgical phases was assigned to each individual frame, analogous to the Cholec80 dataset (Twinanda et al., 2017) following the annotation protocol in Appendix A. The seven phases were preparation (P0), calot triangle dissection (P1), clipping and cutting (P2), gallbladder dissection (P3), gallbladder packaging (P4), cleaning and coagulation (P5) and gallbladder retraction (P6). The phases did not necessarily occur in a fixed order.

2.2.2. Action

The surgical action is the functional component of the activity a surgeon performs within a phase. Action was annotated as performed following the annotation protocol in the appendix, if any of the four actions grasp (A0), hold (A1), cut (A2) or clip (A3) occurred. Additionally, the performer of the action (organizational component) was

annotated as the left hand of the surgeon, right hand of the surgeon or hand of the assistant.

2.2.3. Instrument

Instrument presence detection is important for surgical workflow analysis because it correlates with the current surgical phase. A total of 21 instruments (plus “undefined instrument shaft”) of different types were annotated and additionally grouped into the seven categories grasper (IC0), clipper (IC1), coagulation instruments (IC2), scissors (IC3), suction-irrigation (IC4), specimen bag (IC5), and stapler (IC6). Because more than one instrument may be present at the same time, annotation of instrument visibility was performed separately for each of the 21 instrument types and for the challenge metrics were computed separately per instrument category. Furthermore, in different surgical centers, instruments by different vendors were used, which increases the representativeness of this dataset. The stapler was not present in the test dataset. For instrument presence, an instrument was annotated visible as soon as its characteristic instrument tip appeared in the image. The annotation continued when the tip disappeared later and only the shaft of the instrument remained visible. If the instrument shaft entered the field of view of the camera without its tip having been visible before, it was referred to as the "undefined instrument shaft", because even a human annotator would not be able to recognize a particular instrument due to the identically looking shafts. Three exceptions to this rule were the suction-irrigation, stapler and the clipper categories, as these instruments have characteristic shafts. Fig. 1 shows sample images of the instruments from the dataset.

2.2.4. Skill

To assess the surgical skill, the videos were scored using the modified Global Operative Assessment of Laparoscopic Skills (GOALS). It has been validated for video assessment of laparoscopic skills, including the five domains depth perception (S1), bimanual dexterity (S2), efficiency (S3), tissue handling (S4) and autonomy (Vassiliou et al., 2005). The item "autonomy" was omitted in our study, because a valid assessment based solely on intraabdominal video alone is not possible without information about what was spoken during the operation or how much assistance was provided by a senior surgeon. The difficulty of the operation (S5) was additionally annotated based on Chang’s adaptation of the GOALS-score (Chang et al., 2007) . Here, parameters such as inflammatory signs, adhesions and individual anatomical conditions were used to objectify the assessment of the skill. Thus, the skill assessment in this study included five ranking components. Skill was annotated for the complete operation and additionally for phases calot triangle dissection (P1) and gallbladder dissection (P3).

3. Methods

3.1. EndoVis challenge 2019

Based on our dataset, 12 international teams trained and evaluated their algorithms during the EndoVis challenge 2019 within the sub-challenge for “Surgical Workflow and Skill Analysis” hosted in conjunction with the Medical Image Computing and Computer Assisted Intervention conference (MICCAI, https://endovissub-workflowandskill.grand-challenge.org/). The aim of this sub-challenge was to investigate the current state of the art on surgical workflow analysis and skill assessment in laparoscopic cholecystectomy on one comprehensive dataset. Specifically, the aims were (1) surgical phase recognition with high precision and recall, (2) surgical action recognition with high precision and recall, (3) surgical instrument presence detection with high precision and recall as well as (4) surgical skill assessment with a

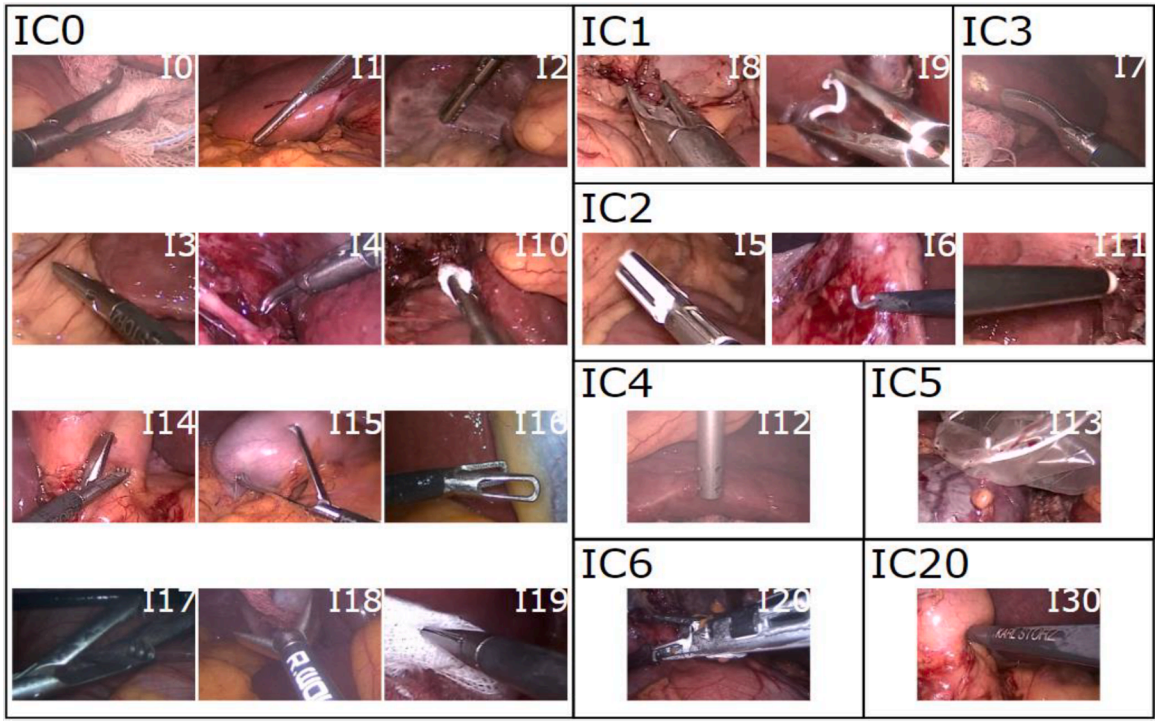

Fig. 1. Instruments in the HeiChole benchmark. Examples of all 21 surgical instruments plus undefined instrument shaft present in the HeiChole benchmark arranged according to the eight categories grasper (IC0), clipper (IC1), coagulation instruments (IC2), scissors (IC3), suction-irrigation (IC4), specimen bag (IC5), staper (IC6), undefined instrument shaft (IC20).
Instruments are curved atraumatic grasper (I0), toothed grasper (I1), fenestrated toothed grasper (I2), atraumatic grasper (I3), overholt (I4), LigaSure (I5), electric hook (I6), scissors (I7), clip-applier metal (I8), clip-applier Hem-O-Lok (I9), swab grasper (I10), Argon beamer (I11), suction-irrigation (I12), specimen bag (I13), tiger mouth forceps (I14), claw forceps (I15), atraumatic grasper short (I16), crocodile grasper (I17), flat grasper (I18), pointed grasper (I19), stapler (I20) and undefined instrument shaft (I30). Numbers I21 to I29 have been reserved for future additions.

low mean absolute error. Before acceptance as a MICCAI challenge, the challenge protocol underwent a peer review process.

Participants were invited to submit a Docker image and a description of the used method(s). For the submission process, participants had to register for the challenge on Synapse (HeiChole Benchmark Website, 2022), upload their Docker container(s) to that project and then submit the appropriate version(s) to the challenge. The container(s) had to implement an interface that took a video as input, preprocessed it if necessary, computed the appropriate challenge results, and output these in CSV file(s). Only full submissions, i.e. with no results missing, for each task were considered.

The Docker images were not and will not be shared with any third party by the organizers. Each team could choose to provide their source code, though they were not required to. Only automatic methods, i.e. not needing any human input during runtime, were accepted. Participants were encouraged to provide results for recognition of surgical phase, action and instrument as well as skill assessment, but it was not required to submit in all categories and participants were free to provide results for a subset. To reduce the complexity of the challenge, not every annotation of the dataset described above was used for the challenge. The action recognition did not include the recognition of the performer of an action. The instrument presence detection did not include the category undefined instrument shaft (IC20).

Participants were free to use third party public data to augment the provided training data. Submissions were “online analysis only”, i.e. methods were not allowed to use information from future frames. We chose “online analysis only” because from a clinical perspective our aim with this work was to develop methods usable for context-aware intraoperative assistance and cognitive robotics. These intraoperative assistance systems only have information about current and past video, but not the whole procedure. In contrast, this is not necessary for skill assessment, because this will not generate an intraoperative assistance but feedback after surgery. However, “online analysis only” was not specifically enforced, because it would have placed many restrictions onto the interfaces to the model and would have complicated the evaluation process greatly. Thus, it was trusted that the participants would adhere to the code of honor. In the case of skill assessment, an entire video could be used as input, because the skill was also annotated for video as a whole.

The challenge committee members were M. Wagner, S. Bodenstedt, A. Kisilenko, H. Kenngott, L. Maier-Hein, S. Speidel and B. Müller-Stich with affiliations as stated in the authors list. The award policy was to award a prize to each winner of one of the four tasks, if at least three participants entered a submission. Members of the organizing institutes were allowed to participate, but were not eligible for awards.

All participants agreed for their results and method description to be published in the challenge paper before participating in the challenge by sending a signed agreement form before access to the challenge data was granted. Apart from that all members of all participating teams were offered coauthorship to this manuscript according to the challenge protocol. For the one participant that did not respond to approve the manuscript, coauthorship was removed, but the person was mentioned in acknowledgements according to standard publication ethics.

On May 31st 2019, the first part of the training dataset consisting of 12 videos was published, followed by the second part, also consisting of 12 videos, on August 15th 2019. With the second part, the organisers’ evaluation software scripts for computing metrics and rankings were provided. The evaluation and submission period of the Docker containers was between October 1st and 7th 2019. The challenge day was on October 13th 2019. On this day, the results of all teams were presented during the EndoVis challenge meeting at the MICCAI in Shenzhen, China. Before publication of the joined paper, no results were allowed to be published.

3.2. Participating teams

The following sections provide a detailed description of the algorithms of the participating teams in alphabetical order. In addition, Table 1 gives an overview of the different methods sorted by teams, because we encouraged the teams to submit multi-task approaches, i.e. methods that produce a solution for more than one problem from the given input, e.g. a deep learning model that has two outputs, one for phase and one for instrument recognition. To allow for easier comparison between different methods for one task, we provide three separate tables Tables C1, C2, C3 Table for phase, action and instrument recognition in Appendix C.

3.2.1. Team CAMI-SIAT (Phase)

The *CAMI-SIAT* team proposed a method for determining surgical phases based on Pseudo-3D residual networks (ResNet) (Qiu et al., 2017). Through the usage of a 3D convolutional network, temporal information from previous frames in an operation could be utilized directly to determine the current surgical phase. Further, they hypothesized that fusing the predictions of the Pseudo-3D residual network with prior knowledge would improve performance. For this, they determined the probability of occurrence of each phase at a given time point in the operations from the training data and applied them to the output of the network for the final prediction. No additional data was used for pre-training.

3.2.2. Team CAMMA (Phase & action)

For the phase recognition, the *CAMMA* team utilized two different methods for image feature extraction in parallel: Inflated-3D (I3D) (Carreira and Zisserman, 2017), a 3D, and an Inception-ResNet (Szegedy et al., 2016), a 2D convolutional network. The aim of mixing 3D and 2D convolutions was to capture fine-grained temporal dynamics from the data. The convolutional networks were followed by 3 long-short-term memory (LSTM) units, one for each preceding feature extractor and one for the combined features. The predictions of the 3 LSTMs were then merged via majority-voting. During training, the binary instrument data was also used as an additional task for enhancing the results of the workflow recognition. ImageNet and Kinetics were used for pre-training. *CAMMA* submitted two models for phase recognition, one with and one without pre-training. Both models were analyzed separately.

For the action recognition, team *CAMMA* built on a ResNet (He et al., 2015) for feature extraction, which was extended with a convolutional LSTM to take temporal information into account (Nwoye et al., 2019). Following the LSTM, a combination of a convolutional layer and a fully-connected layer were utilized for higher-level reasoning on the spatio-temporal features for surgical action recognition. ResNet was pretrained on ImageNet.

3.2.3. Team CareSyntax (Instrument & skill)

For the instrument presence detection, team *CareSyntax* utilized the approach outlined in (Vardazaryan et al., 2018) , which combined a ResNet (He et al., 2015) with additional convolutional layers, so called localization maps, that helped to map features corresponding to a classification to their spatial coordinates. A spatial pooling was then utilized to determine which instrument classes were currently in use. ImageNet and Cholec80 were used for pre-training.

For the skill assessment team *CareSyntax* utilized a method based on (Funke et al., 2019) . The method relied on a 3D-ResNet (Wang et al., 2016) for feature extraction. The method divided a given video into multiple segments, each segment was then fed into the 3D-ResNet. To concatenate the results of the segments, team *CareSyntax* used a fully-connected layer. The final scores were then computed by rounding the output of the network to the nearest integer. Kinetics was used for pre-training.

Table 1
Overview of algorithms. The submissions of the teams participating in the EndoVis challenge are presented with components of their machine learning methods for the specific tasks. 3.3. Statistical Analysis.

Team	Task(s)	Multi-Task	Basic architecture	(Additional) temporal component	Output component	Post-processing	Pretraining	Data augmentation	Loss function (s)	Optimizer
CAMI-SIAT	Phase	No	Pseudo-3D Residual Network (Qiu et al., 2017)	None (3D architecture)	Output of the network is fused with the prior probability of each surgical phase	None	None	RGB shift, brightness and contrast changes, drop-out of frames	Binary cross-entropy loss	Adam (Kingma and Ba, 2017)
CAMMA	Phase	Yes (Instrument recognition was used as an auxiliary task)	Parallel I3D(Carreira and Zisserman, 2017) and Inception- ResNet (Szegedy et al., 2016)	3 LSTMs (192 , 512 and 512 units)	Majority voting to aggregate the outputs of the three LSTMs	None	Inception-ResNet pretrained on ImageNet (Russakovsky et al., 2015) and i3D pretrained on Kinetics (Kay et al., 2017)	None	Binary cross-entropy loss	RMSProp (Hinton et al., 2012)
	Action	No	ResNet-50 (He et al., 2015)	Convolutional LSTM (512 units)	A combination of convolutional layer and fully-connected layer connected to the LSTM	None	ResNet pretrained on ImageNet	None	Binary cross-entropy loss	Momentum Optimizer
CareSyntax	Instrument	No	ResNet-18 (He et al., 2015)	None	Convolutional layers, spatial pooling, fully connected layer	None	ImageNet and Cholec80 (Twinanda et al., 2017)	Rotation and horizontal flip	Weighted cross-entropy loss	SGD (Kiefer and Wolfowitz, 1952)
	Skill	No	3D ResNet-18	None (3D architecture)	Concatenation of results using fully-connected layer	None	Kinetics	Resizing with central cropping, random horizontal flip	Mean squared error	SGD (Kiefer and Wolfowitz, 1952)
CUHK	Phase	Yes	ResNet-50 ()	LSTM (512 units), elapsed time as input	Fully-connected layers connected to LSTM and instrument output					
Phase & Action: Fully-connected layers connected to LSTM	Median filter on preceding frames, Prior Knowledge Inference (Jin et al., 2018)	ResNet pretrained on ImageNet, all pretrained with Cholec80	Cropping, flipping, mirroring and color jitter	Categorical cross-entropy loss	SGD (Kiefer and Wolfowitz, 1952)					
	Instrument			None	Fully-connected layer				Binary cross-entropy loss	
HIKVision	Phase	Yes	ResNet-50 ()	LSTM	Fully-connected layer per task, connected to the LSTM	Prior Knowledge Inference (Jin et al., 2018)	ResNet pretrained on ImageNet	Random crop, rotation and flip	Categorical cross-entropy loss	SGD (Kiefer and Wolfowitz, 1952)
	Instrument								Binary cross-entropy loss	
IGITech	Phase	No	ResNet-50 ()	None	Support Vector Machine	None	ResNet pretrained on ImageNet	Random translation, rotation, flip	Binary cross-entropy loss	SGD (Kiefer and Wolfowitz, 1952)
	Instrument				Fully-connected layer					
	Action	Yes		None		None				

(*continued on next page*)

Table 1 (*continued*)

Team	Task(s)	Multi-Task	Basic architecture	(Additional) temporal component	Output component	Post-processing	Pretraining	Data augmentation	Loss function (s)	Optimizer
Konica Minolta	Instrument		ResNeXt-101 (Xie et al., 2017) with Squeeze-and-Excitation block (Hu et al., 2018)		Fully-connected layer per task		ResNeXt pretrained on ImageNet	Random translation, rotation, resizing, horizontal flip and contrast changes	Binary cross-entropy loss	SGD (Kiefer and Wolfowitz, 1952)
MEVIS	Phase	Yes	ResNet-50 ()	LSTM (512 units)	Fully-connected layer per task, connected to the LSTM	None	ResNet pretrained on ImageNet and Cholec80	Random crop and horizontal flip	Categorical cross-entropy loss, Binary cross-entropy loss	Adam (Kingma and Ba, 2017)
NCT	Phase	Yes	ResNet-50	LSTM (512 units)	Fully-connected layer per task, connected to the LSTM	None	ResNet pretrained on ImageNet, all pretrained with Cholec80	None	Categorical cross-entropy loss	Adam (Kingma and Ba, 2017)
	Action								Dice loss	
	Instrument									
VIE-PKU	Phase	Yes	Parallel ResNet-101 () and I3D (Carreira and Zisserman, 2017)	None (3D architecture)	Fully-connected layer connected to features	Prior Knowledge Inference (Jin et al., 2018), Inter- and intra-task correlation	ResNet and I3D pretrained on ImageNet, I3D pretrained on Kinetics	None	Binary cross-entropy loss	Adam (Kingma and Ba, 2017)
	Action								Weighted Binary cross-entropy loss	
	Instrument									
Wintegral	Phase	Yes (Single and multi-task models)	ResNet-50 ()	None	Results from different models are aggregated, one regressor per task	None	ResNet pretrained on ImageNet	Contrast changes, color jitter, center crop	Categorical cross-entropy loss	Adam (Kingma and Ba, 2017)
	Action								Binary cross-entropy loss	
	Instrument									

3.2.4. Team CUHK (Phase & instrument)

To recognize phase and instrument, team *CUHK* implemented a multi-task approach. A ResNet (Qiu et al., 2017) was used to extract visual features from a given laparoscopic image. The features were used to a) determine which laparoscopic instruments were currently visible via a fully-connected layer and b) to determine the current surgical phase via a LSTM. For the phase recognition task, the elapsed time, normalized via the average duration of all surgeries in the training dataset, was concatenated to the feature representation. The output of the network was then post-processed: a median filter was applied to phase predictions with low probability and the PKI strategy (Jin et al., 2018), which takes phase order and consistency into account, was utilized to detect and correct impossible changes in phases. ResNet was pretrained on ImageNet, all networks were pretrained with Cholec80.

3.2.5. Team HIKVision (Phase & instrument)

Team *HIKVision* utilized a ResNet (Qiu et al., 2017) for feature extraction. The ResNet was first trained for multi-task recognition of phase and instruments. The fully-connected layers for these two tasks were then replaced by a LSTM and again two output layers for phase and instrument recognition. To take phase order and consistency into consideration, PKI (Jin et al., 2018) was used to post-process the output. ResNet was pretrained on ImageNet.

3.2.6. Team IGITech (Phase & instrument)

A ResNet (Qiu et al., 2017) was used by team *IGITech* to extract visual features from a given laparoscopic frame. A fully-connected layer was used to determine which laparoscopic instruments were located in a video frame and a support vector machine was used to recognize the current phase. ResNet was pretrained on ImageNet.

3.2.7. Team Konica Minolta (Action & instrument)

Team *Konica Minolta* used a 101-layer ResNeXt (Xie et al., 2017) with Squeeze-and-Excitation block (Hu et al., 2018) as their base network. They attached a fully-connected layer for the instrument presence detection and one for the action recognition to the network. Furthermore, a third layer was added for detecting instrument super-classes. Here, one super-class generally consisted of a combination of instruments that regularly occur together, e.g. grasper and scissors. ResNeXt was pretrained on ImageNet.

3.2.8. Team MEVIS (Phase)

Team *MEVIS* used a ResNet (Qiu et al., 2017) for extracting visual features from the laparoscopic video frames. A LSTM was used to incorporate information from past frames. For each task, a fully-connected layer was added to the LSTM. During training, frames at the beginning of a video clip were given a weight of zero to allow the network a "warm-up phase". ResNet was pretrained on ImageNet and Cholec80. In addition, video-clips from 16 Cholec80 videos were added during training to increase the dataset and video-clips were sampled stratified based on the phase-labels.

3.2.9. Team NCT (Phase, action & instrument)

Team *NCT* used a multi-task approach for determining from a laparoscopic video the phase, which instruments were visible and what actions were being performed. The approach used a ResNet (Qiu et al., 2017) for feature extraction, a LSTM was used to propagate information from past frames along the temporal axis and used the information for predicting current frames. For each task, i.e. phase, action and instrument presence detection) a fully-connected layer was added to the LSTM. ResNet was pretrained on ImageNet, all networks were pretrained with Cholec80.

3.2.10. Team VIE-PKU (Phase, action & instrument)

The approach of team *VIE-PKU* combined two approaches for feature extraction. A 2D architecture, ResNet (Qiu et al., 2017), and a 3D architecture, 3D (Carreira and Zisserman, 2017), were run in parallel to capture both spatial and temporal features. Starting from the features, three separate branches emerged, one branch built upon the features to predict the current surgical action, a second branch predicted which surgical instruments were currently visible. A third branch combined the features with the action and instrument predictions to finally determine the current surgical phase. The phase results were then post-processed using PKI (Jin et al., 2018) and also using learned intra- and inter-class probabilities. ResNet and I3D were pretrained on ImageNet, I3D was also pretrained on Kinetics.

3.2.11. Team Wintegral (Phase, action & instrument). Team *Wintegral* combined several different types of models. They trained single-task models for each task, three multi-task models that combined phase recognition with instrument type detection, instrument category presence detection and action recognition respectively, as well as binary one-vs-all models that focused on a single class. For all models, a ResNet (Qiu et al., 2017) was used as a basis. The results of the different models were then aggregated and one predictor for each task was used to compute the final result. All models were pretrained on ImageNet.

The properties of the dataset were analyzed with descriptive statistics mean and standard deviation (SD). For the phases, inter-rater-agreement before consensus was calculated using Fleiss' kappa (Fleiss, 1971), where values of kappa between 0.81 and 1.00 can be considered almost perfect. To calculate results of the EndoVis challenge for the tasks of phase, action and instrument recognition, the F1-score $F1 = 2.\frac{precision \cdot recall}{precision + recall}$ over all classes was computed. The final ranking was calculated by averaging the F1-scores per video. We selected the F1-score as it takes both false positives and negatives into account. We decided to average over all classes and all videos so that each class and video would have the same weight in the final score regardless of occurrence rate and length.

For skill assessment, the mean absolute error $MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{Y}_i - Y_i|$ over all criteria was evaluated and ranked with n being the total number of samples, which is 9 in our case. We decided to use the MAE for skill assessment, as it gives a good impression on how far predictions are on average from the reference.

To analyze the ranking stability, bootstrapping methods were applied as suggested in (Maier-Hein et al., 2018). The rankings were recalculated over 1000 bootstrap datasets drawn from the test set together with a 95% confidence interval to assess the variability in ranks against small perturbations. This means, in all three tasks, 9 videos were used for testing. For every task and video, different phases, instruments and actions were defined which are counted as one case each. For example, seven phases were defined for which the metric values per video were calculated. Therefore, this resulted in 63 cases for the phase task, 54 cases for the instrument task and 36 cases for the action task. For every task, new bootstrap datasets were created with the same number of cases (63, 54 or 36 cases, depending on the task) by sampling with replacement without taking correlations into account. With this procedure, a case may disappear from the newly created bootstrap dataset or appear multiple times. This way small perturbations were simulated in the datasets to check how the rankings changed with those "new" datasets. This was done completely random and repeated until 1000 bootstrap datasets have been produced per task without specific stratification by hospital. We computed 95% bootstrap confidence intervals across bootstrap samples, which ranged from the 2.5% to the 97.5% quantile of the bootstrap distribution.

Kendall's tau (Kendall, 1938) correlation coefficient was used to quantify the changes between the original and the bootstrap rankings for all three tasks. It is ranged between -1 and 1, where a value of 1 indicates an unchanged ranking and -1 a reversed ranking.

Finally, pairwise Wilcoxon signed rank tests with a 5% alpha level and adjustment according to Holm for multiple testing were computed between each algorithm pair. The tests were used to check for

algorithms that were significantly superior in terms of performance compared to their competitors. The results of this analysis are independent from the bootstrap analysis.

The statistical analysis described above was performed using the *challengeR* toolkit, version 1.0.1, that was presented in (Wiesenfarth et al., 2021).

For the performance analysis across methods, the relationships between specific design components of the algorithms and their overall performance were investigated univariately. Accordingly, we analyzed the components multitask, Basic architecture, temporal component, post-processing, data augmentation, loss function(s), and optimizer. These components were chosen based on our structured algorithm overview in Table 1 that we designed according to what in our opinion were the most influential components of the algorithm. Furthermore, only those components were analyzed that were different between the teams. These relationships were investigated visually for each possible phase (P0-P6), instrument (IC0-IC6), and action (A0-A3) by plotting scatterplots of F1-scores against each variable, with the points of the scatterplot coloured to show the corresponding algorithm/team. For the phases and instrument tasks, the R package brms (Bürkner, 2017) was used to fit 9 two-part zero-inflated beta mixed models analyzing the influence of each component on the F1-score. This method is based on a mixture distribution and simultaneously fits two models with the same predictors, i.e. a logistic model evaluating the proportion of zero F1-scores and a beta model for the non-zero F1-scores. Each component was included as a fixed effect alongside a random algorithm/team effect and another random effect for phase or instrument. No possible interactions between the design features were investigated because the data is not fully factorial, i.e., not all 192 possible combinations of features (6 binary variables and one with three options) were represented in the data. Only 14 combinations were observed and 7 of these occurred only once, i.e. in a single algorithm/team. Statistical code (R session information) can be found in the R markdown files in Appendix E.

4. Results

In this section we will give a quantitative overview of the dataset including the annotations, followed by the results of the EndoVis challenge 2019. Together they build the foundation for the HeiChole benchmark that will serve as a validated reference benchmark.

4.1. Dataset for surgical workflow and skill analysis

We introduced a novel dataset with 33 laparoscopic cholecystectomy videos comprising a publicly available training dataset of 24 videos and an unpublished test dataset of 9 videos. Together all 33 videos had a total operation time of 22 h (40.04 min ± 18.06 min per video) from three surgical centers including annotation of 250 transitions between seven surgical phases, 5514 occurrences of four surgical actions, 6980 occurrences of 21 surgical instruments from seven categories and 495 skill classifications in five dimensions.

4.1.1. Phase

There was a high variability of the individual phase lengths and occurrences as visualized by Fig. 2 for the training dataset only. Taking training and test dataset together, in 28 videos (84.8%), all seven phases occurred. Each operation began with phase preparation (P0) and ended with phase gallbladder retraction (P6, 84.8%) or cleaning and coagulation (P5, 15.2%). Fig. 3 gives an overview of transition probabilities between phases for the whole dataset. Phase cleaning and coagulation (P5) was omitted in five videos (15.1%). The average number of phase transitions per video was 7.6 (± 1.8 SD). The mean phase length in the whole dataset ranged from 1.4 min ± 1.2 min (gallbladder retraction, P6) to 17.2 min ± 9.6 min (calot triangle dissection, P3). Inter-rater agreement before consensus as calculated with Fleiss' kappa per video was 0.92 ± 0.06 for the training dataset and 0.95 ± 0.04 for the test dataset. Here, video 1 was an outlier with 0.71 and all other videos had 0.86 or higher.

4.1.2. Action

For the whole dataset a total of 5514 actions were annotated, with an average of 167.1 (± 93.6 SD) actions per operation. Grasp (A0) was by number of occurrences the most common action ($n = 2830$ in total, $n = 28.6 \pm 34.6$ per operation, mean total action length per operation 9.9 $s \pm 11.3$ s), followed by hold (A1) less often, but much longer in total ($n = 2124$ in total, $n = 21.5 \pm 25.9$ per operation, mean total action length per operation 974.2 $s \pm 879.0$ s). Less common as well as shorter in time were cut (A2, $n = 315$ in total, $n = 9.6 \pm 9.1$ per operation, mean total action length per operation 7.7 $s \pm 10.6$ s) and clip (A3, $n = 245$ in total, $n = 7.4 \pm 3.3$ per operation, mean total action length per operation 9.6 $s \pm 5.4$ s). Whereas the actions grasp (A0) and hold (A1) occurred throughout the complete operation, the actions clip (A3) and cut (A2) were mainly conducted during phase clipping and cutting (P2). The left hand of the surgeon, as well as the assistant hand performed only grasp

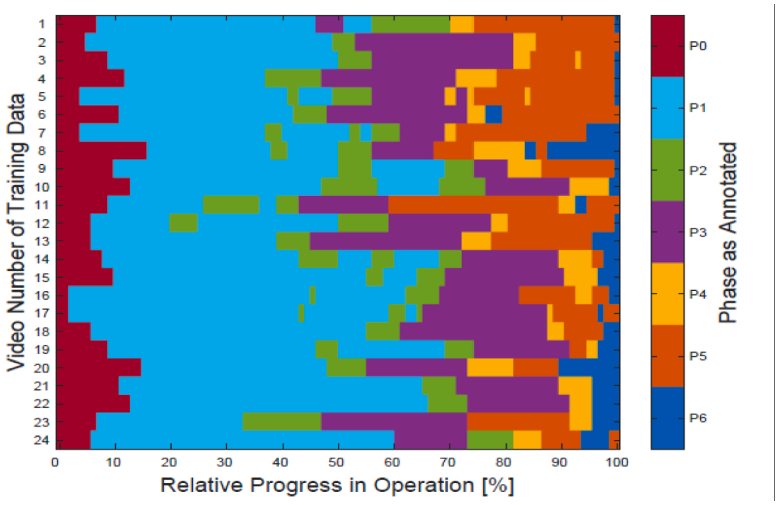

Fig. 2. Phase distribution in the training dataset. Operation duration is normalized to 100% and phase annotations are displayed in increments of 1%. Phases displayed are: preparation (P0), calot triangle dissection (P1), clipping and cutting (P2), gallbladder dissection (P3), gallbladder packaging (P4), cleaning and coagulation (P5) and gallbladder retraction (P6).

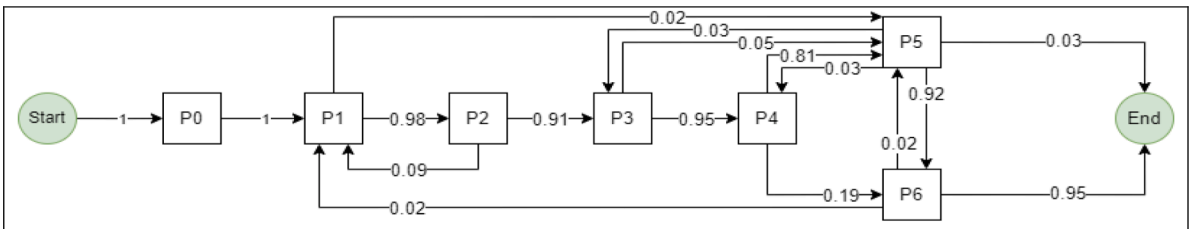

Fig. 3. Graph representation of the phases and their possible transitions. Phase transition probabilities were calculated based on the whole dataset (training and test dataset together).

(A0) and hold (A1). Cut (A2) and clip (A3) were performed by the surgeon's right hand only. Mean total action length per procedure of hold (A1) was 1497.8 *s* ± 620.6 s (left hand of the surgeon), 68.6 *s* ± 56.6 s (right hand of the surgeon) and 1356.3 *s* ± 830.1 s (assistant hand).

4.1.3. Instrument

We found a total of 6980 instrument occurrences in the whole dataset, with an average of 211.5 (± 134.0 SD) instrument occurrences per operation. The average instrument category presence per operation was highest for the grasper category (IC0) with 25.1 min ± 12.3 min, followed by the coagulation instruments category (IC2) with 18.1 min ± 8.4 min, specimen bag (IC5) with 4.2 min ± 3.4 min, the suction-irrigation category (IC4) with 3.8 min ± 6.7 min, the clipper category (IC1) with 1.5 min ± 0.9 min, scissors category (IC3) with 1.1 min ± 1.4 min and stapler category (IC6) with 0.03 min ± 0.16 min. Fig. 9 illustrates the variation in instrument category presence depending on the progress of the operation for the test dataset. For example, the grasper category (IC0) is almost continuously present in the test dataset, followed by the coagulation instruments category (IC2). The categories of clipper (IC1) and scissors (IC3) are with a small proportion in the middle third of the surgeries, which corresponds with the phase clipping and cutting (P2). In contrast, the category specimen bag (IC5) has a high presence in the last third of each operation, which corresponds to the phase gallbladder packaging (P4). In the beginning of the operation, no instruments are visible, which corresponds with the preparation phase (P0), as this only includes trocar insertion and visual inspection of the abdomen. Furthermore, according to the annotation rules, the appearance of the first instrument marks the beginning of the next phase.

4.1.4. Skill

The skill assessment based on the ranking components ranged between the medium grade 3 and the best grade 5 (Fig. 4 for the training dataset only). The grades 1 and 2 have never been given. The dataset contains surgeries of each level of difficulty.

4.1.5. Differences between centers

For the reference annotation all phases and actions occurred in all the centers. The instrument presence differed per phase depending on the center, in which the operation was performed. For example, the argon beamer (I11, average instrument presence per operation 29.83 *s* ± 61.05 s SD) was used exclusively at Heidelberg University Hospital, whereas the crocodile forceps (I17, average instrument presence per surgery 29.72 *s* ± 43.85 s) was used only at Salem Hospital and GRN-hospital Sinsheim.

The cases analyzed at Heidelberg University Hospital were rated to be more difficult than the cases at the smaller hospitals. There were 2 of 15 cases at Heidelberg University Hospital rated with the highest difficulty grade 5, another 2 cases with a grade 4. Only 3 cases were rated with the easiest difficulty 1. In contrast, grade 5 was not assigned to cases from either Salem Hospital or GRN-hospital Sinsheim. 8 of Salem's 15 cases were given a grade 1, as were all 3 cases from Sinsheim. Only one case from Salem was rated 4.

4.2. EndoVis challenge 2019

A comparison of the performance of the algorithms in the individual recognition tasks is presented in Fig. 5. A ranking of the algorithms together with ranking uncertainty is presented in Fig. 6. Appendix D gives a detailed overview of the challenge results per team per video of

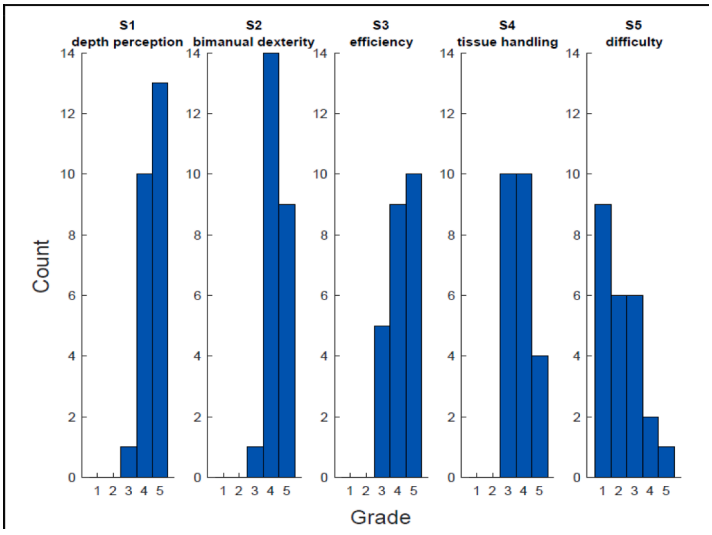

Fig. 4. Distribution of annotated grades for surgical skill and operation difficulty in training dataset. The histograms present the annotated reference distribution of surgical skill grades within the training dataset of the HeiChole benchmark. Displayed are the skill components depth perception (S1), bimanual dexterity (S2), efficiency (S3), tissue handling (S4) as well as the degree of difficulty (S5) of the operation.

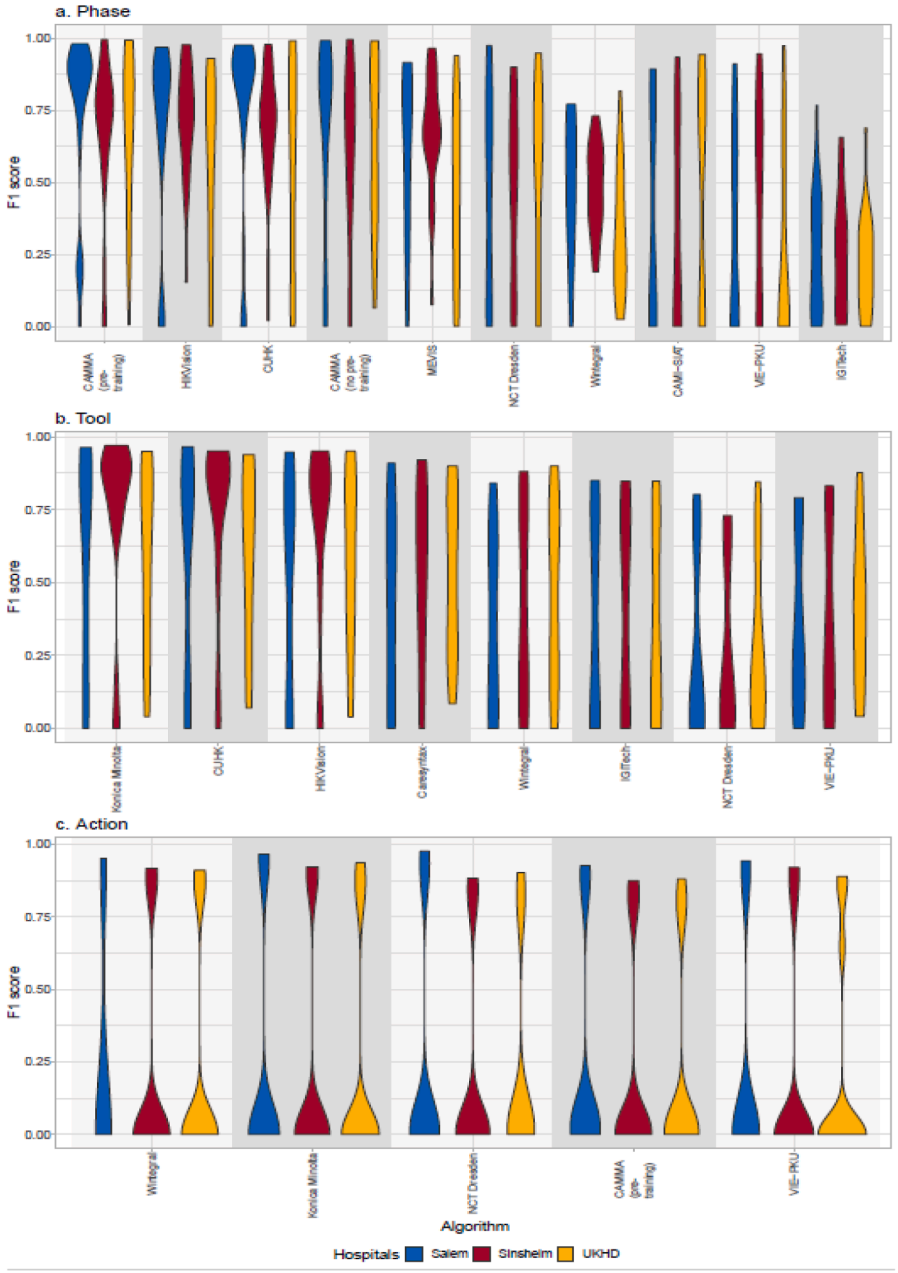

Fig. 5. Visualization of raw metric values. Violin plots representing the F1-scores for each algorithm for the phase, instrument and action task. The scores are shown separately for each hospital from which the videos were taken. Salem Hospital is blue, Heidelberg University Hospital (UKHD) is red, GRN-hospital Sinsheim is yellow.

the test dataset. However, only results that do not reveal too much information about the test dataset were performed, thereby preventing boost of algorithm performance by tailoring to the test data in future submissions (Maier-Hein et al., 2018). Thus, no metrics are published from which original annotations could be derieved (e.g. predicted probabilities). Fig. 10 presents an overview of algorithms performance across methods grouped by algorithm components. Figures include results for two results for team CAMMA (with and without pre-training), which were submitted after the challenge deadline and were thus not considered for the challenge awards.

4.2.1. Phase

The results achieved in the task of phase recognition are depicted in Fig. 5. For phase recognition the algorithms of HIKVision and CUHK, the winning teams participating in the EndoVis challenge, achieved a F1-score up to 65.4% and 65.0%. Since the difference between the two best methods of HIKVision and CUHK was so small and the second best team achieved superior detection results in most of the videos, both were declared the winner of the challenge and received the prize. As an example, Fig. 7 shows a confusion matrix of the different phases for HIKVision and all participants. The highest recognition rates were achieved for phase preparation (P0, 79.4%) and calot triangle dissection (P1, 82.8%). The lowest recognition rates were achieved for phases of gallbladder packaging (P4, 50.2%) and cleaning and coagulation (P5, 54.8%).

When including submissions that were received after the challenge deadline (CAMMA with and without pre-training), the best performing team CAMMA (with pre-training) and the runner-up HIKVision achieved significantly superior results to the algorithms from rank 5 to 10. No significant superiority was found for them compared to algorithms

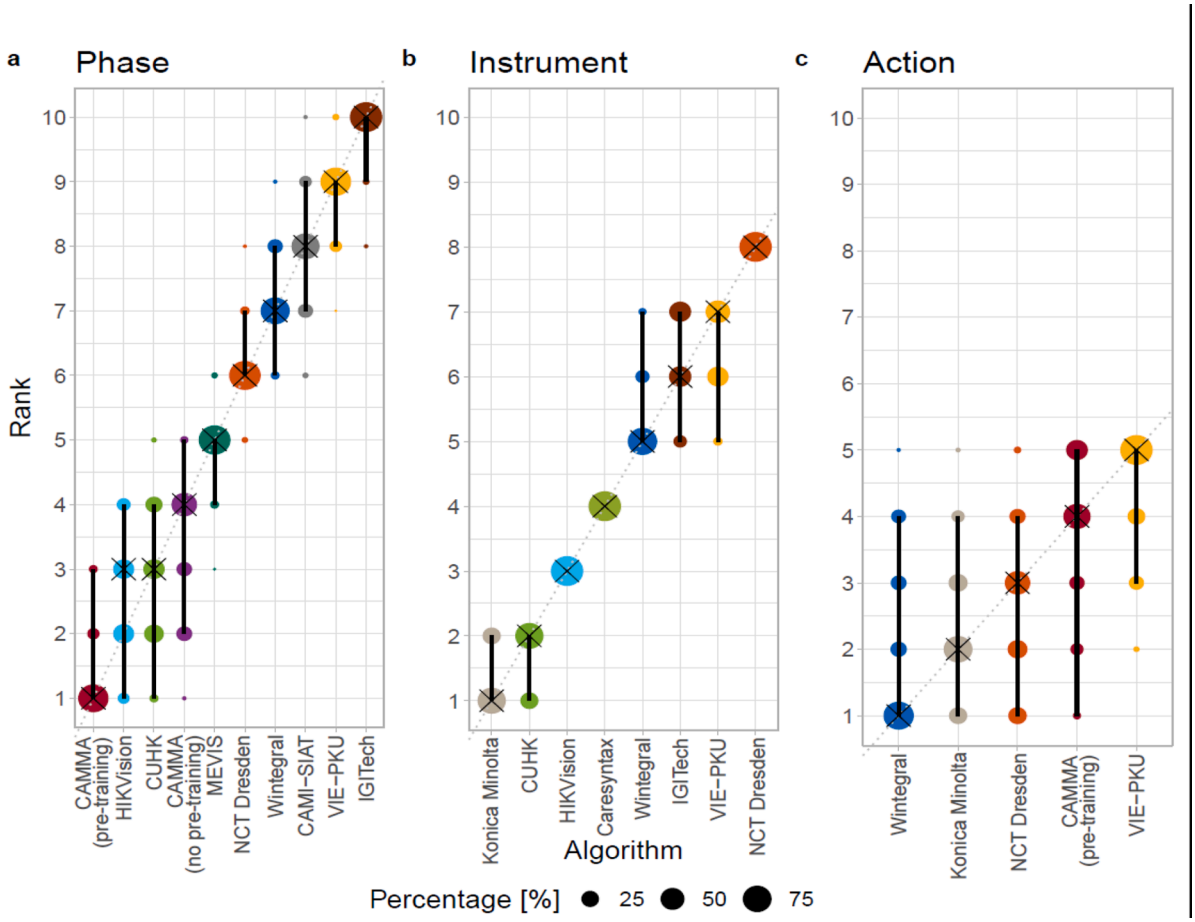

Fig. 6. Ranking uncertainty. Blob plots visualizing ranking uncertainty for (a) the phase, (b) the instrument and (c) the action task. The radius of the blobs represents the frequency of a rank achieved by an algorithm for 1000 bootstrap samples. The median rank is displayed by a black cross. The 95% confidence intervals of these 1000 bootstrap samples are shown by a black line and are whole-numbered quantiles of the whole-numbered ranks.

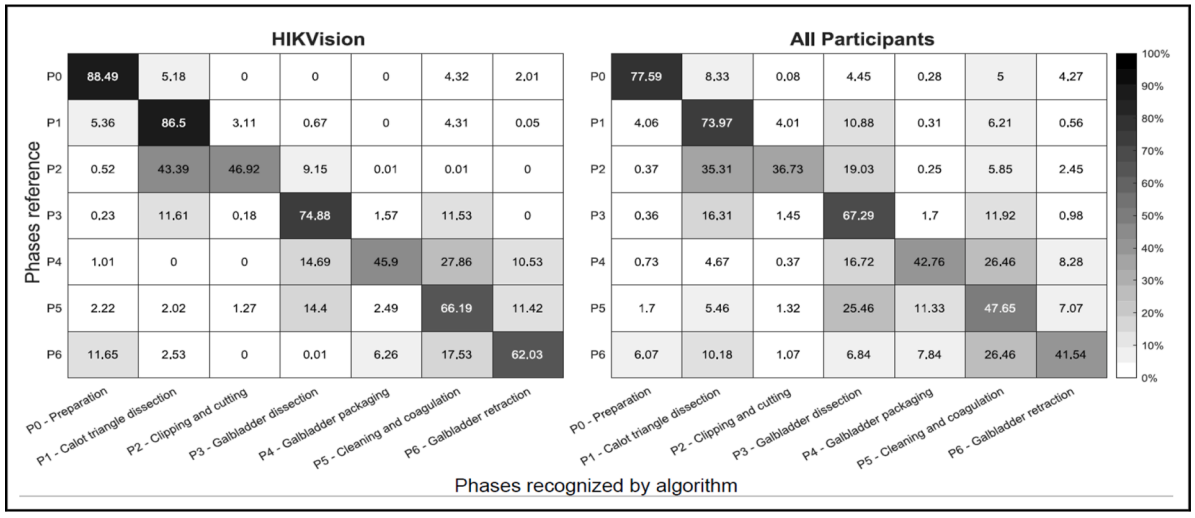

Fig. 7. Phase recognition results. Left: Phase recognition results for one of the challenge winners, HIKVision, for all test videos. Right: Results of all algorithms for all test videos. Due to rounding errors values in a row may not exactly add up to 100% per row as they should.

CUHK (rank 3) and CAMMA without pre-training (rank 4). This is also reflected by the ranking uncertainty analysis performed by using bootstrapping strategies. The frequency of ranks for each algorithm achieved across 1000 bootstrap samples together with the 95% confidence interval is illustrated in Fig. 6. It can be seen that the ranking is relatively stable, with the second and third algorithm being very close together. This is further supported by the mean Kendall's tau value for all bootstrap samples of 0.93 (median: 0.91, interquartile range (IQR): (0.91,0.96)).

4.2.2. Action

The task of action recognition was completed with the lowest recognition rates (Fig. 5). An F1 score of only 23.3% was reached by the best team Wintegral. The action grasp (A0) occurred during 1.2% (mean of all operations) of an operation, whereas the participants recognized it during 1.1% (mean recognition of all teams) of the operation. In comparison, the action hold (A1) occurred during 73.3% of an operation, whereas the participants recognized it during 83.2% of the operation. The actions cut (A2) and clip (A3) occurred least frequently, each during 0.4% of an operation. Cut (A2) was recognized during 1.0% and clip (A3) during 0.3% of the operation. It is apparent that all challenge algorithms almost exclusively recognize hold (A1), whether it occurred or not. Fig. 8 shows the exemplary results for the recognition of the action hold (A1) of all teams frame by frame.

None of the algorithms showed significant differences in the performance metric scores. Analyzing the ranking uncertainty revealed that the rankings were not as clearly separable as for the phase task (Fig. 6). The first three algorithms were closely together, therefore often exchanging their ranks. This is also reflected in a lower mean Kendall's tau of 0.66 (median: 0.60, IQR: (0.60,0.80)).

4.2.3. Instrument

Konica Minolta, the best team in the instrument presence detection task, achieved a F1-score of 63.8%. Their algorithm recognized 51,872 instrument occurrences in the test dataset, with an average of 5763.56 $\pm$ 4310.01 instrument occurences per operation. Compared to the reference, the algorithm detected a similar relative instrument presence within the operation progress, except for minor differences. For example, noticeable differences are within 10–35% of the operation progress, where the scissors (IC3) and specimen bag (IC5) were detected less frequently (Fig. 9). The mean results of all participants also showed considerable differences, especially in the presence detection of scissors (IC3) and specimen bag (IC5). Specimen bag (IC5) was detected continuously during the operation progress, whereas scissors (IC3) were detected less frequently than in the reference.

For the instrument task, the rankings were very stable across all bootstrap samples (Fig. 6). Ranks 6 and 7 were often interchanged and the first two algorithms were close together. Besides this, the ranking was quite clear with a mean Kendall's tau of 0.93 (median: 0.93, IQR: (0.93,1.00)). This was further shown by the statistical analysis. The winning team (Konica Minolta) and the runner-up (CUHK) both were superior over all other algorithms. The same trend could be seen for the other algorithms, most of them showing significant effects for their following ranks.

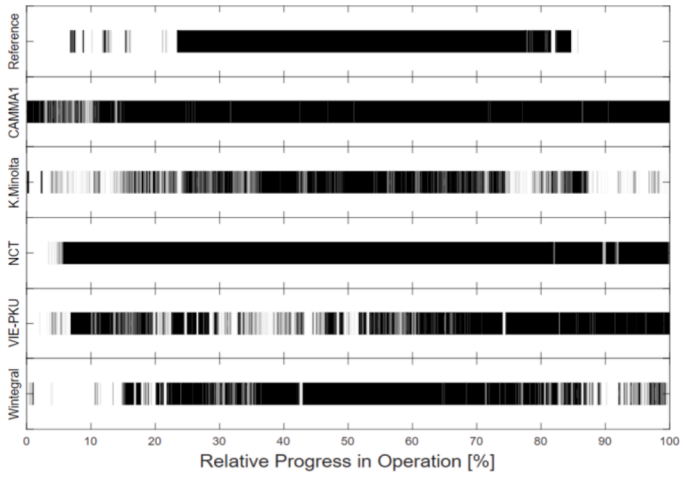

Fig. 8. Action recognition results. Comparison of all challenge algorithm results for recognition of action hold (A1) for test video HeiChole-25. The x-axis represents the relative time progress in the operation.

4.2.4. Skill

The average absolute error for skill assessment of the only participating team CareSyntax was 0.78 ($n = 1$ team). For each surgical video, a score of 4 was given for the ranking components depth perception (S1), bimanual dexterity (S2) and efficiency, and a score of 3 for tissue handling (S4). The difficulty (S5) was rated with a score of 2 for each video.

4.2.5. Performance analysis across methods

Fig. 10 gives an overview of the performance analysis across methods for the algorithm components multitask, basic architecture, temporal component, post processing, data augmentation, loss function, and optimizer as of Table 1, separated for phase recognition and instrument presence detection. The algorithm component pre-training is not depicted, because the only team not using it was CAMI-SIAT for phase recognition and therefore the influence of this variable cannot be evaluated. Furthermore, team CAMMA submitted results of the same algorithm for phase recognition with and without pretraining (Figs. 5 and 6) that allow for evaluating the effect of this component while keeping all other factors constant.

When interpreting Fig. 10, it is, however, important to note, that the plots from the descriptive analysis seem to suggest an effect as there are clear differences in median values. However, the results of the fitted models do not provide sufficient or conclusive evidence for this hypothesis. This is the case for phase recognition and instrument presence detection for all algorithm components with the exception of using a temporal component for phase recognition. Here, the results of the two-part models shows that including the temporal component improves the F1-score in the cases when F1-score is not zero (odd's ratio: 3.46 95% confidence interval CI [1.73,6.75]). This result is slightly difficult to interpret because the estimate for the logistic part of the model suggests that this feature also increases the probability of having an F1 score of zero (odd's ratio 6.62, 95%-CI [0.28, 188.67]. However, given that the estimate for the second part has a very large 95%-confidence interval, there is insufficient evidence for this part of the model and we cannot conclude that a temporal component can improve phase recognition.

The plots for action recognition are available in Appendix E only and were omitted from the manuscript as little insights could be drawn because of the low overall performance and action grasp being constantly predicted better than the others. Furthermore, Appendix E includes the underlying code, as well as separate figures for all algorithm components for phases, instruments, and actions.

4.3. HeiChole benchmark & online leaderboard

The results of the EndoVis challenge demonstrate a comparative validation of algorithms based on our comprehensive multi task dataset. New methods can be compared to these results by testing their algorithms with our dataset to allow for a validated reference benchmark. The online leaderboard is available on the challenge website on Synapse (HeiChole Benchmark Website, 2022). The training dataset will be published on Synapse simultaneously with this publication. However, we decided not to publish the test dataset to avoid boost of algorithm performance by tailoring to the test data (Maier-Hein et al., 2018). To use the dataset for future testing of novel algorithms against this benchmark, the following approach was implemented. To have their algorithm tested, a team must:

a Register on Synapse to register for the challenge on www.synapse.org/heichole (HeiChole Benchmark Website, 2022). As an option, a team can be created or an existing team can be joined.
b Access the training data.
c Train their algorithm on the provided training dataset.
d Submit the trained algorithm to HeiCholeOrganizers@synapse.org.

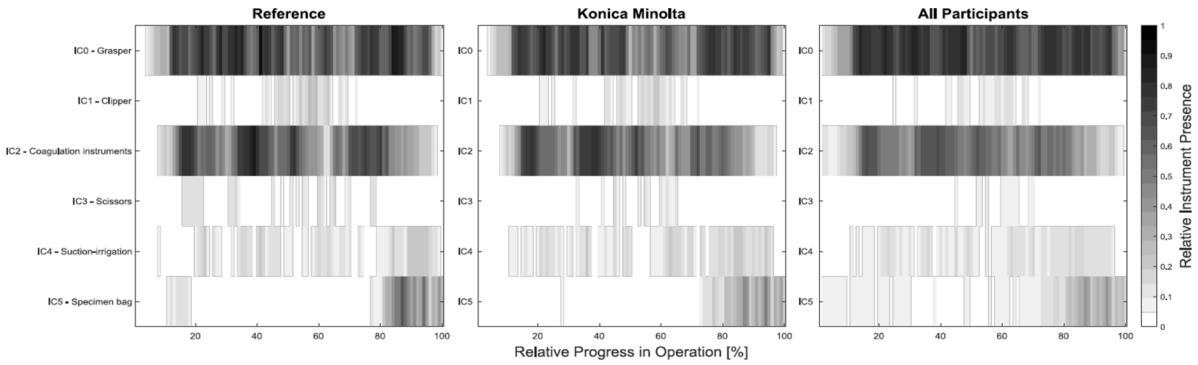

Fig. 9. Instrument presence detection results. For the test dataset this figure compares reference annotation, averaged results of all challenge algorithms and the best challenge algorithm (Konica Minolta) for the different instrument categories IC0 to IC5 (IC6 was not present in the test dataset). The horizontal axis of the graphs represent the relative time progress in the operation.

We will then evaluate the algorithm using the test dataset, report the results to the team, and publish the results in our online leaderboard available on the challenge website on Synapse upon request of the submitting team. To compute a performance rank the entries will be sorted, as each task consists of a scalar metric. The current leaderboard as of the publication of this study is shown in Table 2.

Furthermore, researchers are encouraged to not only submit novel algorithms, but also scripts for alternative metrics for evaluating the algorithms. We will then decide, whether this metric discloses too much information about the test data. If it doesn't, we will calculate the metric and report the results on the challenge website where appropriate.

5. Discussion

5.1. HeiChole benchmark

Our study introduces the HeiChole benchmark by presenting a novel open dataset for surgical workflow and skill analysis together with the results of the EndoVis challenge 2019, sub-challenge for surgical workflow and skill analysis, and its comparative validation of ML algorithms for automatic phase, instrument, action and skill recognition in laparoscopic cholecystectomy. In the following sections we will discuss achievements and limitations of the dataset and the performance of the algorithms for each task. Finally, we will outline future research directions for surgical workflow and skill analysis.

5.2. Dataset quality

One achievement of the dataset is a reliable reference annotation reflecting the variability of surgical video data in different hospitals for the commonly performed operation laparoscopic cholecystectomy. The varying resolutions and frame rates of the recorded videos illustrate the high variability of the recording of even a standard operation within different centers. Moreover, as described earlier, even the instruments used in the dataset vary between centers and to a certain extent even within one of the centers. This underlines the necessity to train recognition algorithms on datasets of various hospitals in order to increase the performance and thus the applicability and generalizability. The phase annotation was based on the established Cholec80 dataset (Twinanda et al., 2017). This addresses the problem of high heterogeneity of phase definitions reported by different groups (Garrow et al., 2020). Furthermore, we present the first dataset with surgical action annotated frame-by-frame in a clinical setup. Previously, for example in the JIGSAWS dataset, gestures were annotated only in an experimental setup (Ahmidi et al., 2017). In addition, the dataset is supplemented by a skill assessment based on the established GOALS score (Vassiliou et al., 2005). The reliability of phase and skill annotation is enhanced by multiple raters who annotated according to precise rules in an annotation protocol and resolved disagreement by consensus. Finally, the dataset is published to make it transparent, open and reusable for the scientific community.

Despite these accomplishments, there are also some limitations to the dataset. First of all, a limitation of reference annotation is the difficulty in sufficiently addressing interindividual differences in patient anatomy by annotation rules. Within the HeiChole benchmark the phase clipping and cutting (P2) was annotated for only one specific artery and one cystic duct, but not for additional vessels such as small veins. Consequently, the HeiChole benchmark does not reflect the potential diversity of the vascular anatomy as previously described (Andall et al., 2016) in order to limit the complexity of the recognition tasks. Furthermore, consistent action and instrument annotation was difficult when the performing instrument was pivoted out of the camera's view for example by repeatedly disappearing from camera view during the action hold (A1) or the instrument was no longer visible due to smoke generated by coagulation. To address this, the reliability of annotations, for example of instrument presence or electric coagulation, could be improved by the additional acquisition of medical device sensor data (Maier-Hein et al., 2021).

The effects of different resolutions and framerates on the performance of ML algorithms are difficult to foresee. While a higher framerate could probably be beneficial, increased resolution could introduce more noise to the processed image, thereby decreasing the algorithms' performance. Furthermore resolution differences can also act as confounders. In contrast to this, it is also possible that the image data added at a higher resolution is beneficial. Further research is needed to assess these effects accurately. Consequently, algorithm performance should not be directly compared using datasets with different parameters such as resolution and frame rate without further reflection on the effects of the parameters (Roß et al., 2021).

5.3. Surgical phase recognition

In comparison to the previously achieved average precision of 91% for automatic phase recognition on the Cholec80 dataset (Twinanda et al., 2017), the best participating team in this EndoVis challenge reached a F1-score of 65%. One possible reason for the lower results is that Cholec80 had a higher rigidity in the phase order with being linear from P1 to P7 except for variability in the order of P5 and P6 after completion of P4 as well as mutual transitions between P5 and P6 before start of P7. The added variability in our dataset, as reflected by the visualization of phase transitions probabilities in Fig. 3, may have rendered the phase recognition more complex. Furthermore, phases such as clipping and cutting (P2) and gallbladder packaging (P4) proved to be more challenging to recognize than others, because instruments typical for these phases, such as scissors for P2 or the specimen bag for P4, were also briefly used in previous or subsequent phases. In surgical

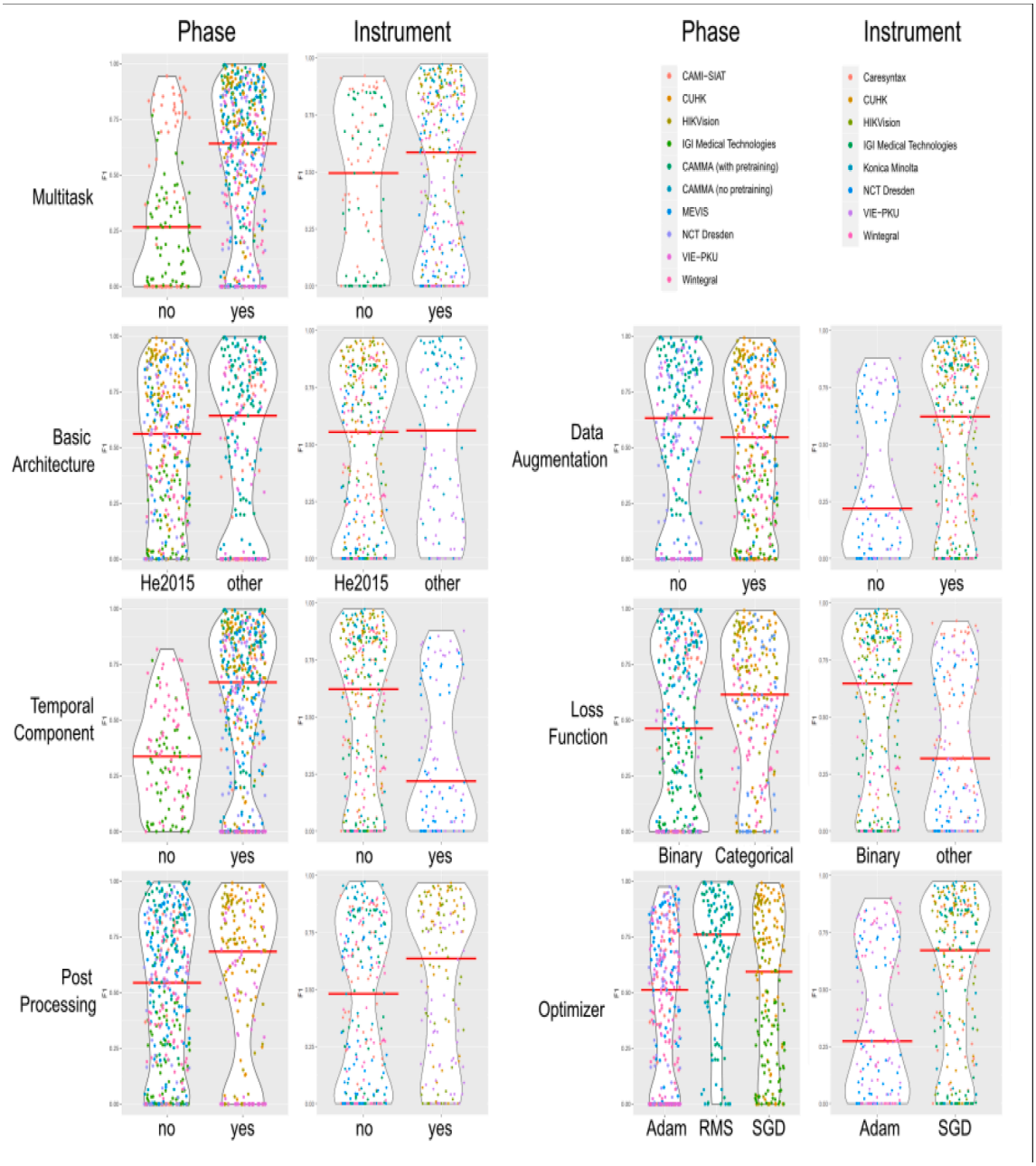

Fig. 10. Performance analysis across methods. This plotmatrix gives an overview of results for phase recognition and instrument presence detection depending on algorithm components as describe in detail in Section 3.2 and Table 1. Dots represent F1-values per team and phase (P0-P6) or instrument category (IC0 – IC6). Plots for action recognition are displayed in Appendix E only, because results are contorted by difference in performance between action grasp and the other actions (see 4.2.2. Action). Importantly, univariate analysis could not confirm visual differences.

reality, there are many variations in the workflow, especially for the more difficult cases that at the same time could benefit more from the help of computer assistance. Thus, the reflection of variations is important for the future clinical translation of the results. Another important factor in improving the results is increasing the size of the open dataset to enhance the training of ML algorithms. For example, phase recognition by ML on a dataset of 1243 laparoscopic cholecystectomies was published, but only hyperparameters, not the training data were made publicly available (Bar et al., 2020).

From a technical aspect, it is interesting to note that the top three teams (CAMMA, CUHK and HIKVision) all used a combination of CNNs with a recurrent network, here in all cases a LSTM. While it can be expected that these methods outperform approaches that do not take temporal information into consideration (e.g. Wintegral), it is of interest to note that the recurrent approaches also generally outperform methods based on 3D convolutions. This could be explained by the fact that, at least in theory, the approaches utilizing a recurrent architecture can recall information from any previously seen frame, while the hindsight of the 3D convolution-based networks is limited by the size of the convolutional kernels used and the amount of layers to more of a local scope. The top-performing team, CAMMA, actually utilizes and merges these global and local scopes, showing that both have their merit in phase recognition. In the performance analysis across methods the descriptive analysis (Fig. 10) gives the impression of highly improved results when using multitask learning, temporal component, and RMSProp as an optimizer, whereas ResNet (He et al., 2015), post processing, and choosing categorical cross-entropy loss (instead of binary) result in only little improvement. On the other hand, data augmentation

Table 2
Leaderboard: Rankings for the phase, instrument and action task. The table shows the ranking as of the publication of this study for the submissions of all participants according to the task. Participants marked with * have entered after the submission deadline of the EndoVis challenge.

Phase recognition task	Action recognition task							
	Instrument presence detection task							
Rank	Participant	average f1-score [%]	Rank	Participant	average f1-score [%]	Rank	Participant	average f1-score [%]
1	CAMMA (pre-training)*	68.78	1	Wintegral	23.28	1	Konica Minolta	63.82
2	HIKVision	65.38	2	Konica Minolta	22.83	2	CUHK	62.95
3	CUHK	64.98	3	NCT Dresden	22.62	3	HIKVision	58.20
4	CAMMA (no pre-training)*	63.60	4	CAMMA (pre-training)*	22.10	4	Caresyntax	50.13
5	MEVIS	57.30	5	VIE-PKU	21.75	5	Wintegral	41.59
6	NCT Dresden	49.00				6	IGI Medical Technologies	38.86
7	Wintegral	42.47				7	VIE-PKU	38.47
8	CAMI-SIAT	38.65				8	NCT Dresden	27.45
9	VIE-PKU	33.29						
10	IGI Medical Technologies	23.93						

seems to actually reduce performance for phase recognition. However, univariate analysis confirmed a difference only for the use of a temporal component and only if the result was not zero. This is also true for the use of pre-training that we compared based on both submissions of team CAMMA (Figs. 5 and 6).

For multicenter comparison it can be observed that the top methods performed best on the data collected from Salem and worst on the data collected from UKHD. The results on the data from Sinsheim, which was not included in the training data, are generally between those for the other two centers. This can be explained by the fact that the cases from UKHD tended to be the more complicated of a university hospital and thus have a higher variance in surgical progression. This generally makes recognizing the correct phase more complicated, while the other two centers tended to treat more standard cases. This observation leads us to conclude that, while it is important to include data from different centers in datasets, because tools and surgical guidelines tend to vary, it is at least equally important that the data should reflect the variance in patient anatomy and case difficulty.

5.4. Surgical action recognition

The algorithm of the leading team Wintegral has almost exclusively detected the action hold (A1), which is the most common action in the dataset and led to a class imbalance. Detailed analysis of the performance within the HeiChole-25 video demonstrates that this action was assigned to frames that were annotated as such in the reference, but also to frames that were assigned to other actions or where no action at all had taken place. This observation is consistent with the percentual average of the recognized actions over the entire test dataset.

Thus, the comparably high performance of phase recognition could not be reproduced for the more difficult recognition task of surgical actions. This may have different reasons. In this study, surgical actions during a cholecystectomy were annotated and recognized for the first time in the literature. Neither we nor the challenge teams could utilize reference results or training datasets of other research groups and thus broke new ground. Additionally, analogous to the instrument presence, several, mostly brief and subtle, actions may occur during the same frame. This complicates the recognition process. The results may surely be optimized by further research and larger datasets for algorithm training.

While the participating teams all used varying methods for the action detection, ranging from vanilla feed-forward neural networks to 3D convolutional network and to recurrent networks, the resulting performances were quite similar. The submitted methods were all able to similarly well detect the action “Hold” (F1 scores ranging from 0.84 to 0.91), but could not reliably detect any other action. The average F1 scores for each action apart from “Hold” for each method was well below 0.03. Since “Hold” was the most common action in the dataset, we conclude that the current training dataset did not adequately mirror the variance in the other actions. For this reason we did not perform a univariate analysis of performance across methods for action recognition.

5.5. Instrument presence detection

Overall, the task of instrument presence detection can be as challenging as phase recognition. In addition, several instrument categories can be used within a single frame. Also the detection of the small instrument tips is impeded by the non-static camera, quickly changing perspectives and the resulting motion blur, whereas phase recognition can utilize information from the whole surgical scene including the patient anatomy.

Furthermore, the analysis of the instrument presence distribution in the phases within the reference test dataset highlights that certain instrument categories were used more frequently during certain phases according to their functionality leading to a class imbalance. This conclusion can serve as a reference point for further research on automatic workflow recognition, but should be supported by other sources of information, such as real-time sensor data, as variations may occur. The instrument category grasper (IC0) is the only instrument used in phases P1 through P6 with a high proportional presence. This is little surprising, since the instruments of this category are universally applicable for trivial grasping as well as for blunt dissection. The other instruments also fit, according to their function, into the phases in which their highest proportional presence was detected. For example, the category scissors (IC3) are mainly used during clipping and cutting (P2), when they are required to cut the vessels, and to a lesser extent during calot triangle dissection (P1), for the dissection of highly adherent tissues. Similarly, the instruments from the category clipper (IC1) were mainly used during clipping and cutting (P2), with small variations to the previous and subsequent phases, for example, for clipping varying vascular supply to the gallbladder.

The performance analysis across methods showed that for instrument presence detection, the temporal component does not seem to be as important as for tasks such as phase recognition. The highest performing method (Konica Minolta) for example did not include any temporal information at all, instead opting to explicitly learn the co-occurrences of different types of surgical instruments. Data augmentation also seems to help increase accuracy on the given task, as it can be noted that the top 5 methods all used some form of data augmentation,

while the bottom 3 methods did not. The direct comparison (Fig. 10) furthermore gives the impression that SGD as an optimizer (instead of Adam), binary cross-entropy loss (instead of categorical or a combination), and post-processing are correlated with higher performance. However, for none of the algorithm components there was a difference in the univariate analysis that can be explained by the choice of that component.

Regarding multicenter comparison, it is interesting to note that most submitted methods actually performed best on data from Sinsheim, which was not represented in the training data, and very closely followed by Salem. The make of the tools used in Sinsheim are identical to the ones in Salem, meaning that models that perform well on Salem data should generalize to Sinsheim. In the videos from UKHD, a much larger variety of instruments from a different vendor are used, increasing the difficulty of determining the right class.

5.6. Skill assessment

The results of the challenge algorithm for skill assessment do not sufficiently reflect the reference, because every video was repeatedly rated with the same score for each skill component. However, these results are not representative as only one team participated in this task. Here, the difficulty in skill assessment is to provide a holistic evaluation considering many aspects of the video. For this reason, the reference annotation was also performed by two independent raters in order to achieve more objective results. The raters did not use the full range of the scores in the final grading. The lack of exploitation of the lower grade scores was already noted in earlier works on surgical skill rating (Doyle et al., 2007). It can be attributed either to the best possible effort of the surgeons and thus the selection of the data, or an error of central tendency, in which the raters distribute average ratings regardless of performance.

5.7. Multitask learning

As an important characteristic of our challenge we provided multiple recognition tasks on the same dataset. We did this based on the assumption that the tasks may interact, e.g. the phase recognition could be improved by the joint action or instrument recognition. Together, they will form a more comprehensive surgical scene understanding. Indeed, most of the teams (except IGITech and CAMI-Siat) used a multi-task approach, i.e. learned tasks simultaneously. Our visual analysis of the algorithms components (Fig. 10) gives the impression of a better performance with multi-task learning. However, from the univariate analysis we could not show that multitask actually makes a difference. Thus, it is unclear whether the difference in the descriptive statistics is a mere selection bias confounded by the teams. Whereas the lack of a comparative validation on the level of algorithm components and design choices is a limitation of our study, the investigation of e.g. the similar algorithms with and without multitask learning would have made the design of this challenge probably too confusing. Thus, this can be a matter of future of investigation on our benchmark when researchers decide to investigate this matter with our open dataset and the online-leaderboard. Furthermore, this is an interesting idea for future challenges.

5.8. Reducing bias in the challenge

The introduction and validation of the HeiChole benchmark succeeded by means of the challenge design of the EndoVis challenge. This allowed the comparison of the results of several different algorithms submitted by international teams, demonstrating the usability of the HeiChole benchmark as a validation tool for phase, instrument, action, and skill recognition. As biomedical challenges become increasingly important for the evaluation and validation of methods for surgical data science, quality control is of utmost relevance to ensure reproducibility and comparability of the results (Maier-Hein et al., 2018). For example, this was addressed in the EndoVis challenge by providing detailed information about the data collection and processing in response to the challenge design questionnaire required by MICCAI 2019 for submission. In order to avoid possible boost of algorithm performance by tailoring to the test data, the test dataset and the corresponding reference annotation was not published. This prevents any manipulation of the algorithm performance, so that all (future) teams participate under equal conditions.

Furthermore, there are some technical parameters to consider, which affect the challenge performance of the teams. Thus, the effects of different resolutions and framerates on the performance of ML algorithms are difficult to foresee. While a higher framerate could probably be beneficial, increased resolution could introduce more noise to the processed image, thereby decreasing the algorithms' performance. In contrast to this, it is also possible that the image data added at a higher resolution is beneficial. Further research is needed to assess these effects accurately. Consequently, challenge algorithm performance should not be directly compared using datasets with different parameters such as resolution and frame rate without further reflection on the effects of the parameters.

Finally, it is important to acknowledge that the plots shown in this paper should only be considered as an approximate representation of the data shown. Comparing videos with different numbers of frames makes it necessary to evaluate the progress of the operations in relative values, so rounding is necessary to achieve an integer number of frames in a given percentage range. Therefore, at the end of each operation, a number of frames is being omitted in the visualization. This number of frames is always less than one percent. These problems are inevitable when trying to visualize large and inhomogeneous datasets, however, they should be considered when interpreting the challenge results.

5.9. Future research directions

In future studies, the applicability of automatic phase and action recognition, instrument presence detection and skill assessment should be investigated for more complex surgical procedures, such as esophageal or pancreatic surgery for cancer. Larger datasets are essential to improve the performance of the algorithms, ideally with addition of medical device sensor data and anesthesiological data from vital sign monitoring and drug administration to complement manual reference annotation. Also, to speed up annotation processes the development of time and cost effective annotation tools should be realized. The ML algorithms should incorporate surgical knowledge, such as correlation of certain instruments and phases.

Regarding the challenge design, future challenges may already integrate performance analysis across methods a priori, for example by asking teams to submit two algorithms with a specific component being changed that is a current matter of scientific investigation.

6. Conclusion

Surgical workflow and skill analysis are promising technologies to support the surgical team, but there is still room for improvement, as shown by our state of the art comparison of ML algorithms on the novel HeiChole benchmark. The continued creation of open, high-quality datasets is of utmost importance in order to allow the development of accurate and robust ML algorithms as a foundation for artificial intelligence and cognitive robots in surgery.

User notes

The dataset was published under a Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license on Synapse (HeiChole Benchmark Website, 2022), which means that it will be publicly available for non-commercial usage. Should you wish to use or refer to

this dataset, you must cite this paper. The licensing of new creations must use the exact same terms as in the current version of the dataset.

CRediT authorship contribution statement

M. Wagner, A. Kisilenko: Conceptualization, Methodology, Validation, Writing - Original Draft, Writing - Review & Editing, Formal analysis, Data Curation, Investigation, Resources, Project administration. **S. Bodenstedt:** Conceptualization, Methodology, Validation, Writing - Original Draft, Writing - Review & Editing, Formal analysis, Data Curation, Visualization, Data Analysis, Software, Resources, Project administration. **D. Tran, P. Heger, L. Mündermann, F. Nickel, M. von Frankenberg, F. Mathis-Ullrich:** Resources, Writing - Review & Editing, Data Curation. **D.M. Lubotsky, A. Reinke, C. Reid, A. Kopp-Schneider:** Visualization, Formal analysis, Resources, Writing - Review & Editing. **B. Müller, T. Davitashvili, M. Capek:** Data Curation, Writing - Review & Editing. **T. Yu, A. Vardazaryan, C.I. Nwoye, N. Padoy, X. Liu, E.J. Lee, C. Disch, H. Meine, T. Xia, F. Jia, S. Kondo, W. Reiter, Y. Jin, Y. Long, M. Jiang, Q. Dou, P. A. Heng, I. Twick, K. Kirtac, E. Hosgor, J.L. Bolmgren, M. Stenzel, B. von Siemens, Long Zhao MSc, Zhenxiao Ge Msc, Haiming Sun MD, Di Xie PhD, Mengqi Guo, Daochang Liu**: Data Analysis, Software, Writing - Review & Editing. **H.G. Kenngott, L. Maier-Hein, S. Speidel, B.P. Müller-Stich:** Conceptualization, Writing - Review & Editing, Supervision, Resources, Funding acquisition.

Declaration of Competing Interest

M. Wagner, B.-P. Müller-Stich, S. Speidel and S. Bodenstedt worked with medical device manufacturer KARL STORZ SE & Co. KG in the projects "InnOPlan", funded by the German Federal Ministry of Economic Affairs and Energy (grant number BMWI 01MD15002E) and "Surgomics", funded by the German Federal Ministry of Health (grant number BMG 2520DAT82D and BMG 2520DAT82A). Lars Mündermann is an employee of KARL STORZ SE & Co. KG. A. Reinke works with the Helmholtz Imaging Platform (HIP), a platform of the Helmholtz Incubator on Information and Data Science. S. Kondo was an employee of Konica Minolta Inc. when this work was done. Wolfgang Reiter is an employee of Wintegral GmbH, a subsidiary of medical device manufacturer Richard Wolf GmbH. I. Twick, K. Kirtac, E. Hosgor, J. Lindström Bolmgren, M. Stenzel and B. von Siemens are employees of Caresyntax GmbH. Felix Nickel received travel support for conference participation as well as equipment provided for laparoscopic surgery courses by KARL STORZ SE & Co. KG, Johnson & Johnson, Intuitive Surgical, Cambridge Medical Robotics, and Medtronic. The other authors have no conflicts of interest.

Data availability

Data is published on www.synapse.org/heichole.

Acknowledgements

We thank T. Nguyen (team IGITech) for participation in the challenge. This person was not listed as a coauthor, because of not responding to the requested authors' confirmation, but signed an agreement to publish the results when participating in the challenge.

Funding

This work was supported by the National Center for Tumor Diseases (NCT) Heidelberg within the Cancer-Therapy-Program „Surgical Oncology", by the German Federal Ministry of Economic Affairs and Energy within project "InnOPlan" (grant number BMWI 01MD15002E), by the German Federal Ministry of Health within project "Surgomics" (grant number BMG 2520DAT82D), 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), by the German Academic Exchange Service (DAAD) with a scholarship for T. Davitashvili for studying medicine in Germany (Scholarship program ID: deutsche Auslandsschulen, 2017 (57314589)), by the Helmholtz Imaging Platform (HIP), a platform of the Helmholtz Incubator on Information and Data Science, by Agence Nationale pour la Recherche (ANR-16-CE33-0009, ANR-10-IAHU-02) and Banque Publique d'Investissement (BPI CONDOR), by the Richard und Annemarie Wolf-Stiftung, by the Guangdong Key Area Research and Development Program (2020B010165004), by the Shenzhen Key Basic Science Program (JCYJ20180507182437217), by the Hong Kong RGC TRS Project No. T42-409/18-R, and by the Clinical Medicine Plus X-Young Scholars Project of Peking University

None of the funding sources had influence on study design, the collection, analysis and interpretation of data, the writing of the report or the decision to submit the article for publication.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2023.102770.

Appendices

Appendix A

Annotation rules

In the following, additional details on the annotation protocol for phase, action, instrument, and skill annotation as described in the "materials & methods" section of the main article is presented. The annotations are at least two frames long, as the video annotation research tool Anvil does not allow the independent annotation of a single frame.

Phase annotation rules

The phase annotation includes seven phases: preparation (P0), calot triangle dissection (P1), clipping and cutting (P2), gallbladder dissection (P3), gallbladder packaging (P4), cleaning and coagulation (P5) and gallbladder retraction (P6). The phases do not necessarily occur in a fixed order.

The preparation phase (P0) begins as soon as the camera is inserted into the abdomen for the first time and the optical trocar is no longer visible in the image. This phase includes orientation in the patient's abdomen and placement of additional trocars for instrument insertion.

Calot triangle dissection (P1) begins as soon as an instrument appears in the image. The phase includes the dissection of connective tissue and fat around the gallbladder and the nearby abdominal cavity to reach the calot triangle and the preparation of the cystic artery and the cystic duct. One main cystic artery and one cystic duct were assumed for each video to reduce complexity, so that clipping and cutting (P2) was annotated for only one specific artery and one cystic duct, but not for additional vessels.

The phase clipping and cutting (P2) begins with the appearance of a clipper, which clips the cystic artery or the cystic duct before cutting. The phase does not begin when a clipper appears to stop bleeding or to clip other vessels. This phase may switch with the previous phase (P1). The change is defined by the appearance of a preparation instrument, for example the electric hook or the overholt, which actually dissects. The change does not take place if an instrument that is not dissecting appears in the image. In case that the electric hook or scissors start dissecting immediately after cutting the vessels, the phase P1 begins as soon as tissue is cut or cauterized.

Gallbladder dissection (P3) begins after the clipping and cutting (P2) as soon as the preparation scissors or the electric hook touches the gallbladder. The phase also begins with accidental contact, without

cutting intention. It includes the clipping of accidentally injured vessels and accessory biliary arteries which are not defined as the main cystic vessels.

The gallbladder packaging phase (P4) begins as soon as the specimen bag enters the picture. This phase includes the recovery of the gallbladder as well as any gallstones. In the case of collecting spilled gallstones after the gallbladder has been packed in the specimen bag, the phase is maintained as long as the stones are placed in the bag and it is not closed. This phase ends with the complete closure of the bag or if the focus is diverted from the specimen bag to start the following phase.

Cleaning and coagulation (P5) begins as soon as the focus of the camera is averted from the specimen bag to check for tissue damage, coagulation begins using electricity, or drainage enters the picture. If the surgical site is checked and no cleaning or coagulation takes place, the phase does not begin. If the retraction of the gallbladder takes place before cleaning and coagulation, the phase begins as soon as the liver falls into the camera focus.

The gallbladder retraction phase (P6) begins when the specimen bag is last grasped to remove it from the abdomen.

The end of the operation is defined by the optical trocar taking 50 or more percent of the image as the camera is pulled out of the abdomen for the last time.

Action annotation rules

The action annotation includes the four actions grasp (A0), hold (A1), cut (A2) and clip (A3). In connection with this, the performer of the action was annotated as the left hand of the surgeon, right hand of the surgeon or hand of the assistant.

The annotation of the action grasp (A0) started as soon as the grasper began to close the instrument and ended with its complete closure. It did not matter whether the grasper grasped tissue or closed without grasping tissue, for example, when the surgeon missed the tissue. Thus, the angle of closure of the instrument at the end of the action varied depending on the thickness of the grasped tissue between the two tips of the grasper. This action is only performed by instruments with the name grasper (I0–3, I10, I16, I18, I19) or forceps (I14, I15, I17). This action is not performed by the overholt (I4), since its primary function in laparoscopic cholecystectomy is to separate tissue bluntly by spreading.

The annotation of the action hold (A1) started after the action grasp, if the grasper successfully grasped tissue. It ended as soon as the grasper began to release its grip or when tissue started to slip out of the grip. In this context the end of the action hold is not always apparent, for example when the instrument releases gripped tissue outside of the camera view, is concealed by smoke or is pulled out of the abdomen while holding, as is often the case with the retraction of the specimen bag in the final phase. For this reason, the started action is annotated, even if the performing instrument is no longer visible. To recognize the end of the action correctly, the tension of the tissue must be considered. For example, if the assistant lifts the gallbladder until the performed instrument can no longer be seen, it is necessary to look for a decrease in tension due to the gallbladder tissue slipping out of the grip and to recognize the associated need for regrasp. In case of pulling the instrument out of the abdominal cavity while holding something, for example the specimen bag, gallstones or misplaced clips, the action hold is annotated until the held object disappears from the camera view and is pulled out of the abdomen.

The annotation of the action cut (A2) started as soon as the cutting instrument began to close around tissue and ended with the complete closure of the cutting instrument. In contrast to action grasp (A0), closing a cutting instrument without tissue was not annotated as cutting. This action also includes the tearing of the tissue by the scissors.

The annotation of the action clip (A3) started as soon as the clipper beganto close around tissue and ended with the clipper beginning to release its grip after the application of a clip.

Instrument annotation rules

21 performing instruments were annotated and additionally divided into the seven categories grasper (IC0), clipper (IC1), coagulation instruments (IC2), scissors (IC3), suction-irrigation (IC4), specimen bag (IC5) and stapler (IC6).

For instrument presence, an instrument was considered visible as soon as its characteristic instrument tip appeared in the image. The annotation continues when the tip disappears later and only the shaft of the instrument remains visible. An example is the disappearance of the instrument tip of the electric hook behind tissue during dissection. Most importantly, the shaft should be clearly associated with an instrument tip.

If the instrument shaft enters the picture without its tip having been visible before, it is referred to as the undefined instrument shaft (I30), because even a human annotator would have difficulties recognizing a particular instrument due to the identical looking shaft. Three exceptions are the suction-irrigation (I12), stapler (I20) and the clippers (I8, I9), as these instruments have characteristic shafts.

One special case for the annotation of instrument presence is the concealment of an instrument by smoke resulting from coagulation. If the instrument tip and shaft are completely or partially obscured for a short time, the instrument is annotated as long as it was not removed from the camera view in the meantime.

Skill annotation rules

As outlined in the methods section, the skill assessment was conducted using a modified GOALS score. It has been validated for video assessment of laparoscopic skills, including the five domains depth perception (S1), bimanual dexterity (S2), efficiency (S3), tissue handling (S4) and autonomy (Vassiliou et al., 2005). The item "autonomy" was omitted in our study, because a valid assessment based solely on intra-abdominal video alone is not possible without information about what was spoken during the operation or how much assistance was provided by a senior surgeon. The difficulty of the operation (S5) was additionally annotated based on Chang's adaptation of the GOALS-score (Chang et al., 2007). Here, parameters such as inflammatory signs, adhesions and individual anatomical conditions were used to objectify the assessment of the skill. Thus, the skill assessment in this study included five ranking components. Skill was annotated for the complete operation and additionally for phases calot triangle dissection (P1) and gallbladder dissection (P3).

Appendix B. Challenge design protocol

The challenge design protocol as submitted before the challenge to MICCAI is attached as a separate pdf document in the supplementary material.

Appendix C. Machine Learning Methods of participating teams

Tables displaying machine learning methods of participating teams sorted by task are available in the supplementary material. PDF can be found in the supplementary material.

Appendix D. Detailed challenge results

Detailed challenge results include precision, recall, f1 score, and specificity per teams and per video and per phase, action, and instrument category, respectively. Thus, data is provided in separate spreadsheet files. The phase spreadsheet furthermore includes Fleiss' kappa for inter-rater agreement before consensus. Spreadsheets can be found in the supplementary material

Appendix E. Performance analysis across methods

Code of statistical analysis, as well as separate figures for all algorithm components for phases, instruments, and actions are included as separate html-documents in the supplementary material.

References

Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B.B., Zappella, L., Khudanpur, S., Vidal, R., Hager, G.D., 2017. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Trans. Biomed. Eng. 64, 2025–2041. https://doi.org/10.1109/TBME.2016.2647680.

Aksamentov, I., Twinanda, A.P., Mutter, D., Marescaux, J., Padoy, N., 2017. Deep Neural Networks Predict Remaining Surgery Duration from Cholecystectomy Videos. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 586–593. https://doi.org/10.1007/978-3-319-66185-8_66.

Al Hajj, H., Lamard, M., Conze, P.-H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., Sahu, M., Galdran, A., Araújo, T., Vo, D.M., Panda, C., Dahiya, N., Kondo, S., Bian, Z., Vahdat, A., Bialopetravičius, J., Flouty, E., Qiu, C., Dill, S., Mukhopadhyay, A., Costa, P., Aresta, G., Ramamurthy, S., Lee, S.-W., Campilho, A., Zachow, S., Xia, S., Conjeti, S., Stoyanov, D., Armaitis, J., Heng, P.-A., Macready, W.G., Cochener, B., Quellec, G., 2019. CATARACTS: challenge on automatic tool annotation for cataRACT surgery. Med. Image Anal. 52, 24–41. https://doi.org/10.1016/j.media.2018.11.008.

Andall, R.G., Matusz, P., du Plessis, M., Ward, R., Tubbs, R.S., Loukas, M., 2016. The clinical anatomy of cystic artery variations: a review of over 9800 cases. Surg. Radiol. Anat. 38, 529–539. https://doi.org/10.1007/s00276-015-1600-y.

Bar, O., Neimark, D., Zohar, M., Hager, G.D., Girshick, R., Fried, G.M., Wolf, T., Asselmann, D., 2020. Impact of data on generalization of AI for surgical intelligence applications. Sci. Rep. 10, 22208. https://doi.org/10.1038/s41598-020-79173-6.

Bodenstedt, S., Wagner, M., Mündermann, L., Kenngott, H., Müller-Stich, B., Breucha, M., Mees, S.T., Weitz, J., Speidel, S., 2019. Prediction of laparoscopic procedure duration using unlabeled, multimodal sensor data. Int. J. Comput. Assist. Radiol. Surg. 14, 1089–1095. https://doi.org/10.1007/s11548-019-01966-6.

Bürkner, P.-C., 2017. **brms** : an *R* package for Bayesian multilevel models using *Stan*. J. Stat. Softw. 80 https://doi.org/10.18637/jss.v080.i01.

Carreira, J., Zisserman, A., 2017. Quo Vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI. IEEE, pp. 4724–4733. https://doi.org/10.1109/CVPR.2017.502.

Chang, L., Hogle, N.J., Moore, B.B., Graham, M.J., Sinanan, M.N., Bailey, R., Fowler, D. L., 2007. Reliable assessment of laparoscopic performance in the operating room using videotape analysis. Surg. Innov. 14, 122–126. https://doi.org/10.1177/1553350607301742.

Doyle, J.D., Webber, E.M., Sidhu, R.S., 2007. A universal global rating scale for the evaluation of technical skills in the operating room. Am. J. Surg. 193, 551–555. https://doi.org/10.1016/j.amjsurg.2007.02.003.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76, 378–382. https://doi.org/10.1037/h0031619.

Funke, I., Mees, S.T., Weitz, J., Speidel, S., 2019. Video-based surgical skill assessment using 3D convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. 14, 1217–1225. https://doi.org/10.1007/s11548-019-01995-1.

Garrow, C.R., Kowalewski, K.-F., Li, L., Wagner, M., Schmidt, M.W., Engelhardt, S., Hashimoto, D.A., Kenngott, H.G., Bodenstedt, S., Speidel, S., Müller-Stich, B.P., Nickel, F., 2020. Machine learning for surgical phase recognition: a systematic review. Ann. Surg. https://doi.org/10.1097/SLA.0000000000004425.

Greenberg, C.C., Ghousseini, H.N., Pavuluri Quamme, S.R., Beasley, H.L., Frasier, L.L., Brys, N.A., Dombrowski, J.C., Wiegmann, D.A., 2018. A statewide surgical coaching program provides opportunity for continuous professional development. Ann. Surg. 267, 868–873. https://doi.org/10.1097/SLA.0000000000002341.

Hashimoto, D.A., Rosman, G., Witkowski, E.R., Stafford, C., Navarette-Welton, A.J., Rattner, D.W., Lillemoe, K.D., Rus, D.L., Meireles, O.R., 2019. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. Ann. Surg. 270, 414–421. https://doi.org/10.1097/SLA.0000000000003460.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. ArXiv151203385 Cs.

HeiChole Benchmark Website, 2022. www.synapse.org/heichole [WWW Document]. 10.7303/syn18824884.

Hinton, G., Srivastava, N., Swersky, K., 2012. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT. IEEE, pp. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.

Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.-W., Heng, P.-A., 2018. SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans. Med. Imaging 37, 1114–1126. https://doi.org/10.1109/TMI.2017.2787657.

Katić, D., Wekerle, A.-L., Görtler, J., Spengler, P., Bodenstedt, S., Röhl, S., Suwelack, S., Kenngott, H.G., Wagner, M., Müller-Stich, B.P., Dillmann, R., Speidel, S., 2013. Context-aware Augmented Reality in laparoscopic surgery. Comput. Med. Imaging Graph. 37, 174–182. https://doi.org/10.1016/j.compmedimag.2013.03.003.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A., 2017. The kinetics human action video dataset.

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81. https://doi.org/10.2307/2332226.

Kiefer, J., Wolfowitz, J., 1952. Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. 23, 462–466. https://doi.org/10.1214/aoms/1177729392.

Kingma, D.P., Ba, J., 2017. Adam: a method for stochastic optimization. ArXiv14126980 Cs.

Kipp, M., 2014. ANVIL. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.024.

Korndorffer, J.R., Hawn, M.T., Spain, D.A., Knowlton, L.M., Azagury, D.E., Nassar, A.K., Lau, J.N., Arnow, K.D., Trickey, A.W., Pugh, C.M., 2020. Situating artificial intelligence in surgery: a focus on disease severity. Ann. Surg. 272, 523–528. https://doi.org/10.1097/SLA.0000000000004207.

Lalys, F., Jannin, P., 2014. Surgical process modelling: a review. Int. J. Comput. Assist. Radiol. Surg. 9, 495–511. https://doi.org/10.1007/s11548-013-0940-5.

Loukas, C., 2018. Video content analysis of surgical procedures. Surg. Endosc. 32, 553–568. https://doi.org/10.1007/s00464-017-5878-1.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., Feldmann, C., Frangi, A.F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B.A., März, K., Maier, O., Maier-Hein, K., Menze, B.H., Müller, H., Neher, P.F., Niessen, W., Rajpoot, N., Sharp, G.C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A.A., van der Sommen, F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., Kopp-Schneider, A., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. 9, 5217. https://doi.org/10.1038/s41467-018-07619-7.

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., Nakawala, H., Park, A., Pugh, C., Stoyanov, D., Vedula, S.S., Cleary, K., Fichtinger, G., Forestier, G., Gibaud, B., Grantcharov, T., Hashizume, M., Heckmann-Nötzel, D., Kenngott, H.G., Kikinis, R., Mündermann, L., Navab, N., Onogur, S., Roß, T., Sznitman, R., Taylor, R.H., Tizabi, M.D., Wagner, M., Hager, G.D., Neumuth, T., Padoy, N., Collins, J., Gockel, I., Goedeke, J., Hashimoto, D.A., Joyeux, L., Lam, K., Leff, D.R., Madani, A., Marcus, H. J., Meireles, O., Seitel, A., Teber, D., Ückert, F., Müller-Stich, B.P., Jannin, P., Speidel, S., 2022. Surgical data science - from concepts toward clinical translation. Med. Image Anal. 76, 102306 https://doi.org/10.1016/j.media.2021.102306.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., 2020. BIAS: transparent reporting of biomedical image analysis challenges. Med. Image Anal. 66, 101796 https://doi.org/10.1016/j.media.2020.101796.

Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., Hashizume, M., Katic, D., Kenngott, H., Kranzfelder, M., Malpani, A., März, K., Neumuth, T., Padoy, N., Pugh, C., Schoch, N., Stoyanov, D., Taylor, R., Wagner, M., Hager, G.D., Jannin, P., 2017. Surgical data science for next-generation interventions. Nat. Biomed. Eng. 1, 691–696. https://doi.org/10.1038/s41551-017-0132-7.

Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P.M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Kisilenko, A., Müller, B., Davitashvili, T., Capek, M., Tizabi, M.D., Eisenmann, M., Adler, T.J., Gröhl, J., Schellenberg, M., Seidlitz, S., Lai, T.Y.E., Pekdemir, B., Roethlingshoefer, V., Both, F., Bittel, S., Mengler, M., Mündermann, L., Apitz, M., Kopp-Schneider, A., Speidel, S., Nickel, F., Probst, P., Kenngott, H.G., Müller-Stich, B.P., 2021. Heidelberg colorectal data set for surgical data science in the sensor operating room. Sci. Data 8, 101. https://doi.org/10.1038/s41597-021-00882-2.

Mascagni, P., Alapatt, D., Urade, T., Vardazaryan, A., Mutter, D., Marescaux, J., Costamagna, G., Dallemagne, B., Padoy, N., 2021. A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy. Ann. Surg. Publish Ahead of Print. https://doi.org/10.1097/SLA.0000000000004736.

Meireles, O.R., Rosman, G., Altieri, M.S., Carin, L., Hager, G., Madani, A., Padoy, N., Pugh, C.M., Sylla, P., Ward, T.M., Hashimoto, D.A., The SAGES Video Annotation for AI Working Groups, 2021. SAGES consensus recommendations on an annotation framework for surgical video. Surg. Endosc. https://doi.org/10.1007/s00464-021-08578-9.

Neumuth, T., Jannin, P., Strauss, G., Meixensberger, J., Burgert, O., 2009. Validation of knowledge acquisition for surgical process models. J. Am. Med. Inform. Assoc. 16, 72–80. https://doi.org/10.1197/jamia.M2748.

Nwoye, C.I., Mutter, D., Marescaux, J., Padoy, N., 2019. Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. Int. J. Comput. Assist. Radiol. Surg. 14, 1059–1067. https://doi.org/10.1007/s11548-019-01958-6.

Qiu, Z., Yao, T., Mei, T., 2017. Learning spatio-temporal representation with pseudo-3D residual networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). Presented at the 2017 IEEE International Conference on Computer Vision (ICCV). Venice. IEEE, pp. 5534–5542. https://doi.org/10.1109/ICCV.2017.590.

Roß, T., Bruno, P., Reinke, A., Wiesenfarth, M., Koeppel, L., Full, P.M., Pekdemir, B., Godau, P., Trofimova, D., Isensee, F., Moccia, S., Calimeri, F., Müller-Stich, B.P., Kopp-Schneider, A., Maier-Hein, L., 2021. How can we learn (more) from challenges? A statistical approach to driving future algorithm development. ArXiv 210609302 Cs.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., 2016. Inception-v4, inception-ResNet and the impact of residual connections on learning. ArXiv160207261 Cs.

Tanzi, L., Piazzolla, P., Vezzetti, E., 2020. Intraoperative surgery room management: a deep learning perspective. Int. J. Med. Robot. https://doi.org/10.1002/rcs.2136.
Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25, 44–56. https://doi.org/10.1038/s41591-018-0300-7.
Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N., 2017. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging 36, 86–97. https://doi.org/10.1109/TMI.2016.2593957.
Taylor, Z., Balocco, S., Sznitman, R., Martel, A., Maier-Hein, L., Duong, L., Zahnd, G., Demirci, S., Albarqouni, S., Lee, S.-L., Moriconi, S., Cheplygina, V., Mateus, D., Trucco, E. Vardazaryan, A., Mutter, D., Marescaux, J., Padoy, N., 2018. Weakly-supervised learning for tool localization in laparoscopic videos. In: Stoyanov, D., Granger, E., Jannin, P. (Eds.), Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 169–179. https://doi.org/10.1007/978-3-030-01364-6_19.
Vassiliou, M.C., Feldman, L.S., Andrew, C.G., Bergman, S., Leffondré, K., Stanbridge, D., Fried, G.M., 2005. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am. J. Surg. 190, 107–113. https://doi.org/10.1016/j.amjsurg.2005.04.004.
Vedula, S.S., Ishii, M., Hager, G.D., 2017. Objective assessment of surgical technical skill and competency in the operating room. Annu. Rev. Biomed. Eng. 19, 301–325. https://doi.org/10.1146/annurev-bioeng-071516-044435.
Vercauteren, T., Unberath, M., Padoy, N., Navab, N., 2020. CAI4CAI: the rise of contextual artificial intelligence in computer-assisted interventions. Proc. IEEE 108, 198–214. 10.1109/JPROC.2019.2946993.
Wagner, M., Bihlmaier, A., Kenngott, H.G., Mietkowski, P., Scheikl, P.M., Bodenstedt, S., Schiepe-Tiska, A., Vetter, J., Nickel, F., Speidel, S., Wörn, H., Mathis-Ullrich, F., Müller-Stich, B.P., 2021. A learning robot for cognitive camera control in minimally invasive surgery. Surg. Endosc. 35, 5365–5374. https://doi.org/10.1007/s00464-021-08509-8.
Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L., 2016. Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 20–36. https://doi.org/10.1007/978-3-319-46484-8_2.
Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci. Rep. 11, 2369. https://doi.org/10.1038/s41598-021-82017-6.
Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI. IEEE, pp. 5987–5995. https://doi.org/10.1109/CVPR.2017.634.

Martin Wagner is a general surgeon at Heidelberg University Hospital. He is head of the research group on artificial intelligence and cognitive robotics within the division of minimally invasive and robot-assisted surgery in the department of surgery. His group brings together surgeons, medical students, computer scientists, roboticists, and designers to create novel solutions for clinical problems in surgery. Together with cooperation partners from academia and industry they seek to leverage the potential of machine learning in helping surgeons to choose the best treatment option for the individual patient and perform the surgical treatment in the best possible way.

Patrick Heger is a general surgeon at Heidelberg University Hospital. He is a scientist and trial physician at the clinical study center (KSC) and takes part in the conduction of clinical trials at the Department of Surgery. Furthermore, he is the principal investigator of several trials there. Additionally, he is a member of the Study Center of the German Society of Surgery (SDGC).

Lars Mündermann is an employee at KARL STORZ. He is head of the data assisted solutions group within corporate research & technology. His group works on the evaluation and assessment of new technologies for computer-assisted surgery.

David M.Lubotsky studied at the LMU in Munich and received his BSc degree in physics. Currently he is a medical student at Heidelberg University and research assistant within the department of general, visceral and transplantation surgery of Heidelberg University Hospital.

Annika Reinke studied applied mathematics at the University of Lübeck, Germany, with a focus on medical image analysis. In 2017 she joined the division of Computer Assisted Medical Interventions at the German Cancer Research Center (DKFZ) to work on scientific benchmarking and validation of AI algorithms, since 2020 as her PhD topic. Having published disruptive findings on biomedical image analysis challenges in Nature Communications, she is a founding member of the initiative of Biomedical Image Analysis ChallengeS (BIAS). She further serves as an active member in several working groups with focus on open science, which led to several high-ranked publications.

Xinyang Liu is a Staff Scientist at Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital of Washington DC. He received his PhD degree from Florida State University in 2010. He worked as a Postdoctoral Fellow at Johns Hopkins Hospital and Brigham and Women's Hospital. His research interests include medical augmented reality, computer assisted surgery and medical image analysis.

Hans Meine received a PhD from the University of Hamburg in 2008 for his fundamental research on image segmentation methods at the department for computer science. Since 2011, he is applying image analysis to various medical image modalities at Fraunhofer MEVIS, where he is nowadays coordinating image analysis and deep learning developments across projects. From 2015–2021, he had a second position at the University of Bremen for research and teaching.

Satoshi Kondo is an associate professor at Muroran Institute of Technology, Japan. He received his B.S., M.S. and Ph.D degrees at Osaka Prefecture University in 1990, 1992 and 2005, respectively. He was with Panasonic Corporation and KonicaMinolta Inc. His research interests are in the fields of computer vision.

Long Zhao is a researcher at Hikvision Research Institute. His research interests focus on unsupervised learning, weakly supervised learning and domain adaptation.

Zhenxiao Ge is an algorithm engineer at Bytedance. His research interests focus on unsupervised learning and anti-risking. This work was done when he was a research intern in Hikvision Research Institute.

Haiming Sun is a researcher at Hhikvision Research Institute. His research interests focus on developing machine-learning methods for anomaly detection.

Di Xie currently is a principal research manager at Hikvision Research Institute. His research interests include computer vision, video understanding and deep neural network optimization.

Mengqi Guo is a Ph.D. student at National University of Singapore. His research interests include computer vision and deep learning. This work was done when he was a research intern in Peking University.

Daochang Liu is a Ph.D. student at Peking University currently. His research interests include video understanding and surgical data science.

Franziska Mathis-Ullrich is Assistant Professor for Medical Robotics at the Karlsruhe Institute of Technology (KIT), Germany. Her research focus is on cognition controlled robotics for minimally-invasive surgery. She received her B.Sc. (2009), M.Sc. (2012) and Ph.D. (2017) from ETH Zurich. Since 2019, she is head of the Health Robotics and Automation Laboratory at KIT. Franziska Mathis-Ullrich has received multiple academic awards (IEEE ICRA Best Paper Award, 2014; IEEE BioRob Best Student Paper Award, 2016; winner of ICRA Microassembly Challenge, 2014 & 2015). In 2017 she made it onto the prestigious Forbes "30 under 30″ list.

Lena Maier-Hein is a full professor at Heidelberg University (Germany) and affiliated professor to LKSK institute of St. Michael's Hospital (Toronto, Canada). At the German Cancer Research Center (DKFZ) she is managing director of the "Data Science and Digital Oncology" cross-topic program and head of the division Computer Assisted Medical Interventions (CAMI). Her research concentrates on machine learning-based biomedical image analysis with a specific focus on surgical data science, computational biophotonics and validation of machine learning algorithms.

Stefanie Speidel is a full professor for "Translational Surgical Oncology" at the National Center for Tumor Diseases (NCT) Dresden and one of the speakers of the DFG Cluster of Excellence "center for Tactile Internet with Human-in-the-Loop (CeTI)". Her current research interests focus on machine learning for computer- and robot-assisted surgery in the context of the future digital operating room.

Sebastian Bodenstedt is a postdoctoral researcher at the National Center for Tumor Diseases (NCT) Dresden and member of the center for Tactile Internet with Human-in-the-Loop (CeTI)". His current research interests focus on developing machine-learning and computer-vision methods for the surgical environment.