scientific data

Check for updates

OPEN

DATA DESCRIPTOR

LapEx: A new multimodal dataset for context recognition and practice assessment in laparoscopic surgery

Arthur Derathé[1], Fabian Reche[1,2], Sylvain Guy[1], Katia Charrière[3], Bertrand Trilling[1,2], Pierre Jannin[4], Alexandre Moreau-Gaudry[1,3], Bernard Gibaud[4] & Sandrine Voros[1] ✉

In Surgical Data Science (SDS), there is an increasing demand for large, realistic annotated datasets to facilitate the development of machine learning techniques. However, in laparoscopic surgery, most publicly available datasets focus on low-granularity procedural annotations (such as phases or steps) and image segmentation of instruments or specific organs, often using animal models that lack clinical realism. Furthermore, annotation variability is seldom evaluated. In this work, we compiled 30 sleeve gastrectomy procedures and performed three levels of annotations for a specific step of this procedure (the fundus dissection): a procedural annotation of fine-grained activities, a semantic segmentation of the laparoscopic images, and the assessment of a surgical skill, specifically the quality of exposition of the surgical scene. We also conducted a comprehensive annotation variability analysis, highlighting the complexity of these tasks and providing a baseline for evaluating machine learning models. The dataset is publicly available and serves as a valuable resource for advancing SDS research.

Background & Summary

Patients, hospitals and insurance companies share a common expectation for safer outcomes[1]. This calls for more explicit, objective and quantitative monitoring of surgical interventions in order to provide context-aware assistance to surgeons, especially at the beginning of their learning curve. Surgical Data Science (SDS) is a new research domain which aims at providing solutions to professionals of interventional medicine for decision support, context-aware assistance and training[2].

Video of the operative field is the main source of data acquired intraoperatively in the operating room regarding the surgical process followed by the surgeon. However, there is few findable, accessible, interoperable and reusable data, which remains a major limitation for SDS[3]. The access to such raw video data is conditioned by a close partnership with clinicians. It is also constrained by personal data protection regulations, by the low digitization and the low structuring of clinical data, as well as by the high variability in the processes of care and of data management in hospitals[2]. Sharing surgical video data empowers SDS research projects by, for example, enabling comparison of different methodological approaches on realistic and varied datasets or enabling collective annotation efforts.

In SDS, there is indeed a need for annotated data especially for deep learning techniques. For instance, recordings of surgeries can be annotated with procedural, skill and image-level annotations. Procedural annotation describes the steps of a procedure at different levels of granularity (e.g., phases, steps, activities)[4]. Skill annotation characterizes the different surgeons using criteria, such as the OSATS score[5]. Image-level annotation may be instruments/ organs localization with bounding boxes[6], semantic segmentation[7] or instance segmentation[8].

Several publicly available surgical video or image datasets include one or two distinct types of annotations. Endovis 2018[9] includes image-level annotations for the segmentation of robotic nephrectomies, The Dresden Surgical Anatomy Dataset[10] provides semantic segmentation of abdominal organs and vessel structures. MISAW[11] provides 27 sequences of robotic micro-surgical anastomosis on artificial blood vessels with kinematics and procedural annotations at three levels of granularity (phases, steps, activities). Cholec80[12] provides procedural annotations of phases and instrument presence annotations of 80 cholecystectomies, while CholecT50[13]

[1]Univ. Grenoble Alpes, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, INSERM, TIMC, 38000, Grenoble, France. [2]Department of digestive surgery, CHU de Grenoble, Grenoble, France. [3]Clinical Investigation Center - Innovative Technology, CHU de Grenoble, Grenoble, France. [4]Université Rennes, INSERM, LTSI - UMR S 1099, 35000, Rennes, France. ✉e-mail: Sandrine.Voros@univ-grenoble-alpes.fr

offers 50 videos of cholecystectomies with procedural annotations of phases and activities, 5 of which contain bounding boxes over the instruments' annotations. The recent Heichole2021[14] dataset offers both procedural (phases/actions) and instance segmentation of 24 cholecystectomy procedures. To our knowledge, only two datasets of training activities performed on pelvitrainers setups, JIGSAWS and PETRAW, and one dataset of cholecystectomy procedures, Chole80-CVS, provide three distinct types of annotations. JIGSAWS[15] provides kinematic, surgical activity and skill (modified OSATS score) annotations for robotic suturing, knot-tying and needle-passing tasks (5 repetitions of each exercise performed by 8 distinct surgeons). PETRAW[16] provides kinematic, procedural and segmentation annotation for 100 robotic peg-transfer tasks; Chole80-CVS[17] adds a skill annotation on top of the Chole80 dataset, with the assessment of a Critical View of Safety score during two specific phases of the procedure. Finally, thanks to the public availability of the Chole80 dataset, another group released CholecSeg8k[18] with a semantic annotation of ~8 thousand frames from 17 surgeries of Chole80.

In this paper, we introduce and share a new dataset, LAParoscopic EXpertise or LapEx[19], which offers a multimodal annotation of laparoscopic sleeve gastrectomy procedures. Sleeve gastrectomy is the most commonly performed weight loss surgery (~380 000 cases/year worldwide). It involves resecting a portion of the stomach to promote weight loss by removing the section that produces ghrelin, the hormone that stimulates appetite, and by restricting food intake. Obesity is a major public health epidemic with over 4 million people dying each year from being overweight, and in some cases, weight loss surgeries are recommended. They are associated with 50% reduction in all-cause mortality among obese adults.

The LapEx[19] multimodal annotation is composed of a procedural, an image-based and a skill related annotation task. By describing these different aspects of the surgery, we make available different information that we assume to complement and enhance each other:

- The procedural annotation describes the activities of the practicing surgeon and the assistant at the granularity level of surgical activities.
- The skill annotation is twofold:
  - a clinical criterion, the exposure quality of the scene. It was performed by the expert surgeon involved in the study,
  - the operating surgeon is identified by an index (two expert surgeons performed the recorded procedures).
- A full scene laparoscopic segmentation is also provided, with 11 instruments/organs classes.

Methods

Overview. We provide a new multimodal dataset named LapEx: LAParoscopic EXpertise[19]. This dataset includes 30 video recordings of sleeve gastrectomies and annotations exclusively during the step of the Fundus dissection. This choice was done in accordance with our clinical partner since this step:

- is a key step in sleeve gastrectomy to ensure a complete longitudinal resection of the stomach. In this part of the stomach are the most numerous ghrelin-producing cells that will stimulate the appetite. An incomplete resection of the fundus will result in less effective weight loss and subsequent dilation with weight gain. A dissection and resection too close to the gastric-oesophageal junction will expose the patient to a high risk of gastric leaks.
- is a classical one but with a huge disparity among the various levels of practice of the surgeons in laparoscopy. Therefore, it is a particularly interesting step regarding teaching and training of surgeons.

Two left-handed senior surgeons at the Grenoble University Hospital, France, helped by one assistant (whose handedness was not known), performed 15 surgeries each.

Our dataset contains:

- 30 anonymized videos of sleeve gastrectomies recorded at 25 frame per second and performed by two senior surgeons at the Grenoble University Hospital, France, helped by one assistant; 15 surgeries per surgeon.
- procedural annotations with quadruplet (actor – which hand is concerned by the action, surgical instrument used to perform the action, verb representing the action, target – anatomical structure on which the action is performed) describing the actions performed by both the surgeon and the assistant, during the dissection of the Fundus surgical step.
- skill annotations consisting of:
  - the identification of the practising surgeon (index $\in [0, 1]$ for each surgery)
  - the "quality of exposure" assessed along the dissection of Fundus step (index $\in [1, 3]$ manually annotated by an expert surgeon). The assessment of the quality of exposure was performed at particular moments during the dissection of the Fundus surgical step, resulting in 735 "quality of exposure assessments" for the whole dataset.
- a full scene segmentation on 735 images with 11 organs/instruments classes

The full dataset is available within the French Open Science repository recherche.data.gouv, and is described in more details below.

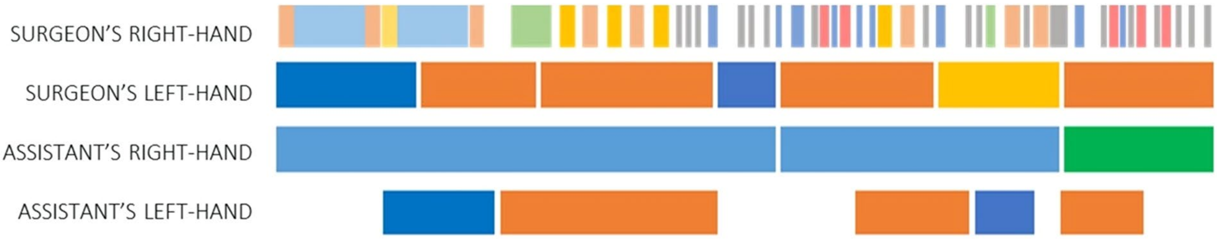

Fig. 1 Visual representation of LapEx's procedural annotation. Each colour corresponds to an activity.

		# of activities		Activities duration (in s.)	
Actor type	**# of activity labels**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
Assistant left hand	3	5.00	3.65	82.44	170.91
Assistant right hand	4	1.77	1.74	285,05	197.89
Surgeon left hand	14	13.57	8.48	34.01	39.73
Surgeon right hand	24	58.53	20.34	5,97	4.96

Table 1. Statistics on activities performed per surgery for the various actors in the LapEx dataset.

Clinical data acquisition protocol. 30 sleeve gastrectomy videos were obtained in the frame of a monocentric retrospective study on medical data. Study ethics approval was obtained on 24 May 2018 (CECIC 342 Rhône-Alpes-Auvergne, Clermont-Ferrand, IRB 5891). All participants were informed of the purpose of the research. Data contained in the database (i.e., the videos) were then de-identified: after inspection and processing to ensure that the videos did not contain identifying information (e.g., out-of-body frames were removed), and each video was associated to a unique ID. As the videos were de-identified, in accordance with French and European regulations, patient consent for open publication of the data was waived. A contract was then established between the Grenoble University Hospital and the involved research laboratories to make the data available to researchers for the creation of the annotation, the development of SDS methods[20,21]. Since the de-identified database was obtained from health data and since it still deals with individual data of patient, the sharing of the database is governed by a Data Transfer Agreement (DTA_BDD_LapEx.pdf) provided with the database.

Procedural annotation. We provide procedural annotations at the fine granularity level of "activities". An activity is usually represented as a triplet < instrument; verb; target > describing how a surgical instrument performed an action/verb on a surgical target[4]. As several instruments can be manipulated at the same time, we considered the activities performed by the two actors (surgeon and assistant) operating together in the surgical video with a quadruplet < actor; instrument; verb; target > for each hand of each actor.

The vocabulary of our quadruplet is aligned on the OntoSPM ontology[22]:

- actor ∈ [Surgeon right-hand; Surgeon left-hand; Assistant right-hand; Assistant left hand]
- instrument ∈ [flat grasper; electrothermal bipolar forceps; liver retractor]
- verb ∈ [holding; pushing; sealing and dividing; pulling; retracting; coagulating; idle]
- "Idle" corresponds to no action of the instrument
- target ∈ [stomach; ligament; liver; compress; lipom of his; adhesion; meso-gastro posterior; spleen; idle].

The meso-gastro posterior corresponds to the attachment of the posterior surface of the stomach to the small epiplonium. "Idle" corresponds to an activity where no anatomical structure is touched by an instrument.

The procedural annotation task was conducted using the b < > com [Annotate] software[23], with the expert surgeon, highly experienced in sleeve gastrectomy (having performed over 2,000 procedures), taking the lead. The surgeon was responsible for annotating all activities performed simultaneously by the four hands during the dissection of the fundus. Assisting throughout the process, the project's lead PhD student aimed to gain a comprehensive understanding of the surgical procedure, develop an annotation vocabulary aligned with the OntoSPM ontology, question the expert's decisions when uncertainties arose, and provide support in using the annotation software. This collaboration resulted in a multidimensional sequence, with four quadruplet labels defined at each point in time (Fig. 1).

For each actor, we estimated the number of types of activities (activity labels), the mean and standard deviation of the number of activities and the mean and standard deviation of the duration of activity samples. These simple statistics are provided in Table 1.

We also provide histograms of the activities triplets occurrences (Fig. 2) and histograms of the occurrences the instrument (Fig. 3, left), verb (Fig. 3, middle) and target (Fig. 3, right) components of the activity triplets. These histograms highlight huge class imbalances in the dataset.

For further insight, we studied the procedural annotation variability by annotating a surgery one year after the initial segmentation. This study is described in subsection 1 of the Technical Validation section.

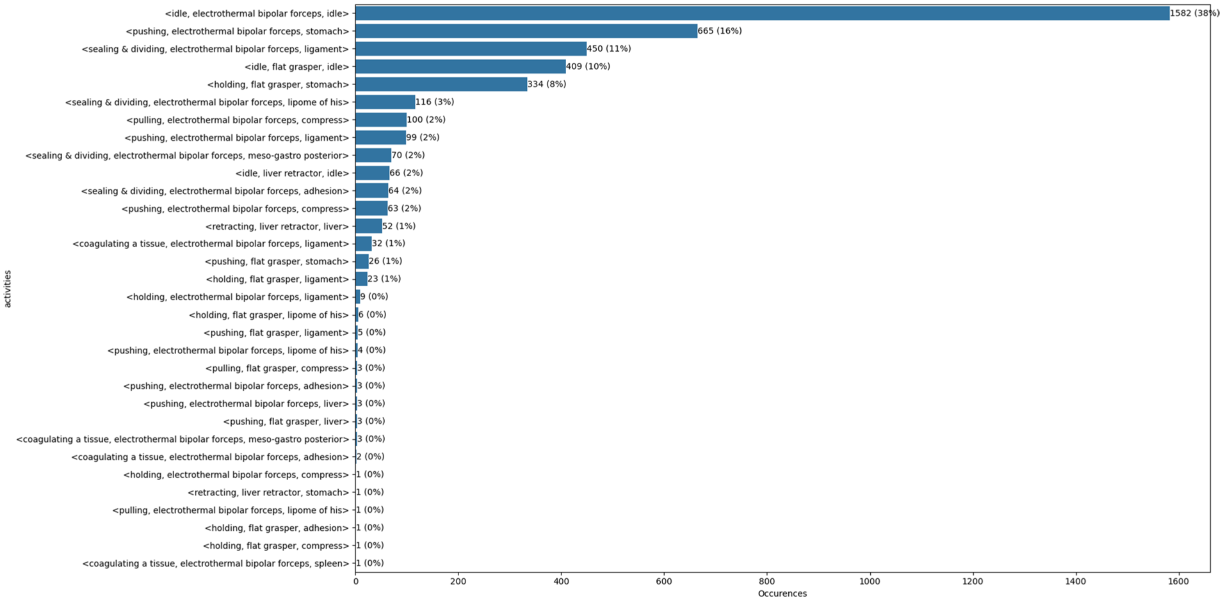

Fig. 2 Histogram of the occurrences of the activity triplets in the procedural annotation of LapEx.

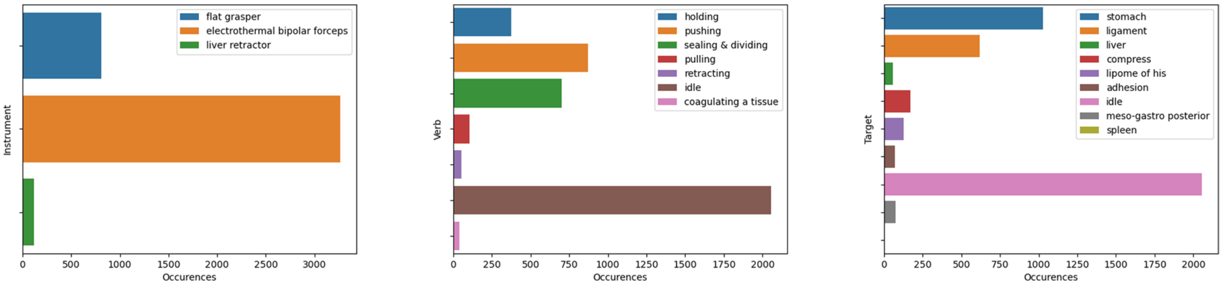

Fig. 3 Histograms of the occurrences of the tool (left), verb (middle) and target (right) components of the activity triplets in the procedural annotation of LapEx.

Skill annotation.

Profile of surgical practice. Each surgeon has a distinct practice profile shaped by their experience, skills, and sensitivities. To capture the individual profiles, we we assigned a unique identifier to each senior operating surgeon.

"Quality of exposure" annotation. In laparoscopic surgeries, the surgeon's vision is entirely dependent on the endoscopic video, which provides a limited field of view. To handle these vision constraints, the surgeon must skilfully manage the surgical space and optimize access to the surgical target. In collaboration with our clinical partner, and based on the clinical literature, we defined the quality of exposure of a surgical target as the as the degree to which the target is adequately exposed in relation to the surgical process and the specific objectives for that target. When the target's exposure is good, the surgeon can focus fully on the procedure's goals. When exposure becomes suboptimal, the surgeon improves it by modifying the relative positions of organs and instruments. Thus, we defined the "quality of exposure" as the surgeon's ability to assess the state of the target's exposure and to make adjustments as needed to achieve the surgical objectives.

The quality of exposure annotation was performed alongside the procedural annotation task, by the same surgeon, scientist, and software as described in Section 3. For each activity in the procedural annotation where the verb in the activity triplet was "sealing and dividing" or "coagulating", the final image of the activity was recorded and annotated to assess the quality of exposure of the surgical scene. These two activities were selected because they are dissection gestures that typically alter the quality of exposure. The main objective during the fundus dissection is to expose the stomach, and to do so the surgeon performed dissection activities mainly with his/her right hand (99% of the dissection activities), and the target structures of the right-hand dissection activities were the ligament (66%), the lipome of His (15%), the meso-gastro posterior (10%) and adhesions (9%).

The quality of exposure was on a three-level scale:

- 1 = good quality
- 2 = satisfying quality but exposure could be improved
- 3 = unsatisfying quality

The annotation process resulted in the annotation of the quality of exposure for 735 frames for the whole dataset.

For further insight, we studied the "Quality of exposure" annotation inter-rater variability by asking an expert digestive surgeon, not involved in the study, to annotated the quality of exposure independently. This study is described in subsection 2 of the Technical Validation section.

Semantic segmentation annotation.

The same 735 images mentioned in section 4.2 were also segmented pixel-wise to identify the visible objects in the scene. The list of classes as well as some images of our dataset with their ground truth segmentations are displayed on Fig. 4.

The segmentation task was carried out by three scientists: the project's lead PhD student, a postdoctoral fellow, and a senior researcher, all of whom were experienced with laparoscopic images. The task was performed using the dedicated software CamiTK[24]. Before beginning the segmentation, the postdoctoral fellow and senior researcher were introduced to the clinical procedure, shown several segmentation examples to help them identify the relevant structures, and trained on the segmentation tool. After this training, the workload was divided among the three scientists based on their availability until the task was completed:

- Operator 1 (PhD Student): 390 segmented images from 15 videos (videos # 14, 15, 18–30),
- Operator 2 (Postdoctoral Fellow): 269 segmented images from 12 videos (videos # 1–10, 16 and 17)
- Operator 3 (Senior Scientist): 80 segmented images from 3 videos (videos #11–13).

Figure 5 highlights huge class imbalances in LapEx[19]: some classes are present in few images, and some represent a very tiny surface in the images (which leads to pixel-wise class imbalance).

For further insight, we studied the inter-operator variability of the segmentation task: three annotators (scientists accustomed to laparoscopic images) segmented the same 36 images extracted from the dataset. These images were chosen as to represent the variability of the dataset. This study is detailed in subsection 3 of the Technical Validation section.

Data Records

The LapEx dataset[19] can be downloaded here: https://doi.org/10.57745/1F0UBU.

The dataset is shared under the Data Transfer Agreement DTA_BDD_LapEx.pdf provided at the root of the dataset directory. In order to download the dataset, the downloader must request an access to the data and agree to comply with the terms of the data transfer agreement. The "read_me" text files describing the contents of the dataset the Data Transfer Agreement can be previewed before requesting access.

It contains two directories; each compressed as tar.gz files

The first directory, **LapEx_dataset** is composed of:

1. a **README_Lapex.txt** file describing its contents
2. 30 sub-folders, one per surgery.

Each surgery sub-folder is composed of:

- A .csv file called **activities_index.csv** containing the procedural annotation of the surgery, as a list of activities performed by the surgical actors during the dissection of the fundus step of the procedure, with the following shape:

  Actor_idx, Start timestamp, End timestamp, Verb_idx, Instrument_indx, Target_idx
  Activity #1
  Activity#2
  …
  Activity #N

  Each annotated activity is a line of the CSV file which is described by its actor, its start and end timestamps, and by its triplet <verb;instrument;target>. Timestamps are given in milliseconds and are synchronized with the timestamps of the JPG files in directory "frames". The verb, instrument and target items are integer labels which corresponding concepts are listed in the **LapEx_dataset/metadata** folder.

- A .csv file called **scores.csv** containing the skill annotation of the surgery, which was performed at the end of each activity with verb "sealing & dividing" or "coagulating a tissue", with the following shape:

  Timestamp, Quality of exposure, Surgeon's Idx
  Exposure #1
  Exposure #2
  …
  Exposure #M

  Timestamps are given in milliseconds and are synchronized with the timestamps of the jpeg files in directory "frames". Quality of exposure is categorized as follows:

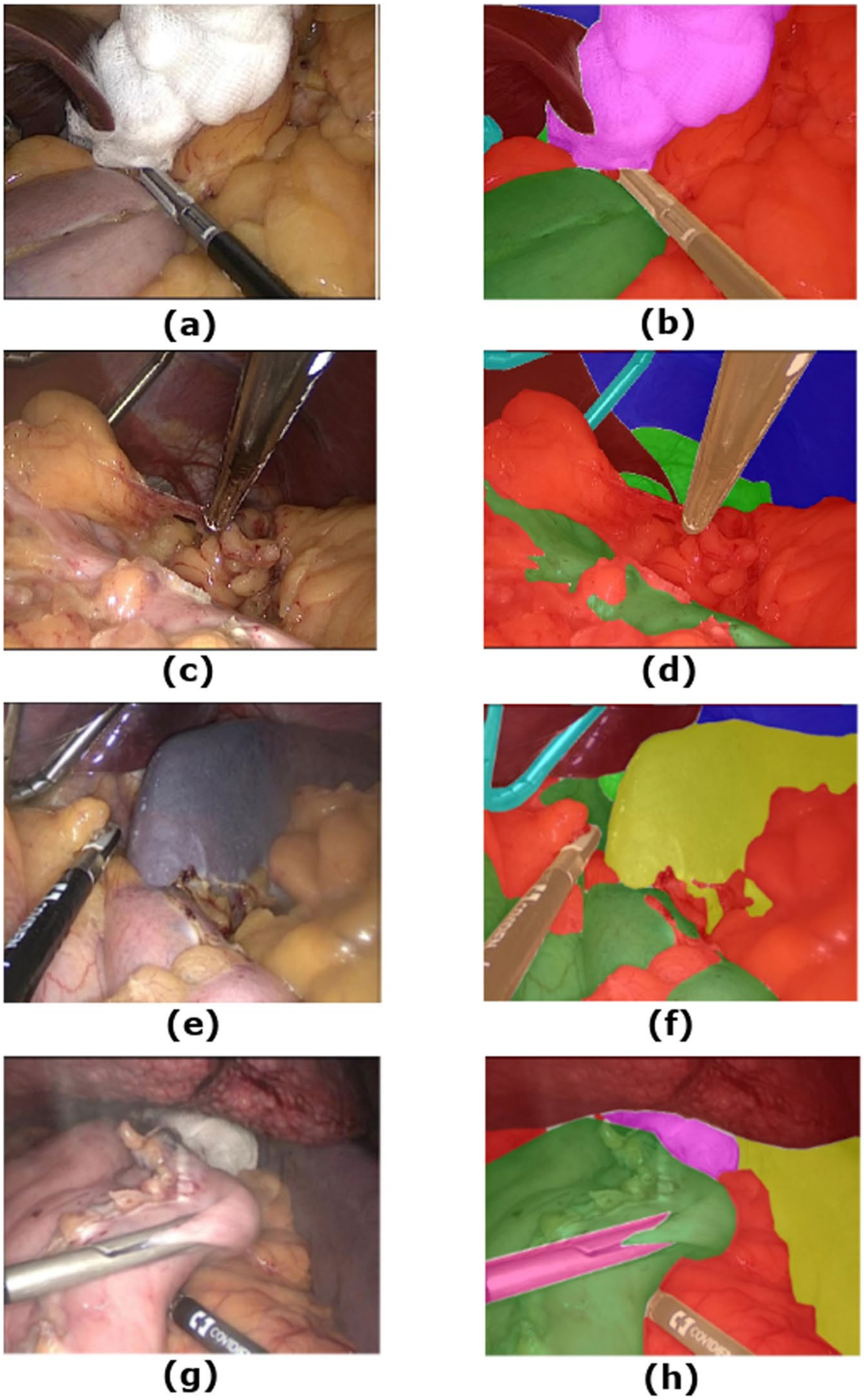

Fig. 4 Examples of images (left column) and their corresponding semantic segmentations (right column) in the LapEx dataset. Classes: adipose tissue (red), stomach (forest green), electrothermal bipolar forceps (orange), compress (pink), liver retractor (light blue), flat grasper (light pink), spleen (yellow), diaphragm (light green), liver (brown), abdomen (blue).

  - 1: good quality
  - 2: acceptable quality
  - 3: non-acceptable quality

- A **/frames** directory containing a sequence of 720×576 RGB jpeg images numbered with their timeframe in milliseconds, corresponding to the dissection of the fundus step of the surgery.
- A **/seg** directory containing the semantic annotation of the surgery as a list of 720×576 grey-level jpeg images, numbered with their timeframe in milliseconds followed by the suffix "_seg". Only a sub-part of the images

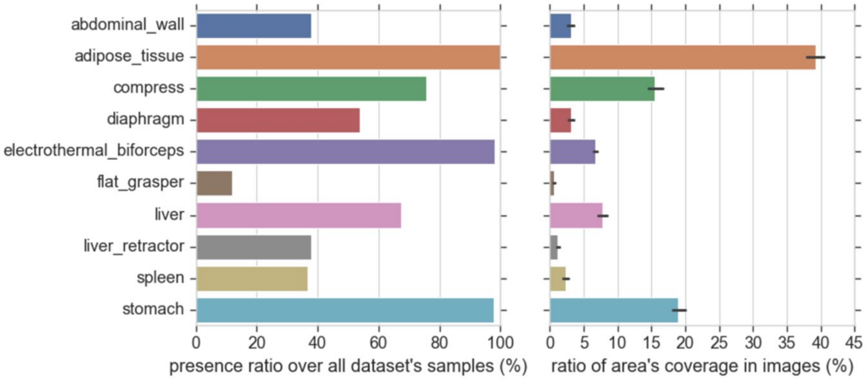

Fig. 5 Class imbalance in LapEx dataset. Left: ratio of the number of frames in which a class is present over the total number of frames. Right: ratio of the area of a class divided by the total area of all images.

of the **/frames** folder are semantically annotated, those images corresponding to "sealing and dividing" or "coagulating a tissue" activities during the dissection of the fundus step.

3. A **/metadata** directory containing metadata about the procedural annotation and the segmentation:

   - File actor.csv contains the dictionary linking the actor textual description to their integer labels used in the activites_index.csv files.
   - Files "instruments.csv", "verb.csv" and "target.csv" contain dictionaries linking the textual descriptions of the activity components to their integer labels used in the triplet <verb;instrument;target> of the "activities_index.csv" files for each procedure.
   - File "segmented_objet.csv" contains the exhaustive list of the segmented objects with their associated grey levels used for identifying the objects in the semantic segmentation in **/seg** directory.
   - File recommended_split.txt contains a train/test split recommendation which enables a fair repartition of videos in both splits, with at least one quality score of 3 and videos performed by each surgeon.

   The second directory, **LapEx_technical_validation_data** contains

1. a **README_Lapex_Technical_Validation.txt** file describing its contents
2. a tar.gz compressed file with three sub_folders:

   - **/intra_op_varia**bility_proc contains the data used for the evaluation of the intra-operator variability in the procedural annotation. It consists of two.csv files with the activity annotations of the same surgery, one year apart. These files follow the same syntax as the activities.csv files of the LapEx dataset. The corresponding surgical video and subtitle files are also provided.
   - /inter_op_variability_seg contains the data used for the evaluation of the inter-operator variability in the segmentation annotation. It contains:

     - the video used for the validation
     - one subfolder OrigFrames with the extracted frames from the video that were annotated by three annotators. The frames are stored as 720×576 grey-level png images.
     - three subfolders, each one containing the list of segmentations performed by one annotator, as 720×576 grey-level png, following the same syntax as the /seg directory of the LapEx dataset.

   - **/inter_op_variability_quality_expo** contains the data used for the evaluation of the inter-operator variability in the segmentation annotation. It contains a.csv files with the quality annotation of a sub-part of the dataset (80 frames) by the original annotator and a second, independent annotator.

Technical Validation

Validation of the procedural annotation.

In order to assess the intra-operator variability of the procedural annotation, the scientist who was involved in this task with the clinical partner segmented one surgery of the dataset on year after the initial segmentation performed jointly with the expert surgeon. We provide both corresponding procedural annotation files and the associated surgical video in the dataset repository.

Figure 6 shows a graphical illustration of the two annotations. We note that long activities (Surgeon left hand and Assistant right hand) are almost perfectly recognized, while there is a bigger difference in annotations for the very short activities (Surgeon right hand). Most of the differences seem to come from the annotations of the target component of the activities.

This qualitative analysis is confirmed by a statistical analysis using the Cohen's Kappa coefficient κ[25], a statistic classically used to assess intra- or inter- annotator variability. For the Surgeon left hand's annotations

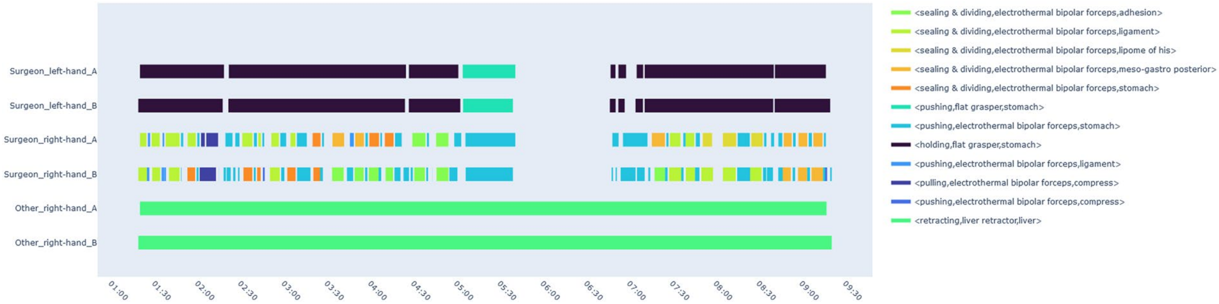

Fig. 6 Illustration of the variability of the procedural annotation of Lapex. Operator A is the expert surgeon who annotated the full dataset. Operator B is the scientist (trained by the expert surgeon) who performed alone the annotation one year later.

		Instrument component	Verb component	Target component	Full activity triplet
Surgeon right hand	Cohen's kappa κ	0.67	0.73	0.57	0.57
	Corresponding agreement level	Substantial ($0.41<\kappa<0.60$)	Substantial ($0.61<\kappa<0.80$)	Moderate ($0.41<\kappa<0.60$)	Moderate ($0.41<\kappa<0.60$)

Table 2. Agreement between the two annotators (surgeon and scientist vs. scientist alone one year later) for the activities' annotation. The annotations are compared using the Cohen's kappa coefficient (Scikit Library implementation).

$\kappa = 0.96$ and for the Assistant right hand's annotations $\kappa = 0.99$ (almost perfect agreement, $0.81<\kappa<1$). For the surgeon right hand's annotations, $\kappa = 0.57$ (moderate agreement for $0.41<\kappa<0.60$). For the Surgeon right hand's annotations, we also analysed the agreement between annotators for the individual components of the activity (Table 2). The majority of the disagreements come from the annotation of the target: the second annotator was a scientist and although he was trained by the expert surgeon, he sometimes lacked the anatomical knowledge to differentiate anatomical structures (e.g., the meso-gastro posterior and the ligaments).

Validation of the quality of exposure annotation. To assess the inter-rater variability in the annotation of the quality of exposure of the surgical scene, 80 samples (40 surgeries from surgeon 0, 40 from surgeon 1) were randomly selected in the set of the 735 annotated frames. After being shown a tutorial, an expert digestive surgeon, from the same clinical centre, but not involved in the study was given access to the videos and the annotation timestamps and was asked to judge the quality of exposure. The quality of exposure scores being ordinal data, we used the Cohen's Kappa coefficient κ with quadratic weights (Scikit library implementation). We obtained $\kappa =$ 0.513 (moderate agreement) with a higher agreement for ratings of surgeon 1 ($\kappa = 0.63$ or substantial agreement) than surgeon 0 ($\kappa = 0.39$ or fair agreement).

Validation of the segmentation annotation. In order to evaluate the consistency in the semantic annotation three annotators segmented the same 36 images extracted from the dataset. Annotator AD was the PhD student who also annotated most of the images of the LapEx dataset. Annotator SV was the senior scientist who annotated a few images of the LapEx dataset. Annotator MC was a new annotator (Masters Student). These images were chosen as to represent the variability of the dataset, and as to ensure that all the possible structures were segmented at least once. We extracted short video clips containing these images from the surgery videos, in order to provide helpful context for the annotation task. We then combined the clips to form a video, which is provided in the dataset repository. We compared the segmentation of the annotators using the mean Intersection over Union (mIoU) metric, a classical metric used for the evaluation of semantic segmentation. The mIoU was computed as described in[26]: a confusion matrix was computed for all the of each annotator pair using the Python scikit-learn library and was used to compute the per-class IoU over the validation dataset. The mIoU of each annotator pair was then computed as the mean over all classes of the per-class IoU. Computing the IoU on the whole dataset rather than image-by-image enables handling missing labels.

The best mIoU of 0.79 obtained for the pair AD/SV, followed by a mIoU of 0.77 for the pair AD/MC and a mIoU of 0.66 for the pair MC/SV. Figure 7 illustrates the similarity and overlap of segmentations between the different operators by displaying the distribution of mIoU values across all classes when the operators are compared two by two. Operator AD has a higher mIoU with operators MC and SV than operators MC and SV have with each other and is therefore more stable in the annotations.

Figure 8 displays the distribution of IoU between operators for each segmented object class. The classes on which annotators agree the most are the compress (average per-class IoU=94%), the abdominal wall (IoU=93%), and the surgical instruments (electrothermal biforceps: IoU=89% and liver retractor: IoU=88%). On the contrary, the three classes with the lowest level of agreement are the spleen (34%) and the diaphragm (49%) and the stomach (60%).

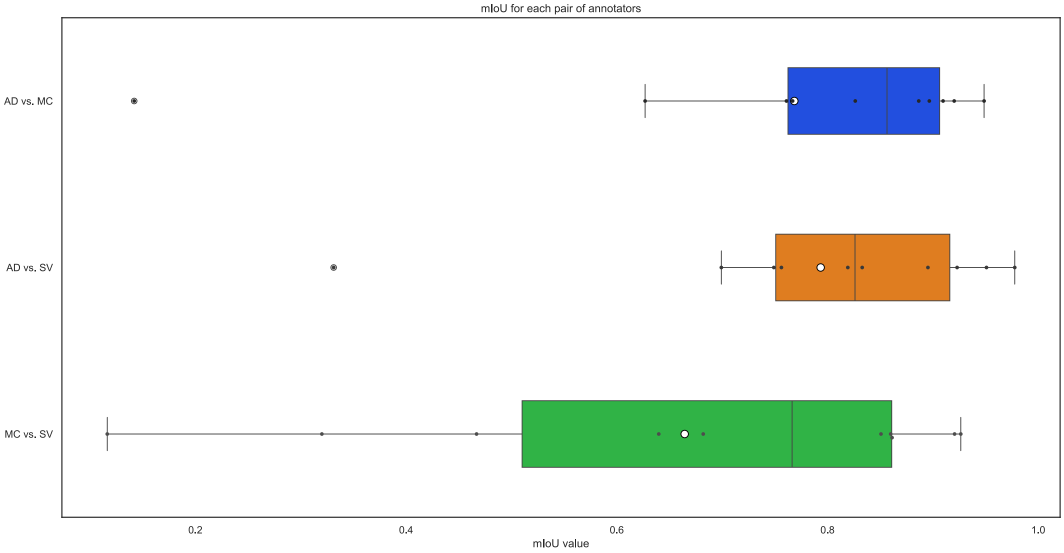

Fig. 7 Inter-annotator variability in the semantic segmentation of Lapex. Distribution of the mean Intersection over Union (mIoU) between two annotators, averaged on all the segmented structures. The black bars in the moustache plots correspond to the median values, the white dots to the mean values.

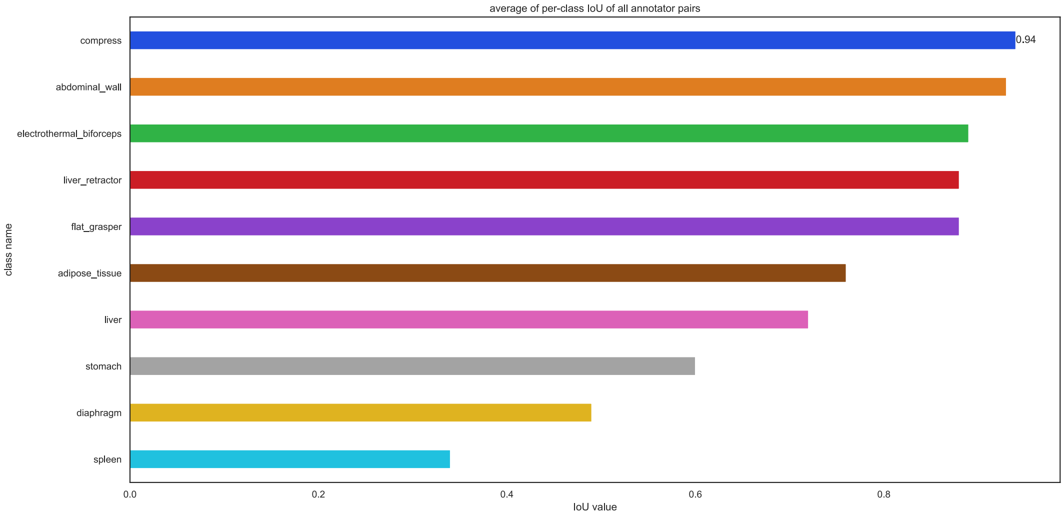

Fig. 8 Inter-annotator variability in the semantic segmentation of Lapex viewed per segmented object class (per-class IoU averaged over all operator pairs).

The low IoU for the spleen can be explained by the fact that this organ is rarely present and barely visible in the images as it is hidden behind other structures. We explain the low IoU for the diaphragm by the challenge of delimiting this structure from the abdominal cavity, as their frontier is very fuzzy. Indeed, the diaphragm is visually a whiter portion of the abdominal cavity, and its delimitation is made by observing the surrounding abdominal tissue. As the images can be more and less zoomed-in during the surgery, making a consistent delineation is complicated. Finally, the stomach IoU is also low because some annotators considered that adipose tissue on the stomach belonged to the stomach while other annotators did not.

Similarly, we provide the distribution the mean Hausdorff distance between contours for each annotator pair (Fig. 9) and the mean Hausdorff distances between contours over annotator pairs for each segmented object class (Fig. 10).

As suspected from our analysis of the IoU errors, the highest Hausdorff distances are obtained for the stomach and adipose tissue. The Hausdorff distances for the spleen and the diaphragm are much lower because they occupy a much smaller portion of the images. However, qualitatively these distances are reasonable, as they are lower than 20 pixels for most of the organs and the inter-rater pairs. Figure 9b provides as qualitative interpretation for this error by providing the 5 mm diameter instrument in pixels.

Figure 11 illustrates these issues (on frame #11000 of the semantic segmentation validation video).

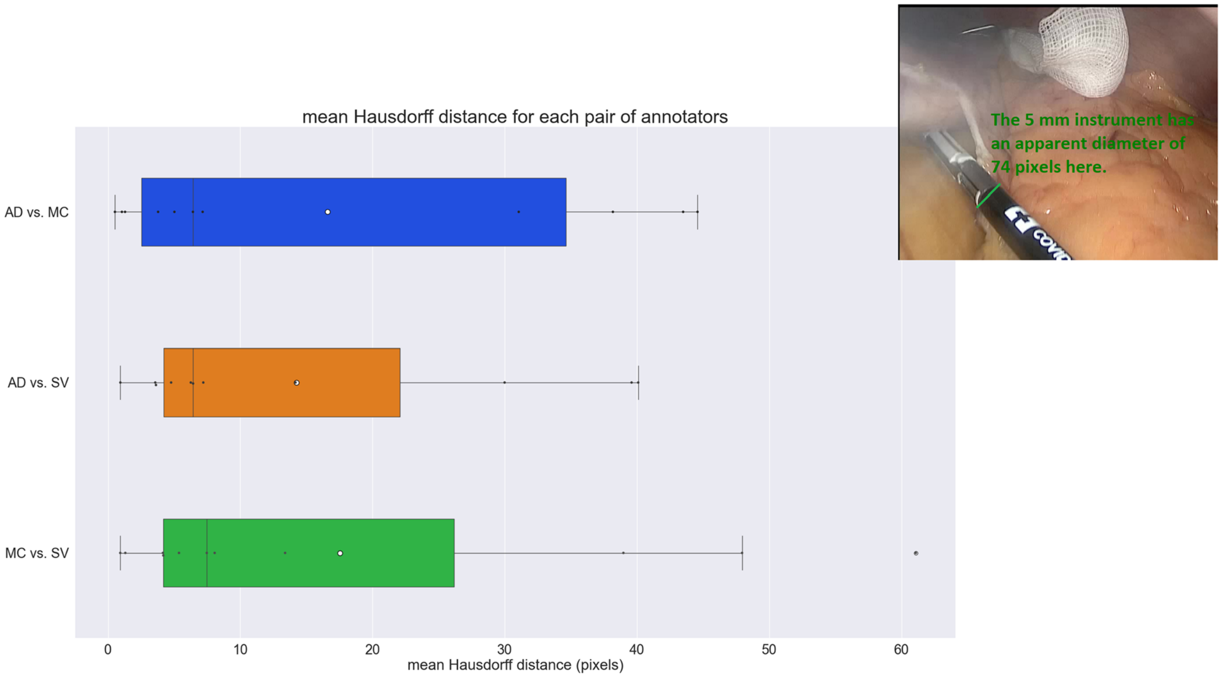

Fig. 9 Left, inter-annotator variability in the semantic segmentation of Lapex. Distribution of the Hausdorff distance between the contours of two annotators, averaged on all the segmented structures. The black bars in the moustache plots correspond to the median values, the white dots to the mean values. Right, example of the size in pixels of a visible instrument for comparison purposes.

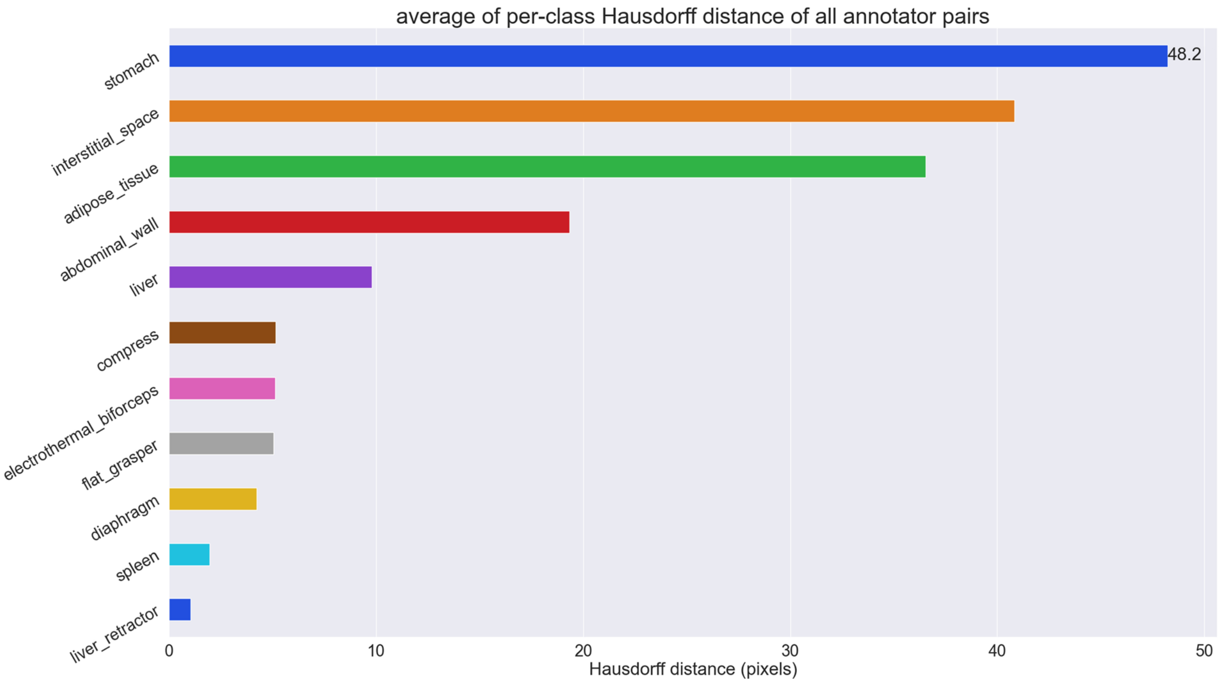

Fig. 10 Inter-annotator variability in the semantic segmentation of Lapex viewed per segmented object class (per-class Hausdorff distance averaged over all operator pairs).

Usage Notes

Dataset access and restrictions. The LapEx dataset[19] can be downloaded here: https://doi.org/10.57745/1F0UBU.

The dataset is shared under the Data Transfer Agreement DTA_BDD_LapEx.pdf provided at the root of the dataset directory. In order to download the dataset, first sign up for an account at recherche.data.gouv.fr, then select the restricted datasets to be requested, and then select 'Request Access', which will redirect to a summary

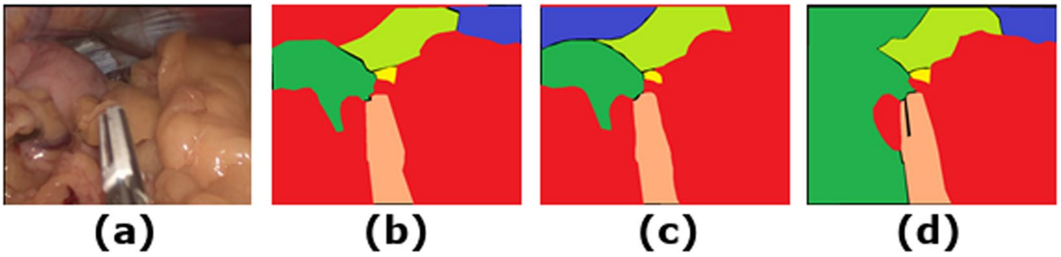

Fig. 11 Illustrations of the challenge of segmenting some organs consistently. (**a**) original frame of a video of the validation set, (**b**) annotation from AD, (**c**) annotation from MC, (**d**) annotation from SV. As fatty tissue (red) covers the stomach (green), there is a confusion between the two classes. The visible part of the spleen (yellow) is very small. The diaphragm (light green) is also hard to delimit as it merges with the abdominal cavity.

of the data usage agreement that can be accepted. After acceptance of the agreement, the corresponding author will receive an email notification and will grant access to the dataset (please count a maximum of 2 weeks delay) to the user. The "read_me" text files describing the contents of the dataset and the full Data Transfer Agreement can be previewed before requesting access.

Dataset limitations.

Dataset contents. LapEx is the first publicly available, annotated dataset for sleeve gastrectomy featuring multimodal annotations. Developing such a dataset is a labour-intensive process, particularly due to the need for expert surgeons' involvement. We aim for this dataset to facilitate the evaluation of novel methods for surgical data science (SDS) and to support their generalization across various surgical procedures. Additionally, its public accessibility invites collaborative efforts to enhance and expand the LapEx annotations, mirroring the success of the Chole8K dataset, which enriched Chole80 with instrument segmentations.

The dataset has several limitations: it was derived from recordings of surgeries conducted at a single centre by two expert surgeons, which already exhibit noticeable differences in practice. These variations are likely to become even more pronounced when comparing surgeons across hospitals and countries[27]. Additionally, the dataset includes a relatively small number of surgeries (30) and focuses exclusively on a single step of the surgical procedure—the Dissection of the Fundus—for annotations. The quality of the exposure annotation, while guided by expert judgment, remains somewhat subjective and was assessed only at specific moments within this surgical step. Furthermore, we limited semantic annotation to these specific moments due to the significant time required for manual image segmentation.

Dataset technical validation. As highlighted in a recent review on annotation for surgical process model analysis[28], studies that assess the variability of procedural annotations are extremely rare—only one out of 34 selected studies in the review addressed this issue. This variability should be considered when evaluating the performance of computational models, but it is seldom assessed. For example, a state-of-the-art model for fine-grained activity recognition[29] achieved a mean average precision of 0.42 on a dataset annotated by a single annotator. Given this level of precision, the scarcity of datasets with activity annotations, and the lack of a baseline for annotation variability, Lapex[19] can be valuable for evaluating the generalization capabilities of computational networks and analysing prediction accuracy in relation to annotation variability. In our study, the second annotator (a scientist), who had been trained by the expert surgeon a year earlier, still showed only moderate agreement with the primary annotator due to differences in annotating the target components of very short activities. We can assume that annotation variability would be reduced if another expert surgeon annotated the dataset, and that annotation variability would decrease if another expert surgeon were to annotate the dataset. This also highlights the importance of having procedural annotations performed by, or in the presence of, expert surgeons to minimize errors—particularly those related to accurately identifying the start and end of events. This problem is especially significant for the annotation of activities, which are short and quickly follow one another. As we were unable to carry out a real study of intra-operator variability in the presence of the surgeon, the actual level of variability of the scientist-surgeon pair remains unknown. Performing an inter-operator variability validation with several expert surgeons, ideally from different hospitals, would be interesting. However, it would require significant availability from highly skilled surgeons.

Regarding skill annotation, we provided an estimation of the inter-operator variability of the "quality of exposure" score. It should be noted that the annotation was performed by an expert digestive surgeon albeit not a specialist in bariatric surgery. We were surprised to observe that the agreement between annotators varied significantly depending on the performing surgeon. To our knowledge, the only similar work[17] did not provide an inter-operator variability assessment for their proposed clinical quality criteria.

Finally, our validation of inter-operator variability in segmentations highlights the challenge of achieving consensus on the definitions of anatomical structures observed by a moving camera with a limited field of view. Some structures, such as surgical instruments or well-defined and visible organs, can be segmented reliably. However, further work is needed within the community to establish agreed-upon definitions of anatomical structures and their spatial relationships, as a few structures were extremely challenging to segment (e.g., the diaphragm, the spleen and the stomach because it was surrounded by adipose tissue). The annotators, who were scientists trained to understand surgical scenes, achieved over 77% mIoU when one of the two compared annotators was the lead scientist involved in the study, and the segmentations were completely independent. For comparison, the recent Dresden Surgical Anatomy dataset[10] compared individual annotations from three medical students to a fused annotation manually corrected by an expert, achieving a 75–97% mIoU consensus. In their dataset, unusable images (e.g., those where an organ was completely hidden or blurry) were excluded, whereas we annotated all images regardless of their quality.

Code availability

No code is available.

Received: 26 February 2024; Accepted: 5 February 2025;
Published online: 26 February 2025

References

1. Weiser, T. G. *et al.* An estimation of the global volume of surgery: a modelling strategy based on available data. *Lancet* **372**, 139–144 (2008).
2. Maier-Hein, L., Vedula, S.S., Speidel, S. et al. Surgical data science for next-generation interventions. *Nat Biomed Eng* **1**, 691–696 https://doi.org/10.1038/s41551-017-0132-7 (2017).
3. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3** (2016).
4. Neumuth, T. *et al.* Structured recording of intraoperative surgical workflows. **6145**, 54–65 https://doi.org/10.1117/12.653462 (2006).
5. Martin, J. A. *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery* **84**, 273–278 (1997).
6. Jin, A. *et al.* Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* **2018-January**, 691–699 (2018).
7. Shvets, A. A., Rakhlin, A., Kalinin, A. A. & Iglovikov, V. I. Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018* 624–628 https://doi.org/10.1109/ICMLA.2018.00100 (2019).
8. González, C., Bravo-Sánchez, L. & Arbelaez, P. ISINet: An Instance-Based Approach for Surgical Instrument Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **12263 LNCS**, 595–605 (2020).
9. Endovis Grand Challenge. https://endovissub-instrument.grand-challenge.org/.
10. Carstens, M. *et al.* the Dresden Surgical anatomy Dataset for abdominal Organ Segmentation in Surgical Data Science. https://doi.org/10.1038/s41597-022-01719-2.
11. Huaulmé, A. *et al.* MIcro-surgical anastomose workflow recognition challenge report. *Comput Methods Programs Biomed* **212**, 106452 (2021).
12. Twinanda, A. P. *et al.* EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Trans Med Imaging* **36**, 86–97 (2017).
13. CholecT50. https://github.com/CAMMA-public/cholect50/blob/master/docs/README-Downloads.md.
14. HeiChole. https://www.synapse.org/#!Synapse:syn25101790.
15. JIGSAWS. https://cirl.lcsr.jhu.edu/research/hmm/datasets/jigsaws_release/.
16. PETRAW. https://www.synapse.org/#!Synapse:syn25147789/wiki/.
17. Ríos, M. S. *et al.* Cholec80-CVS: An open dataset with an evaluation of Strasberg's critical view of safety for AI. *Sci Data* **10**, 194 (2023).
18. Hong, W.-Y. *et al.* CholecSeg8k: A Semantic Segmentation Dataset for Laparoscopic Cholecystectomy Based on Cholec80. (2020).
19. Derathé, A. *et al.* LapEx: LAParoscopic EXpertise. *Lapex: A new multimodal dataset for context recognition and practice assessment in laparoscopic surgery* https://doi.org/10.57745/1F0UBU. Recherche Data Gouv (2025).
20. Derathé, A. *et al.* Predicting the quality of surgical exposure using spatial and procedural features from laparoscopic videos. *Int J Comput Assist Radiol Surg* **15** (2020).
21. Derathé, A. *et al.* Explaining a model predicting quality of surgical practice: a first presentation to and review by clinical experts. *Int J Comput Assist Radiol Surg* **16** (2021).
22. Gibaud, B. *et al.* Toward a standard ontology of surgical process models. *Int J Comput Assist Radiol Surg* **13**, 1397–1408 (2018).
23. Annotate. https://b-com.com/en/bcom-surgery-workflow-toolbox-annotate.
24. Fouard, C., Deram, A., Keraval, Y. & Promayon, E. CamiTK: A Modular Framework Integrating Visualization, Image Processing and Biomechanical Modeling. *Studies in Mechanobiology, Tissue Engineering and Biomaterials* **11**, 323–354 (2012).
25. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* **20**, 37–46 (1960).
26. Long, J., Shelhamer, E. & Darrell, T. Fully Convolutional Networks for Semantic Segmentation.
27. Huaulmé, A. *et al.* Distinguishing surgical behavior by sequential pattern discovery. *J Biomed Inform* **67**, 34–41 (2017).
28. Nyangoh Timoh, K. *et al.* A systematic review of annotation for surgical process model analysis in minimally invasive surgery based on video. *Surg Endosc* **37**, 4298 (2023).
29. Nwoye, C. I. *et al.* Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Med Image Anal* **78**, 102433 (2022).

Acknowledgements

This work was supported by French state funds managed by the ANR within the Investissements d'Avenir Programme: Labex CAMI under reference ANR-11-LABX-0004 and MIAI @ Grenoble Alpes, under reference ANR-19-P3IA-0003. Data acquisition and sharing is a time-consuming process: we express our warmest thanks to the surgeons of the digestive surgery department of the Grenoble University Hospital involved in the study, the CIC-IT clinical research associates Emilie Chipon and Isabelle Boudry, and the IRT b< >com for the provision of the Surgery Workflow Toolbox [annotate] software used in this study.

Author contributions

A.D., F.R., P.J., A.M.G., B.G., S.V. conceived the database. A.D., F.R. were involved in the generation of the annotations. F.R., B.T., K.C., S.V. performed annotations for the for the inter-annotator validation studies, and more specifically, F.R. and B.T. performed the expert annotations of the quality of exposure score. A.D., S.G., S.V. were also involved in the drafting of the work.

Competing interests

SV and AMG hold a patent related to the computer-based analysis of surgical procedures (EP2197384B1, US9649169B2, WO2009027279A2). Other authors have no competing interests to declare.

Additional information

Correspondence and requests for materials should be addressed to S.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.