Endoscopic Vision Challenge 2023: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Endoscopic Vision Challenge 2023

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 coloscopy and surgical training. These datasets can be used for both de novo development as well as validation of methods. EndoVis organizes highprofile 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 five 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 ...

2023

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: SIMS: Surgical Instrument Multi-Domain Segmentation Challenge

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.

Automatic detection and tracking of surgical tools in endoscopic video footage is a prerequisite for many computer- and robot-assisted surgery systems with a multitude of possible applications such as the assessment of surgical performance, workflow analysis, augmented reality overlays or cognitive surgical robots. Especially in the field of cognitive robotics, initial testing of prototypes may be performed in several stages including ex-vivo and in-vivo animal operations before in-human trials may take place. This poses a unique challenge in algorithm development for tool recognition, as both the amount and types of tools used in initial training stages and the anatomy of humans and animals may differ significantly. We present a challenge dataset containing segmentation masks of moveable instruments parts in both human and porcine data from laparoscopic and robotic cholecystectomies with the goal to solve instrument recognition in a multitude of different settings. Information on instrument type is provided in the form of multi instance bounding boxes. The challenge consists of four tasks: the two tasks segmentation of instruments parts with or without type classification will be performed with or without knowledge of the image domain (in-vivo laparoscopic human, in-vivo robotic human, ex-vivo laparoscopic porcine, ex-vivo robotic porcine). To allow for analysis of specific error sources, the dataset will also include information about image quality on a frame-wise level.

Keywords

List the primary keywords that characterize the task.

Instrument segmentation,Instrument classification, Laparoscopic image analysis, Surgical Data Science, Cognitive Surgical Robotics

ORGANIZATION

Organizers

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

NCT Dresden: Sebastian Bodenstedt, Alexander Jenke, Stefanie Speidel
Heidelberg University Hospital: Martin Wagner, Marie Daum, Ala Tabibian

b) Provide information on the primary contact person.

Sebastian Bodenstedt (Sebastian.Bodenstedt@nct-dresden.de)
Martin Wagner (Martin.Wagner@med.uni-heidelberg.de)

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

Repeated 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

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

https://www.synapse.org/sims2023

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.

We will provide several awards, cash awards will depend on the availability of sponsoring.

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.

Results of all teams will be first presented at the Endoscopic Vision Challenge meeting at MICCAI 2023. Afterwards, the information will be made available to all participating teams. The results will be made publically available in the form of a joined journal paper.

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 in the total number of submission, at least two authors of each team will be listed as authors in

alphabetical order on the joined challenge paper in addition to the organizers and data annotators. Before publication of the joined paper, no results may be 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.

Docker container on the Synapse platform. Link to submission instructions: https://www.synapse.org/sims2023

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.

Only the final submission for each team will be evaluated. To allow for sanity checks, the organizers will provide the participants results on data selected from the training dataset.

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 data (1st part): 1st of May 2023
Release of training data (2nd part): 1st of July 2023
Start of evaluation: 1st of September 2023
Submission deadline: 1st of October 2023
Challenge Day: Day of Endovis 2023
(8th or 12th of October 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).

As the data consists of anonymized laparoscopic videos, 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.

The script(s) for computing metrics and rankings will be made available with the second training data set.

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

To participate in the challenge, each team has to submit a Docker image capable of producing results on the testing examples. The Docker images will not be shared by the organizers. Participants are encouraged, but not required to make their code available as open source.

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.

As sponsoring is still to be determined, no information regarding conflicts of interest can be provided. Only the organizers and some members of their institutions will have access to the test case 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

Training, Assistance, Surgery, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation, Localization, Classification

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 or robotically assisted cholecystectomy or porcine gallbladders undergoing cholecystectomy during surgical training or research.

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 Heidelberg or one of its partner hospital (Salem), patients undergoing robotically assisted esophagectomy at Heidelberg University Hospital, ex vivo data obtained from a laparoscopic or robotic surgical training center at Heidelberg University Hospital with porcine organs.

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).

Segmentation masks of surgical instrument parts, frame-wise bounding boxes with information on instrument type, frame-wise tags on image quality

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.

Abdomen shown in laparoscopic or robotic video data, Porcine organs shown in laparoscopic or robotic training environment.

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.

Surgical instruments

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 algorithms are optimized for accuracy across multiple visual domains of minimally invasive cholecystectomy.

The challenge consists of four tasks: the two tasks segmentation of instruments parts with or without type classification will be performed with or without knowledge of the image domain. Teams may participate in one or more of the tasks.

Segmentation of instrument parts
Here we aim to find algorithms that can segment the different parts of surgical instruments with a high amount of accuracy measured by the average DICE coefficient and normalized Hausdorff distance over all classes
Segmentation of instrument parts with classification of instrument types Here we aim to find algorithms that can segment the different parts of surgical instruments with a high amount of accuracy measured by the average DICE coefficient and normalized Hausdorff distance over all classes and in addition can assign the segmented parts to an instrument instance and correctly determine the type of instrument with a high amount of accuracy as measure by the mean average precision (MAP)

The four domains are:
1) In-vivo human laparoscopic cholecystectomy

2) In-vivo human robotic cholecystectomy
3) Ex-vivo porcine laparoscopic cholecystectomy
4) Ex-vivo porcine robotic cholecystectomy

For the tasks with knowledge of the image domain, we will provide that knowledge as a input for the docker container. For the tasks without knowledge of the image domain, only the image will be provided to the docker container to test whether an algorithm rules all domains (a so called the-one-ring-algorithm).

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 varying types of laparoscopic cameras

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

Video collected during routine laparoscopic and robotically assisted surgeries at participating surgical centers, as well as recorded during surgical training exercises with ex vivo porcine organs.

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.

Heidelberg University Hospital, Heidelberg, Germany
Hospital Salem, Heidelberg, Germany

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 of human surgical cases (laparoscopic cholecystectomy, robotic esophagectomy). Ex vivo porcine surgeries performed by either medical students, surgical residents or surgical specialists.

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 video of a single minimally invasive gallbladder removal, either performed laparoscopically or robotically as an in-vivo human surgery or ex-vivo porcine training case. Per video, between 50 and 100 frames

will be sampled at equidistant intervals and annotated with segmentation masks of surgical instruments. To address the problem of data imbalance, the equidistantly sampled frames will be supplemented with manual selection and annotation of additional frames of less commonly used instrument categories until a minimum of 10 frames per category per domain is reached. The segmentation masks will be split into instrument parts (i.e. jaws, wrist, shaft) and multi-instance instrument type labels of at least 36 instruments in 8 categories are provided. Additionally, the dataset contains frame-wise tag based information on image quality (lighting conditions, distortion, blood/smoke). Test cases are identical to training cases, but neither raw data nor annotations will be provided for the test cases.

The dataset will contain 8 training cases and 2 test cases for each modality (in-vivo laparoscopic human, in-vivo robotic human, ex-vivo laparoscopic porcine, ex-vivo robotic porcine).
In the first part of the release 50% of the training data will be released.

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

500 images training and 50 Images test per Domain, i.e. 2000 images training and 200 Images test in total.

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 chosen due to annotation effort with a 80% / 20% split for training and test data.

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.

Test data:
No further characteristics.
The distribution of classes in the data is the real-world distribution.

Training data:
To address the problem of data imbalance, the equidistantly sampled frames will be supplemented with manual selection and annotation of additional frames of less commonly used instrument categories until a minimum of 10 frames per category per domain is reached.

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.

Annotation will be performed by a medical student. A surgical resident will function in a supervisory role. Additionally a subset of the annotations will be performed by another medical student out of a group of five to calculate inter-rater-variability as a human baseline.

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.

We started our annotation protocol based on Ross et al. (ROBUST-MIS) but with changes. Please find here our annotation protocol:

Instrument segmentation:
Instrument is defined as an elongated rigid object put into the patient and manipulated directly from outside the patient.
Non-rigid tubes, bandages, compresses, needles (are indirectly manipulated through another instrument), coagulation sponges and metal and plastic clips are excluded.
Each pixel can only belong to exactly one structure. Particularly, based on the matter (liquid or solid, smoke doesn't count) that can be seen first along the line of sight of the endoscope, the label would be established. As a consequence multiple contours are possible for a single instrument that is covered by another matter, instrument, blood or tissue, which crosses the instruments outer contours. This rule is also valid for transparent instruments.
Many medical instruments present holes. A hole is made up of pixels that do not belong to instrument itself but are either completely surrounded by pixels of the same instrument or are completely surrounded by pixels of one instrument and the margin of the image where we come to the conclusion, based on the video context, that the instrument would close the hole outside the image. Holes are not considered a part of the instrument and are excluded from the segmentation mask.
If there is a text overlay, this should be ignored, but Image overlays are not considered as a part of the instrument.
As laparoscopic instruments have no movable wrists, they are segmented into two parts, Head and Shaft, whilst DaVinci instruments are segmented into three parts, Head, Wrist and Shaft.

Image quality:
Image quality is assessed as tags frame-wise on a global level and includes the following characteristics:
Is the overall image quality good?
Is the image subjectively too bright?
Is the image subjectively too dark?
Is there dirt on the camera lens?

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.

The annotators are medical students who have studied medicine for at least one year. All students underwent annotational training and are supervised by a surgical resident.

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.

NA

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 training and test data consists of videos recorded from the endoscopic video feed during surgery and compressed using MPEG-4. Each video was preprocessed to remove
frames taken from outside the abdominal cavity, in order to ensure the anonymity of patient and surgical staff.

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.

Annotation errors include poor lighting conditions (over-/underexposure), motion artifacts, distortion or instruments covered by blood, smoke or tissue. Instruments that are only partially visible could be misclassified by raters, which is why the 'unknown instrument' category was established. Interrater annotator variability will be addressed as described above. We do not expect relevant differences between the training and test cases.

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

NA

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)

For segmentation: The DICE similarity coefficient and Normalized Surface Distance (NSD) will be used for ranking. For instrument type classification: The mean average precision (MAP) will be used as described for the Pascal VOC dataset.

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

The DICE coefficient (and the positive correlated IoU) is a metric used commonly in the segmentation and the surgical workflow community and is generally an accepted metric. For segmentation tasks the NSD complements the DICE coefficient as it examines contour while the DICE coefficient examines the volume overlap. MAP is also a commonly used metric, e.g. used in the Pascal VOC challenge.

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 DICE coefficient and Hausdorf distance will be computed for every class in each selected image in a video and then averaged over each video and then over all videos. We will then compute two ranks, one for each metric. The rank of each team will then be determined through its average rank. For the combination of the segmentation and classification challenge, we will compute the rank for each metric and compute the average rank to decide the leader board.

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

As each team will have to submit a docker image for evaluation, missing cases should not occur. If a class is not present in a case/image but as detected, the DICE coefficient for that class will be set to 0.

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

We decided to first average over each video to make sure that each case is weighted equally, even though the videos vary in length.

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.

Boot-strapping and a Wilcoxon signed rank test will be performed to determine the stability of the rankings and the significance of the differences in methods.

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

Bootstrapping was identified by Maier-Hein et al. as an appropriate tool to determine rank variability.

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.

NA

TASK: PitVis: Surgical workflow and instrument recognition in endonasal 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.

Motivation: The pituitary gland, found just off the base of the brain, is commonly known as “the master gland”, performing essential functions required for sustaining human life. Clinically relevant tumours that have grown on the pituitary gland have an estimated prevalence of 1 in 1000 of the population, and if left untreated can be life limiting. The “gold standard” treatment is endoscopic pituitary surgery, where the tumour is directly removed by entering through a nostril. This surgery is particularly challenging due to the small working space which limits both vision and instrument manoeuvrability, and thus can lead to poor surgical technique causing adverse outcomes for the patient. Computer-assisted intervention can help overcome these challenges by providing guidance for senior surgeons and operative staff during surgery, and for junior surgeons during training. Challenge:
• We propose a challenge to recognise the surgical steps and instruments during endoscopic pituitary surgery, and in doing so, aid with decision making during surgery. There are three sub-tasks for this challenge: (1) 27-steps multilabel recognition; (2) 23-instruments multi-label recognition; (3) 27-steps and 23-instruments multi-task classification. These steps and instruments were identified in an international consensus Delphi study (10.1007/s11102-021-01162-3).
Data:
• Participants will be provided with a training dataset of endoscopic pituitary surgery videos with both step and instrument ground-truth annotations provided. The evaluation on unseen test
data will be performed through docker submissions. The testing data will not be released to participants. This dataset will be the first large scale dataset of endoscopic pituitary surgery videos that will be made publicly available along with the post-challenge analysis publication.

Keywords

List the primary keywords that characterize the task.

Endoscopy; Pituitary; Surgery; Classification; Multi-task learning

ORGANIZATION

Organizers

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

Adrito Das [1], Yitong Zhang [1], Sophia Bano [1], Francisco Vasconcelos [1], Dimitris Psychogyios [1], Danyal Z. Khan [1,2], Hani J. Marcus [1,2], Danail Stoyanov [1]:

[1] Wellcome/EPSRC Centre for Interventional and Surgical
Sciences (WEISS), University College London, London, UK [2] UCL Queen Square Institute of Neurology, University College London, United Kingdom.

b) Provide information on the primary contact person.

Adrito Das: adrito.das.20@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).

https://endovis.grand-challenge.org/ sub-challenge site TBA

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.

We will offer certificates and cash awards to the top two teams. Cash awards will be subject to the availability of funds from the sponsors. Contacts will be made for taking sponsors on board.

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.

All docker submissions will be evaluated on an unseen test dataset. The test dataset will not be made available to the participants. This data will be acquired using the same procedure as the training dataset. The submitted results will be announced on the day of the challenge sorted in descending performance metric values. The results for all teams will be announced during the EndoVis Challenge at MICCAI2023. 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-authors (max 2 authors per team) of the joint publication. The joint journal is intended to be published within 10 months of the challenge. The participating teams can publish their methods separately but only after the joint journal publication. The embargo time will be 12 months. Submission with performances below a given simple baseline result will not be included in the final 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.

Submission instructions with a complete step-by-step docker submission instructions will be posted on the subchallenge website and will be sent to the registered participants via email. Each team must submit the running code as docker container via synapse. The submission 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.

Multiple submissions will be allowed and will be validated on a small subset of the training dataset to ensure correctness of the docker container. For the final evaluation on the unseen test dataset, only the last submitted container will be used.

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: 3rd April 2023
Training data release: 17th April 2023
Team registration open: 14th August 2023
Docker submission deadline: 11th September 2023
Methodology report submission: 11th September 2023
Challenge Day: Day of Endovis 2023
(8th or 12th of October 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).

All the data used in the training will be published for research and educational purpose after the challenge joint paper. The data will be made openly available via our affiliated institution data server. We already have the ethics in place for releasing fully anonymised data in this domain for research purposes only. This study has been registered with the governance committee of the National Hospital for Neurology and Neurosurgery, and all patients have provided written informed 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.

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 organisers will have access to the test data and 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

Training, Assistance, Surgery, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Classification.

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.

Target cohort is the same for the challenge cohort. It is intended that developed algorithms could be applied to

real clinical settings.

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

Clinically acquired in endoscopic pituitary videos.

Imaging modality(ies)

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

Pituitary surgery videos captured using a 4mm endonasal 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).

No further information other than image data 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.

Endoscopic pituitary videos using the transsphenoidal approach, via the sphenoid sinus, a form of skull-base surgery, where the pituitary tumour is exposed.

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.

Surgical navigation of pituitary tumour surgery.

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: • Mean-average of the F1-score and edit-score: (1) step recognition; (2) instrument recognition; (3) mean-averaged step and instrument recognition. • Edit score evaluates the classification accuracy of each segment penalising poor classification ordering (equation 12 http://colinlea.com/docs/pdf/2016_ICRA_CLea.pdf)

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).

Hopkins Telescope with AIDA storage system (Karl Storz Endoscopy, United Kingdom) for video recordings

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

The dataset captures real data from live surgeries, all from a single centre.

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.

National Hospital for Neurology and Neurosurgery, Queens Square, London, United Kingdom.

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).

Expert surgeons performed the surgery and acquired the videos. The dataset was acquired during real surgical procedures and was not specifically designed for this challenge.

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.

Each video (training/validation/testing) represents one video of endoscopic pituitary surgery. Ground truth annotations will be provided for all sub-tasks at 1 frame per second. Each video will be provided in full at 720p.

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

All three subtasks have the same training/validation/testing cases.
• Training: 25 videos sampled with ground-truth labels (at 1 frame per second). Each video is approximately 70 mins long.
• Validation: Participants can choose their own training-validation split from the training data.
• Testing: 25 videos sampled at 1 frame per second with ground-truth labels. Test data will not be released but will be evaluated using the submitted docker containers.

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

These numbers are chosen to balance a good trade-off in annotation effort while introducing sufficient visual diversity. Moreover, this procedure is not prevalent unlike other endoscopic procedures and only limited data is available. Moreover, not all recordings are suitable for annotation and algorithm development due to some having large portions of missing steps.

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.

None

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.

All videos (training/validation/testing) are uploaded to the TouchSurgery (https://www.touchsurgery.com/) platform for manual annotation. (1) Two expert surgeons annotated the 27-steps of each video, giving timestamps for each step. (2) An annotation company (anolytics, https://www.anolytics.ai/) followed by quality control from an expert surgeon and an academic researcher was used to annotate the 23-instruments of each video at 1 frame per second.

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.

All videos (training/validation/testing) were given the same annotation framework. (1) Annotators were instructed to familiarise themselves with the 27-steps before annotation. (2) Annotators were instructed to familiarise themselves with the 23-instruments before annotation.

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.

All videos (training/validation/testing) used the same annotators. Clinical experts: in neurosurgical training with academic backgrounds. Academic researcher: Third year of academic experience specifically working on endoscopic pituitary training. Annotation company: internal quality control with domain knowledge experts.

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.

All videos (training/validation/testing) used the same methods. Merging of annotations was not performed, but quality control between annotators to verify correct annotations was.

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.

All videos (training/validation/testing) use the same methods. Videos are compressed to 720p.

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.

All videos (training/validation/testing) use the same methods. Videos are compressed to 720p (mp4).

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

All videos (training/validation/testing) use the same methods. Not applicable. All annotators and experts thoroughly review the annotations before agreeing on the final annotation.

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)

• F1 score ensures both precision and recall are balanced and therefore the better use case for the algorithm within surgical practice. It is also a standard metric for recognition tasks.
• Edit score ensures a prediction is not rapidly changing, as this would not translate to clinical settings well.

Ensemble methods are permitted. Submitted dockers will be evaluated on a single 32 GB GPU (NVIDIA DGX-2 Tesla V100 Tensor Core), and a docker that fails to run within this GPU limit will be considered invalid. Additionally, a docker must run with a maximum interference runtime of half the length of each test video or it will be considered invalid.

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

F1 score ensures both precision and recall are balanced and therefore the better use case for the algorithm within surgical practice. It is also a standard metric for recognition tasks.
• Edit score ensures a prediction is not rapidly changing, as this would not translate to clinical settings well.

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.

(1) highest mean-averaged F1-score and edit-score.
(2) highest mean-averaged F1-score and edit-score.
(3) highest mean-averaged F1-score and edit-score.

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.

F1-score and edit-score are standard evaluation metrics used in video recognition and segmentation tasks.

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.

Intended post challenge.

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

Intended post challenge.

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.

Other metrics, including: AUROC (frame-level); balanced-accuracy (frame-level); sensitivity (frame-level); specificity (frame-level); confusion matrix (frame-level); and event ratio (video segmentation metric), will be provided in the final challenge write-up.

TASK: SurgToolLoc: Surgical tool localization and keypoint detection 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 or key points 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 localize tools and the corresponding key-points in video frames. 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), Conor Perreault, 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

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

https://endovis.grand-challenge.org/ sub-challenge site TBA

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.

No additional data 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.

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 lalso 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 grand challenge. 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.

There will be a preliminary testing phase where participants will be able to evaluate their algorithm containers on a smaller test set. For the final testing phase, teams will only be allowed 2 runs and the best run will be counted for the results.

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 March 2023;
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.

The evaluation container github repo will be made public.

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

It will be encouraged to have all participating teams' algorithm containers uploaded and made available publicly on github. However, we will accept private submissions as well

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.

Sposorship/funding will be done by Intuitive Surgical primarily. The organizers who are affiliated with Intuitive will have access to the test set 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.

Task category(ies)

State the task category(ies).

Examples:

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

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.

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/keypoints (in testing data).

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

NA

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 porcince 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 and key points 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)).

Precision, Accuracy.

Additional points: The assessment will be done using mean average precision over multiple intersection-overunion (IOU) values for bounding box detections and over different object keypoint similarity thresholds for keypoint detection. Both these metrics are standard for COCO dataset

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)).

NA

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, videos will be clipped to 30s each and only tool presence labels will be provided. For the testing set, the videos will be longer and of variable length and bounding boxes/tool keypoints 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. Each case can range from 1-5 hrs. For the final dataset, we will plan to divide the cases into 30s sub clips for training while keep the original video lengths for testing set

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 10+ 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+ annotaters) 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.

NA

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. Similarly, mean average precision (mAP) for different object keypoint similarity thresholds 0.50:0.05:0.95 will be used for keypoint detection 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 thresholds, this metric provides a more thorough evaluation of tool localization/keypoint detection 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 (e.g mAP IoU 0.5:0.05:0.95 for bounding box detection) - 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 prediction models.

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: Syn-ISS: Synthetic data for Instrument Segmentation 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.

Instrument segmentation and workflow analysis for surgical procedures has transformative applications for surgical care provided to patients and their care providers, i.e., the surgeons. Surgical data science research for computer-assisted interventions has seen a significant growth over the past decade with global collaboratives and multi-institutional datasets being developed. A common limitation noted by the surgical data science community is the size of datasets and the resources needed to generate training data at scale for building reliable and highperforming machine learning models. Beyond unsupervised and self-supervised approaches another solution within the broader machine learning community has been a growing volume of literature in the use of synthetic data (simulation) for training algorithms than can be applied to real world data. Synthetic data has multiple benefits like free groundtruth, scale, rare events, anatomical variations, etc. A first step towards proving the validity of using synthetic data for real world applications is to demonstrate the feasibility within the simulation world itself. Our proposed challenge is to train machine learning methods for instrument segmentation and activity recognition using synthetic data and test their performance within the synthetic data itself. That is, the challenge participants will be provided datasets from a virtual reality simulator for surgical tasks and procedures. Their submissions will be tested on datasets from the same virtual reality simulator tasks and procedures. The evidence generated from this challenge's results will inform future work using synthetic data for real world applications. The datasets made available from this challenge can accelerate research in this domain of transfer learning from simulation to the real world and provide a stepping stone for future challenges that are driving closer to the real world applications.

Keywords

List the primary keywords that characterize the task.

Workflow Analysis, Activity Recognition, Virtual Reality, Instrument Segmentation, Synthetic Data, Surgical Education, Surgical Data Science

ORGANIZATION

Organizers

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

Anand Malpani:Surgical Science, Kimberly Glock:Surgical Science

b) Provide information on the primary contact person.

Anand Malpani (anand.malpani@surgicalscience.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.

synapse.org

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

https://endovis.grand-challenge.org/ sub-challenge site TBA

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.

We are planning to award monetary prizes. Details are 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 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).

All participating teams that reveal their identity can nominate members of their team as co-authors for the challenge publication. The method description submitted by the participant will be used in challenge publication. Personal data of the participant will include their names, affiliation, and contact addresses. References used in the method's description may be published in the challenge results as well. Participating teams may publish their own results separately with an explicit allowance from the challenge organizers once the challenge publication has been accepted for 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.

Submission instructions will be posted to the website and sent via email. Results will be submitted via a docker container through the Synapse platform.

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.

Test data will not be released to participants. They may choose to evaluate their algorithms using the training 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

Publish challenge website (April 2023);
Registration opens (May 2023);
Training dataset released (June 2023);
Test dataset released (August 2023);
Submissions due (September 2023);
MICCAI 2023 challenge day (October 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).

Not applicable.

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 test dataset.

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

We will ask participants to submit their code in a docker container. The participants will be encourage to make their code publicly 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.

Surgical Science will be sponsorship/funding the challenge. The organizers who are affiliated with Surgical Science will have access to the test case labels and will conduct evaluation of submissions.

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

Training, Education, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Classification, Detection, Segmentation

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 and advanced surgical skills training using virtual reality performed by users of varying experience levels with the simulator.

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

The challenge cohort is the same as the target cohort.

Imaging modality(ies)

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

Laparoscopic 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).

Spatial object segmentation labels

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

NA

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.

Simulated laparoscopic video data. In case of basic skills the simulated scene could be a dry skills lab setting. In case of advanced skills the simulated scene could be the pelvic or abdominal cavity of humans.

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.

Location of the laparoscopic instruments in the surgical scene

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: Find instrument segmentation algorithm with high performance on metrics specified in Parameter 26.

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 hardware platform for the dataset is a virtual reality simulator manufactured by Surgical Science (challenge organizers). Video capture software is used to acquire the dataset.

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

Video images are recorded showing the simulated surgical scene. After the scene is recorded, it is then ready for the user to view on the simulator monitor device.

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.

Dataset is captured on a Surgical Science virtual reality simulator at the Surgical Science office.

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 experience level of study participants will range from beginners to experts.

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 is an image of the surgical simulated scene. Images come with annotation of the various objects of interest. The training and test cases are composed of the same set of information / data.

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

The dataset will comprise of 8000 or more cases. The approximate split of the data will be: training cases: 60%, validation cases: 20%, test cases: 20%.

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

The total number is an estimate based on previous datasets collected. A minimum of 10 users performing the task 10 times will achieve these numbers. The specific proportions are 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.

The dataset is limited by the realism provided by the simulation graphics and physics. The cases will be balanced for the different users as well as their experience with the simulators. set. The cases from the same user will not appear in both training and test sets.

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.

The annotations for the segmentation are machine generated through the virtual reality simulation. There is no need for human annotation. The simulator software knows what pixels in the image a particular object is rendered to. These provide highly accurate groundtruth for the challenge tasks.

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.

Not applicable

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.

Not applicable

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.

The image data for the segmentation task is provided as is, i.e., no pre-processing is involved in it.

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.

As the annotations are machine generated there is no concern for inter- and intra-annotator variability. There is a possibility of corner cases where the machine generated labels for the segmenation task are incorrect. Human verification of a subset of the dataset chosen at random can reduce the chances of such errors in the segmentation task.

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)

A typical image segmentation metric like Intersection over Union (IoU) or Dice coefficient will be used. Additionally, a metric that quantifies the extent of the mismatch between the prediction mask and groundtruth mask like the Hausdorff distance will be used. The multiple metrics will be combined to generate an overall rank as described in Item 27.

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

These metrics have been used extensively in the past challenges like SAR-RARP80, HeiSurf, PETRAW, Cholec-Triplet

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 algorithms will be ranked using an aggregated value over all test cases. The ranking of the algorithm will be made by ordering the metric values. The result for each case in the segmentation challenge would be an aggregation over the number of objects present in the case.

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

Since the participants will submit their scripts via docker most of the test cases should have a corresponding result. If the result is missing those cases will be considered misclassified.

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

This is based on previous year's 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 will compare the performance of the various submissions using statitical methods. The analysis will be done using Python or R. The specific test to be used will be determined later. It will be one of the typical tests like ANOVA, Mann Whitney U, etc.

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

These are standard tests to compare paired data, i.e., predictions for the same test cases by various algorithms

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.

NA

TASK: SurgRIPE: Surgical Robot Instrument Pose Estimation

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.

Robot-assisted Minimally Invasive Surgery (RMIS) has evolved significantly in the last decades thanks to the advances in Artificial Intelligence (AI) and surgical robotics such as the da Vinci platform, which provide surgical assistance through enhanced visualization and feedback control. Pose estimation of surgical instrument has become an important task in RMIS. Although external devices including depth camera, kinematic information, electromagnetic trackers can be used for accurate estimation, they are not practical in in-vivo surgery which is limited to the space and hardware setup. Some vision-based method will use external markers to provide extra visual feature for detection. A limitation of these methods is that the marker must be always visible in the camera's Field of View and is sensitive to the background variation and occlusion. In addition, every marker attachment require complicated calibration. In this case, vision-based markless tool tracking methods provide a practical and cost-effective approach to tool tracking without requiring any modifications on the hardware setup or the attachment of external markers.The pose estimation task has been well studied in the computer vision area, focusing on natural scenes i.e. LineMOD [1], YCB-Video [2]. However, datasets for the same problem in surgical scenes are still lacking due to some challenges in surgical scenarios. Unlike the large size of item in natural scene, the surgical instruments are too small to put on the marker board for ground truth acquisition as those in [1][2]. In addition, the high precision requirement limits the usage of depth camera or other external hardware for data collection. To address these problems, we propose a novel pipeline to provide accurate and consistent ground truth for surgical instrument pose estimation. Da Vinci Si endoscopic stereo camera is also used to provide highquality image.In this challenge, multiple instrument whose CAD model is publicly available online will be used for data collection. SurgTrack3D provides a large set of videos clips there were collected from surgical scenarios. The dataset includes the video automatically labeled with ground truth pose and segmentation mask, which can be easily used for validation and training. The segmentation mask is projected by the 3D model of the instrument automatically given the camera parameters. We use the CAD model provided by the Intuitive. If necessary, bounding box of each instrument can also be generated given the ground truth pose and CAD model. Participants can also do some sampling on the 3D model for better data processing. Additionally, occlusion items attached with marker will be used to generate occlusion in the test image. It is used to test the method's robustness under occlusion.The participants will be given the input image, segmentation mask and CAD model of instruments. The proposed methods need to estimation the pose of the instrument with different background and light conditions in the surgical scenes. Pre-trained models and external datasets are free to use. ADD metric, AUC-ADD curve, translation and rotation error will be evaluated but the ADD score will be the primary metric, which is the most common and comprehensive evaluation method of pose estimation task.

[1] Hinterstoisser et al.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes, ACCV 2012[2] Xiang et al.: PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes, RSS 2018

Keywords

List the primary keywords that characterize the task.

Pose Estimation;Tool Tracking; Endoscopy; Laparoscopy; Robotic surgery; Computer Assisted Interventions

ORGANIZATION

Organizers

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

Haozheng Xu(haozheng.xu19@imperial.ac.uk),
Chi Xu(chi.xu20@imperial.ac.uk),
Baoru Huang(baoru.huang18@imperial.ac.uk),
Stamatia Giannarou: stamatia.giannarou@imperial.ac.uk
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.

endovis.grand-challenge.org

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

https://endovis.grand-challenge.org/ sub-challenge site TBA

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.

The award will consist of a diploma. Currently, we do not have money prizes to offer, but we will try to find a sponsor until the challenge.

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.

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, SurgTrack3D data and its tools (labeling and 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. 2 - The final validation estimation result. For organizers to check that the participant's methods is running properly.

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.

A validation dataset will be released for evaluation. For final test, a subset with occlusion and a subset without occlusion will be tested.

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 2023 onwards - Participants can register online
15th of March 2023 - The training and validation datasets will be released (the test set remains hidden)
15th of September 2023 – Submission date
12nd of October 2023 – Results released (last day of MICCAI)
13rd of October 2023 – Test set released

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).

No alive human or animal tissue is involved in this study so thics approval is not 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.

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 evaluate metrics will be added along with the dataset itself. Common evaluation metrics are used including ADD(-S), translation error, rotation error, etc

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

Assistance, Surgery, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Tracking, Segmentation

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 organ phantoms.

Imaging modality(ies)

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

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).

CAD Model of instrument

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 RGB 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.

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, Robustness.

Additional points: High accurate surgical instrument pose estimation for endoscopic image with high generalizability and robustness to varation and occlusion.

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 stereo laparoscope of a da Vinci Si surgical system, EndoWrist Instruments with CAD model, including Large Needle Driver, Prograsp Forceps, etc.

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

A 3d-printed holder will be attached on the instruments for ground truth. The laparoscope take images of the instruments for pose estimation

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.

Imperial College London, Hamlyn Center

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).

da Vinci Si surgical system

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.

Training and test cases both represent a RGB image of one instrument. A case provides a raw RGB image with Ground Truth Instrument Pose, segmentation mask and 3D model.

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

20000 images for training, 2000 images for validation, 4000 images for test.

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

This dataset contains 5 instruments. Variation of light condition, brightness, background and occlusion will be added. Video Clip will be recorded for training. For generalisation test, test image will be captured spearately with high variation.

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.

Sample with occlusion will be added in the test dataset, therefore the number of test image is larger than validation dataset.

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 automatically labelled with a pattern in the image. Transormfation between the pattern the instrument was manually calculated.

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.

None

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.

All the data was annotated. Only 2 organizer of the challenge manually calculate the transformation between pattern and instrument.

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.

None

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 Image is undistorted with camera parameters we calibrated.

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.

The ground truth pose is acquired by the pattern attached on the instrument shaft. The transformation between the shaft and the pattern may vary if the attachment is changed. Due the the careful caliration, the error is less than 0.2 mm.

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)

Average Distance (ADD) distance, translation error and rotation error, Area under curve given ADD threshold

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

We choose ADD metric since it is the most common evaluation metric in the pose estimation task. Since the size of the object will affect the value of error, AUC curve can comprehensively evaluate the pose estimation with different size. Translation and rotation error is highly important in surgical tool tracking scenario.

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 rankings will be determined by the ADD score.

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 fair and popular metrics.

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.

Test dataset consists of two subset, one contains no occlusion and another one contains occlusion. The weight of the method performance under occlusion will be over weighted due to it's more challenging and important.

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

Refer to 27c.

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.