Multicentric Exploration of Tool Annotation in Robotic Surgery: Presenting a Starting Guideline for Surgical Artificial Intelligence Projects

Pieter De Backer, MD, MScEng[1,3,4,5], Jennifer A. Eckhoff, MD[2], Jente Simoens, MScEng[1], Dolores T. Müller, MD[2], Charlotte Allaeys[3], Heleen Creemers[3], Amélie Hallemeesch[3], Kenzo Mestdagh[3], Charles Van Praet, MD, PhD[5], Charlotte Debbaut, MScEng, PhD[4], Karel Decaestecker, MD, PhD[5], Christiane J Bruns, MD[2], Ozanan Meireles, MD, FACS[6], Alexandre Mottrie, MD, PhD[1,7], Hans F. Fuchs, MD, FACS[2]

Abstract

Background Artificial Intelligence (AI) holds tremendous potential to reduce surgical risks and improve surgical assessment. Machine Learning, a subfield of AI, relies on video and image data, where annotations provide veracity about the desired target features. Yet, methodological annotation explorations are limited to date. Here, we provide an exploratory analysis of the requirements and methods of instrument annotation in a multi-institutional team from two specialized AI centers and compile a structured manual for future AI projects focusing on instrument detection.
Methods We developed a bottom-up approach for team annotation of robotic instruments in robot-assisted partial nephrectomy (RAPN), after which it was validated in robot-assisted minimally invasive esophagectomy (RAMIE). Furthermore, instrument annotation methods were evaluated for their use in Machine Learning algorithms. Overall, we evaluated the efficiency and transferability of the proposed team approach and quantified performance metrics (e.g. time per frame required for each annotation modality) between RAPN and RAMIE.
Results The proposed annotation methodology was transferrable between both RAPN and RAMIE. The bottom-up approach of annotation management and training resulted in accurate annotations and demonstrated efficiency in annotating large datasets and diverse annotator groups. The average annotation time for RAPN for pixel annotation ranged from 4.49 to 12.6 minutes per image; for vector annotation this was decreased to 2.92 minutes. Similar ranges of pixel annotation times were denoted for RAMIE. Lastly, we elaborate on common pitfalls encountered throughout the annotation process.
Conclusions We propose a successful bottom-up approach for annotator team composition, applicable to any annotation project. Our results set the foundation to start AI projects for instrument detection, segmentation and pose estimation. Due to the immense annotation burden resulting from spatial instrumental annotation, further analysis into sampling frequency and annotation detail needs to be conducted.

Key Words Artificial Intelligence, Computer Vision, Supervised Machine Learning, Annotation, Instrument Segmentation

Corresponding Author: Pieter De Backer
Address: Proefhoevestraat 12, 9090 Melle
Telephone: 00329 334 69 26
Email: pieter.de.backer@orsi.be
Funding information: Flanders Innovation & Entrepreneurship Agency (reference HBC.2020.2252)
Author Affiliation:
1) ORSI Academy, Belgium
2) Robotic Innovation Laboratory, University Hospital Cologne, Department of General, Visceral, Tumor and Transplantsurgery, Germany
3) Department of Human Structure and Repair, Faculty of Medicine and Health Sciences, Ghent University, Belgium
4) IBiTech-Biommeda, Faculty of Engineering and Architecture, and CRIG, Ghent University, Belgium
5) Ghent University Hospital, Urology, Belgium
6) Massachusetts General Hospital, Surgical Artificial Intelligence and Innovation Laboratory, Boston, USA
7) OLV Hospital Aalst-Asse-Ninove, Urology, Belgium

Introduction

In recent years, the application of Artificial Intelligence (AI) techniques in the automated analysis of surgical video data has surged. More specifically, subfields of AI such as Computer Vision (CV) and Machine Learning (ML) are increasingly being applied to surgical videos. A systematic literature search for the key words "Computer Vision" and "Surgery" in Pubmed as well as Arxiv, a more commonly used platform in computer science, illustrates a continuous uprise in scientific publications in the field with a 50 fold increase between 2010 and 2021 (Fig. 1). Automated analysis of surgical workflow promises intraoperative guidance, objective risk assessment and overall increase of surgical safety. One major building block for surgical video and image analysis through CV and ML is the detection and pose estimation of surgical instruments(1). Knowing and foreseeing the position, orientation and action of an instrument provides a ML model with insights in the surgeon's intent, allowing for more thorough comprehension of surgical workflow and potentially prevention of harmful tool-tissue interaction.

The spatial and temporal patterns, provided by moving instruments, are detected by ML models. State of the art ML models rely largely on supervised learning(2), meaning that the models learn to interpret video and imaging data on the basis of previously internalized, labeled data containing annotations of target features, such as instruments, organs, surgical phases or actions. Recent ML algorithms have even demonstrated good accuracy with respect to identification of operative steps and surgical tools in various procedures(3). Besides providing visual information about current surgical actions, instrument segmentation also lends itself to CV analysis of surgical video data by giving clear boundaries and contrast compared to surrounding anatomical structures and tissue, which should also be interpreted. As such, labeling or annotating instruments is generally considered a good starting point for any new AI projects focused on surgical analysis and can serve as the basis for temporal as well as spatial analysis of surgical procedures.

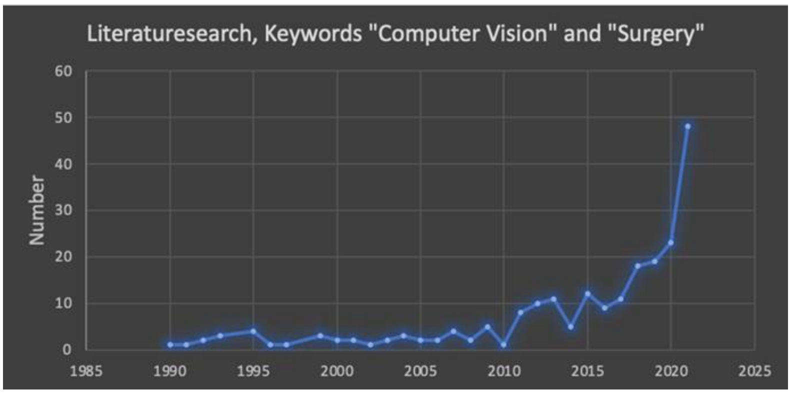

Figure 1: Literature search in Pubmed for the key words "Computer Vision" and Surgery"

In spatial analysis of surgical procedures, two major types of annotation methodologies need to be differentiated. Firstly, during pixel segmentation each pixel of the image is assigned to a certain class, or in this case instrument, by precise delineation of areas of interest(4). Segmentation is one of the most popular image processing tasks(5), as it provides the most precise delineation of the surgical scene. However, it does not necessarily provide information on the position of the instrument and the ongoing action. In contrast, the second type of spatial annotation, called vector annotation, specifies the orientation of the main part of a target feature in an image. This yields information such as an instrument pose and can be performed through coarse spatial delineations denoting key points as a simplified version of a full instrument.

In temporal analysis of surgical procedures, time points or tags for certain predefined events and phases are annotated. Previous explorations of annotation requirements and proposals for annotation guidelines are quite coarse, lacking specific step-by-step information for AI laboratories or researchers new to the field. Clear guidelines proposing homogenous approaches to instrument annotation are missing. Previous work on the generation of standardized guidelines was initiated by the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) AI Taskforce(2). In a Delphi Consensus, general questions around surgical video and imaging annotation, such as general characteristics of an annotation framework were addressed by an expert panel. While this presents a fundamental first step towards unified methodology, questions remain about what to annotate specifically, who the annotators should be, how to efficiently construct an annotation team and how to commence an AI project. Hence, this work sets out to address some of these questions in a "how to" manner. It demonstrates a feasible approach for instrument annotation along the SAGES consensus recommendations, aiming to assist surgical research groups novel in the field of AI. With regard to spatial annotations, the SAGES consensus recommends a hierarchical approach to increasing detail. For tool annotation this means increasing granularity ranging from the type of instrument, function, manufacturer and potentially functional subunits of the instrument (i.e. frontal edge of cautery hook rather than just hook). Therefore, we analyze the feasibility of instrument annotation in great detail and describe encountered pitfalls in order to contribute to potential future refinements of this framework.

Due to the previously observed low inter-annotator variability and good concordance between annotations of laymen and medical experts(6), instrument annotation is highly suitable for team annotations and multi-institutional collaborations. This facilitates the generation of large annotated datasets, required for supervised ML, and eases the annotation burden for individuals. Nevertheless, currently available annotated laparoscopic and robotic surgical video datasets focused on tool segmentation remain rather small and are limited to routine procedures, such as laparoscopic cholecystectomy(7), gastric bypass, sleeve gastrectomy and colectomy. This limits the diversity of annotated tools to the few examples present in these standard procedures and restricts transferability of annotations to other more complex procedures containing a wider range of instruments. Additionally, instrument segmentation has been concentrating mainly on laparoscopic procedures, while robotic procedures are finding their way into surgical practice with equal or even superior outcomes. As such, robotic surgery is clearly gaining significance and prevalence regarding desired AI applications. However, up to date existing robotic datasets tend to be limited to ex-vivo experiments. We suspect that the limited diversity of annotated datasets is not only due to shortage of qualified, medical annotators but that it is also related to missing guidelines governing annotation methods(8). Furthermore, the limited clinical information and datasets openly accessible to most engineers regarding the target procedures results in repetitive utilization of existing datasets rather than expanding the field towards novel approaches. This results in limited generalizability and transferability of the current CV techniques.

In this work, we present a comprehensive framework for robotic tool annotation, that may serve as groundwork for the start of any robotic surgical AI project on surgical workflow analysis evolving around instrument detection. This includes a detailed outline of a possible annotator team to decrease annotation burden for experienced surgeons. We demonstrate lessons learned from annotating >10.000 images manually by novel as well as expert annotators in two specialized AI centers and propose solutions to overcome said obstacles. To show generalizability, we apply the annotation methods developed during annotation of instruments in robot-assisted partial nephrectomy (RAPN) to robot-assisted minimally invasive esophagectomy (RAMIE). This may serve as a detailed guide for instrument annotation for any researcher interested in CV based automated tool detection.

Materials and Methods

Data collection

We analyzed 82 videos of robotic-assisted partial nephrectomy videos (RAPN) from two Belgian tertiary referral centers (OLV Hospital Aalst and Ghent University Hospitals) collected between 05/2018 and 10/2021. Videos were collected from the Intuitive Si, X and Xi robots (Intuitive Surgical™, California, USA) using Hauppauge PVR Rocket recorders (Hauppauge Computer-Works GmbH - Mönchengladbach, Germany) and a MAQUET Tegris system (Getinge Ag – Göteborg, Sweden). Videos were captured in 3 different resolutions: 720x576 pixels (p) resolution at 25.00 frames per second (fps) (VOB format), HD ready resolution (1280x720p) at 58.94 fps and Full HD resolution (1920x1080p) at 30.00fps (both mp4 format).

In a multicentric collaborative effort, with the goal of exploring transferability of annotation methodology and add to the diversity of the dataset, additionally, 94 videos of robotic-assisted minimally invasive esophagectomy were collected in a German specialized upper GI Cancer center (University Hospital of Cologne, Department for General, Visceral, Tumor and Transplant Surgery) between 01/2020 and 07/2021. Videos were collected from the Intuitive Xi robots (Intuitive Surgical™, California, USA) using Medicapture HD USB300 Recorders (MediCapture Inc.- Pennsylvania USA) in HD resolution (1280 x 1024p) at 29.97 fps (mp4 format) Out of the 94 obtained RAMIE videos, only 15 were selected for annotation. Exclusion criteria were video quality issues (blurred images, incomplete video data), deviations from the standardized operative protocol resulting in use of different instruments and adverse events (minor bleeding, rupture of suture). For this transfer feasibility study, the annotation focused on frames extracted from the anastomotic reconstruction phase, as defined by Fuchs et al(9), where the starting point of the operative phase is defined as when the hook touches the esophagus first and the endpoint when the needle is removed after suturing in of the circular stapler. This was decided due to the significance of the anastomotic phase in postoperative course and presence of various additional instruments throughout this reconstruction phase.

Data Preprocessing

Preprocessing of the videos included patient deidentification and removal of images containing patient clues inside the videos, as well as joining of multiple video segments to one consecutive video. Finally, for tool annotation purposes, video frames, meaning still images, were extracted at a framerate of 0.5 Hz for 5 videos, 0.1 Hz for 1 video and 0.05 Hz for 76 videos, meaning one still image was being captured every 2, 10 and 20 seconds of video respectively. The images were saved as uncompressed png-files resulting in a total count of 27597 RAPN images. Of these, 23598 images were completely annotated, resulting in 56 completely annotated procedures (16622 images). This slicing and video reformatting, as well as the frame sampling, was conducted using freely available open source software ffmpeg (FFmpeg Developers (2016) - available from http://ffmpeg.org/). These images were subsequently uploaded to a professional annotation platform.

Annotation Methods

Spatial annotation of robotic instruments in RAPN and RAMIE were performed in the online annotation platform SuperAnnotate (Sunnyvale, CA, USA) by both ORSI Academy, Belgium and the Robotic Innovation Laboratory, University Hospital Cologne, Germany. Apart from SuperAnnotate, we also tested the freeware software Computer Vision Annotation Tool (CVAT) (https://github.com/openvinotoolkit/cvat).

Annotation methods included pixel segmentation and vector annotation to label positions and orientations of the instruments. Figure 2 gives an overview of the explored necessary granularity of instrument segmentation through pixel annotation. An example of vector annotations is given in Figure 3.

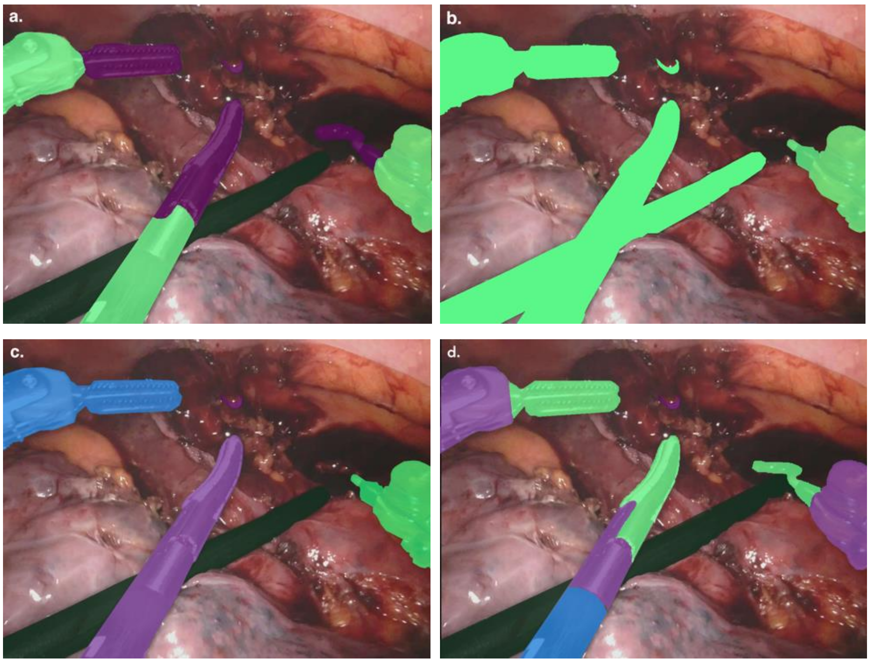

Figure 2: Pixel Annotation in RAMIE at different levels of granularity : (a) segmentation of tip and shaft (b) binary segmentation (instruments vs background) (c) instance segmentation of full instruments (d) segmentation of all articulations

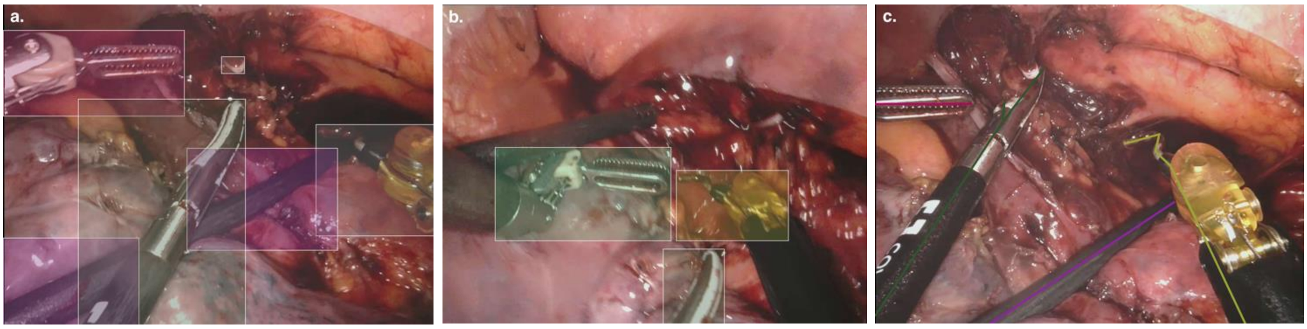

Figure 3: Use of Vector Annotation in RAMIE (a) Bounding Boxes delineating the whole instrument (b) Bounding boxes delineate the tip of the instruments (c) Vector Annotations provide pose estimations

- Pixel Segmentation
  During pixel segmentation the edges of each instrument were manually delineated and later assigned to corresponding classes. All pixel segmentations were performed using polygon tools. Tools facilitating segmentation like watershed algorithms(10) were explored and abandoned, as they were found to introduce impreciseness into the dataset due to less crisp side edges or imperfections at the instrument tips.

- Vector Annotation
  Vector Annotation was applied to the same surgical frames used for pixel annotation and entails both rectangular bounding boxes and wireframe pose estimation.

Bounding boxes were applied to frame the outer edges of each instrument, not accounting for the exact borders and subparts of the instruments. The bounding boxes could either include the entire visible part of the instrument or just the functional tip.

Next, wireframes or vectors were created to indicate the position of the individual sub-elements of the robotic instruments and clearly differentiate the transition points between these sub-elements. This was considered particularly important for pose estimation of robotic instruments to account for the various angles and degrees of freedom. Vectors were placed from the start to endpoint of each sub compartment of an instrument, the transitions between the vectors illustrated an angle in the instrument or a different functional unit.

Down Sampling of Frames for Annotation Purposes

1 minute of a 60 fps video contains 3600 images. As such it is impossible to annotate every single image and the annotation load should be balanced against the loss of detail. Taking into consideration that the ideal sampling frequency required for training of ML models is yet to be determined, we provide a quantitative analysis for instrument annotation by down sampling of video frames. We resampled RAPN frames, sampled at periods of 2 seconds to sampling periods of 4, 6, 8, 10, 20, 30 and 40 seconds. Per time interval, we calculate the intersection over union (IoU) for all RAPN procedures as a measure to quantify at which frequency the instruments are sufficiently differently positioned from the previous one. Figure 4 shows a correlation between the extracted framerate and the position change of the instruments, it also shows the calculation of IoU(11).

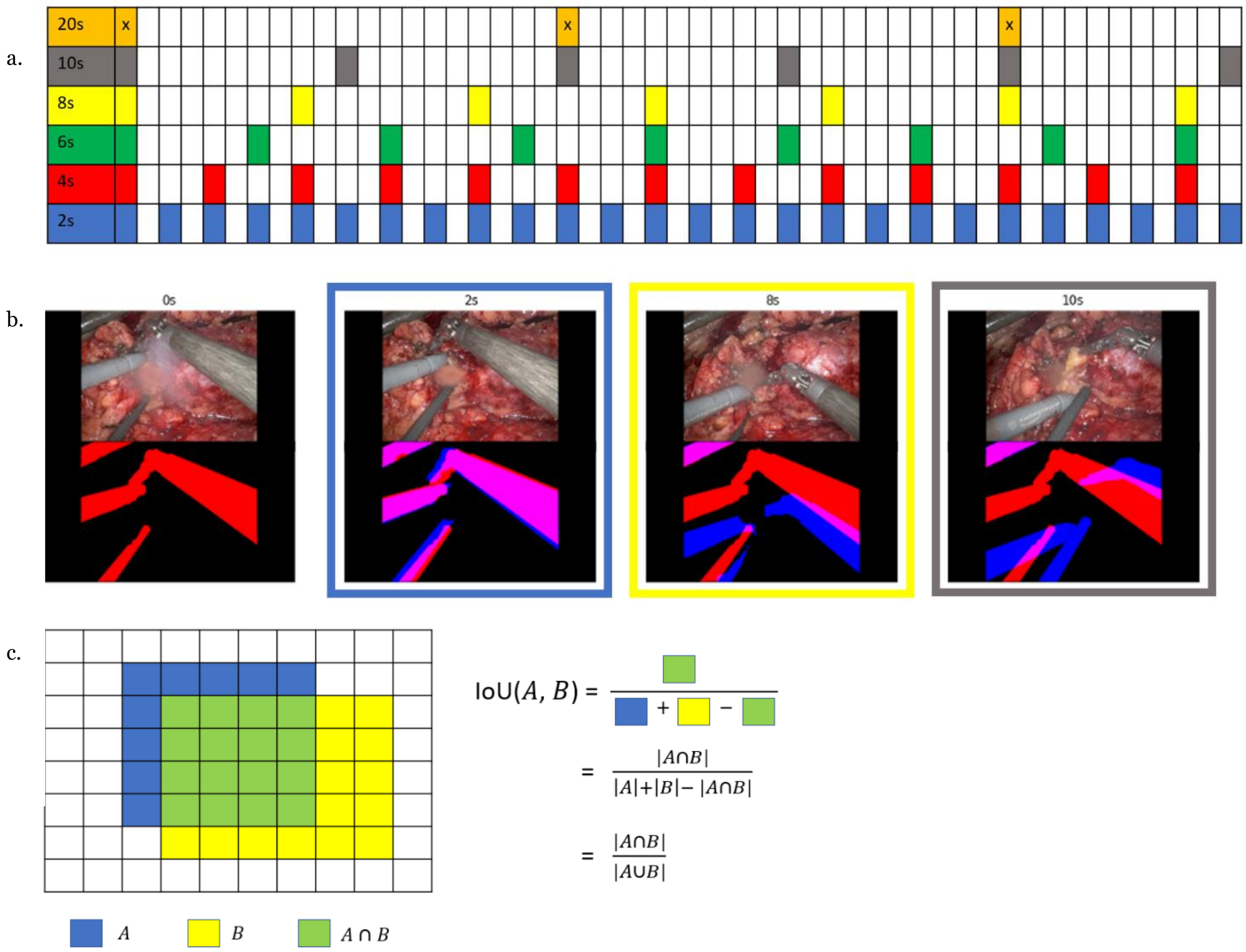
a.
20s
x
x
x
10s
8s
6s
4s
2s
b.
0s
2s
8s
10s
c.
$\mathrm{IoU}(A,B) = \frac{🟩}{🟦 + 🟨 - 🟩}$
$= \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$
$= \frac{|A \cap B|}{|A \cup B|}$
A
B
A ∩ B

Figure 4: Correlation between (a) Sampling Rate and (b) position change of robotic instruments as compared to the previous positions; (c) Intersection over Union calculation

Annotation Team Composition and RAPN annotation

Due to the previously described low inter annotator variability in tool annotation and the large workload of two robotic datasets, we composed a hierarchical team structure (Fig. 5). The team was led by one annotation-experienced supervisor, in our case a surgical resident, who served as the supervisor for a "Quality-Assessment (QA) Team" as well as an "Annotator (A) Team". The supervisor took responsibility for data collection and data handling as well as close communication with an engineering team, concerning the requirements on annotations for ML purposes. Besides that, the supervisor's roles were (1) conducting annotation training among the QA Team, (2) coordination of the annotation process including guidance of the individual annotators in the QA Team, and (3) ensuring the legal and ethical integrity of the data handling by coordinating with the involved ethical and regulatory committees.

Additionally, parts of the annotation workload was performed by professional annotation teams with no medical expertise (denoted as 'laypeople' in Figure 5). This served as measure to decrease annotation burden and helped to explore and assess the feasibility of our proposed hierarchical annotation team composition.

The QA Team consisted of 4 medical students with equivalent levels of annotation expertise who previously worked in different annotation packages. The QA Team's role is to (1) pass on the annotation training and annotation guidance to the A-Team and (2) perform a quality check on every frame annotated by the A-Team. The A-Team in turn consists of a group of medical students (n=33), responsible for (1) performing spatial annotations of instruments and (2) overseeing and checking laymen workforce's annotations. The laymen workforce is part of professional organization, and thus were already proficient in the software packages at hand, but they had no insight in the labeling methods or requirements. Troubleshooting, questions or doubts concerning the annotation process and technical obstacles were handled in a layered "bottom up fashion" were the A-Team consults the QA team, which then consults the supervisor in turn. 10489 out of 16622 images were pre-annotated by the laymen workforce to investigate the impact of annotators with no medical experience on performing instrument annotation with regard to accuracy and time saving.

Both the laymen and A-Team were provided with a list of instruments required to annotate, and performed a test set annotation round on which they received feedback. The test set was composed by the supervisor with respect to adequate image variability and possible bottlenecks or questions that arise during annotation. Both groups were asked to do every annotation fully manually, (i.e. without the help of intelligent algorithms, e.g. watershed algorithms etc). All images were quality checked in SuperAnnotate (Sunnyvale, CA, USA). 10489 images were pre-annotated by laymen, 22% of these images were pre-annotated in CVAT (https://github.com/openvinotoolkit/cvat), while the remaining 78% was pre-annotated in SuperAnnotate.

For pre-annotation, 2 schemes were tested. In a first scheme, the laymen workforce only annotated instruments as binary segmentations, without indicating instrument names (see Figure 2b). In a second experiment, laymen also added the instrument name, and the A-Team provided quality control on the annotated instruments, as well as added additional small items like needles, wires, gauzes, etc. Subsequently all annotations of both schemes were quality checked by the QA Team.

Annotation Training

For the A-Team, annotation training consisted of having every annotator in the A-Team annotate a predefined test set of sample frames to overcome the suspected learning curve during annotation of robotic instruments. The A-Team was immediately trained to discern the different instruments, clips, wires, hemostatic agents, etc. after a hands-on instrument training session combined with an introduction to the annotation software package. Laymen workforce also received a test set, with a focus on robotic and laparoscopic instruments. They were already proficient in the labeling software at hand and only received further feedback or instructions if the test set they delivered back was insufficiently precise.

Annotation Process

Every annotator within the A-Team was required to perform between 50 and 72h of accumulated annotation activity in the annotation platform over the course of 4 weeks, divided over segmentation and vector annotation. There was no predefined split between the different annotation types per annotator. The A-Team was given new images dependent on the general progress of the annotation process. Every annotator in the A-Team received a fixed amount of images which are submitted to the QA team when the annotator believes them to be finished. Incorrect images are sent back to the A-Team by the QA team, accompanied by feedback so that they can be adjusted properly. The A-Team annotators can view their performed working hours inside the software package, which allows for a sense of self-evaluation and benchmarking. The focus of the annotation process was primarily on pixel annotation, attaining a total of 11431 images. In this identical dataset, 3896 vector annotations were finished by the A-Team. After finishing the annotation work packages, all A-Team annotators filled in an online questionnaire about their annotation impressions.

Annotation of RAMIE

The above described bottom up approach for annotation of robotic instruments was then transferred to a separate annotator group in the Robotic Innovation Laboratory for annotation of RAMIE. Annotation training was performed in the same manner as described above. The primary goals of this project extension was the assessment of transferability of the annotation methodology established throughout this instrument annotation project, the validation of the established annotator team structure and the quantification of the times needed by novel annotators to perform pixel and vector annotation of robotic instruments in RAMIE after implementation of a pre-established annotation

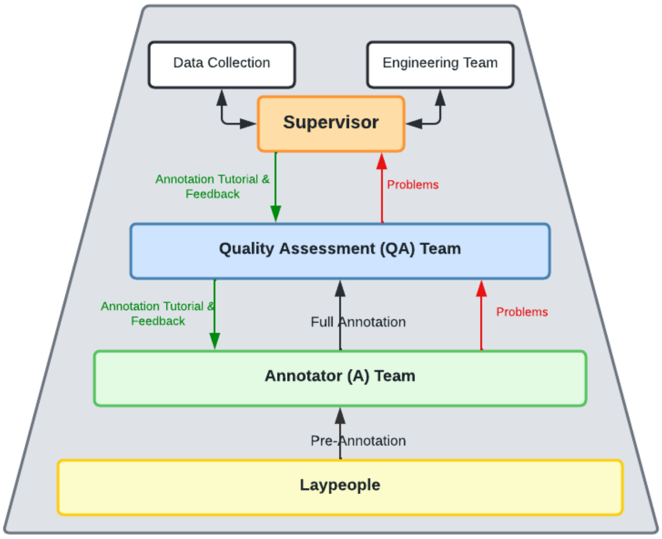

Figure 5: Overview of the proposed hierarchical team composition for scaling data annotation.

guide. Within the RAMIE video dataset, 15 videos of the anastomotic phase were selected as suitable for annotation. With a mean duration of 27.39 min of the anastomotic phase, an average of 82 frames per video were extracted at a sampling period of 20 seconds. This would result in a total of 1232 frames for all 15 videos. Pixel as well as vector annotations of 183 images were performed within the SuperAnnotate software. The Annotator Team consisted of 3 medical students, the QA Team of 2 surgical residents with little annotation experience and the supervisor was an attending for upper GI surgery.

Results

In the following section we describe the observations made throughout the spatial annotation of robotic instruments. We illustrate our findings with regard to the ideal sampling frequency for instrument pose estimation. We evaluate the annotator team composition for our hierarchical bottom up approach. Furthermore, we display the average time spent on the spatial annotation of robotic instruments in each individual frame of RAPN and RAMIE and elaborate on general pitfalls.

Dataset Instrument Distribution

We accumulated a total of different instruments throughout the whole RAPN dataset. We discern between robotic and laparoscopic instruments and other non-organic objects such as bulldog clamps or hemolock clips. A list of all different instruments and objects for pixel and vector annotation can be found in the appendix. Figure 6a gives an overview of the top 5 instruments present in the RAPN dataset.

Sampling Frequency

To determine sufficient change in instrument positioning for two consecutive frames, the intersection over union (IoU) was calculated. Figure 6b-d shows the correlation between IoU and sampling period. Expectedly, the IoU decreases as the sampling period increases. As the procedure progresses and the camera angle changes, the instruments indeed have higher chances of being positioned at completely different angles and locations. This results in a higher variety of diverse frames, leading to more varied training data available for ML models. However, sampling periods should not be too large as to not lose too much surgical detail. Figure 6b shows the pattern of decreasing overlap (IoU) when all instruments are considered together as in a binary segmentation (Figure 2b). When investigating single instruments (monopolar curved scissors and large needle driver) on a frame per frame basis, we denote similar trends. We see a decrease down to 8 seconds, after which the IoU slowly start to plateau. We choose a sampling period of 20 seconds (0.05 Hz) as for some instruments like the large needle driver (Figure 6d), the IoU still experiences quite a drop. After periods of 20 seconds, all curves tend to flatten, meaning the instruments are equally different positioned.

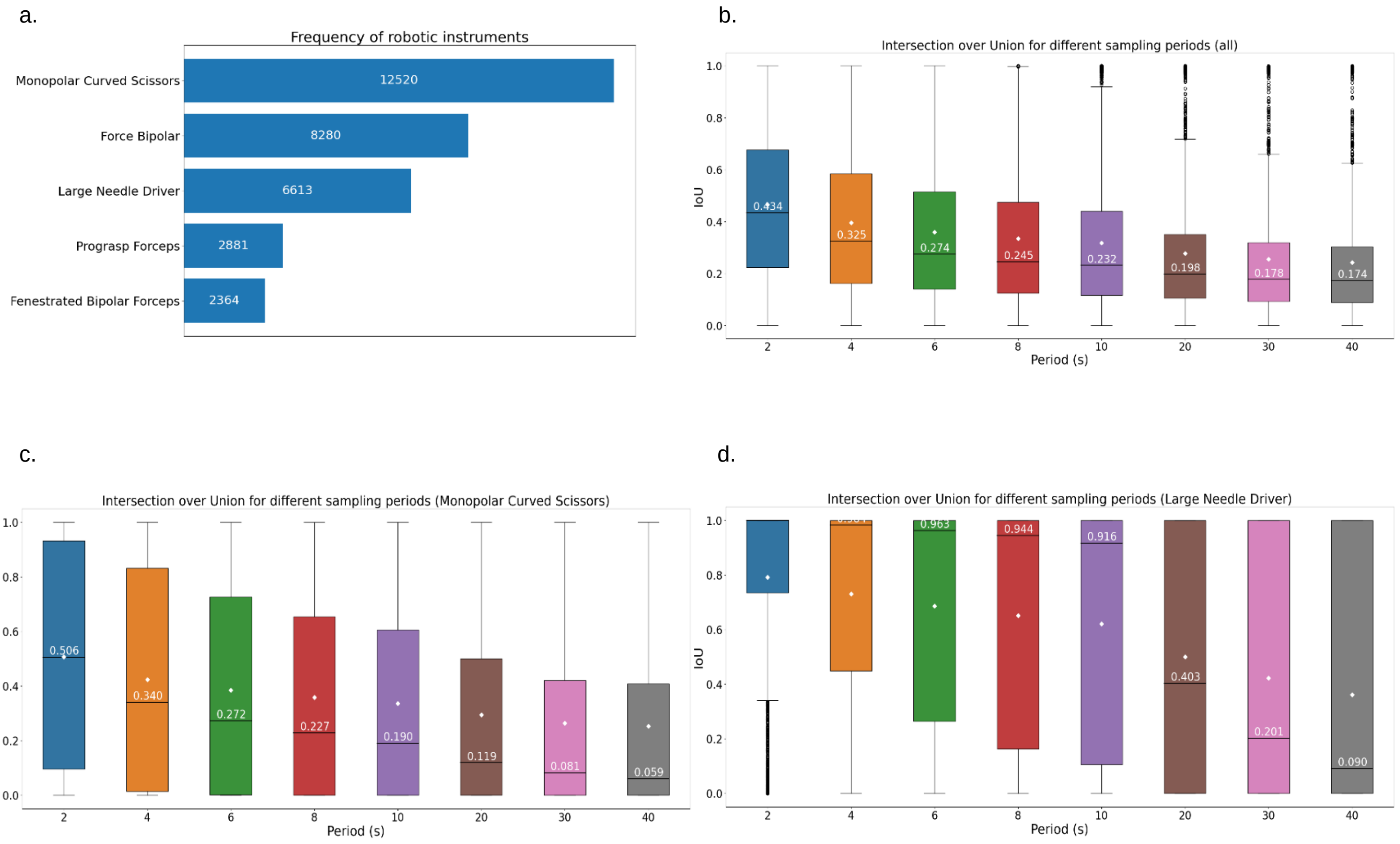

Figure 6: (a.) Distribution of top 5 annotated robotic instruments. (b.) Boxplot diagram of the IoU of all instruments compared between frames for different sampling frequencies, as in binary segmentation - Figure 2.b. (c.) Boxplot diagram of the IoU of monopolar curved scissors, the most prevalent class, for different sampling frequencies. (d.) Boxplot diagram of the IoU of large needle driver, the third most prevalent class, for different sampling frequencies. The white dots inside the boxplots denote the mean IoU for that period.

Annotator Team and Annotation Time

The deployed hierarchical, bottom-up approach for annotator team composition proved to be feasible and effective. Additionally, the pyramid based approach proved to be a cost-effective way to scale-up annotation efforts. Figure 7 provides a flowchart of our different experiments for all completely annotated procedures.

Out of the 16622 annotated frames, 6025 images were annotated by medical students in the A-Team without any prior annotation, with a subsequent quality check, resulting in the overall longest annotation time of 12.6 minutes per image. 108 images without prior annotation were directly annotated by the QA Team, resulting in the overall second fastest annotation time of 5.17 minutes per complete image.

The fastest throughput time per image was achieved for the images which underwent pre-annotation by laymen and which were subsequently adjusted and checked by the QA Team, resulting in a total annotation time of 4.49 minutes per image.

The 10489 images which where pre-annotated by a professional workforce, showed that professional teams tend to work faster, however, a thorough quality check which corrects certain classes or adds certain details by the A or QA Team is required. We do note that, bringing in a professional workforce of laymen with no medical background, approximately halves the required time per annotation compared to annotation from scratch by students.

When looking at different types of pre-annotation by these laymen (not depicted in Figure 7), the setup in which pre-annotation consisted of binary segmentation as in Figure 2.b., subsequently took the QA Team 2.23 minutes per image on average to complete. In the second setup, pre-annotation was performed as in Figure 2.c, in which annotators now also assign the instrument class. On average, the QA Team took 1.66 minute per image on average to complete.

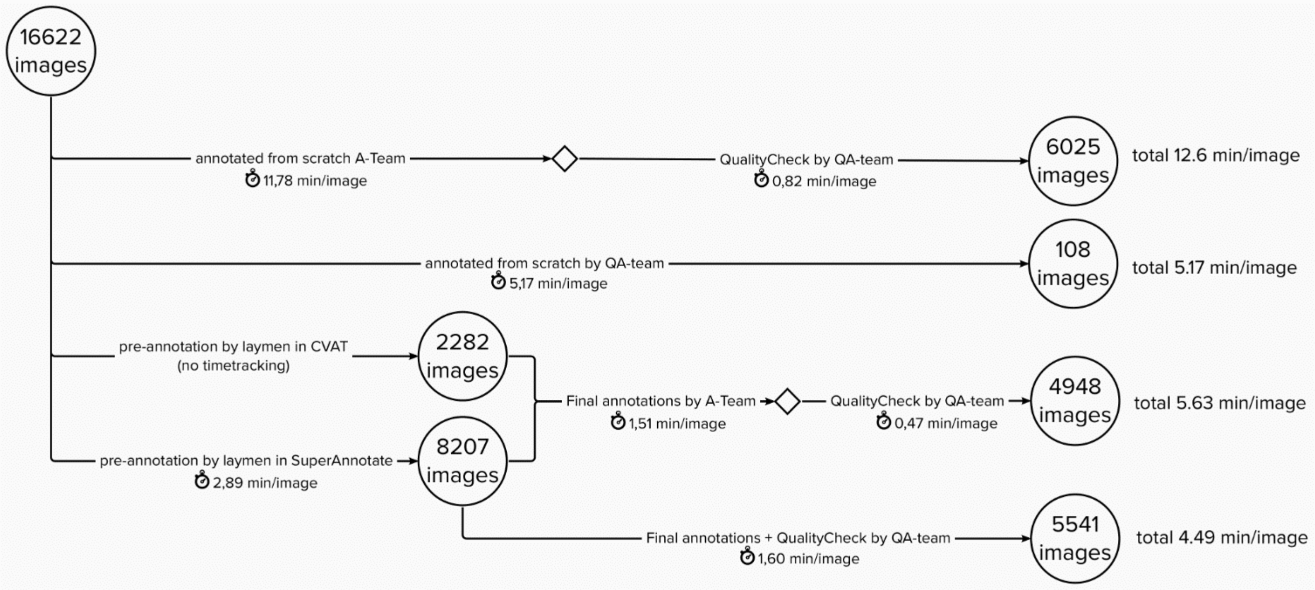

Figure 7: Flowchart of different annotation times for annotation pipelines.

Learning curve for annotation

Within the A-Team, 29 medical students completed their assigned working package, 4 students dropped out. All participating 33 medical students filled in a questionnaire afterwards.

21% (7 students) felt they needed 0-4 hours to master the instrument segmentation, 52% (17 students) estimated 4-8h would be necessary to master the instrument segmentation, and 27% (9 students) needed 8-12h of annotation before achieving a sense of mastery. No participants indicated needing more than 12 hours of annotation to achieve a sense of mastery.

A large variation was observed between annotation times of different A-Team members. Plotting average annotation times versus number of completed images indeed revealed a trend of decreasing annotation times as more images were annotated. Figure 8 shows average pixel segmentation times plotted against the number of completed images. We note a clear decrease with increasing numbers of completed images, which tends to plateau around 500 images. Even at the fastest annotation rate of 5.17 minutes per finished instrument segmentation, the annotation of 500 images corresponds to approximately 43 hours. As such, there is a clear discrepancy between the objective assessment and the subjective feeling a novel annotator.

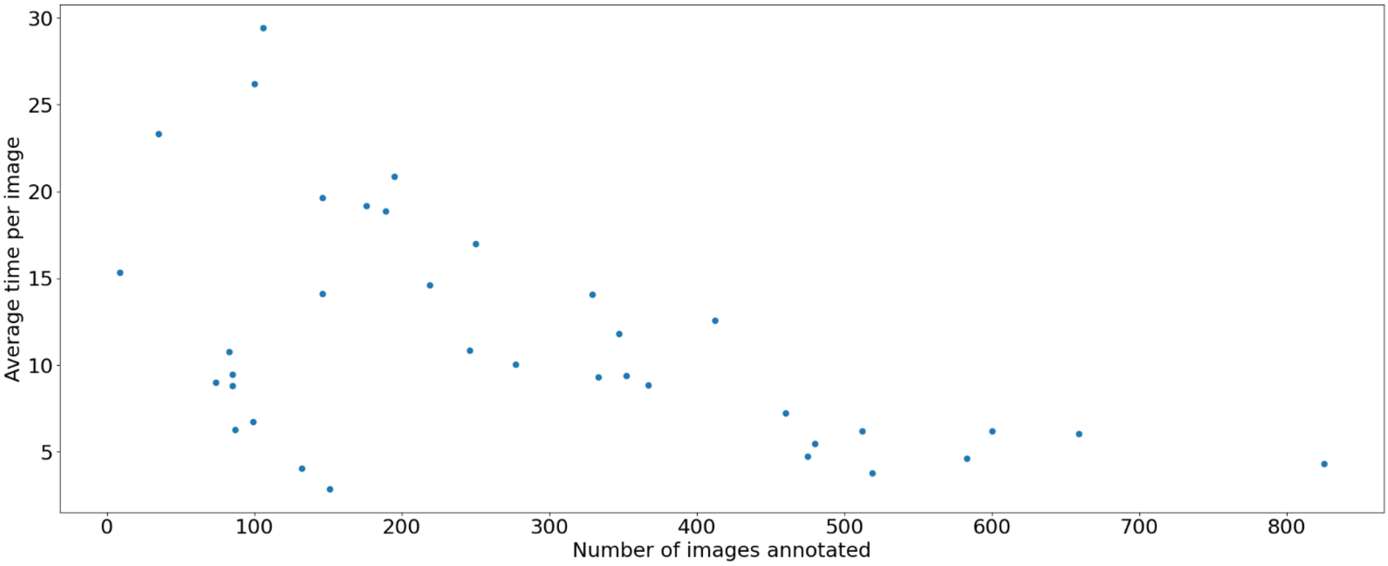

Figure 8: Annotation times per images versus number of annotated images. All images were annotated from scratch.

Pixel Annotation

Pixel Annotation was evaluated for different levels of granularity. Figure 9 shows the different levels of detail deployed during semantic segmentation through pixel annotation. For figure a, b and c, each color represents a different part of the instrument. This means that figure 9a has a total of 9 classes (3 classes for each instrument) and figure 9b consists of 6 classes (2 classes per instrument). Figure 9c shows a binary segmentation, separating instruments from the background, and thus consists of only one class: “instruments”. Figure 9d shows the most commonly used approach, where each instrument is assigned an individual class, which can be considered the most time efficient. We note that every form of pixel annotation can be derived from figure 9a by simply merging the different classes per instrument. While this proved to be the most precise form of pixel annotation, the time investment is higher. Hence, we chose to perform the multiclass segmentation as depicted in Figure 9d for the entire dataset.

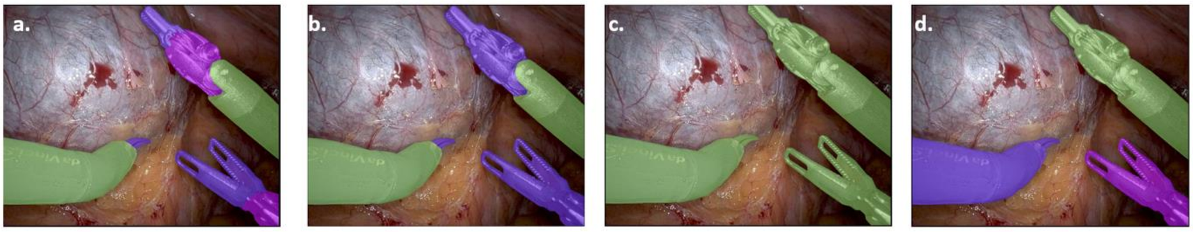

Figure 9: Different levels of surgical instrument segmentation (a) segmentation of the articulations (b) segmentation of tip and shaft (c) binary segmentation (instruments vs background)(d) multiple instances segmentation

On a pixel level, we chose a very precise delineation of all visible objects. The pixel annotation of graspers can be seen in figure 10a. Figure 10b. shows the spearing out of fatty tissue at the tip of the bipolar grasper and Figure 10c shows the clear differentiation of tissue within the endobag in contrast to the surrounding bag. An important observation made throughout the pixel annotation was to focus annotation on what is seen, rather than annotation based on interpretation of previous frames. For example, graspers may be visible behind an endobag or inside a transparent trocar. Here, we choose to delineate the item closest to the camera, but speared out the others as the purpose for instrument detection is to comprehend surgical workflow and detect harmful tool tissue interaction. We also decided to annotate camera stains (Fig. 10b), as object detection models may falsely recognize them as instruments. While this approach may seem overly time consuming, we correlated the granularity of performed annotation with ML model detection of present instruments.

When training segmentation algorithms on the detailed pixel annotation that we performed, we demonstrate a very exact detection of the edges and orifices.

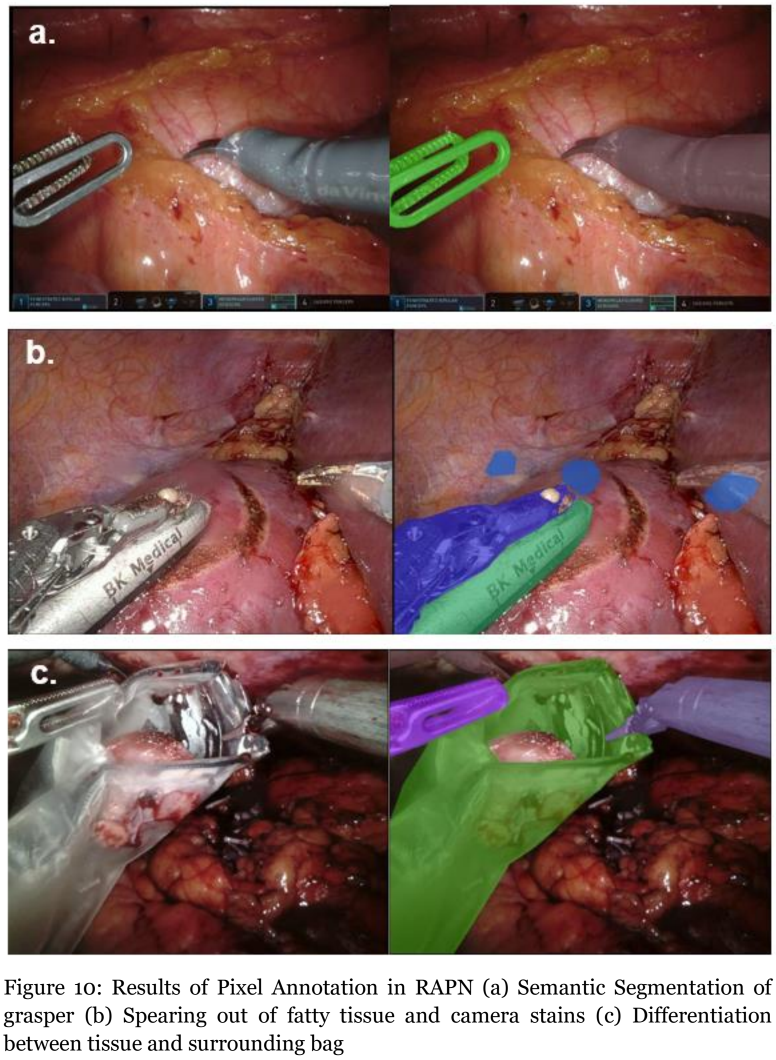

Figure 10: Results of Pixel Annotation in RAPN (a) Semantic Segmentation of grasper (b) Spearing out of fatty tissue and camera stains (c) Differentiation between tissue and surrounding bag

Figure 11 shows results of a machine learning model (more specifically a deep learning model named DeepLabV3+(12)) performing semantic segmentation on unseen images, trained on this RAPN dataset. We denote the clear need of precise annotation. The model is able to detect the orifices inside the bipolar forceps, something which would have been impossible if these orifices would have never been annotated.

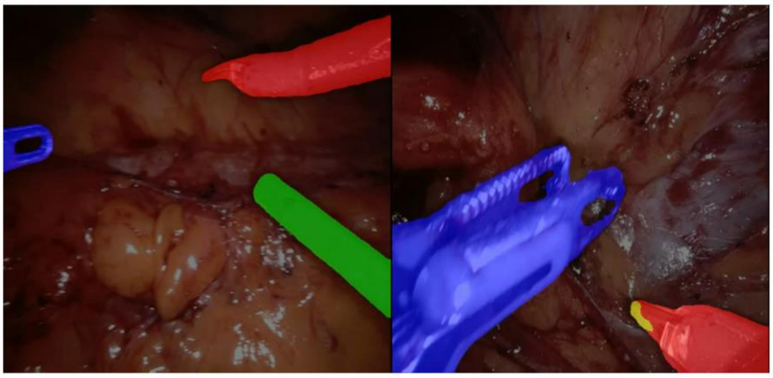

Figure 11: Results of DeepLabV3+ model doing semantic segmentation on new unseen images. We denote how the orifice of the force bipolar is correctly predicted, due to very precise input data of this model.

Another problem encountered was dealing with multiple instances of the same class, e.g. with hemolock clips. Figure 12 shows a surgical scene after renorrhaphy with 12 hemolock clips. When identical items are present in the image, we found it important to assign these to different classes. As can be seen in Figure 12a, when all clips have the same class and therefore same color it is hard to distinguish the positioning of the clips. Applying different labels and colors for each individual clip, as seen in Figure 12b, allows for easy differentiation. This is a good example of the difference between 'instance segmentation' (Fig. 12a) and 'semantic segmentation' (Fig. 12b). This might be important for algorithms evaluating the safety of clip placement but also in bleeding detection to differentiate individual objects from a bleeding vessel. The same applies to other surgical items that can be present multiple times, e.g. vessel loops, needles and sutures or bulldog clamps.

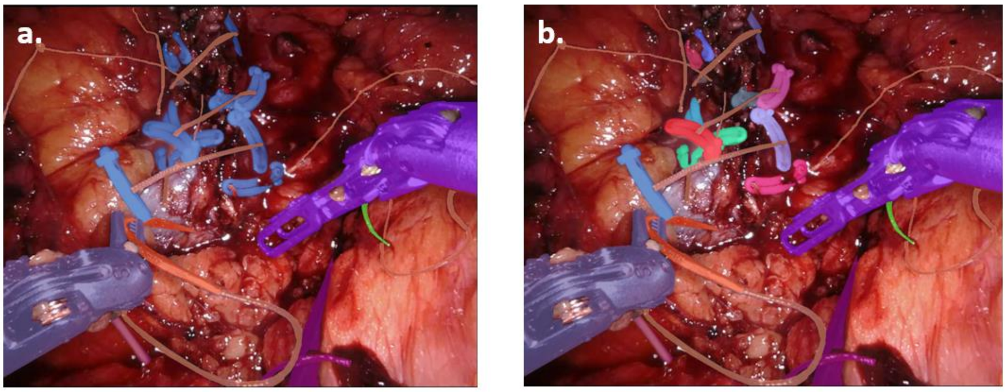

Figure 12: Results for pixel annotation for smaller objects (e.g. clips) (a) semantic segmentation (b) instance segmentation

Vector Annotation

In addition to various levels of detail in pixel annotation we evaluated different inclusion criteria for bounding boxes and vector annotation. In Figure 13a, the entire visible part of the instrument is included in a bounding box assigned to an individual class.

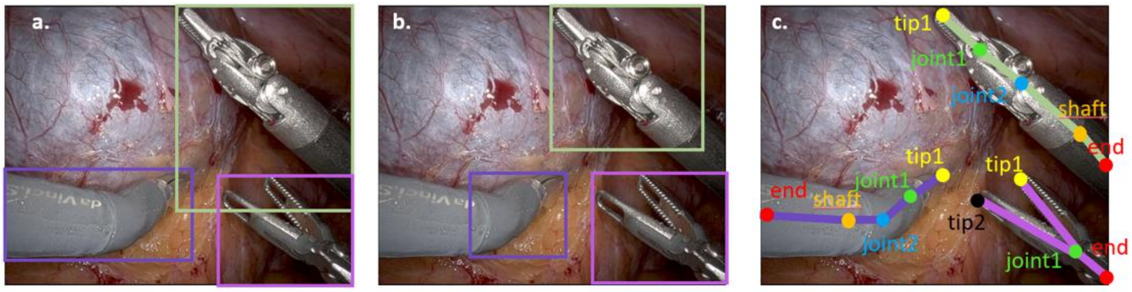

Figure 13: Bounding Box and Vector annotation of robotic instruments. (a) bounding boxes delineating the whole instrument (b) Bounding boxes delineate the tip of the instruments (c) Vector Annotation for pose estimation: separate points indicate the entry point in the image ('end'), the start of the instrument tip ('shaft'), both joints of the robotic instrument ('joint1' and 'joint2'), as well as the instrument tips ('tip1' and 'tip2', or 'tip1' if the instrument is closed)

Bounding boxes offer an easy way to locate an instrument inside the surgical scene and the annotation is very time efficient. However it is also a very coarse annotation not allowing for conclusion on where the exact tip of the instrument is located within the operating field. If bounding boxes overlap, it is also difficult to interpret how instruments are interacting (Fig. 13a). As shown in Figure 14b, bounding boxes can also be derived automatically from segmentations. We know exactly which pixel correspond to each instrument, as such we also know the exact rectangle which will delineate all of these pixels correctly. Bounding boxes are frequently used in applications such as autonomous driving(13) where items move very fast and the main goal is instant detection of items. In contrast to surgery, it is often of less importance in autonomous driving what happens inside the bounding box. In surgery, you can only derive actions by looking what happens inside the bounding box. Therefore, we also explored a more precise way of delineating the whole instrument, by focusing our bounding box only on the tip of the instrument, as seen in Figure 13b. While we performed an increase in granularity throughout pixel annotation (Fig. 2c), the decrease in bounding boxes appeared more effective for the vector annotation part.

Figure 13c shows the results of our final wireframe definition for vector annotation. During the process, we noted it was far more useful to indicate different points along the instruments rather than defining lines. Rather than gathering a collection of vectors along the axis of the instrument without knowing what the start and ending point of the vector signifies, we started to denote the different hinges of the instruments, and derive the vectors as the simple lines between these points. As such figure 13c demonstrates how adding simple points or subdivisions of the instruments using vector annotation can derive a fully functional wireframe, as compared to Figure 3c in which only lines are present. For our final convention, the point where the instruments enter the frame was noted as 'end', the location of both 'joints' were indicated and the position of the 'tips' were separately annotated when they could be seen. The 'shaft' noted the transition between instrument the head of the instrument and the shaft.

This allows detailed differentiation of visible and occluded parts of the instrument and derivation of the exact functional subunits of the instrument interacting with tissue. This wireframe information can subsequently be combined with a segmentation mask to show if the occlusion is due to another instrument, arising from overlaying tissue or because parts of the instrument are located outside of the camera view. Wireframes also provide information on the instrument orientation and give more fine grained details such as opening or closing of graspers, certain instrument angulations etc.

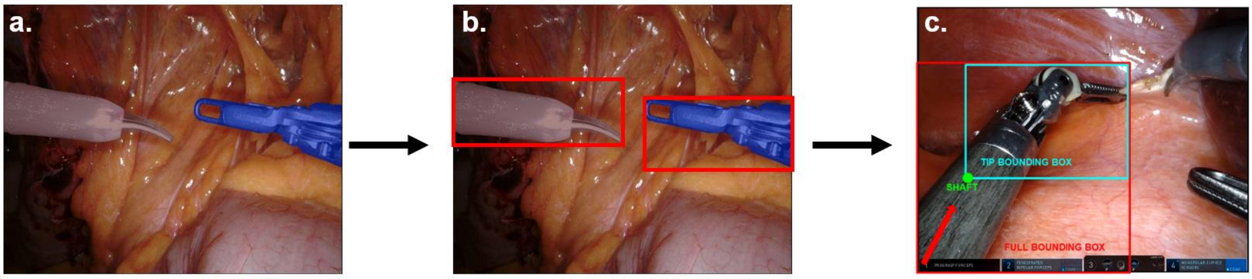

Figure 14: Results of automated deduction of bounding box position from instrument segmentation

Transferability of Annotation Workflow to RAMIE

The transfer of the previously established hierarchical annotation team approach was perceived well at the Robotic Innovation Laboratory at University Hospital Cologne by participating medical students as well as surgical residents. Even though the annotator team was significantly smaller (a total of 5 annotators compared to 38 annotators without a layman annotation workforce) the annotation framework was successfully transferred to the second annotation site and completed by novel as well as experienced annotators. To enhance the feedback structure, there was close collaboration with ORSI academy to address difficulties along the annotation process. Average annotation time for pixel annotation was 6 minutes per frame and average annotation time for vector annotation was 3 minutes per frame. With an average total duration of 196 minutes for the thoracic part of RAMIE the annotation of all frames at sampling periods of 20 seconds, the total estimated time for annotation would result in 14.7 annotation hours per procedure. Figure 15 shows the results of pixel segmentation for RAMIE. Annotators reported that the annotation accuracy was most likely affected by camera aspects, such as presence of smoke and additional annotation burden by delineating information irrelevant to future model detection, such as display information as seen in figure 15c and 15d and the bottom of the image and alongside the instruments.

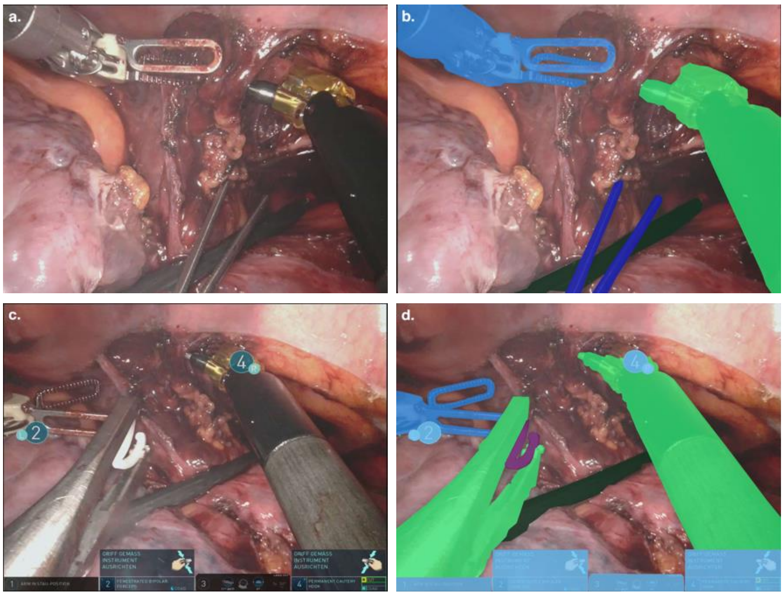

Figure 15: Results of pixel annotation in RAMIE (a + b) surgical scene without distraction (c+d) surgical scene requiring annotation of display information

Figure 16 shows the results of combined bounding box and vector annotation in RAMIE. It is notable that the cautery hook was perceived the most difficult for vector annotation due to the changing angle effecting the orientation of the functional tip (Fig. 16c).

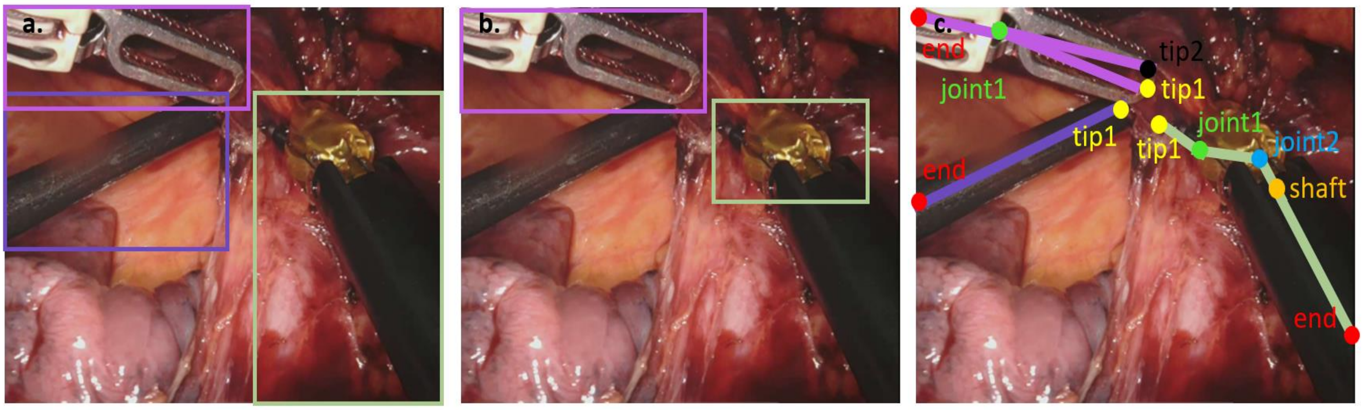

Figure 16: Results of bounding box and vector annotation in RAMIE contrasting the different angles of the cautery hook requiring alteration of annotation and resulting in higher diversity of frames. Note the absence of the tip bounding box for the suction, as there is no articulating or rotating end.

Encountered Pitfalls

We provide frequently encountered pitfalls for new annotators. A detailed description of all encountered pitfalls throughout the annotation process is given in the appendix. Figure 17 for example shows common errors while using watershed techniques, in which images are preprocessed and an automated edge detection is plotted over the image. The first column shows the suggested delineation by automated edge detection (a, d) in an effort to speed up annotation. The corresponding segmentation by clicking these predefined 'superpixels' generated by watershed algorithms(10) are shown in the second column (b,e). The third column (c,f) shows the annotation when performed fully manual. We see that in figure c and f, the tip is immediately segmented correctly and the stem is annotated much more precise by means of a straight line.

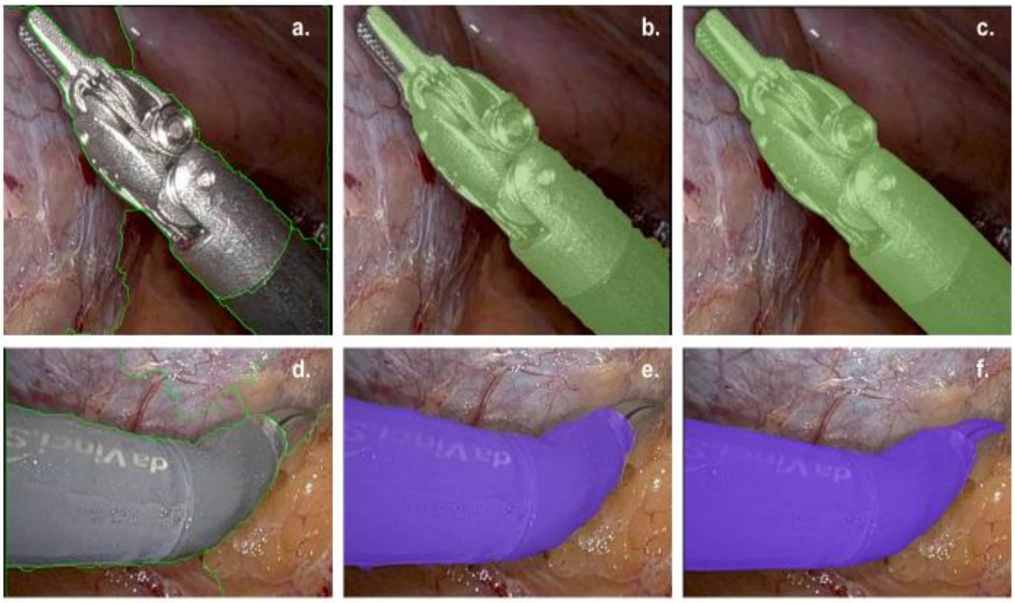

Figure 17: Decrease in annotation accuracy when using watershed techniques (a+d): Initial proposal of similar regions to speed up annotation by clicking them. (b+e): Pixel annotations resulting from the use of these regions (c+f) Correct segmentation, performed by annotating an image fully manual.

Other common errors in pixel annotation was failure to respect the edges of the image, assigning wrong classes to instruments, forgetting to assign classes to instruments and inconsistent delineation of the edges of instruments. These errors show that image annotation is a delicate work that requires concentration and practice and may point towards annotation expertise and phenomena such as annotation fatigue.

To our experience, marking instruments outside the picture, forgetting to annotate objects or to assign classes were often the result of inattention and inaccuracy, whereas incorrectly assigning classes to instruments is mainly due to lack of knowledge with the various (robotic) instruments and objects. One solution we propose is to consult previous or following pictures for clarification, e.g. when an instrument is unrecognizable on a picture, because the full tip and joints are hidden behind tissue. An example is shown in Figure 18, where the left upper instrument is in fact a large needle driver. Because their stem is dark grayish, it can be mistaken for a suction or another robotic instrument with the same stem.

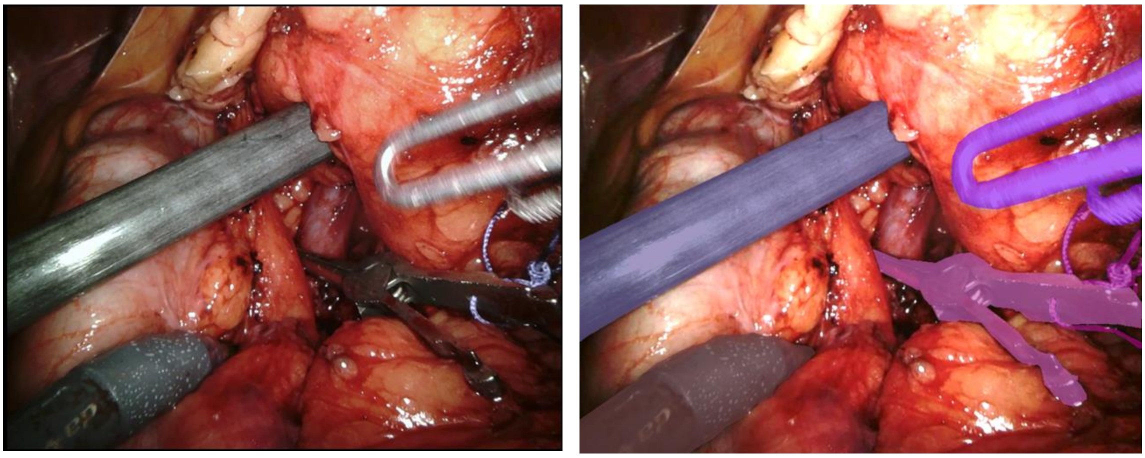

Figure 18: Example of incorrect identification of instruments, which may be solved by consulting chronologically later or earlier frames

Inaccurate annotation of the instrument edges by new annotators was quickly addressable through guidance from the QA-team and practice. This feedback loop and guidance does not only increases the precision of the annotations but also increases the efficiency of the annotator. The errors mentioned above improve as the annotators become more experienced with annotating, further confirming our hypothesis on a clear learning curve effect on the different tools and the annotation program. The QA-team provides feedback on a per image basis, by returning the images to the annotator with accompanying feedback on how to solve the error and avoid it in the future. This active feedback speeds up the learning process of the annotators which results in smoother and more qualitative annotations. After this, the annotator addresses the feedback, resubmits the image for quality check and if correctly annotated, the QA-team changes the image status to completed. If only minor edits are necessary, the QA-team can make these adjustments themselves.

Discussion

Supervised machine learning compromises the majority of applications of artificial intelligence in surgery today. Arguably the performance of a model is just as good as the underlying information the model was trained on. As Meireles et al lay the ground stone for surgical video annotation in the SAGES consensus(2), a key message was that annotation, both spatial as well as temporal, should increase in granularity to ultimately account for all facts compromising the surgical field of view that are perceived by the surgeon. This study is the first of its kind, exploring not only the required detail of spatial instrument annotation but also describe encountered pitfalls. Besides passing on detailed information about first steps of instrument annotation, we demonstrate that our annotation methodology, annotation teaching and team composition is transferable across two highly complex robotic operations, RAPN and RAMIE, and across institutions.

Another key message of this work is that annotation efficiency can be significantly increased by a hierarchical annotation team structure. Arguably, the bottom up approach requires a large workforce, however it also significantly decreases the annotation burden for clinicians. Allan et al(14) noted that a good rule of thumb for starting instrument annotation is to annotate all non-biological objects, as these are also easily identifiable by laymen, and leaving the more detailed discrimination of subunits of instruments and functional properties to medical students or clinicians. Ward et al pointed out that a major obstacle in time-consuming annotation, such detailed as pixel annotation, is the time spent by clinicians on annotations rather than designating that time to patient care and clinical work(8). Additionally, annotator selection was pointed out to be one of the big challenges in surgical video annotation. The balance between clinical and annotation expertise is crucial for meaningful annotations(8). An alternative successful top down approach may consist of a medical team, e.g. of medical students and surgical residents, who perform a very coarse indication of the instrument location and class, after which laymen perform a precise segmentation starting from this annotation(6). Nonetheless, this requires experienced clinicians to check and adjust the laymen images, making the overall time saved questionable. We found annotation outsourcing to be very feasible and efficient but conclude that annotation outsourcing as well as the following supervision should be based on image count and not on hours performed. Given this fact, future research should focus on average time spent in annotating a set of images, rather than annotating a certain amount of time and seeing how many images can be done. The extent to which pre-annotations by laypeople speed up the process, and what qualifications annotators should possess has to be further investigated.

In this exploratory study, we performed much more detailed pixel annotation, compared to previous datasets(15,16), including the semantic segmentation of individual clips and annotation of camera stains to provide instrument detection algorithms with information to rule out false detection of such stains as relevant objects. This approach is also confirmed in previous large laparoscopic datasets(15,16). As lens stains are very frequent during any surgical procedure and have a high impact on visibility and edge detection of relevant instruments, we consider annotation of artefacts, like lens stains, to be equally important as delineating the target objects themselves.

In comparison to time consuming and very detailed pixel annotation, bounding boxes display the disadvantage of information loss. In surgical procedures where small movements may cause huge damage, the detection of the full instruments through bounding boxes does not appear to be helpful. The tip of the instrument provides most information about the position of the instrument, its differentiation towards others tools and the tool-tissue interaction. As such, we found tip-bounding box annotation to be most effective, in combination with vector annotation. After all, smaller boxes mean more exact localization of functional instruments. Throughout the 2015 EndoVis Challenge(17), a time-efficient strategy of separating the tip from the rest of the instrument was used in contrast to the very time consuming segmentation of the whole instrument as used throughout this paper. Overall, we propose a combination of annotation methods. In surgical scenes where wide range movements are dominant and a lower detection accuracy by machine learning models may be acceptable, bounding box annotation may be sufficient, whereas in fine-tuned dissection phases detailed pixel annotation and vector annotation is advisable.

The main question arising from this exploration is the effectiveness of annotation and appropriateness of performed annotations for machine learning applications. Which annotation modality and annotator team composition should be chosen for instrument detection algorithms is the first question every AI project should address. The second question should be how many frames have to be annotated for reliable model training and consecutive model validation, which brings up the question of appropriate sampling frequencies on model training. When analyzing complex and lengthy procedures, it is unfeasible to annotate every single frame. For example, annotating every frame of a one hour long video recorded at 30 fps, requires annotating 108.000 images. As such frames should only be sampled every few seconds in consideration of the balance between accounting for all relevant information throughout the procedure and avoiding annotation of identical scenes. Supervised computer vision models perform best when given a high variation of images, however very little is known on the sampling frequency required for detection and prediction of tool position. Another frequently addressed problem in data science, specifically medical datasets, is the class imbalance problem, that arises from low variability in surgical scenes, for example due to unchanged field of view and rare change in instruments(18). This makes annotation projects focused on more than one procedure all the more relevant.

Going beyond sampling periods of 20 seconds would imply losing other relevant temporal surgical tool information, which might be relevant for action detection or tool-tissue interaction. These aforementioned items are all considered next steps in the field of surgical data science.

The need for a dedicated quality team is evident, as they provide high quality annotations, and have shown to be the fastest, both in annotation from scratch as in annotation and quality assurance of images pre-annotated by of laymen. With this comes the insight that there is a minimum number of annotations to be performed to achieve a decent performance. We found this to be more than 500 completed annotations in images from scratch.

Next to this, large variations in annotation times exist if you train people completely new to the use of annotation platforms. We suspect motivational factors to influence annotation times and note that assigning a predefined set of images works better than assigning working hours. Nevertheless, we found that for different robotic procedures, the expected time is around 4.5-6 minutes per pixel annotation per image and these times half when addressing vector annotations. This allows dedicated research teams to estimate the required workload to do these tasks.

This work focused heavily on the annotation of instruments. Future work should focus on the annotation of soft tissues to study and quantify tool-tissue interactions, as well as optimal sampling periods, which may differ as the soft tissues might be less prone to movements as compared to the instruments which are in constant movement.

To conclude, the validation of any annotation framework requires the exploration of other feasible, effective and well generalizable annotation methodologies. Consensus can only be reached through diverse exchange of experience. Defining exact quantitative and qualitative benchmarks for spatial as well as temporal annotation is crucial to insure reliable applications of machine learning to surgery.

Acknowledgements
Francesco Cisternino, Federica Ferraguti, Joni Dambre, Saar Vermijs, Margot Troch, Matthias Boeykens, Esma Bensiali

Disclosures

Pieter De Backer: no conflicts of interest or financial ties to disclose, funding by the Flanders Innovation & Entrepreneurship Agency - HBC.2020.2252
Jennifer A. Eckhoff: no conflicts of interest or financial ties to disclose, Educational Funding Olympus Corporation, Japan
Jente Simoens, Dolores T. Müller, Charlotte Allaeys, Heleen Creemers, Amélie Hallemeesch, Kenzo Mestdagh, Charles Van Praet, Charlotte Debbaut, Karel Decaestecker, Christiane J Bruns, Ozanan Meireles, Alexandre Mottrie, Hans F. Fuchs: no conflicts of interest or financial ties to disclose.

References

1. Shvets AA, Rakhlin A, Kalinin AA, Iglovikov VI. Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). 2018. p. 624–8.

2. Meireles OR, Rosman G, Altieri MS, Carin L, Hager G, Madani A, et al. SAGES consensus recommendations on an annotation framework for surgical video. Surg Endosc. 2021 Sep;35(9):4918–29.

3. Ward TM, Mascagni P, Ban Y, Rosman G, Padoy N, Meireles O, et al. Computer vision in surgery. Surgery. 2021 May;169(5):1253–6.

4. Hashimoto DA, Rosman G, Meireles OR. Artificial Intelligence in Surgery: Understanding the Role of AI in Surgical Practice. McGraw-Hill Education; 2021. 432 p.

5. Maier-Hein L, Eisenmann M, Reinke A, Onogur S, Stankovic M, Scholz P, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun. 2018 Dec 6;9(1):5217.

6. Maier-Hein L, Mersmann S, Kondermann D, Bodenstedt S, Sanchez A, Stock C, et al. Can masses of non-experts train highly accurate image classifiers? A Crowdsourcing Approach to Instrument Segmentation in Laparoscopic Images. Med Image Comput Comput Assist Interv. 2014;17(Pt 2):438–45.

7. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. IEEE Trans Med Imaging. 2017 Jan;36(1):86–97.

8. Ward TM, Fer DM, Ban Y, Rosman G, Meireles OR, Hashimoto DA. Challenges in surgical video annotation. Comput Assist Surg (Abingdon). 2021 Dec;26(1):58–68.

9. Fuchs HF, Müller DT, Leers JM, Schröder W, Bruns CJ. Modular step-up approach to robot-assisted transthoracic esophagectomy-experience of a German high volume center. Transl Gastroenterol Hepatol. 2019 Aug 23;4:62.

10. Kornilov AS, Safonov IV. An Overview of Watershed Algorithm Implementations in Open Source Libraries. Journal of Imaging. 2018 Oct 20;4(10):123.

11. Reinke A, Tizabi MD, Sudre CH, Eisenmann M, Rädsch T, Baumgartner M, et al. Common Limitations of Image Processing Metrics: A Picture Story [Internet]. arXiv [eess.IV]. 2021. Available from: http://arxiv.org/abs/2104.05642

12. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. arXiv [csCV] [Internet]. 2018 Feb 7; Available from: https://arxiv.org/abs/1802.02611

13. Danzer A, Griebel T, Bach M, Dietmayer K. 2D Car Detection in Radar Data with PointNets. 2019 Oct;61–6.

14. Allan M, Kondo S, Bodenstedt S, Leger S, Kadkhodamohammadi R, Luengo I, et al. 2018 Robotic Scene Segmentation Challenge [Internet]. arXiv [cs.CV]. 2020. Available from: http://arxiv.org/abs/2001.11190

15. Roß T, Reinke A, Full PM, Wagner M, Kenngott H, Apitz M, et al. Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge. Med Image Anal. 2021 May;70:101920.

16. Maier-Hein L, Wagner M, Ross T, Reinke A, Bodenstedt S, Full PM, et al. Heidelberg Colorectal Data Set for Surgical Data Science in the Sensor Operating Room [Internet]. arXiv [cs.CV]. 2020. Available from: http://arxiv.org/abs/2005.03501

17. Allan M, Shvets A, Kurmann T, Zhang Z, Duggal R, Su Y-H, et al. 2017 Robotic Instrument Segmentation Challenge [Internet]. arXiv [cs.CV]. 2019. Available from: http://arxiv.org/abs/1902.06426

18. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. Int J Mach Learn Comput. 2013;224–8.