

# Endoscopic Vision Challenge 2022: Structured description of the challenge design

## CHALLENGE ORGANIZATION

### Title

Use the title to convey the essential information on the challenge mission.

Endoscopic Vision Challenge 2022

### Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

EndoVis

### Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

With the advent of artificial intelligence as key technology in modern medicine, surgical data science (SDS) promises to improve the quality and value of the particular domain of interventional healthcare through capturing, organization, analysis, and modeling of data, thus creating benefit for both patients and medical staff. Holistic SDS concepts span the topics of context-aware perception in and beyond the operating room, data interpretation and real-time assistance or decision support. At the same time, minimally invasive surgery using cameras to observe the internal anatomy has become the state-of-the-art approach to many surgical procedures. Contributing to the key aspect of perception, endoscopic vision thus constitutes a central component of SDS and computer-assisted interventions.

From this arises the necessity for high-quality common datasets that allow the scientific community to perform comparative benchmarking and validation of endoscopic vision algorithms. With EndoVis, we present you a large collection of publicly accessible datasets comprising various computer vision tasks (classification, segmentation, detection, localization,...) and subdisciplines ranging from laparoscopy to colonoscopy and surgical training. These datasets can be used for both de novo development as well as validation of methods. EndoVis organizes high-profile international challenges for the comparative validation of endoscopic vision algorithms that focus on different problems each year at MICCAI, thus representing a major driving force of advancements in the field. This year we propose 5 different sub-challenges under the umbrella of EndoVis.

### Challenge keywords

List the primary keywords that characterize the challenge.

Surgical Vision, Endoscopy, Classification, Detection, Segmentation

### Year

The challenge will take place in ...

2022

## **FURTHER INFORMATION FOR MICCAI ORGANIZERS**

### **Workshop**

If the challenge is part of a workshop, please indicate the workshop.

none

### **Duration**

How long does the challenge take?

Full day.

### **Expected number of participants**

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

60 (based on numbers from previous EndoVis challenges)

### **Publication and future plans**

Please indicate if you plan to coordinate a publication of the challenge results.

The joint publication will be coordinated by the particular sub-challenge organizers.

### **Space and hardware requirements**

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

depends on the specific sub-challenges, e.g. DREAM/synapse platform for example

## **TASK: SurgToolLoc - Endoscopic surgical tool localization by leveraging tool presence labels**

### **SUMMARY**

#### **Abstract**

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

The ability to detect and track surgical instruments in endoscopic video can enable numerous transformational interventions. These include assessing surgical performance, efficiencies, tool choreographies, tool use, and other operational or logistical aspects of OR resource planning among other applications. However, the annotations needed to train machine learning models to robustly identify and localize surgical tools are hard to obtain. Annotating bounding boxes around surgical instruments, frame-by-frame in video, is time consuming and needs to be repeated for a wide variety of surgeries to capture the range of possible surgical tools. Moreover, ongoing annotator training is needed to stay up to date with surgical instrument innovation. However, in robot-assisted surgery, timestamps of instrument installation and removal events associated with instrument names can be programmatically harvested from the system log, providing proxy annotations for tool presence in the video feed. In this challenge, we invite the surgical data science community to leverage automatically extracted tool presence data from instrument installation and removal events as weak labels to train machine learning models to detect and localize tools in video frames with bounding boxes. The ability to use only tool presence labels to localize tools would significantly reduce the annotation workload needed to train robust tool detection, localization, and tracking models.

#### **Keywords**

List the primary keywords that characterize the task.

weak learning; object detection; object localization;

### **ORGANIZATION**

#### **Organizers**

a) Provide information on the organizing team (names and affiliations).

Aneeq Zia (Intuitive Surgical), Xi Liu (Intuitive Surgical), Kiran Bhattacharyya (Intuitive Surgical), Ziheng Wang (Intuitive Surgical), Max Berniker (Intuitive Surgical), Anthony Jarc (Intuitive Surgical)

b) Provide information on the primary contact person.

Aneeq Zia (aneeq.zia@intusurg.com)

#### **Life cycle type**

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place. Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

**One time event with fixed submission deadline.**

### **Challenge venue and platform**

a) Report the event (e.g. conference) that is associated with the challenge (if any).

**MICCAI.**

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

**grand-challenge.org and synapse.org**

c) Provide the URL for the challenge website (if any).

**<https://endovis.grand-challenge.org/>, Sub-challenge site TBD**

### **Participation policies**

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

**Fully automatic.**

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

**Open to publicly available data including pre-trained nets. No privately prepared annotations are allowed to be used for training.**

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

**May participate but not eligible for awards and not listed in leaderboard.**

d) Define the award policy. In particular, provide details with respect to challenge prizes.

**3 monetary prizes for 1st, 2nd, and 3rd place. Exact amounts TBD**

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

**Top three performing methods will be announced publicly and posted on the website.**

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

**The organizers will publish a challenge paper within six months after the challenge. Following which, the participating teams can publish their own results from the challenge citing the challenge paper. Possibility of a**

combined publication amongst the participating teams/organization team will also be discussed after the challenge.

### **Submission method**

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Submission instructions will be posted to the website and sent via email. Results will be submitted via a docker container through Synapse. Teams will be required to identify bounding box locations and tool names for the test set as the output of their model. Specific directions on the format of this output will be provided during the challenge.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

The participants will not be allowed to evaluate their algorithms before submission - only one final submission per team.

### **Challenge schedule**

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Release of training cases in April 2022; registration closing Aug 2022; submission date September 2022; release of results at MICCAI 2022

### **Ethics approval**

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

An existing Western IRB will be used

### **Data usage agreement**

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC ND.

### **Code availability**

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

open source on challenge site. A sample of tool detection evaluation code can be found in 2021 SimSurgSkill challenge: <https://www.synapse.org/#>

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

open and private code submission will be accepted

### **Conflicts of interest**

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Sponsorship/funding will be done by Intuitive Surgical primarily. The organizers who are affiliated with Intuitive will perform testing.

## **MISSION OF THE CHALLENGE**

### **Field(s) of application**

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research.

### **Task category(ies)**

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

**Classification; Detection; Localization; Tracking**

### **Cohorts**

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Basic surgical tasks performed on porcine model by trainees during robotic surgical training. Tasks include suturing of different styles (1-hand, 2-hand, running), and dissection performed on various anatomy (uterine horn, rectal vein/artery, etc.). Tools include (but are not limited to) graspers, needle drivers, scissors, staplers, clip appliers, and energy instruments.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Basic surgical tasks performed on porcine model by trainees during basic robotic surgical training

### **Imaging modality(ies)**

Specify the imaging technique(s) applied in the challenge.

Single channel of endoscopic video

### **Context information**

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

The video clips will come with ground truth tool presence labels (in training data) and tool bounding boxes (in testing data)

b) ... to the patient in general (e.g. sex, medical history).

n/a

## Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

**The data will be acquired from basic tasks being performed on a porcine model using a da Vinci Xi or Si system**

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

**Prediction of tool bounding boxes utilizing only tool presence labels**

## Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy, Precision.

**Additional points: The assessment will be done using mean average precision over multiple intersection-over-union (IOU) values for bounding box detections**

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

**The Intuitive Data Recorder (IDR) will be used to capture video at 720p and 30fps from one channel of the endoscope on da Vinci Xi or Si system.**

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

n/a

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

**Data will be collected at Intuitive Surgical training labs**

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Experience of study participants will mostly be beginners (early in their learning curve) with a few experts (practicing surgeons) if possible

### **Training and test case characteristics**

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

A case in this challenge will comprise of a video of a surgical task being performed on a porcine model. For the training cases, only tool presence labels will be provided. For the testing set, bounding boxes for surgical tools will be annotated for evaluations.

b) State the total number of training, validation and test cases.

We will have 100+ cases for training and 50+ for testing. We will ensure variability in the dataset through the variety of tasks completed on the porcine model on different anatomy.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The numbers indicated were kept keeping in mind data collection technicalities and to provide enough data to the participants for developing meaningful models.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

We will try to ensure that the dataset has a balanced range of different tools within the training and testing set. We expect our dataset to have around 4-5 unique tool labels with unequal distribution across classes as some tools occur much more often than other (e.g. needle driver)

### **Annotation characteristics**

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

We will use a crowd (5+ annotators) to annotate tool bounding boxes. The annotations will not be redundant as bounding box annotations are not that subjective. This will also allow us to achieve the large scale of annotated dataset required for this challenge.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

For test set annotation, the crowd-sourced annotators were already trained and experienced in spatial annotation for surgical tools. Each frame will be annotated then reviewed by the annotation team to ensure quality. Bounding box labels will be placed around the surgical tools along with an object ID for object tracking. Additional tool classification label, such as left or right side will also be annotated.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

**Annotators will have significant experience in labelling bounding boxes for surgical tools.**

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

n/a

### **Data pre-processing method(s)**

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

**Raw video frames will not be altered**

### **Sources of error**

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

**Image annotation will only be needed for the test set. Main sources of error would include the bounding box not being 'tight' around the tool. Its hard to estimate the error quantitatively but we don't expect it to be more than 5%.**

b) In an analogous manner, describe and quantify other relevant sources of error.

The tool presence labels will be generated using the events stream from the da Vinci system. There is a possibility of a dropped event that can cause error in the training tool presence labels. However, we do not expect this to happen frequently.

## **ASSESSMENT METHODS**

### **Metric(s)**

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

**Mean average precision (mAP) for different intersection over union (IoU) values 0.50:0.05:0.95 will be used to assess performance of tool bounding box prediction algorithms**

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

This is a standard metric used for bounding box prediction algorithms (and is also the COCO primary challenge metric). By varying the IoU thresholds, this metric provides a more thorough evaluation of tool localization accuracy

### **Ranking method(s)**

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

The performance rank will be based on the rank of the evaluation metric (mAP IoU 0.5:0.05:0.95) - the higher the value of this metric, the higher the ranking of that team will be.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Missing results will be penalized and no score will be given for those cases

c) Justify why the described ranking scheme(s) was/were used.

Using the standard metric being used within the object detection research seems like the right way to rank teams. The metric tests the algorithms for detection of objects of different sizes which will be useful in differentiating high and low performing teams

### **Statistical analyses**

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Standard statistical methods to test for significance in results like t-test, ANOVA etc will be used

b) Justify why the described statistical method(s) was/were used.

The mentioned statistical methods are fairly standard and used extensively in literature to test for statistical significance of regression mod

### **Further analyses**

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

No further analysis will be performed

## **TASK: CholecTriplet2022 - Surgical Action Triplet Detection and Localization**

### **SUMMARY**

#### **Abstract**

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Formalizing surgical activities as triplets of the used instruments, actions performed, and target anatomies acted upon provides a better, comprehensive and fine-grained modeling of surgical activities. Automatic recognition of these triplet activities directly from surgical videos would facilitate the development of intra-operative decision support systems that are more helpful, especially for safety, in the operating room (OR). Our previous EndoVIS challenge, CholecTriplet2021, and existing works on surgical action triplet recognition tackles this as a multi-label classification of all possible <instrument, verb, target> combinations. For better clinical utility, real-time modeling of tool-tissue interaction will go beyond determining the presence of these action triplets, to also include estimating their locations in each video frame. Hence, this challenge extends our previous challenge on action triplet recognition to also include bounding box localization of the regions of action triplets. For the lack of spatially annotated dataset and to exploit large dataset without expensive and tedious annotation effort, this challenge focuses on the development and evaluation of weakly-supervised approaches for bounding box localization of the instruments performing the actions on CholecT50 dataset. Participants will develop and compete with algorithms to recognize action triplets as well as localize their region of likelihood in laparoscopic videos without the use of spatial annotation during training. For an in-depth analysis of the participating methods, the challenge will assess three subtasks in a single submission: (1) correct triplet recognition, (2) correct instrument's bounding box localization and (3) correct pairing of box-triplet detection. These three elements need to be produced by the same model or linked algorithms processed on a single run. This novel challenge investigates the state-of-the-art on surgical fine-grained activity detection, weak supervised learning of action location, and will strengthen this new promising research direction on surgical fine-grained activity modeling in computer-assisted surgery.

#### **Keywords**

List the primary keywords that characterize the task.

Surgical activity detection, surgical action detection, action triplet, tool-tissue interaction, CholecT50, deep learning, laparoscopic video, medical image analysis

### **ORGANIZATION**

#### **Organizers**

a) Provide information on the organizing team (names and affiliations).

Chinedu Nwoye, Deepak Alapatt, Aditya Murali, Saurav Sharma, Armine Vardazaryan, Nicolas Padoy (CAMMA Lab, University of Strasbourg & IHU Strasbourg)

b) Provide information on the primary contact person.

Chinedu Nwoye (nwoye@unistra.fr)

## Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place. Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

**One-time event with a fixed submission deadline.** If a certain number of submissions (at least 6) is reached at that time, a joint paper will be written and the dataset will be made public. Otherwise, the challenge will remain open until a sufficient number of submissions are reached.

## Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

<https://cholectriplet2022.grand-challenge.org/>

## Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

**Fully automatic.**

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

**Training is not restricted to the data provided for the challenge, publicly available data, including open, pre-trained networks, may be used.**

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

**May participate but not eligible for awards and not listed in leaderboard.**

d) Define the award policy. In particular, provide details with respect to challenge prizes.

**The winner of the sub-challenge will be awarded a prize, if at least 3 teams submit a result for the task. Then, depending on the number of teams in the sub-challenge, a maximum of 2 runner-ups can also be awarded a prize.**

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

The results of all teams will be first presented during the Endoscopic Vision Challenge meeting at MICCAI 2022. Afterwards, the information will be made available to all participating teams. The results will be made publicly available in the form of a joint publication.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

Depending on the number of submissions, top-N performing teams will co-author the joint publication. Every selected teams can submit at most 2 qualifying authors. The sub-challenge organizers determine the order of the authors in a joint challenge paper. Participants are allowed to publish their own results separately only after a publication of a joint challenge paper (expected by end of 2022). All participating teams/institutions will be acknowledged in the joint publication.

### **Submission method**

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

The participating team will submit a docker container with their codes that will be benchmarked internally by the organizers on the unseen test data. A valid docker submission is one that jointly produces outputs for the three sub-tasks. Submission instructions and a template docker in the required submission format with specific input/output protocol will be provided on the link: <https://cholectriplet2022.grand-challenge.org/submission>

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Evaluation metrics will be provided for participants to evaluate their algorithm on their chosen validation data split. During the validation phase, participants would conduct self-evaluation on the validation data for sanity check using their docker image. The docker image would not be submitted to the challenge organizers at the validation stage. Formats and guides for self-validation will be provided to the participants at the link: <https://cholectriplet2022.grand-challenge.org/validation>

Participants will be required to submit the output of their successful validated docker for confirmation that their development method follows the submission guideline. During the final submission stage, the participants will submit their final docker for evaluation. Only the last submission for each team before the deadline will be evaluated for the challenge ranking.

### **Challenge schedule**

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

April 1st: Release of first training dataset and the script for the evaluation metrics

May 1st: Launch of slack interaction forum and colab code blog

Jun 1st: Release of docker submission template

Jun 15th: Self-validation phase for testing docker containers

Jul 1st: Submission of self-validation report deadline

Jul 15th: Release of second training dataset

Sep 1st: Docker, Presentation & Report Submission deadline (11:59pm GMT)

Sep 18th/22th: Challenge Day

### Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

As the data consists of anonymized laparoscopic videos (i.e. no meta data identifying patient or surgical staff member is contained and all frames depicting something outside the abdominal cavity have been removed) no ethics approval is required.

### Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC.

**Additional comments:** During the course of the challenge, the data may only be used to prepare challenge submissions, no other uses are permitted. Once the data has been published after the challenge, it will be released, most likely under CC BY-NC.

### Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The script(s) for computing evaluation metrics will be made available to the participating teams at the link:  
<https://cholectriplet2022.grand-challenge.org/evaluations>

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

To participate in the challenge, each team must submit a docker image capable of producing results on the testing examples. A docker image template will not be shared by the organizers. Each team can choose to provide their source code, though they are not required to. Only a paper describing their method is required.

### **Conflicts of interest**

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Awards for the challenge will most likely be sponsored by IHU Strasbourg. Computational infrastructure will probably be provided by NVIDIA. We will provide details of the sponsorship two months before the conference. Only the organizers of the challenge will have access to the labels.

## **MISSION OF THE CHALLENGE**

### **Field(s) of application**

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research.

Additional points: Surgical Workflow Analysis

Surgical action recognition

### **Task category(ies)**

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

**Classification:** surgical action triplet recognition

**Localization:** surgical action triplet localization

**Detection:** surgical action triplet detection

### **Cohorts**

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

**Patients undergoing laparoscopic cholecystectomy.**

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

**Patients undergoing laparoscopic cholecystectomy at the University Hospital of Strasbourg, France.**

### **Imaging modality(ies)**

Specify the imaging technique(s) applied in the challenge.

**Laparoscopic video stream**

### **Context information**

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

**none**

b) ... to the patient in general (e.g. sex, medical history).

**none**

## Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

### Videos from laparoscopic cholecystectomies

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

### Localizing and recognizing surgical tool-tissue interactions in laparoscopic videos

## Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy, Precision.

Additional points: Action-triplet recognition: algorithm with high average precision.

Surgical instrument localization: algorithm with high average precision.

Action-triplet detection: algorithm with high average precision for joint recognition and localization.

From the predicted action triplet labels, the evaluation software will be able to extract and assess also the average precisions for the correct triplet's components (such as instrument, verb, target, instrument-verb, instrument-target, etc.) .

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Recordings from laparoscopes at University Hospital of Strasbourg, France.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Videos start at the first insertion of the laparoscope into the patient and stop with the last removal of the laparoscope. Frames that were recorded outside the patient's body have been censored (removed entirely).

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

**University Hospital of Strasbourg, France.**

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

**Surgeons.**

### **Training and test case characteristics**

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

One case is a single laparoscopic intervention. A video of the entire operation will be provided for each case. Each case has an action triplet annotation (100 classes, each frame has a 1D binary vector, each entry indicating if the corresponding action is being performed (1) or not (0)).

Additionally, each case is also provided with supplementary annotations for: Instruments, verbs and targets. All the supplementary annotations follow the same format as the triplet annotation case. Both training and testing cases are annotated with the same parameter. However, the test set will provide bounding box labels over the surgical instruments for spatial detection evaluation.

b) State the total number of training, validation and test cases.

45 training videos and 5 test videos. We provide the cases as videos because participants may want to also exploit the temporal information. On average, a video contains 2.08K frames. Participants can split the training videos into training and validation sets on their own volition.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The total number was determined by the annotation effort, the total number of test cases was chosen to maximize the ability to generalize and evaluate while maintaining a large enough training set. The test cases are chosen from videos not in the public domain.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

**The distribution of classes in the data is the real-world distribution.**

## Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Two surgeons annotated the data. The first surgeon annotated 40 videos and the second surgeon annotated 10 videos. Where there is ambiguity, label mediation is provided by a third clinician. The label ambiguity arises during class-label normalization, label super-classing, and clinical relevance rating of the triplet labels, which were used for the selection of the top 100 triplet classes.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

Surgeons were given a list of items to annotate and were also educated on the use of the annotation software. Additionally, there is an internal annotation document developed by the surgeons as a guide for the annotation (it is however not planned to make it public).

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Surgical experts involved in both clinical practice and research carried out the annotation. Their research field is image guided surgery and their field of practice is digestive and endocrine surgery. They have 16 and 20 years experience respectively.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Label disagreements between annotators are solved through mediation.

## Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Frames outside the abdominal cavity were removed. Participants are free to apply any further preprocessing if that improves their algorithm.

## Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Possible source error:

Disagreement in the labels super classing (estimated 1-3 subclasses)

Disagreement in the beginning and termination of action (est. 0.08 – 1.00 secs)

Disagreement in action similarity (estimated 1-5 frames per video)

Possibility of unrepresented class in training/testing data (max. 2/100)

Possibility of incorrect component class in training/testing data (max. 0.08%)

Possibility of bounding box border ambiguity (est. 0.5%)

b) In an analogous manner, describe and quantify other relevant sources of error.

none

## ASSESSMENT METHODS

### Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Action triplet detection will be assessed by mean average precision (mAP). Three inter-linked subtasks will be assessed as follows:

- (1) Triplet Recognition: Correct triplet ID classification is assessed by mean AP under precision-recall curve, as done in the previous CholecTriplet2021 challenge.
- (2) Instrument Localization: Correct bounding box localization of the instrument performing the action is assessed by detection mean AP. A detection is assigned a true positive (TP) if the degree of overlap between a predicted bounding box and the ground truth exceeds a certain threshold and the instrument ID is correct.
- (3) Triplet Detection: Correct pairing of the box-triplet. This metric will assess the correctness of the associated action triplet to every localized instrument. Here, a prediction is considered as TP if the predicted triplet ID is correct, assigned to the right instrument involved in the tool-tissue interaction, which must be localized at a minimum IoU with the ground-truth bounding box. Here, It is important to note that the bounding box detection would be evaluated with the triplet association ID, not the tool/organ ID. This means that the ID of a box over the same instrument-target can change based on the interaction between the two.

These three sub-tasks are considered as a single challenge: a complete submission must produce the three outputs in a single docker run to be considered for evaluation.

For the primary evaluation, the AP metric will be computed at a threshold of 0.5.

First, per category AP is computed over all frames in a video,

The video AP is obtained by averaging the category APs across all the test videos,

And final mAP is obtained as the overall average of the video category AP.

Additionally, we will assess the triplet recognition AP for the individual components as a secondary metric.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The mAP combines both precision and recall and is more useful than accuracy given uneven class distribution. The precision score gives a balanced confidence for the reliance of the algorithm in surgical procedure.

### Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

The performance metrics (triplet detection mAP [recognition, localization, recognition + localization]) produces a

scalar value which will be used to rank the model performance specifically in descending order for each subtask.

In the case of a tile:

For recognition: we would evaluate the recognition mAP of the components of the triplets namely: instrument-verb, instrument-target, instrument, verb, and target, in that order.

For localization: we would increasingly evaluate at higher IoU threshold until a decider is obtained.

b) Describe the method(s) used to manage submissions with missing results on test cases.

Only full submissions for the task will be considered.

c) Justify why the described ranking scheme(s) was/were used.

The ranking of the metrics is based on the clinical relevance and task difficulty.

### **Statistical analyses**

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

The level of significance in differences in metrics and model ranking stability will be analyzed using the Wilcoxon signed-rank test.

We will also examine the performance of each model on each test video.

b) Justify why the described statistical method(s) was/were used.

Wilcoxon signed-rank test is a non-parametric statistical hypothesis test to determine whether the median difference between two sets of observation is significant especially if the differences between pairs of data are non-normally distributed.

Also, analyzing the algorithm's performance on each test case and then on the whole will help to understand how the data label skewness affects the task learning and reveals the strength of each algorithm given a peculiar data distribution.

### **Further analyses**

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

For the joint challenge paper, combining algorithms strengths and/or model ensemble may be evaluated.

Furthermore, future direction of further works will be proposed following the algorithm's strengths and weaknesses.

## **TASK: SAR-RARP50 - Instrumentation segmentation and Action Recognition on robotic Radical Prostatectomy**

### **SUMMARY**

#### **Abstract**

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Surgical tool segmentation and action recognition are fundamental building blocks in many computer assisted intervention applications ranging from skills assessment to automated surgical procedures and improved surgery planning. Nowadays, learning-based approaches for these tasks usually outperform classical methods but rely on large amounts of high-quality annotated data. The translation on real surgical procedures is however limited by the data availability. Moreover, publicly available datasets are mostly focused on a single task and do not allow the development of multitask approaches.

To tackle the data scarcity issue, we release the first publicly available in-vivo dataset including 50 suturing segments extracted from Robotic Assisted Radical Prostatectomy (RARP) procedures, with labels for instrumentation segmentation and action recognition.

#### **Keywords**

List the primary keywords that characterize the task.

Surgical action recognition, Surgical instrument segmentation, Prostatectomy, Multitask learning

### **ORGANIZATION**

#### **Organizers**

a) Provide information on the organizing team (names and affiliations).

Beatrice Van Amsterdam, Dimitris Psychogios, Emanuele Colleoni, Danail Stoyanov: University College London

b) Provide information on the primary contact person.

Beatrice Van Amsterdam (beatrice.amsterdam.18@ucl.ac.uk)

#### **Life cycle type**

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place. Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

One time event with fixed submission deadline.

### Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

[synapse.org](https://synapse.org)

c) Provide the URL for the challenge website (if any).

<https://endovis.grand-challenge.org/>, Sub-challenge site TBD

### Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

**Participants will be allowed to use only the provided training data in addition to publicly available datasets (e.g. Imagenet, COCO, KINETICS, JIGSAWS, EndoVis datasets)**

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

**Data will not be released to members of our institute, thus we do not plan any restriction for such people.**

d) Define the award policy. In particular, provide details with respect to challenge prizes.

**We are currently discussing with our industrial partners about a possible money prize.**

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

**We plan to announce results from all participating teams. We are going to do further analysis for top-performing methods. Members of top-performing methods are going to be invited as authors to the challenge paper publication.**

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

**For each team invited to the publication, we will allow up to three authors.**

### Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

**All participating teams will need to provide a docker image to run inference on the test set and produce results in the specified submission format.**

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

**The organizers will not provide any validation service, participants can qualitatively assess the performance of their algorithms using the unlabeled test data**

### **Challenge schedule**

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

We plan two cycles of training data release: 1) half of the training data in the first half of June 2) the rest of the training data in the middle of July. The test data will become available in the middle of August. Participants will be able to join the challenge until one week before the deadline. The submission deadline is going to be on the second week of September. The workshop day will be the same as EndoVis and results will get published during the workshop-challenge day

### **Ethics approval**

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

**All data are collected under local ethics with patient consent**

### **Data usage agreement**

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC ND.

### **Code availability**

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

**We will provide the evaluation code with the release of the evaluation data.**

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

**Participants will containerize their inference scripts and weight in a docker. The organizers will not make participants' code or pre-trained models available.**

### **Conflicts of interest**

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

**We are discussing sponsorships and possible money prizes with our industry partners. The only party that will have access to the surgical instrumentation test labels is the annotation service we use to generate the ground truth data**

## **MISSION OF THE CHALLENGE**

### **Field(s) of application**

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

**Research.**

**Additional points: Intervention assistance, Intervention follow-up and Training**

### **Task category(ies)**

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

**Surgical instrumentation Segmentation and Action recognition**

### **Cohorts**

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

**Male patients with prostate cancer who undergo Robot Assisted Radical Prostatectomy**

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

**We believe that the challenge cohort coincides with the target cohort**

### **Imaging modality(ies)**

Specify the imaging technique(s) applied in the challenge.

**RGB images were collected using a stereo endoscope**

### **Context information**

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

**There will not be any additional information other than the target labels**

b) ... to the patient in general (e.g. sex, medical history).

**No patient information will be released along with the data**

## Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

**Pelvic anatomy is shown in laparoscopic video data**

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

**Submitted algorithms have to estimate segmentation masks for all the medical instruments( surgical tools, needles, threads, etc) within the scene and/or estimate temporal annotations of performed actions**

## Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy.

Additional points: The challenge will host three different sub-tasks, with each sub-task evaluated individually. Action recognition, algorithms should optimize the detection of all action classes. For the segmentation task, algorithms should optimize for segmentation accuracy. Lastly, for the multitask subchallenge, participants should optimize their multitask learning method to perform well in the objectives of the first two sub-tasks simultaneously.

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

**The video sequences were acquired using a DaVinci Si robot equipped with a stereo endoscope and data were captured using a dVLogger device.**

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

**Data logging started after the endoscope was within the patient abdomen, recording at 60 frames per second at 1080i resolution**

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

**All videos were acquired at University College Hospital at Westmoreland Street, London, UK**

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

**The participating surgeons have a different level of experience, ranging from surgical registrar to an expert consultant. The robot used among all the cases was a Da Vinci Si**

### **Training and test case characteristics**

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

**For the recognition task, each video corresponds to a different case and is annotated temporally with action labels.**

**For the segmentation task, each image corresponds to a case. Images come with annotation for nine different classes corresponding to tool parts, and other instrumentation (needles, threads, etc)**

b) State the total number of training, validation and test cases.

**For the action recognition task, we have 40 cases included in the training/validation set and 10 cases in the test set.**

**For the segmentation task, we currently have annotated 6 videos corresponding to around 1000 test cases**

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

**The total number of cases was restricted by data availability while we choose train/test proportions based on previous challenges.**

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

**For the segmentation dataset which includes 9 labeled classes, cases almost always include 5 of the classes and the rest 4 classes are sporadically present. For the action recognition task, action distribution is unbalanced. In both cases data unbalance is present because data came from real operations**

### **Annotation characteristics**

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

For the action recognition task, all the cases have been manually annotated. For the segmentation task, cases were annotated by a paid annotation service (<https://humansintheloop.org/>) and double-checked by the organizers. The annotation protocol is the same across validation, training, and test sets.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

For the segmentation task, annotators were given the following requirements: no overlapping classes, the annotation must follow the same annotation format as the 2018 EndoVis Robotic Scene Segmentation Challenge (<https://endovissub2018-roboticscenese Segmentation.grand-challenge.org/home/>). For the action recognition task, annotators were instructed to assign only one class per frame, choosing from a list of predefined actions. The action list was decided in collaboration with an expert surgeon.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

For the action recognition task, annotations were manually generated by an engineer with experience in surgical action recognition. The segmentation task cases were manually annotated by non-medical professional annotators and annotations were validated by the organizers.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

not applicable

### **Data pre-processing method(s)**

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Before annotating either task, videos were deinterlaced and stereo channels were synchronized in time along with kinematic information. The same processing was applied across all training, validation, and train splits.

### **Sources of error**

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

For action recognition, possible error sources may include imprecision in the gesture boundary or action ambiguities linked to non-standard surgical gestures and the particularities of each surgeon's technique. For segmentation, labels may be imprecise in cases where surgical instrumentation is not fully visible or in situations where vignetting or noise is present. For both tasks, false annotations due to human error may exist.

b) In an analogous manner, describe and quantify other relevant sources of error.

not applicable

### **ASSESSMENT METHODS**

## Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Segmentation algorithms will be assessed based on the square root of mean Intersection over Union (mIoU) score multiplied by Normalized Surface Dice (NSD). For the action recognition task, algorithms will be ranked based on the weighted average of per-class f1 score. For the multitask sub-challenge, participants will be ranked based on the multiplication of the scores achieved in each category by their multitask method.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

For action recognition task we chose the F1 score as it have been proven to be valid evaluation metrics in previous EndoVis challenges (e.g. SurgVisDom challenge). For the segmentation challenge we considered IoU, one of the most used metrics to evaluate such models. IoU does not account for true negatives, that would give an erroneous estimation of the prediction when the background/foreground ratio diverges from 1. On the other hand, NSD assess how the predicted segmentation contour is close to the ground truth one. We are taking the square root of the two scores to not suppress their contribution when computing the multitask metric that follows. For the multitask sub-challenge we believe multiplying the F1 score and the proposed segmentation metric, penalises methods that are fine-tuned to focus on one task over the other, thus promoting multitask learning.

## Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

For all three tasks, we compute metrics for each video and then we rank participants based on the mean performance of their algorithms across all videos. The winning submission for each challenge will be the one with the highest overall score

b) Describe the method(s) used to manage submissions with missing results on test cases.

Since participants are asked to submit a docker container to produce results in our test set, it shouldn't be possible for the submission to omit results for individual frames/videos. In the unlikely scenario that frames are omitted, those frames will be treated as misclassifications.

c) Justify why the described ranking scheme(s) was/were used.

The participants will be ranked based on the performance of their algorithms for each sub-task similar to previous EndoVis challenges

## Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

We plan to investigate the commonalities between the performance of different approaches through our dataset and try to identify patterns. The analysis will be done in python or MATLAB

b) Justify why the described statistical method(s) was/were used.

It is very difficult to find meaningful metrics and statistics to evaluate multitask approaches. For this reason, we will continue trying to develop a methodology to improve the quality of our analysis.

### **Further analyses**

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

In addition to Item 28-a, depending on the number of submissions we will merge the output of the top submissions and verify if results can be improved via ensembling. Common problems and biases of the submitted investigated and discussed.

## **TASK: SimCol-to-3D: Simulated Colonoscopy data for 3D (scene) reconstruction**

### **SUMMARY**

#### **Abstract**

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Colorectal cancer (CRC) is one of the most common cancers in the world. Polyp removal is an effective CRC screening method; however, navigating through the colon to detect polyps is extremely challenging. A navigation system can help the operator identify colon surfaces that have not been sufficiently screened for polyps or provide a training platform. However, reconstructing the colon from video is an unsolved task. Feature-based methods fail due to self-occlusion, reflective surfaces, lack of features, and deformations. Learning-based methods could be a robust alternative but require accurate scene depth and camera pose predictions.

We propose a challenge to further advances in depth and pose prediction during colonoscopy. We aim to provide a synthetic dataset with ground truth depth and camera pose. The data will be shared with participants of the challenge and at a later stage made publicly available. The aim of the challenge is to accurately predict depth and pose and generalize to the anatomies of different patients. We propose three sub-challenges: (i) depth prediction in simulated colonoscopy, (ii) camera pose estimation in simulated colonoscopy, and (iii) generalization to real video.

#### **Keywords**

List the primary keywords that characterize the task.

Colonoscopy, pose estimation, depth estimation

### **ORGANIZATION**

#### **Organizers**

a) Provide information on the organizing team (names and affiliations).

Anita Rau<sup>1</sup>, Sophia Bano<sup>1</sup>, Yueming Jin<sup>1</sup>, Danail Stoyanov<sup>1</sup>

<sup>1</sup>: Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and Department of Computer Science, University College London, London, UK

b) Provide information on the primary contact person.

Anita Rau (a.rau.16@ucl.ac.uk)

#### **Life cycle type**

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place. Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

**One time event with fixed submission deadline.**

### **Challenge venue and platform**

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

**grand-challenge.org or synapse.org**

c) Provide the URL for the challenge website (if any).

**Part of EndoVis challenge (<https://endovis.grand-challenge.org/>). Sub-challenge site TBD**

### **Participation policies**

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

**Fully automatic.**

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

**Publicly available data is allowed.**

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

**May participate but not eligible for awards and not listed in leaderboard.**

d) Define the award policy. In particular, provide details with respect to challenge prizes.

**There will be cash awards and certificates for the winner and first runner-up in each sub-task provided at least 3 teams submit the results. Cash awards will be subject to the availability of funds from the sponsors. Contacts will be made for taking sponsors onboard.**

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

**The results for all teams will be announced during the EndoVis Challenge at MICCAI2022. The results will also be made publicly available on the sub-challenge website. The submitted results will also be presented publicly in the joint publication.**

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All participating teams will submit a brief methodology report in MICCAI format (no more than 4 pages). Top N teams from each sub-task will be invited to be co-author of the joint publication. Each invited team can nominate at most 2 authors for the joint publication. The joint journal is intended to be published within 8 months of the challenge. The participating teams can publish their methods separately but only after the journal publication. The embargo time will be 10 months.

### Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Submission instructions will be posted on the sub-challenge website and will be sent to the registered participants via email. Each team must submit the running code as docker container via synapse and the results on the test set in the same format as the provided training ground truth annotations for each sub-task. This should also be accompanied by the methodology report.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

The participating team will not be allowed to evaluate their algorithms on the test data. We will not provide results or leaderboard to participants before the challenge day. Only the last submitted docker container, output files and report will be used for the evaluation.

### Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Challenge website and challenge registration opens: 1st April 2022

Training data release: 15th April 2022

Team registration open: 15th August 2022

Test data release: 15th August 2022

Submission deadline: 7th September 2022

Methodology report submission: 7th September 2022 (only valid docker submissions accompanied by output files and methodology report will be evaluated and included on the challenge day)

Spotlight presentation: Challenge day in MICCAI2022

Decision of challenge winners: Challenge day in MICCAI2022

Joint journal submission: March 2023

## Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

The data used is mostly synthetically generated. Part of the data for sub-task 3 comes from clinical colonoscopy which is fully anonymised, hence no ethics approval is required.

## Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC SA.

## Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Organisers evaluations script will be made available via github along with the detailed instructions on docker submission with dummy docker example.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Participating teams are encouraged (but not required), to provide their code as open access.

## Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Only the sub-challenge organisers will have access to the test data labels.

There is no conflict of interest.

We are currently looking for sponsors and we will update the organizers of MICCAI/EndoVis once it has been finalised.

## MISSION OF THE CHALLENGE

### Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research.

Additional points: Intervention assistance, diagnosis, screening, research.

### **Task category(ies)**

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Depth estimation, camera pose estimation.

### **Cohorts**

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

For the sub-task 1 and 2, target cohort is the same for the challenge cohort. For the sub-task 3, target cohort is real patient data.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The data is based on a publicly available colon mesh that was generated from a CT scan of a colon. Based on the

mesh a camera traversal is simulated in Unity. A virtual colonoscope follows a trajectory and RGB images and depth maps are rendered. The trajectory is randomly translated and offset leading to a diverse set of relative camera poses.

### Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Colonoscopy

### Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

No further information other than image data will be provided.

- For sub-task 1 (depth estimation), input images along with the depth map will be provided.
- For the sub-task 2, ground truth camera poses will be provided next to the RGB images.
- For the sub-task 3, no ground truth or additional information will be provided.

b) ... to the patient in general (e.g. sex, medical history).

none

### Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

In the final application the colon is imaged with a monocular endoscope. As ground truth for real colonoscopy is not readily obtainable, in this challenge we focus on synthetic colonoscopy video.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Depth and 3D camera pose estimation using sequential coloscopy images.

### Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy.

Additional points: For subtask 1 we assess the accuracy of predicted depth maps. For subtask 2 we assess the local

translation and rotation error of a prediction. For the third subtask we assess the accuracy of predicted poses and depth maps on real colonoscopic video.

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

For subtasks 1 and 2 the data is simulated in Unity. For subtask 3 data is acquired with a colonoscope during CRC screening.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

The data is based on the CT scans of different patients. The training data is based on a publicly available colon mesh. The test data is derived from two additional CT scans that are not currently publicly available.

For subtask 3 we provide short real colonoscopy video sequences from which it is possible to obtain sensible Structure-from-Motion pseudo labels.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

For subtasks 1 and 2 the data is generated in Unity by the organizers of this challenge. The real data is based on a public dataset.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

The data is simulated from the CT scans of 3 patients. The data generation process simulates a real colonoscope traversing the colon forward and backwards. The camera rotates mainly around the forward pointing axis but also along the horizontal and vertical axes. These camera movements replicate the screening method during real colonoscopy.

### Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

One case is a single endoscopic image. To compare a prediction to the reference a scale factor must be computed. That scale factor is based on an entire trajectory and applied to all cases in the same trajectory. So although a case

is an individual image, the results depend on the other images in the same trajectory.

b) State the total number of training, validation and test cases.

**More than 18,000 training and validation cases (combined). About 6,000 test cases.**

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

**The training data contains 15 trajectories through the same structure. The number of trajectories is chosen to provide sufficiently different relative camera poses. For the test dataset we provide three trajectories each from two CT scans. Three trajectories per mesh provide sufficient information about the generalizability of a method.**

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

**Different from existing synthetic data in colonoscopy, this dataset replicates real camera movement more realistically. It displays mainly forward and backwards movements with rotations around the centreline as induced by the screening technique during real colonoscopy.**

### **Annotation characteristics**

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

**N/A (dataset generated through simulation)**

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

**N/A**

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

**N/A**

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

**N/A**

### **Data pre-processing method(s)**

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

**N/A**

## Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

N/A (No image annotation to be performed)

b) In an analogous manner, describe and quantify other relevant sources of error.

N/a

## ASSESSMENT METHODS

### Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Subtask 1 will be assessed based on the mean square error between predicted and ground truth depth. Subtask 2 will be assessed based on the relative translation error (RTE) which reflects both local translation and rotation error. For both subtasks, we will assess accuracy up to scale, which means a single scale factor is computed based on the whole camera trajectory and applied to all predictions within the same trajectory. The third subtask will be assessed based on the RTE between prediction and the visually verified COLMAP (or other SfM method) result.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

We are using the standard metrics which are commonly used in the literature of depth [a, b] and 3D camera pose [b, c] estimations. For camera pose we focus on the relative translation error as it is more robust. It assesses the local translation and rotation errors as opposed to the more noisy global consistency of a camera trajectory. For the evaluation of real colonoscopy data, in the current clinical setup it is not possible to acquire ground truth, but visually verified COLMAP (structure from motion) can be obtained. Our recent work [under review] demonstrates that this can be used as pseudo-ground truth for computing the relative translation error.

[a] Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[b] Ozyoruk, Kutsev Bengisu, et al. "EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos: [c] Endo-SfMLearner." arXiv preprint arXiv:2006.16670 (2020)

[c] Sturm, Jürgen, et al. "A benchmark for the evaluation of RGB-D SLAM systems." 2012 IEEE/RSJ international conference on intelligent robots and systems. IEEE, 2012.

### Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

Sub-task 1 – depth estimation: The lowest value of the MSE over all test data will be used to declare the winner

and team rankings.

Sub-task 2- pose estimation: The lowest value for the median RTE over all test examples.

Sub-task 3 - pose estimation: The lowest value for the median RTE over all test examples.

b) Describe the method(s) used to manage submissions with missing results on test cases.

**Not allowed. Such submissions will be considered invalid.**

c) Justify why the described ranking scheme(s) was/were used.

n/a

### Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Rank stability analysis will be performed and reported with the challenge results. For the final joint challenge paper, we will assess the results of the participating teams for statistical significance. If the underlying test assumptions are fulfilled (depending on the properties of the submitted data which we can only assess after the challenge deadline) we will use a Student's-T-test, otherwise a Wilcoxon signed-rank test could be more appropriate.

b) Justify why the described statistical method(s) was/were used.

The T-test is one of the most used statistical hypothesis tests to assess if the means of two datasets are significantly different. However, the T-test assumes that means of the two sets are normally distributed. If that is not the case, the Wilcoxon signed-rank test could be more appropriate. The choice of test post challenge will be further justified in the analysis paper.

### Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

**Intended post challenge**

## **TASK: SurgT - a challenge for tissue tracking in surgery**

### **SUMMARY**

#### **Abstract**

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Visual tracking involves following a bounding-box throughout a video sequence. This is a crucial task in Computer-Assisted Interventions (CAI), with a range of applications including soft tissue deformation estimation, lesion tracking, augmented reality and robotic visual servoing. Medical applications require accurate trackers that are robust in challenging conditions prevalent in surgery. There are various types of challenges, such as dealing with specular highlights, deformation of anatomical structures, paucity of reliable visual features. Additionally, challenges due to occlusion; where tracked features are obfuscated due to smoke, blood, overlapping organs, and surgical instruments moving over the tracked bounding-box. Furthermore, challenges due to the changing illumination conditions, camera parameters, image resolution, artifacts from video compression, and other characteristics of camera technology. Hence prior to being utilized in real-world practice, tissue trackers need to be evaluated in a large and diverse dataset that captures all these challenging conditions. The general problem of tracking has been well studied within the computer vision community, focusing on natural scenes i.e. OBT [1] and VOT [2]. However, datasets for the same challenge in surgical scenes are still lacking. Hence, there is a dire need for a large publicly available dataset for benchmarking tissue trackers in surgery, for fueling development in this area and improving surgery. To address this, we propose the SurgT sub-challenge, a new first-of-a-kind collection of tools and datasets for benchmarking tissue trackers in surgery capturing the aforementioned challenges. We believe SurgT will become the standard for validating emerging new visual tracking algorithms in surgery.

For this sub-challenge, we will provide a large set of videos that were collected from surgical datasets available online. The SurgT dataset includes annotated videos, that can be used for validation/testing of the tracking methods, and non-annotated training data, which will allow self-supervised training of deep learning-based methods. We annotated a single bounding-box per video, both on the left and right stereo images. The bounding-box's centroid corresponds to a unique keypoint that is easily distinguishable throughout the surgical video. Each bounding-box is calculated using the following process: Firstly, two centroid positions are defined over the target keypoint, one for the left and one for the right stereo image. Given these two centroids, the disparity is then used to calculate the 3D position of the keypoint, relative to the left stereo camera. Then, a fixed size sphere around that 3D point is projected into the 2D images. This projection defines an ellipse which is used to define the width and height of the surrounding bounding-box. This also ensures that the bounding-box scales appropriately as the target 3D keypoint moves closer or further away from the camera. This was only possible to execute since all the videos of the SurgT dataset are stereo-rectified, with known intrinsic, distortion, and extrinsic camera parameters.

The participants will be given a set of input videos and a bounding-box on the first frame of each video, around the keypoint to be tracked. Proposed methods should track this bounding-box throughout the video sequence using self-supervised solutions, though they are free to use pre-trained models of their choice (supervised and/or self-supervised). Participants are not allowed to manually annotate the provided training data. The participants will be evaluated on three metrics: Accuracy, Precision and Robustness. These metrics will measure the performance of the bounding-box tracking across the entire video sequence. Where accuracy is defined as the average bounding-box

overlap, Precision is defined as the average MAE between the central pixel of the bounding-boxes and Robustness is the number of successfully tracked frames over the sequence.

[1] Y Wu, J Lim, MH Yang. Online object tracking: a benchmark. IEEE Conference on Computer Vision and Pattern Recognition. 2013.

[2] M Kristan, R Pflugfelder, A Leonardis, J Matas, F Porikli, L Cehovin, et al. The Visual Object Tracking VOT2013 Challenge Results. IEEE International Conference on Computer Vision Workshops. 2014.

## **Keywords**

List the primary keywords that characterize the task.

Tracking; Endoscopy; Laparoscopy; Robotic surgery; Computer Assisted Interventions

## **ORGANIZATION**

### **Organizers**

a) Provide information on the organizing team (names and affiliations).

Joao Cartucho, a.weld20@imperial.ac.uk, samyakh.tukra17@imperial.ac.uk, Stamatia Giannarou  
The Hamlyn Centre for Robotic Surgery, Imperial College London, London SW7 2AZ, UK

b) Provide information on the primary contact person.

Stamatia Giannarou (stamatia.giannarou@imperial.ac.uk)

### **Life cycle type**

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place. Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

One time event with fixed submission deadline.

### **Challenge venue and platform**

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

Part of EndoVis challenge (<https://endovis.grand-challenge.org/>). Sub-challenge site TBD

## Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

**Fully automatic.**

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Participants can also use any other publicly available datasets for training their methods excluding videos from: (a) Hamlyn endoscopic dataset; (b) SCARED dataset; (c) EndoVis 2017 dataset. The participants cannot use those datasets (a, b, and c) since some of those videos are used on the test set of the challenge, which is used for getting the final results that rank the participating teams. Similarly, the participants can also use pre-trained neural networks

or pre-existing methods if they are publicly available and were not trained already on the excluded datasets (a, b, c). More information about the excluded datasets can be found in the dataset section.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

**May participate but not eligible for awards and not listed in leaderboard.**

d) Define the award policy. In particular, provide details with respect to challenge prizes.

**We are currently in contact with a number of companies and dealing with sponsorship for the awards.**

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

**The methods and performance results of the participating teams will be made publicly available during the EndoVis Challenge at MICCAI2022.**

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

**All the results and methods will be part of a joint publication to be published at the end of the challenge. All the members of the teams will be made authors of this joint publication. The participating teams may publish their own results separately only after the joint publication being published. Similarly, SurgT data and its tools (SurgTlabeling and SurgT-benchmarking) can only be used for other publications once the joint challenge publication is published.**

## Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Participants must email the organizers:

1 - A link to a Docker container which includes their method, the code, the weights for deep learning-based methods, and the `SurgT\_benchmarking` tool. The SurgT benchmarking tool and the instructions on how to use it is provided here: [https://github.com/Cartucho/SurgT\\_benchmarking](https://github.com/Cartucho/SurgT_benchmarking)

2 - The final validation scores, calculated using `SurgT\_benchmarking` tool. For organizers to check that the participant's methods is running properly.

The organizers will simply (1) download and set-up each team's Docker, (2) add the test set videos and annotations to the `SurgT\_benchmarking` tool, and (3) run the benchmarking code get the final scores. Participants will not get access to the test set until the final results are obtained for all the teams.

Both (a) validation data and (b) a sample method - that runs on that validation data - are provided in our benchmarking tool. The participants will be instructed at the beginning of the challenge to set-up a docker image with the benchmarking tool. Then they will replace the sample method with their own method to guarantee that the results can be easily obtained at the end of the challenge.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

During the challenge participants can only assess their methods on the validation test set which is provided along with the benchmarking code: [https://github.com/Cartucho/SurgT\\_benchmarking](https://github.com/Cartucho/SurgT_benchmarking) . Results on the test set will only be obtained by the organizers after the final submission of the methods.

### Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

From 1st of March 2022 onwards - Participants can register online

15th of March 2022 - The training and validation datasets will be released (the test set remains hidden)

1st of September 2022 – Submission date

18th to 22nd of September 2022 - Results released at MICCAI's challenge day.

23rd of September 2022 – Test set released

After 23rd of September: all data and benchmarking tools available online.

### Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

Ethics approval is not necessary for the data since we are using publicly available data. We have had agreement from the dataset owners to re-use their videos and repurpose the dataset for our tracking challenge.

## Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY.

## Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The organisers' evaluation software is publicly available online at

[https://github.com/Cartucho/SurgT\\_benchmarking](https://github.com/Cartucho/SurgT_benchmarking) . And the code that used to annotate is also available online at: [https://github.com/Cartucho/SurgT\\_labelling](https://github.com/Cartucho/SurgT_labelling)

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Participants must share their code and models at the end of the challenge and make them available online.

## Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Only the organizers will have access to the test data. The participants and other researchers can only access the test data after the end of the challenge. The authors have no conflicts of interest to report.

## MISSION OF THE CHALLENGE

### Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research, Assistance, Surgery.

### **Task category(ies)**

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Tracking.

### **Cohorts**

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

**The target cohort will be tissue tracking in human in vivo during surgery.**

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

**The challenge cohort are animals and humans. General in-vivo/ex-vivo stereo footage.**

### **Imaging modality(ies)**

Specify the imaging technique(s) applied in the challenge.

**Stereo endoscopic RGB.**

### **Context information**

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

**Each video is provided along with an `info.yaml` and a `calibration.yaml` file. The info file includes a description of the type of surgery. The calibration file includes intrinsic, extrinsic and distortion camera parameters.**

b) ... to the patient in general (e.g. sex, medical history).

**none**

## Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

**The target is surgical tissue surfaces, in-vivo human, general endoscopic stereo RGB video data. The challenge does also include in-vivo and ex-vivo porcine footage.**

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

**General surgical tissue surfaces.**

## Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

**Accuracy, Precision, Robustness.**

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Our data is a collection of preexisting publicly available datasets. We have obtained consent to use the datasets for the purpose of this challenge. In total there are 6 videos from EndoVis2017, 9 videos from SCARED, and 7 videos from the Hamlyn dataset. The datasets were captured using multiple versions of da Vinci robots, including the most recent version - da Vinci Xi.

Please refer to the following links for full info:

- SCARED: <https://arxiv.org/abs/2101.01133>
- EndoVis2017: <https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/>
- Hamlyn: <http://hamlyn.doc.ic.ac.uk/vision/>

Note: we have had access to a private set of videos from the EndoVis2017 which are not available online and will be part of the test data.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Please refer to 21a.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

**The Hamlyn Centre, Intuitive Surgical Inc.**

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Please refer to 21a.

### **Training and test case characteristics**

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

One case corresponds to a specific surgery. Each case is composed of multiple 30 seconds videos, which capture different stages of that surgery. In the validation and test set, we have 4 videos per case. In the training set, all the videos per case are provided.

The videos in the validation and test set have been manually annotated. The videos in the training set were not manually annotated. The reason for this is that we propose a challenge that tries to find solutions around the issue of lack of labelled data in the medical domain, i.e the goal of the challenge is to promote the development of self-supervised trackers, or trackers that have been trained on other datasets/domains, including pre-trained models of their choice (supervised and/or self-supervised). Participants are not allowed to annotate the training data.

b) State the total number of training, validation and test cases.

Training: 13 cases

Validation: 3 cases

Test: 6 cases

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

We based our split on the commonly used 70/10/20 split. However, to accommodate for the video length difference and quality of the video and the diversity within the entire dataset, we manually selected the videos to be used for each portion.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

It is important to note that we are providing ground truth for validation as well as for test cases. The validation set will be available to the participants since we believe it is useful for there to be examples of tracking so that the participants have data to validate against whilst developing their algorithms. The validation ground truth should not be used for training. The training data should not be manually labelled.

## Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Each case was labeled using software `SurgT\_labelling` was used to perform this. Only the validation and test set were annotated. This tool can be found publicly available online at: [https://github.com/Cartucho/SurgT\\_labelling](https://github.com/Cartucho/SurgT_labelling). The data was labelled by 4 annotators. 2 of which labelled the data, and the other 2 reviewed the quality of the data.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

For the SurgT challenge, the annotators are never allowed to use interpolation, automation or external intervention to aid their labelling. Only manual labelling was allowed.

The provided instructions were the following: (1.) The annotator should watch the entire video and decide which keypoint will be labelled. We recommend the annotator to choose a keypoint that is easy to label throughout the video; (2) Then, the annotator should classify if the keypoint `is\_visible\_in\_both\_stereo` for all the frame-pairs of the video. (3) Then, in the image-pairs where the keypoint is visible, the annotator should label the keypoint while respecting: (a) The annotator should ensure that the keypoint is mapped accurately and corresponds to the same target in both stereo images; (b) The annotator should also look back at the previous frame in the video sequence to ensure temporal video consistency in labelling; (c) If the keypoint is difficult to label, or there were conflicting opinions between annotators, then the annotator should set `is\_difficult = True` for that image pair. (4) Finally, the annotations are reviewed by another annotator.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

This data was annotated by the 4 organizers of the challenge.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

We do not have multiple annotations.

## Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

The original videos were split into multiple 30 second videos.

## Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Human error, labelling the wrong pixel. The error would not be more than a few pixels. We assume the camera calibrations are correct. We assume that there was instantaneous left and right image capture, no synchronicity issues. Most of the data was captured using the da Vinci systems, therefore in the captured images, there is the presence of PAL artefacts.

b) In an analogous manner, describe and quantify other relevant sources of error.

No other error is expected.

## ASSESSMENT METHODS

### Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

We measure the performance of the algorithms using: Accuracy, Precision and Robustness.

Inspired by the metrics used in the OBT [1] and VOT [2] tracking challenges, we ranked the performance of the algorithms using an average between: (1) Accuracy, (2) Centroid Precision and (3) Robustness. For each frame-pair, where the keypoint is classified as "visible":

#### 1. Accuracy

Accuracy is the overlap between the ground-truth and the predicted bounding-boxes. The overlap is measured using the Intersection over Union (IoU), as this is the standard metric for detection tasks.

#### 2. Centroid Precision

Precision is calculated using the MAE between the central pixel of the ground-truth and the predicted bounding-boxed. We calculate the MAE both in 2D, average pixel distance, and in 3D, average mm distance.

#### 3. Robustness

Robustness is 1 if the Accuracy > 0.1, otherwise is 0.

The final accuracy, precision and robustness score are calculated by averaging the scores on video frames.

The full benchmarking code can be found at: [https://github.com/Cartucho/SurgT\\_benchmarking](https://github.com/Cartucho/SurgT_benchmarking)

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

We have chosen these metrics because of the success of the VOT and the OBT challenge. Which is the state-of-the-art benchmarking tool for trackers, please refer to 26a.

### Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

Similarly to the VOT challenge, we will rank the trackers by computing the EAO (Expected Average Overlap) score, which is calculated using the area under the EAO curve. The EAO curve is the average of the tracker's accuracy at each timestamp over all the videos that the tracker processed.

b) Describe the method(s) used to manage submissions with missing results on test cases.

We will run the final test case, this should not be a problem.

c) Justify why the described ranking scheme(s) was/were used.

These are the fairest metrics to analyze a tracker with.

### Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

We do not intend to do statistical analysis.

b) Justify why the described statistical method(s) was/were used.

NA

### Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

No further analyses will be performed.

### ADDITIONAL POINTS

#### References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.