Endoscopic Vision Challenge 2024 (EndoVis24): Structured description of the challenge design

Remark: This challenge has been slightly modified. All changes are highlighted in red.

CHALLENGE ORGANIZATION

Title

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

Endoscopic Vision Challenge 2024 (EndoVis24)

Challenge acronym

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

EndoVis24

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. EndoVis (http://endovis.org/) organizes highprofile international challenges for the comparative validation of endoscopic vision algorithms that focus on different problems each year at MICCAI, comprising various computer vision tasks (classification, segmentation, detection, localization, etc) and subdisciplines ranging from laparoscopy to coloscopy and surgical training. It acts umbrella for several sub-challenges in this field, this year we propose 8 different sub-challenges within EndoVis.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Surgical Vision, Endoscopy, Classification, Detection, Segmentation

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE 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.

60-80 (based on the number of 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
normal conference infrastructure on the challenge day (beamer, loud speaker, ...)

TASK 1: FedSurg: Federated Learning for Surgical Vision

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.

This federated learning sub-challenge in Surgical Data Science addresses the urgent need to harness the wealth of sensitive surgical data while safeguarding patient privacy and adhering to stringent data governance regulations. As healthcare data is inherently private, centralized sharing raises significant privacy concerns. Decentralized learning methods offer solutions by enabling model training without exposing raw patient data, and instead sharing the model
weights.
In this challenge, we aim to evaluate different methods for federated learning in surgical video analysis based on a novel video dataset of laparoscopic appendectomy videos. In addition to addressing the critical issues of privacy and data governance in surgical data, this federated learning challenge recognizes the pressing need for the classification of laparoscopic appendectomy videos. Laparoscopic appendectomy is a common surgical procedure for removing the appendix, typically performed to treat appendicitis. The classification of these videos is vital for several reasons. Firstly, it allows for the categorization
of surgical techniques, aiding in the identification of best practices and the refinement of surgical skills. Secondly, accurate classification supports the development of automated systems for surgical assistance, providing real-time feedback to surgeons and enhancing surgical outcomes. Lastly, a comprehensive understanding of laparoscopic appendectomy videos contributes to the advancement of surgical education and training programs. By
evaluating different federated learning methods on this specific task, the challenge aims to not only address privacy concerns but also foster advancements in surgical video analysis, ultimately improving patient care and surgical practices.

Keywords

List the primary keywords that characterize the task.

Federated Learning, Laparoscopic Video Analysis, Laparoscopic Surgery, Appendicitis, Appendectomy

ORGANIZATION

Organizers

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

Max Kirchner, Alexander Jenke, Sebastian Bodenstedt, Stefanie Speidel, National Center for Tumor Diseases Dresden, Germany

Fiona Kolbinger, Purdue University, West Lafayette, IN, USA

Oliver Lester Saldanha, Jakob Nikolas Kather, Else Kröner Fresenius Center for Digital Health, Dresden, Germany

b) Provide information on the primary contact person.

Max Kirchner
max.kirchner@nct-dresden.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

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).

N/A

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).

TBA

Participation policies

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

Only automatic methods are allowed.

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.

Only the provided training data and publicly available data, including open,
pre-trained networks, may be used. Due to the federated learning nature of the
challenge, most training data will be kept private. Some examples will be
provided to develop and select the methods used in the federated learning part of the challenge. This provided data may not be used to pretrain the models used in the federated learning part of the challenge.

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.

Members of the organizing institutes may participate in the challenge, but are not eligible for awards.

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 EndoVis challenge at MICCAI 2024. Afterwards, the information will be made available to all participating teams. The results will be made publicly 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).

An embargo time is defined. All contributing members of each participating team will be listed on the joint challenge publication. 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.

Submitting a Docker container, or alternatively a Docker Compose file with additional code, on the Synapse platform, with submission instructions to be provided at a later date (TBA).

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 first working submission after a sanity check 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 of May 2024
Intention to submit:1st of August 2024
Deadline for docker submission (optional short extension if limited
participants): 14th of August 2024
Training phase: 15th of August 2024
Evaluation phase: Starting 30th of September 2024
Challenge Day: Day of Endovis 2024

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).

This study was prospectively reviewed and approved by the Institutional Review Board of the Technical University Dresden, Germany (approval number: BO-EK-332072022, approval date: August 4, 2022). The corresponding study was prospectively registered at the German Registry of Clinical Trials (DRKS, URL: https://drks.de/search/de/trial/DRKS00030874, registration ID DRKS00030874).

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: Data provided in the context of this challenge may not be shared by challenge participants and may not be used for commercial purposes.

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 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 training in a federated learning approach and 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 and members of the participating surgical centers 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

Diagnosis, Intervention assistance, Intervention follow-up, Prognosis, 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.

Human patients undergoing appendectomy for suspected appendicitis

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

Human patients undergoing appendectomy for suspected appendicitis

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).

frames from laparoscopic appendectomies, annotations regarding best view and inflammation level (lab data)

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

patients with indication for appendicitis

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.

Intra-abdominal imaging (laparoscopy)

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.

Classification of the stage of appendicitis in laparoscopic imaging from appendectomy recordings

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)).

Find the best classification algorithm to predict the stage of appendicitis of laparoscopic imaging from appendectomy videos. The algorithm should perform well on two tasks.
- Task 1: The algorithm should perform well on the data of a client which is not part of the training set. The aim is to test the generalization ability of the algorithm.
- Task 2: The provided algorithm should perform well on test data of each single client independently. This allows us to evaluate the adaptation ability of the
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).

Conventional handheld laparoscopes used for laparoscopic surgery

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

Full-length recording of the entire laparoscopic appendectomy, starting at insertion of the camera into the abdominal cavity and finishing at trocar removal from the abdomen)

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

- University Hospital Carl Gustav Carus Dresden, Department of Visceral, Thoracic and Vascular Surgery (Dresden, Germany),
- University Hospital Carl Gustav Carus Dresden, Department of Pediatric Surgery

(Dresden, Germany),
- Diakonissenkrankenhaus (Dresden, Germany),
- Städtisches Klinikum Dresden-Friedrichstadt (Dresden, Germany),
- Krankenhaus St. Joseph-Stift (Dresden, Germany),
- Asklepios-ASB Klinik Radeberg (Radeberg, Germany)
- [possibly: St. Elisabethen-Krankenhaus (Ravensburg, 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).

Surgical teams performing laparoscopic appendectomy

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 stack of images from laparoscopic
appendectomy video.
- Frames temporally centred around an optimal view of the appendix will be extracted using FFmpeg from the original laparoscopic videos.
- The underlying stage of inflammation is annotated and provided as label.
- The participants can either use the set of images or a single image as input.

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

The data collection process has not yet been completed. Currently, 20 to 30 videos per centre are available. It is expected that we will receive a total of 200 to 300 videos from all centres together. The exact number of videos available cannot be estimated in the current state of the data collection process.

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

One centre is reserved completely for testing to measure the generalization ability of the global model.
- Furthermore, 20 % of the data of each client is reserved for testing. The other 80 % are intended for training and validation. The teams can decide how many images they use for training and validation.
- Only 25 % of the training/validation set is published to avoid cheating; to make sure the model is trained in a federated learning setting.

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.

Class distribution will match the actual occurrence rate in clinic

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.

One expert annotator (surgeon independently carrying out laparoscopic
appendectomies) is determining best view and inflammation level, second observer verification

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.

Annotation protocol outlining the classification process in two languages, including both verbal and visual (example images) explanations. Link to annotation protocol:
- https://drive.google.com/file/d/1dzovHQ-Ss70ik1mkSoB5VMEUErAjfZHF/view?usp=sharing

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.

Surgeons independently carrying out laparoscopic appendectomies

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.

Frames centred around an optimal view of the appendix will be extracted using FFmpeg from original laparoscopic videos

Sources of error

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

Misclassification was counteracted by using a detailed annotation protocol with both example images and verbal explanations, as well as a custom graphical user interface for annotation. Therefore, we deem the risk of misclassification limited to borderline cases (i.e., cases between partial and total necrosis of the appendix). This potential source of error equally applies to all centres (i.e., training, validation, and test data).

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)

F1-Score, Expected Cost (EC)

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

- F1-Score: In multi-class classification tasks, the F1-score is commonly used as the primary evaluation metric due to its capability to balance precision and recall. This is important because avoiding classifying stages into the wrong stage and identifying all positive instances are crucial factors. The F1-score helps achieve this by calculating the harmonic mean of precision and recall.
- Expected Cost: It is essential to take ordinal monotonicity as well into account for this use case. That means that we penalize bigger misclassification more in comparison to smaller misclassification.

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 decision to use these metrics and their associated ranks is justified by their ability to capture the key aspects of the problem domain. Additionally, the use of ranks allows for a fairer comparison between teams, as it takes into account the relative performance of each team on each task.

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 training and evaluation, only a full working submission will be considered.
- If the docker container does not work on our infrastructure, we will inform the teams, and they have the chance to fix it.

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

- Task 1 (Generalization): The F1-Score and EC score will be computed for all test cases. We will then compute two ranks, one for each metric. The task rank of each team will then be determined through its average rank.
- Task 2 (Adaptation): The average and the lower percentile of the client's F1-Scores and ECs will be calculated for each participant. By combining both scores, we are able to reflect worst case performance of algorithms and the average performance on the dataset of each client. The task rank will be calculated based on the sum of the rank of each score.
- For the combination of both challenge tasks, we will compute the average rank to decide the leader board.

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. in Why rankings of biomedical image analysis competitions should be interpreted with care 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 2: HiSWA-RLLS: Hierarchical surgical workflow analysis for robotic left lateral sectionectomy

SUMMARY

Abstract

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

Surgical workflow analysis is a fundamental tasks for realizing the intelligent surgery. Understanding the workflow makes it possible to deliver surgeons better situational awareness and in-time suggestions. However, the workflow could be vary due to the flexible instrument options and stylish maneuvers of different expert, increasing the difficulty of recognition. The workflow analysis requires structural interpretations to disentangle the complex procedural information to precisely extract useful, discriminative information from image contexts. Surgical process modeling (SPM) has garnered considerable research interest as a widely used approach for surgical education and analysis protocols. In this challenge, the HiSWA-RLLS dataset curates 50 videos of 50 patients operated by 5 surgeons and annotates a hierarchical workflow, including 4 steps, 12 tasks, and 34 activities represented as the triplet of instruments, actions, and objects. The workflow annotation consists of 3 inter-relations across different annotation types and intra-relations in annotation labels of 6 annotation types. The dataset also contains under-effective annotations which indicate unnecessary stylish behaviors in the clinical environment. The first task is under-effective frame identification. The other task is multi-task surgical workflow analysis. The hierarchical annotations in HiSWA-RLLS are helpful to explore the relational information and surgical style information in the workflow analysis. The rich domain knowledge of surgical procedures provides the potential of implanting domain knowledge within the algorithms. This subset offers new possibilities to derive advanced AI algorithms based on new relational information and recognition tasks.

Keywords

List the primary keywords that characterize the task.

Multi-relation, Multi-task, Surgical workflow analysis

ORGANIZATION

Organizers

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

S.Kevin Zhou, Shang Zhao, Qiyuan Wang, Dai Sun, the School of Biomedical Engineering & Suzhou Institute for Advanced Research,Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), University of Science and Technology of China

b) Provide information on the primary contact person.

S.Kevin Zhou
skevinzhou@ustc.edu.cn

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).

N/A

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

grand-challeng.org

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

TBA

Participation policies

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

Only automatic methods are allowed.

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.

The data used to train algorithms is open to publicly available data including (open) pre-trained nets

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.

Members of the organizing institutes may participate in the challenge but are not eligible for awards

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

The teams of top performing methods will be awarded with a certificate.

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 top three performing methods will be announced publicly at the EndoVis challenge. Participating teams can choose whether the performance results will be made public.

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).

Participating team members qualifies as the authors of the challenge paper; the participating teams can publish their own results after the acceptance of the challenge paper.

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.

Regular submission: Docker container on the Synapse platform. Link to submission instruction will be posted on the webpage.

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.

It is allowed to evaluate algorithms before submitting final results but only the last run is counted for final evaluation. Although we will provide the metric implementation for evaluation, the final submitted results that are evaluated by our organizers will be recognized as the final record.

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 the training data: April 2024 - August 2024
Registration: August 2024
Submission: September 2024
Challenge Day: Day of Endovis 2024

Ethics approval

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

The videos in the challenge dataset has achieved ethics approval in the healthcare centers to be used for research purposes.

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 (Attribution-NonCommercial-ShareAlike)

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 template scripts will be provided.

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

Docker container on the Synapse platform. Template will be released on a github repo. Link to submission instruction and Docker configuration will be posted on the webpage.

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.

TBA

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

- Diagnosis
- Education

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

Research

Task category(ies)

State the task category(ies)

Examples:

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

Classification

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.

robotic left lateral sectionectomy on real patients

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

robotic left lateral sectionectomy on real patients

Imaging modality(ies)

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

Surgical video images

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 annotation files are provided for training data.

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

Patients include both males and females. The patients range in age from 28 to 72 years old.

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 endoscopic 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.

Surgical workflow status in endoscopic images

Assessment aim(s)

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

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

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

- Task 1: Recognition of under-effective frames in surgical workflow given a set of surgical images.
- Task 2: Recognition of hierarchical surgical workflow given a set of surgical images.

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).

Da Vinci Si Robotic Surgical Systems

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

Videos are captured from the streaming box which has an output resolution of 1920x1080, whereas the raw video resolution in Da Vinci Si is 1280x1024.

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

The Faculty of Hepatopancreatobiliary Surgery, the First Medical Center of Chinese PLA General Hospital

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 dataset is collected from the surgeries conducted by expert surgeons.

Training and test case characteristics

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

Examples:

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

Training and test cases both represent endoscopic images of RLLS. There are 3 auxiliary annotations of each triplet action component for effective frames.

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

Training samples are around 200k frames, testing samples are around 100k frames.

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

The dataset is split by experts because the task aims to develop workflow analysis algorithms that are robust to surgical styles across different experts. The RLLS surgical style is unity in diversity. We can fully isolate the stylish information in 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.

The class distribution exactly follows the real-world label distribution in RLLS. Each expert has his preference for selecting instruments and stylish surgical maneuvers. Hierarchical annotations follow the knowledge structure of RLLS surgery education, naturally involving the structural domain knowledge in the annotation.

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 dataset is annotated from 3 experts who have more than 5 years of clinical experience, the most experienced expert has more than 10 years of clinical experience.

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.

The healthcare institution follows the training materials to define the knowledge structure of annotations. The containment relation between step, task, and shall activity should be satisfied. For example, when we define the temporal range of a step, we treat the first frame of the first task and the last frame of the last task in this step as the beginning and the ending frame for the step. The same principle applies to the task and activity as well.

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.

There are 3 expert annotators who have more than 5-year clinical experience. Specifically, the annotator with the most expertise level has more than 10 years of clinical experience.

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.

When encountering annotation discrepancies on a temporal fragment, we first follow majority voting. The annotation is finalized by the most experienced expert annotator when a complete discrepancy occurs. Since RLLS is a standard robotic liver surgery, only few completely distinct labels appeared during the annotation process.

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.

Our challenge dataset provides raw data aims to give the participants a set of practical clinical procedures for workflow analysis. The image data for training are extracted from the surgical videos. All surgical images are extracted by sampling frames with a fixed sampling rate (5fps), and the resolution is resized to 256x320.

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.

Since the annotation is on videos, the start and end for each temporal segment may not be consistent. However, the difference variance is within seconds due to inter-annotator variability which is acceptable.

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

The foggy and highly specular regions in the image content are prone to influence the determination of annotators, which is a practical factor in real-world surgery. These kinds of frames may mislead experts to determine the surgical status.

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)

Accuracy for binary under-effective frame recognition. Mean average precision (mAP) for each individual annotation in task 2 (hiearchical workflow analysis). The overall evaluation for workflow analysis is taking the average of mAPs across the annotation hierarchy

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

The accuracy can represent the identification power of model in a binary classification task which is recognizing the under-effective frames from the surgical video. The mAP denotes the precision of identifying the label of a temporal surgical segment. This metric allows us to estimate the average precision of each category in annotations, which is significant to the dataset that contains biased distribution and subject diversity.

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.

In task 1, the ranking scheme for binary classification is accuracy since the quantities of under-effective frames and effective ones are close. For task two, it is required to evaluate all tasks including step, task, activity, and each component in activity. Since the number of categories for each task is different, the difficulty of tasks is varying. We just weigh all tasks the same so that the average mAP of all major tasks will be a feasible consideration for ranking the overall workflow performance.

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

The submissions without involving corresponding metric evaluation will not be considered for corresponding metric ranking.

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

The ranking will consider the value Accuracy in the task1. The ranking will consider the average of mAPs from each task for challenge task2 (workflow analysis)

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.

Bootstrapping for ranking uncertainty. This would avoid training and testing data has non-overlapped labels.

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

This allows the robust data partition to minimize the possibility of biased label distribution in the training and testing sets. And this will mix the surgical style across videos in the dataset partition.

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 3: PhaKIR: Surgical Procedure Phase Recognition, Keypoint Estimation, and Instrument Instance Segmentat

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.

Accurate and reliable recognition and localization of surgical instruments in endoscopic video recordings is the basis for a variety of applications in computer- and robot-assisted minimally invasive surgery (RAMIS), such as surgical training systems, surgical skill assessment, or autonomous endoscope guidance. The robust handling of real-world conditions such as varying illumination levels, blurred movement of the instruments and the camera, severe sudden bleeding that impairs the field of view, or even unexpected smoke development is an important prerequisite for such procedures. To process the information extracted from the endoscopic images in the best possible way, the inclusion of the context of the operation can be used as a promising possibility, which can be realized, for example, by knowing the current phase of an intervention.
In our subchallenge, we present a dataset for which three tasks are to be performed: Instance segmentation of the surgical instruments, keypoint estimation, and procedure phase recognition. The following annotations are available for this: pixel-accurate instance segmentations of surgical instruments together with their instrument types, of which a total of 20 categories are distinguished, coordinates of relevant instrument keypoints (instrument tip(s), shaft-tip transition, shaft), and a classification of the phases of an intervention into seven different phase categories. Our dataset consists of 13 individual real-world videos of human cholecystectomies ranging from 23 to 60 minutes in duration. The procedures were performed by experienced physicians, and the videos were recorded in three hospitals. In addition to the complete video sequences, we provide annotations in a one-frame-per-second time interval, resulting in approximately 30,000 annotated and 838,000 not annotated frames. In addition to existing datasets, our annotations provide instance segmentations of surgical instruments, relevant keypoints, and intervention phases in one dataset and thus comprehensively cover instrument localization and the context of the operation. We believe that providing our dataset and conducting our subchallenge will contribute to the exploration of new approaches in RAMIS, especially taking temporal information into account, and enrich the community in the field of instrument recognition and phase classification.

Keywords

List the primary keywords that characterize the task.

Endoscopic Vision, Surgical Instruments, Instance Segmentation, Keypoint Estimation, Phase Recognition

ORGANIZATION

Organizers

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

Tobias Rueckert, Christoph Palm, OTH Regensburg
Dirk Wilhelm, Hubertus Feußner, Daniel Rueckert, TU Munich

b) Provide information on the primary contact person.

Tobias Rueckert
tobias.rueckert@oth-regensburg.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

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).

N/A

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

Synapse.org and self-hosted individual challenge website

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

Synapse.org: In Preparation. Individual challenge website: phakir.re-mic.de

Participation policies

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

Only automatic methods are allowed.

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.

Only training data provided by the challenge organizers and publicly available data sets are allowed for training. In addition, networks that have been pre-trained on publicly available datasets may be used.

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

Members of the organizing institutes may participate, but are not eligible for awards.

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

There will be three tasks in the challenge (instance segmentation, keypoint estimation and phase recognition). The top three teams per task will be named at the MICCAI 2024 event, and each participating team will receive a certificate confirming their participation and indicating their rank compared to the other methods.

e) Define the policy for result announcement.

Examples:

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

The results of all teams will be presented at the MICCAI 2024 EndoVis challenge event. Subsequently, as there is no continuously updated leaderboard, the results of the top three teams for each task will be published on the challenge website and all results of all participants will be included in the joint challenge 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 contributing members of each participating team will be listed on the joint challenge publication. The authors of the submitted methods are not allowed to publish any results before the publication of the joint challenge paper. All participating teams will submit a brief methodology report.

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 self-hosted challenge website. The Submission instructions are in preparation, a link will be provided. All participating teams have to submit a brief 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.

We provide a platform for participants to verify the technical functionality of their submitted Docker containers. For the evaluation of the algorithms, only the last submitted version will be considered. An evaluation on the test data will only be performed once per team to prevent iterative adaptation to the test data.

Challenge schedule

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

- the release date(s) of the training cases (if any)
- the registration date/period

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

Challenge website and challenge registration opens: April 2024
Release of training data: May 2024
Release of docker submission guide/evaluation instructions: 1st of August 2024
Submission deadline and registration closing: 15th of September 2024 ; Methodology report deadline: 15th of September 2024
Challenge Day: Day of Endovis 2024

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).

Only anonymized data collected in an ethically approved research project approved by the local ethics committee of the Technical University of Munich (approval code 337/21 S-EB) is 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-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 evaluation docker container together with an evaluation script for calculating the metrics will be made available to participants on the challenge website.

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

Participants are encouraged to make the algorithms and their Docker containers publicly available. However, if this is not desired by individual participants or groups, private submissions will also be accepted.

Conflicts of interest

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

The challenge is not financially sponsored or supported by any company or entity. The data and the annotations are done within the DeepMIC project funded by the Bavarian Research Foundation with the cooperation partners OTH Regensburg, TU Munich and the company AKTORmed. The test data are only accessible to the challenge organizers and will not be made publicly available either during or after the end of the challenge.

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

Intervention assistance, Surgery, Research, Training

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction

- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Segmentation, Localization, Classification, 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.

Patients undergoing laparoscopic cholecystectomy during real surgical interventions.

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

Same as target cohort, with the restriction that the challenge data came from from the University Hospital rechts der Isar, Munich, Germany, the University Hospital Heidelberg, Heidelberg, Germany, and a smaller regional hospital near Munich, Munich, Germany.

Imaging modality(ies)

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

Monocular endoscopic RGB video recordings

Context information

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

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

Frame-wise instance segmentation masks for surgical instrument parts together with the types of the instruments, keypoints for surgical instruments (tip point(s), shaft-tip transition point, shaft-point), surgical phase annotations on a per-frame basis.

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 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.

Surgical instrument instance segmentations and keypoints. Phases of laparoscopic cholecystectomy.

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)).

The challenge consists of three tasks: Instance segmentation and specification of the type of surgical instruments, determination of different keypoints, and prediction of the surgical phase. Depending on the task different aims of the submitted methods are expected: For instance segmentation, the most accurate pixel-wise labeling of the surgical instruments, including the correct classification of the tool types is the objective. For keypoint determination, the localization of the different keypoints (instrument tip(s), shaft-tip transition, shaft) represents the target. Regarding surgical phase recognition, the correct classification of the phase of the intervention for each individual frame is demanded. As the data involves real-world operations on humans, robust prediction of all these aspects is another key requirement for the algorithms.

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 monocular endoscopic cameras.

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

The videos are from real-world human cholecystectomies from three surgical centers and were recorded during the operation

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

The data originates from the University Hospital rechts der Isar, Munich, Germany (9 Videos), the University Hospital Heidelberg, Heidelberg, Germany (2 Videos), and a smaller regional hospital near Munich, Munich, Germany (2 Videos).

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 cholecystectomies were performed by experienced surgeons with many years of professional experience.

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 corresponds to one frame of a video of a cholecystectomy. Regardless of whether the frame belongs to the training or test dataset, there are annotations for the instance segmentation of the surgical instruments and the type of tools, the coordinates of keypoints, and the surgical phase. The cases and annotations of the test data remain with the organizers and will not be released to the participants. One frame per second is annotated in each operation video and only these frames will be included in the evaluation. There are 13 videos in total, of which 9 come from one center and 2 each from the other two hospitals. To ensure a fair distribution, 6 videos from the first center and 1 video from each of the remaining centers are provided as training data.

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

According to the division of the videos into 8 training and 5 test videos, the number of annotated frames corresponds to approx. 18,500 training images and 11,500 test images.

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

The split is the result of an equal division of the videos from the three clinics into training and test datasets. The number of annotated individual images is calculated based on the length of each video.

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 annotation of frames with the same time interval, i.e. one frame per second, is intended to cover all scenarios and ensure realistic conditions.

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 are made by a team of three people, consisting of two medical students and a professional surgeon with many years of experience who acts as annotator and supervisor.

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.

At the beginning of the annotations an annotation protocol was developed in close cooperation with the medical staff, which is currently being adapted to the status of the annotations, i.e. number of videos, frames, instrument classes, etc., and will be made available to the participants of the challenge when the website is published. The most important annotation rules are as follows: The "Computer Vision Annotation Tool (CVAT)" is used for annotation. For instance segmentation, the surgical instruments were outlined with polygons, assigning each surgical tool to one of a total of 20 predefined classes. Only what can be seen visually was annotated, i.e. no assumptions were made for hidden instruments behind organs and tissue or similar. If there are holes in the tools, for example, in the tip, these are seen as part of the instrument and not explicitly omitted. For superimposed instruments, there is a segmentation mask for each tool, i.e. a pixel can belong to several tools, as is usual for instance segmentation tasks. A video is always annotated by one annotator, and to ensure the high quality of the annotations, each annotated video is checked by another member of the team in a crossover procedure where polygons or instrument classes are adjusted if necessary. The annotations of the keypoints are created automatically based on the segmentations and subsequently checked for correctness by the human annotators. For the classification of the surgical phases, a video is divided into 7 parts based on timestamps, with frames between two timestamps belonging to one phase.

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 annotations are made by a team of three people, consisting of two medical students and a professional surgeon with many years of experience who acts as annotator and supervisor.

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.

No merging of annotations is applied.

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 videos show human real-world cholecystectomies and are converted from .MOV to the .MP4 format. These videos will be made available to the participants of the challenge, together with a script to split the videos into individual frames, which can then be utilized by an algorithm. There will be no further pre-processing or manipulation of this raw data

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 were cross-corrected and supervised by a medical professional, no major sources of error are to be expected and inter-annotator variability should not have any effect. In scenarios with poor lighting scenarios, very fast movements of the instruments including recognizable movement artifacts or with a very heavily smeared camera lens, minimally inaccurate segmentation annotations may occur, but we estimate these to be insignificant.

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 the instance segmentation of the surgical instruments, three metrics are employed. For localization, the mean average precision (mAP) is applied, whereby the Mask Intersection over Union (IoU) is used as the localization criterion. The calculation of the mAP is analogous to that of the Common Objects in Context (COCO) dataset [1], i.e., the mAP is computed for IoU thresholds between 0.50 and 0.95 with an interval of 0.05 and the results are averaged for the final mAP [2]. As a per-class counting metric, the F1-score is applied, and the 95% Hausdorff-Distance (HD) serves as the boundary-based metric. For the assignment strategy of predictions to ground truth segmentations, the Hungarian Maximum Matching Algorithm is utilized.
The evaluation of the keypoint accuracy is analogous to the calculation of the COCO mAP, whereby the object keypoint similarity (OKS) is used instead of the mIoU [3]. The OKS is calculated using the Euclidean distance between a predicted and a ground truth point, which is passed through an unnormalized Gaussian distribution where the standard deviation corresponds to the square root of the size of the segmentation area multiplied by a per-keypoint constant. We use the tuned version of the OKS proposed by COCO, which is based on a per-keypoint standard deviation with respect to the object scale and an adjusted constant. A more detailed description of OKS and the tuned version is given in [3]. The classification of the surgical phases is evaluated using the Balanced Accuracy (BA) as the multi-class counting metric, the F1-score as the harmonic mean of precision of recall as the per-class counting metric, and the Area under the Receiver Operating Characteristic Curve (AUROC) as the threshold-based metric.

1: Lin, Tsung-Yi, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C. Lawrence Zitnick. "Microsoft COCO: Common Objects in Context." European Conference on Computer Vision (2014).
2: Common Objects in Context - Detection Evaluation. https://cocodataset.org/detection-eval. Accessed: 14 November 2023.
3: Common Objects in Context - Keypoint Evaluation. https://cocodataset.org/keypoints-eval. Accessed: 14 November 2023.

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

The metrics mAP, F1-score, and 95% HD for instance segmentation of surgical tools were selected based on the publication by Maier-Hein et al. [1]. Furthermore, these metrics are common and frequently used for instance segmentation and are widely accepted in the community as measures for the quality of an algorithm. The COCO variant of mAP also serves to ensure that algorithms whose predictions have a higher IoU benefit more from it and that a higher IoU has a greater impact on the final result. Analogously, the calculation of the mAP based on the OKS for the keypoint estimation task as a measure of the quality of an algorithm should lead to the most accurate possible determination of keypoint coordinates, with similar benefits as mentioned above. The metrics BA, F1-score, and AUROC for the classification of procedure phases were also selected based on [1], which represent widely used metrics for classification tasks and ensure a reliable statement about the quality of a method.
[1] Maier-Hein, L. et al. "Metrics reloaded: Pitfalls and recommendations for image analysis validation." arXiv. org 2206.01653 (2022).

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.

By determining the results per video and averaging them we give equal weighting to all videos, even if they are of different lengths. The same applies to the task of phase classification in videos regarding the varying number of frames in a phase.

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

As participants submit docker containers containing their algorithms, there should be a corresponding result for each test case. However, if a result is missing, the metrics for this case will be set to zero.

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

We calculate the above mentioned metrics for each class in a frame, and average the results across all classes to get a final result for the single frame. We carry out this procedure for all images in a video to get a per-video result and average the results over all videos in the test dataset. For the phase classification task, we additionally average the per-phase results over all phases in a video to ensure that an unbalanced number of frames between phases does not affect the overall result. Since our challenge consists of three different tasks, and participants do not necessarily have to take part in all tasks, we carry out a separate evaluation and obtain three result lists at the end. For each task, we calculate all mentioned metrics and average the corresponding ranks in order to determine the final task-specific rank.

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 apply statistical methods such as bootstrapping and the Wilcoxon signed rank test to determine the stability of the rankings and the significance of the differences between the submitted algorithms. The statistical analysis will be done with a self-coded python script, cross-checked by an experienced mathematician.

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

In [1], Maier-Hein et al. identify these methods as suitable for determining rank variability.
[1] Maier-Hein, L., Eisenmann, M., Reinke, A. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun 9, 5217 (2018). https://doi.org/10.1038/s41467-018-07619-7

Further analyses

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

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

In addition to the metrics mentioned above, the inference runtime of each algorithm is specified in the final paper, which has no influence on the ranking procedure.

TASK 4: SegSTRONG-C: Segmenting Surgical Tools Robustly On Non-adversarially Generated Corruptions

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.

Accurate segmentation of tools in robot-assisted surgery is a foundational aspect of machine perception, facilitating various downstream tasks, including augmented reality feedback. While existing feed-forward network-based methods exhibit excellent performance in the absence of corruption in test cases, the susceptibility to even minor corruptions can significantly impair the model's performance due to the inherent overfitting nature of such networks. This vulnerability becomes particularly consequential in surgical applications, where high-stakes decisions are commonplace. Prior efforts, such as benchmarking methods on ImageNet-C for image classification, have explored the robustness of models by introducing artificial noise to test images. The CaRTS approach introduces a novel pipeline designed to achieve robust segmentation of robot tools against realistic corruptions. To assess algorithm robustness against non-adversarial corruptions, we expand the dataset in CaRTS to address the challenge of robust robot tool segmentation against non-adversarial corruption. Our goal is to encourage algorithms to exhibit robustness to unforeseen yet plausible complications that may arise during surgery, such as smoke, over-bleeding, and low brightness. The training set comprises 14 mock endoscopic surgery sequences without corruption and corresponding binary segmentation masks where "1" represents robot tools and "0" represents tissue background. In the testing set, there are three sequences with non-adversarial corruptions (smoke, over-bleeding, and low brightness) and digital corruption (ImageNet-C). Additionally, three sequences with an alternative background serve as validation. Participants are challenged to train their algorithms solely on uncorrupted sequences while achieving high performance on corrupted ones for the binary robot tool segmentation task. Successfully achieving this non-adversarial robustness in this benchmark is paramount for the translation of research algorithms into real-world applications. It ensures that these algorithms can navigate and perform effectively in the face of unexpected but reasonable complexities encountered during surgical procedures.

Keywords

List the primary keywords that characterize the task.

Non-adversarial Robustness, Surgical Tool Segmentation, Robotics Surgery, Minimally Invasive Surgery, EndoVis.

ORGANIZATION

Organizers

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

Hao Ding, Mathias Unberath, ARCADE Lab, Johns Hopkins University.

b) Provide information on the primary contact person.

Hao Ding
hding15@jhu.edu
Mathias Unberath, mathias@jhu.edu

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).

N/A

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

self-hosted challenge website for data release and results submission.

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

segstrongc.cs.jhu.edu

Participation policies

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

Any method is allowed; We will divide methods into two categories, one with external data, and one without external data. We give rank within categories.

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.

We will provide training data on uncorrupted domain. External data are allowed but are required to be reported for dividing categories.

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.

Restriction: members with prior access to the data are not allowed to participate. Collaborators within the last two years, even without access, are not considered to get any reward.

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

We will name two winners for both categories. Certificates. We are actively working to find sponsors for prizes.

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 will be first presented at the EndoVis challenge at MICCAI 2024. Participants need to agree with the publication of their methods and results. We will make all methods that officially participated in the competition public. A paragraph for describing the methodology is required to be provided for 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).

(1) All members of the participating team are qualified as authors. (2) They are allowed to publish their own results separately but need to cite the challenge. (3) Embargo time: after the initial write-up of the challenge is on arxiv.

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.

We'll use docker container for the result submission and evaluation. Submission instructions will be provided in the evaluation code. the results will be sent to the participating team via email.

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.

We will provide a validation set on one domain and the test set will remain unpublished until the end of the challenge.

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 and validation: May 2024
Registrations deadline: 23th August 2024
Submission deadline: 30th August 2024. (4)
Methodology reports due: 27th September 2024
Challenge Day: Day of Endovis 2024

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. It is not human-subject data.

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; We will make an expanded version of the dataset including additional modalities and all corruptions available in CC BY at a later time point.

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 provide the evaluation code for the validation set on the link: https://github.com/arcadelab/SegStrongC.

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

We will not require teams to make their code 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.

ARCADE Lab Johns Hopkins University has access to the test case labels. The challenge is funded by Johns Hopkins Internal Funds and partly from Multi-scale Medical Robotics Center, CUHK.

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; Non-Adversarial Robustness.

Task category(ies)

State the task category(ies)

Examples:

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

Endoscopic images/videos of the Minimally Invasive Surgery.

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

Set of animal tissues

Imaging modality(ies)

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

Stereo RGB images.

Context information

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

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

Images are collected via dVRK. da Vinci research robot is involved as the robot tool part. The background tissues are all animal tissues.

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

no patients

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 are stereo RGB images collected via dVRK. da Vinci research robot is involved as the robot tool part. The background tissues are all animal tissues collected from local grocery store. The target is to benchmark the method for endoscopicrobot tool segmentation under non-adversarial corruptions.

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.

Algorithm for binary robot tool segmentation in the endoscopic scenario that are robust against non-adversarial corruptions like smoke, bleeding, and low brightness.

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)).

Segment robot tool from stereo images in mock endoscopic surgery and reduce the influence of the non-adversarial corruptions.

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).

Images are collected by the endoscopic camera manufactured by SCHÖLLY along with the image process unit manufactured by Ikegami.

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

Step1. Camera calibration and hand-eye calibration: we perform a camera calibration to get the intrinsics for the left and right cameras of the stereo and a hand-eye calibration to get the transformation from both cameras to the base frame of both Patient Side Manipulator (PSM) 1 and PSM2. Step2. Trajectory Generation: we use the teleoperation feature of da Vinci research kit to manipulate the PSMs to generate trajectories in free space and record the kinematics of the trajectory. Step3. Trajectory Replay and Recording: we replay the same trajectories and record the videos at 10 fps with the same robot configuration under different scenarios, (1) pure dark background samples for ground truth generation, (2) uncorrupted samples with animal tissue background and regular light. (3) non-adversarial corrupted samples: same animal tissue background with lower brightness, blood, and smoke corruption, as well as alternative background corruption.

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

The data are acquired in the robotorium of LCSR, Johns Hopkins University. The platform for collecting the dataset is the da Vinci research kit.

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

The data are acquired by surgical robotics experts who are familiar with operating the da Vinci robot via dVRK. The data is not collected during real surgery.

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 cases represent RGB images with two PSMs and animal tissue background. Test cases represent RGB images with two PSMs and animal tissue background under one of the non-adversarial corruptions. Training and test cases have annotation for binary segmentation masks for where "1" denotes PSMs and "0" denotes background.

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

we have 6600 cases for training, 1800 cases for validation, and 2400 cases for testing.

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

We collect 16 sequences from different robot and camera configurations, each sequence has 300 images for the left camera and 300 images for the right camera for uncorrupted and corrupted samples. All images are collected at a frame rate of 10 fps. We provide 11 sequences of uncorrupted samples for the training, 3 sequences of uncorrupted samples along with alternative background corruption for validation, and 2 sequences with all corrupted samples for the test.

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 provide 11 sequences of uncorrupted samples for the training, three sequences of uncorrupted samples along with alternative background corruption for validation, and two sequences with four kinds of corrupted samples (smoke, over-bleeding, low-brightness, and ImageNet-C) for the test.

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 collected images exclusively for ground-truth generation. These samples are generated by replaying the same trajectory of the corresponding samples. They are collected with an entirely dark background where the only visible area in the image is the PSMs. We perform the ground-truth generation in a semi-automatic procedure with the help of the segment anything model (SAM): Step 1. Prompt generation: we use a traditional background extraction algorithm to generate rough masks for the foreground PSMs. The rough masks are converted to the prompt points and bounding boxes for SAM. Step 2. Automatic generation: we input the prompt along with the image to generate a fine mask for the robot tool. The first two steps are fully automatic. Step 3. Failure case selection: human annotators are involved in examining the generated mask from SAM and selecting the failure cases. We had two annotators examining the same sequences and using the union of their selections as the set of failure cases. Step 4. Manual correction: the selected failure cases from step 3 are refined by the human annotators via using SAM with human prompt input until satisfactory results are achieved. If the annotator had three attempts but did not get satisfactory results, they would draw the contour to manually annotate this sample.

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 failure case selection: only pass the images where the masks and robot tools are perfectly matched. A perfect match means that the mask covers all visible parts of the robot tools and there is no extra mask covering pixels that does not belong to the robot tools. The border of the robot tools and masks should have no distinguishable offset with human perception. For manual annotation: Step 0. You will start with an empty mask for this image. Step 1. Make prompt. draw a bounding box slightly larger than the target area and then pin 2 points on the target. Step 2. Run model: Click "process" to feed the prompt to the SAM and generate the corresponding mask. Step 3. Judgment: If the generated mask combining the existing mask perfected covers the whole robot arm, click "success and finish" and jump to Step 5. If the generated area perfected covers part of the robot arm, click the "success and continue" button jump to step 1, and continue to annotate for the unfinished part. This mask will be combined with the existing mask via "and" operation. If the generated area fails to cover any part of the robot click "fail" and continue to reannotate this target. If fail three consecutive attempts, jump to step 4. Step 4. Drawing: draw the contour of the robot arm manually as the final result. When finished, click "finish". Step 5. Finish: save the results as the ground truth for this sample.

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.

Algorithms: for automatic prompt generation: we use the background subtraction function named BackgroundSubtractorMOG2 and BackgroundSubtractorKNN from the OpenCV Library. We combine the two algorithms by retaining the intersection of the result foreground mask. We use the extreme point of the generated mask to generate a bounding box and random sample points in the mask as prompt points. For automatic mask generation: we use SAM with ViT-L backbone as the predictor. We denoise the generated mask by erasing the connected components with an area less than a threshold. Annotator: the human annotators are experts who are familiar with the robot and segmentation task.

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 data will not be pre-processed.

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 possible error source might come from the (1) failure of the algorithm which will mostly be corrected by the human annotator. The successful rate (passing the human examination) of the automatically generated mask is 97.7%. (2) Drawing error: only very few cases (in total < 20 samples) need manual drawing, most of the failure cases are fixed by manual prompting. (3) Replay offset: the replayed trajectory might have offsets among each replay due to robot control ability.

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)

Dice Similarity Coefficient (DSC) and Normalized Surface Distance(NDS) averaged from different tolerances for the robot tool for multiple corruptions - low brightness, smoke, blood and ImageNet-C corruption.

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

The Dice Similarity Coefficient(DSC) is the standard metric for segmentation. Normalized Surface Distance(NDS) is complementary to DSC where DSC reveals performance more on the chunk area while NDS focuses on the boundary. We evaluate the performance only on the unseen domain to encourage participants to focus on the robustness of the algorithm.

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.

Higher DSC and NDS mean better segmentation in general. We care about the overall performance of the model on the corrupted domain thus we only test on these domains. Since we have more than one test domain, the summation of points given by rank is a reasonable way to have fair importance for each domain. Using rank and ignoring absolute performance difference is to avoid the result from one specific domain and metric dominating others. Since the goal for the participating team is to develop methods that are robust against all corruptions instead of algorithms achieving dominating performance on one domain but failing to generalize to others, The proposed overall ranking method better suits the goal of the challenge.

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

We treat the score of the missing results as zero.

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

We rank algorithms with total scores summed up over 4 test domains - ImageNet-C corruption, bleeding, smoke, and low brightness. Points are given by the rank for each test domain. Rank from 1 to 5 and get points from 5 to 1 if there are in total of 5 participants. Each Domain will have two ranks based on the DSC and NDS 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.

For ranking each domain, we will first sort the method based on the metrics over images. Then, we do a pairwise significance test for adjacent methods to test whether they have statistically significant differences in performance. If yes, we rank them according to the order, otherwise they have the same rank.

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

The mean value indicates the overall performance, and the pairwise significance test can justify whether one method has significant advantage over the other one.

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.

We will further analyses including dots- and boxplots to show the absolute performance of algorithms on the 4 test domains and 2 metrics. We will also do uncertainty analysis using bootstrapping. We will also do a failure case analysis.

TASK 5: STIR: Surgical Tissue Tracking Using the STIR (Surgical Tattoos in Infrared) Dataset

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.

This challenge is proposed to quantify efficient methods for tracking and reconstruction of surgical tissues in videos. This challenge will help enable robust quantification of tracking and mapping methods over many scenes. This is essential for verifying methods for use in image guidance and automation. Datasets that have been developed thus far either use rigid environments (not general), visible markers (visible to algorithms), or require annotators to label salient points in videos after collection (costly).
To address this gap, we would like to use our dataset, Surgical Tattoos in Infrared (STIR, https://dx.doi.org/10.21227/w8g4-g548) for a tissue tracking challenge. We have an additional approx. 50% of unpublished data from the STIR dataset that can be used for this purpose. STIR has labels that are persistent but invisible to visible spectrum algorithms and comprise hundreds of stereo video clips in both in vivo and ex vivo scenes with start and end points labelled in the IR spectrum.
Submissions including 2D tracking, 3D tracking, and dynamic neural radiance fields are welcomed. Submissions must take in a set of input locations along with a video to track and output a final location of points in each video. The challenge will be separated into real-time streaming methods, streaming methods (operate frame-by-frame), and offline methods (can take in the full video).

Keywords

List the primary keywords that characterize the task.

Surgical tracking, Deformable, Reconstruction

ORGANIZATION

Organizers

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

Adam Schmidt (UBC), Tim Salcudean (UBC)
Omid Mohareri (Intuitive), Simon DiMaio (Intuitive)

b) Provide information on the primary contact person.

Adam Schmidt
adamschmidt@ece.ubc.ca

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).

N/A

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

grand-challenge.org for the challenge, IEEE DataPort for dataset hosting

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

TBA

Participation policies

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

Only fully automatic methods are allowed

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

Participants can use any publicly available datasets, or publicly available (with weights) pre-trained neural networks for training their algorithms.

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.

Organization members may participate, but they will not be eligible for awards. They will be exempt from ranking, although listed on the leaderboard without numbering.

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

The challenge will include a financial prize which is dependent on sponsors. Awards will be given for each category with a distribution of 2:1:1 for each category (real-time streaming, streaming, and offline). A team can win multiple awards, and they can have one submission per category

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 results will be announced publicly at the EndoVis Challenge at MICCAI 2024

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 methods and results will be included in a joint publication, with all team members as authors. Participants may publish their challenge results only after the challenge is completed with citation of the STIR dataset.

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.

A link must be provided to a docker container which uses our benchmarking tool which will be available on github (see item 9). For submission, teams will upload their algorithm that they integrate with our evaluation pipeline. The metrics will be evaluated automatically by our evaluation script. An additional container will be provided for submissions in the real-time category which will benchmark the algorithm on a test sequence.

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.

Members can only assess their methods on the openly available validation set. This can be done through the grand challenge website with a live leaderboard. Docker instructions and an example script will be provided for both the results submission and the benchmarking submission.

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

Participant Registration, and initial website release: March 2024
Release validation set: March 15th
Submission: September 15th
Challenge Day: Day of Endovis 2024

Ethics approval

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

The data for STIR was collected on porcine labs in IACUC-approved experiments at Intuitive Surgical in an AAALAC approved facility. There is no human data involved.

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 evaluation code will be made publicly available on github on February 28th along with an example tracker.

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

Participants must share their code and models with the challenge organizers. Participants must agree to publish their code online after the challenge.

Conflicts of interest

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

Sponsorship will be provided by the organizers. Only the organizers will have access to the test data. Participants will not be able to access said data. The authors have no conflicts 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

Intervention planning, Intervention assistance, research

Task category(ies)

State the task category(ies)

Examples:

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

Tracking

Cohorts

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

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

The target cohort is humans with a focus on applications in tracking mapping and diagnostics in surgery. Additional details for any of the following answers can be found at: https://dx.doi.org/10.21227/w8g4-g548

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

The challenge cohort comprises in vivo and ex vivo stereo footage from porcine labs with many different tissue types.

Imaging modality(ies)

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

Stereo RGB videos from a da Vinci Xi endoscope are used for algorithm testing. Labelled infrared stereo pairs are used for quantification.

Context information

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

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

Each stereo video pair includes calibration, and image IR labels as described in https://dx.doi.org/10.21227/w8g4-g548

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 in the abdomen, with final applications being in minimally invasive surgery. The challenge cohort only includes in vivo and ex vivo animal tissues, while the target cohort is human

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.

The challenge has been designed to target performance on deformable surgical tissue of any type.

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)).

We would like to assess robustness and accuracy of algorithms under many different difficult scenes which include discontinuity of movement and tissue occlusion. We additionally will emphasize the importance of computational efficiency.

DATA SETS

Data source(s)

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

For the data provided, the stereo video is captured using a da Vinci Xi endoscope (see https://dx.doi.org/10.21227/w8g4-g548)

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

The data acquisition is detailed in depth at https://dx.doi.org/10.21227/w8g4-g548 A da Vinci Xi endoscope is used to collect calibrated stereo data along with infrared images for the ground truth labels.

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

Intuitive Surgical

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).

Refer to 21a for camera characteristics. As for the level of expertise, this data is collected by clinician engineers with hundreds of hours of experience using da Vinci robots.

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 a video clip with ground truth infrared start and end points. Training and test cases are the same in structure apart from being captured on different scenes. Test cases will only be usable through the grand challenge interface.

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

There are >1000 training cases, and >100 test cases. Validation will have approx. 20 cases.

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

The training and test cases are separated to provide a large amount for training while still leaving a reasonable size for robust evaluation in testing and validation. The total number of cases is limited by the number of labelling experiments we performed.

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 class distributions of training and test are the same since they are sampled equally. We would like to quantify the performance of methods in environments like those they are trained for.

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.

No annotators were used per-se. One person is used for filtering outlier data (IR frames that did not include labels/etc). Test cases will be verified again by a second user for validity.

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.

Detailed instructions and protocol are in the dataset paper: https://dx.doi.org/10.21227/w8g4-g548

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 data was filtered by Adam Schmidt who has worked with da Vinci systems for over 5 years.

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 image filtering methodology that processes Infrared labels into center labels is described in https://dx.doi.org/10.21227/w8g4-g548. Regarding preprocessing, videos in STIR are only preprocessed for camera calibration, and image rectification. The ground-truth segmentations are preprocessed by using OpenCV to locate segments. Further information is available in the dataset methodology publication, here:

https://arxiv.org/abs/2309.16782

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 primary possible error is bulk tissue movement between the IR frame capture and visible light start, as mentioned in the paper at https://dx.doi.org/10.21227/w8g4-g548 . This possible error is the same for each split and should be <1mm.

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)

We will use <delta_avg^x as is introduced in the TAP-Vid Benchmark [3] which is used for evaluation of many modern point trackers [2, 4, 5]. This is chosen rather than multi-object-tracking metrics [1] which focus on multiple object regions. <delta_avg^x measures the fraction of points closer than a specified threshold distance and is averaged over multiple thresholds. Similar to popular methods [2, 3, 4, 5], we choose thresholds of 4, 8, 16, 32, and 64 pixels.

Median trajectory error [5] and mean trajectory error will also be reported, but not used for ranking. 2D tracking methods will be evaluated in pixels, and 3D will be in millimetres.

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

These were chosen as we are interested in physical accuracy for tissue tracking, so metrics such as IOU or association are not useful for points. We desire metrics that are commonly used and standard for point tracking. <delta_avg^x has become the most used metric for this purpose from our survey of the point tracking literature.

[1] J. Luiten et al., HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking, Int J Comput Vis, vol. 129, no. 2, pp. 548-578, Feb. 2021, doi: 10.1007/s11263-020-01375-2.
[2] M. Neoral, J. Serych, and J. Matas, MFT: Long-Term Tracking of Every Pixel, presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 6837-6847. Accessed: Jan. 09, 2024. [Online]. Available: https://openaccess.thecvf.com/content/WACV2024/html/Neoral_MFT_Long-Term_Tracking_of_Every_Pixel_WACV_2024_paper.html
[3] C. Doersch et al., TAP-Vid: A Benchmark for Tracking Any Point in a Video, in Advances in Neural Information

Processing Systems, Dec. 2022, pp. 13610-13626. Accessed: Jan. 09, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/58168e8a92994655d6da3939e7cc0918-Abstract-Datasets_and_Benchmarks.html
[4] Q. Wang et al., Tracking Everything Everywhere All at Once, presented at the Proceedings of the IEEE International Conference on Computer Vision, 2023.
[5] Y. Zheng, A. W. Harley, B. Shen, G. Wetzstein, and L. J. Guibas, PointOdyssey: A Large-Scale Synthetic Dataset for Long-Term Point Tracking, presented at the Proceedings of the IEEE/CVF International

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 ranking by the metric <delta_avg^x is motivated in 26(a). As for the reason for separating methods by real-time, streaming, and offline, we do this to motivate clinical applications. Some applications may desire the highest accuracy albeit slow algorithm, while others need real-time performance. The less efficient streaming methods would still be useful for methods such as SLAM in which separate, but less efficient, threads are used in the mapping pipeline.

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

Submissions with missing results will be excluded from evaluation.

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

Algorithms will be ranked based on their performance on <delta_avg^x. There will be separate rankings for best 2D and 3D algorithms. Aggregation will be performed over all points over all images. That means a video clip with 15 points will contribute 50% more to the metric than one with 10 points.

Statistical analyses

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

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

We plan to perform a statistical analysis to determine a level of uncertainty of metrics such as the median trajectory error and <delta_avg^x by bootstrapping the data samples.

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

Bootstrapping provides a simple way to estimate variance of population level statistics.

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 6: OSS: Open Suturing Skills 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.

Efficient and precise surgical skills are essential in ensuring positive patient outcomes. Whereas machine learning-based surgical skill assessment is gaining traction for minimally invasive techniques, this cannot be said for open surgery skills. Open surgery generally has more degrees of freedom when compared to minimally invasive surgery, making it more difficult to interpret. By continuously providing real-time, data driven, and objective evaluation of surgical
performance, automated skill assessment has the potential to greatly improve surgical skill training.

Keywords

List the primary keywords that characterize the task.

Skill assessment, Open surgery, Suturing, Artificial intelligence

ORGANIZATION

Organizers

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

Hanna Hoffmann, Sebastian Bodenstedt, Stefanie Speidel, National Center for Tumor Diseases Dresden, Germany
Jan Egger, University Hospital Essen
Setareh Bady, Frank Hölzle, Rainer Röhrig, Behrus Puladi, RWTH Aachen

b) Provide information on the primary contact person.

Hanna Hoffmann (hanna.hoffmann@nct-dresden.de)
Behrus Puladi (bpuladi@ukaachen.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

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).

N/A

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).

TBA

Participation policies

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

Only automatic methods allowed

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.

Only the provided training data and publicly available data, including open, pre-trained networks, may be used

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

Members of the organizing institutes may participate, but are not eligible for awards

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

There will be two tasks in the challenge, the winner of each task will be awarded a prize, if at least 3 teams submit a result for the task. If at least 5 teams participate, the runner-up will also be awarded a prize.

e) Define the policy for result announcement.

Examples:

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

Results of all teams will be first presented at the Endoscopic Vision Challenge meeting at MICCAI 2024. Afterwards, the information will be made available to all participating teams. The results will be made publicly available in the form of a joint 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).

All contributing members of each participating team will be listed on the joint challenge 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.

Docker container on the Synapse platform with link to submission instructions

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 of each team will be evaluated

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: May, 1st
Start of evaluation: August, 1st
Submission deadline: September, 1st 11:59pm GMT
Challenge Day: Day of Endovis 2024

Ethics approval

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

The data used were anonymized, and the collection and publication of the data was authorized by informed consent of each subject. The data collection and the conduct of the original study leading to the dataset were approved by the local ethics committee of the University Hospital RWTH Aachen (approval code EK 352/21 and EK 22-329) and registered, including the study protocol, in the German Clinical Trials Register (DRKS00029307).

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

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 release of the training dataset.

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. Each team can choose to provide their source code, though they are not required to. Only a paper describing their method is required.

Conflicts of interest

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

Awards for the challenge will most likely be sponsored by Johnson&Johnson.; Only the organizers of the challenge will have access to the labels of the test data.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening

- Training
- Cross-phase
-

Research, Training, Education

Task category(ies)

State the task category(ies)

Examples:

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

Classification and prediction

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.

Medical and dental students and residents undergoing open surgical suturing training

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

Medical and dental students, surgical residents, and specialist participating in training

Imaging modality(ies)

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

Birds-eye-view 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).

Each video is annotated with a Global Rating Score (GRS)

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

none

Target entity(ies)

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

Videos in a simulated training setting

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 skill

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)).

Corresponding metrics are listed below (parameter 26).
Surgical skill classification task 1: Surgical skill classification algorithm with a low expected cost and a high F1 score, classifying the global rating score (GRS) intro four classes (novice: 8-15, intermediate: 16-23, proficient: 24-31, expert: 32-40)
Surgical skill classification task 2: Surgical skill classification algorithm with a low expected cost and a high F1 score, classifying the five different scores in the eight different objective structured assessment of technical skill (OSATS) categories

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).

Go Pro Hero 5

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

Videos are of approximate 5 min in length each and show participants suturing until the time limit has been reached. Two videos are taken from each participant: once before theoretical training and once after training. In addition, a single video was recorded from surgical residents and specialists to expand the skills spectrum.

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

University Hospital RWTH Aachen, Department of Oral and Maxillofacial Surgery & Institute of Medical Informatics

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).

Medical and dental students, surgical residents, and 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 participant performing one or multiple sutures before the time limit. Each student has a unique identifier and each video as well. Each video is annotated with a Global Rating Score (GRS). Annotations are performed by three raters (A, B, C)

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

A least 314 training cases and 30 test cases.

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

The total number was chosen due to annotation effort, the total number of test cases was chosen to maximize the ability the generalize and evaluate while maintaining a large enough training set.

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 data contains slightly more novice and intermediate cases than expert cases.

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.

To account for interrater variability, three human annotators were involved.

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.

The annotators were trained in the evaluation of the GRS on the basis of several test cases. The GRS is the summary of eight items (respect for tissue, time and motion, instrument handling, suture technique, procedure of surgery and advance planning, knowledge of specific procedure, quality of final product, and overall performance) of the Objective Structured Assessment of Technical Skills (OSATS) and has been used since the late 1990s:
Datta V, Bann S, Mandalia M, et al. The surgical efficiency score: a feasible, reliable, and valid
method of skills assessment. Am J Surg 2006; 192:372-8.
Hatala R, Cook DA, Brydges R, et al. Constructing a validity argument for the Objective
Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. AdvHealth Sci Educ Theory Pract 2015;20:1149-75.

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 raters worked in a blinded fashion and were experienced surgical residents or specialist in oral and maxillofacial surgery with a dual degree in medicine and dentistry.

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.

Annotations from all three raters are given, and it is up to the challenge participants how to merge them for training. For the evaluation, the annotations will be averaged.

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 videos in the dataset will not be pre-processed and will be available in raw form.

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.

Disagreement on rating - inter-rater agreement of >0.8 measured with pairwise Pearson correlation coefficient

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)

- Classification task 1: expected cost and average F1
- Classification task 2: expected cost and average F1

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

The average F1-score for each class was selected as the F1-score combines both precision and recall and is more useful than accuracy given uneven class distribution.
Expected cost chosen to consider ordinality of classes.

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.

As each task is ranked separately, no other scheme seemed sensible

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

Only full submissions for each task will be considered

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

As each task consists of a scalar metric, the entries will be sorted in ascending order. Rankings for individual metrics will be determined given the order and averaged to one rank per task.

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.

For the paper a Wilcoxon Signed Rank test will be performed to determine significance in metrics.
We will also examine whether leaving out the cases for each task with the overall best/worst performance will influence ranking

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

The Wilcoxon Signed Rank test is a nonparametric test that can be used to determine whether
two dependent samples were selected from populations, which cannot be assumed to be normally distributed, having the same distribution

Further analyses

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

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

The submitted methods will be analyzed for common problems and biases.

TASK 7: SegCol: Semantic Segmentation for Tools and Fold Edges in Colonoscopy data

SUMMARY

Abstract

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

Colorectal cancer (CRC) still stands as the second leading cancer-related deaths in the world. Polyp removal is an effective CRC screening method, especially at early stages; however, navigating through the colon to detect polyps is poses several challenges. A navigation system together with structural information can help the operator identify colon surfaces that have not been sufficiently screened for polyps.

Real-time screening of the colon is still a challenging task. Self-navigation solutions are at an early stage of development and themselves are facing several issues. These are commonly due to lack of reliable ground-truth and scene understanding which the operator may face as well.

Therefore, we propose a set of challenges to progress the techniques of camera navigation during colonoscopy by localising frequent anatomical landmarks (fold edges) and objects that modify the visualised scene (retractable tools embedded in the endoscope).
We aim to provide a novel set of manually annotated semantic labels to an expanded version of an existing dataset (EndoMapper). The dataset contains to date 96 videos of complete colonoscopies, and we aim to annotate a subset of frames/clips totalling 3,000 images with pixel-level semantic labels of instruments and folds (2 classes). The training split of labelled data will be shared with participants of the challenge and at a later stage made publicly available. The aim of the challenge is to accurately predict the segmented regions for both tools and folds and generalise to the anatomies of different patients. Moreover, participants will be able to utilise the remaining part of the unlabelled EndoMapper dataset for semi-supervision. We propose two tasks, both focusing on semantic segmentation of tools and folds: (i) Participants design architecture and training methodology, using any data from the labelled training set (ii) Participants design a training data sampling methodology (active learning), subject to a fixed architecture, training methodology, and training budget (maximum number of training frames).

Keywords

List the primary keywords that characterize the task.

Colonoscopy, Semantic segmentation, Active learning

ORGANIZATION

Organizers

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

Rema Daher, Xinwei Ju, Razvan Caramalau, Baoru Huang, Francisco Vasconcelos, Danail Stoyanov
Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and Department of Computer Science, University College London, London, UK

b) Provide information on the primary contact person.

Rema Daher
rema.daher.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).

N/A

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).

TBA

Participation policies

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

Only automatic methods allowed

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.

Sub-task 1: Publicly available data is allowed.
Sub-task 2: Only frames from released test set can be sampled

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 will not be eligible for the awards.

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

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

e) Define the policy for result announcement.

Examples:

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

The results for all teams will be announced during the EndoVis Challenge at
MICCAI2024. The results will also be made publicly available on the sub-challenge website. The submitted results will also be presented publicly in the joint publication.

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

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

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

Submission method

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

Examples:

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

Submission instructions will be posted on the sub-challenge website and will be sent to the registered participants via email. Each team must submit the running code as docker container via synapse. This should also be accompanied by the methodology report.

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

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

Challenge schedule

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

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

Challenge website and challenge registration opens: 1st April 2024
Training data release: 15th April 2024
Team registration open: 15th August 2024
Test data release: 15th August 2024
Submission deadline: 7th September 2024
Methodology report submission: 7th September 2024 (only valid docker submissions accompanied by output files and methodology report will be evaluated and included on the challenge day)
Challenge Day: Day of Endovis 2024

Ethics approval

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

The data for all the tasks come from the EndoMapper dataset, which is fully anonymised and ethically approved. The recordings were made under the ethical approval of the CEICA Ethics Committee (Comité de Ética de la Investigación de la Comunidad Autónoma de Aragón (CEICA), meetings 04/03/2020 acta 05/2020, 23/09/2020 acta 18/2020, 20/04/2022 acta 08/2022 and 16/11/2022 acta 20/2022).

Data usage agreement

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

Examples:

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

CC BY-NC-SA

Code availability

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

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

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

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

Conflicts of interest

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

Only the sub-challenge organisers will have access to the test data labels.
There is no conflict of interest.

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

Intervention assistance, screening, research.

Task category(ies)

State the task category(ies)

Examples:

- Classification

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

Tool segmentation, folds segmentation, active learning

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.

All target cohort is real patient data

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

Challenge cohort is real patient data, from a single hospital (Hospital Clinico
Universitario Lozano Blesa, Zaragoza, Spain)

Imaging modality(ies)

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

Colonoscopy

Context information

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

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

No further information other than image data and semantic labels 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.

Colon imaged in video with a monocular endoscope.

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.

Semantic segmentation of Tools and Colon Folds using colonoscopy images.

Assessment aim(s)

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

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

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

The requirements are the same for subtask 1 and 2. Instrument labels detected with high threshold-dependent metrics (F1-score, Dice) and fold edges detected with high values in metrics appropriate for contours / thin structures, which include threshold-independent metrics (AP,OIS,ODS) and centerline Dice score.

DATA SETS

Data source(s)

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

For all the tasks, the data is acquired with a colonoscope during CRC screening

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

The data is based on the real colonoscopy scans of different patients. It is composed of sequences from over 90 live recorded sessions.

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

The real data is based on a public dataset, Endomapper1 dataset.
1Azagra, P., Sostres, C., Ferrandez, A. et al. Endomapper dataset of complete calibrated endoscopy procedures. Sci Data 10, 671 (2023). https://doi.org/10.1038/s41597-023-02564-7

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

The data has been acquired by trained clinicians to perform colonoscopies. We further sub-sampled the video recording and smaller sequences by targeting the regions of interests, folds and tools.

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 images from videos of full colonoscopies. The annotations are performed in the same way for all the data. The unlabelled pool will be considered for the challenge initially and depending on the selected samples in sub-task 2, its annotations will be used for evaluation.

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

Task 1: The training data includes more than 1,500 training and validation labelled images (combined), and more than 2,000 unlabelled images (they have labels, but they are not shared with participants); The test data includes about 500 images.

Task 2: The participants will design their active learning selection policy using the same training data from task 1 (1,500 labelled and 2,000 unlabelled). The output of this selection policy should be a subset of 500 out of the 2,000 unlabelled images. The organisers will then evaluate their submitted selection by re-training a fixed model with their selected subset (500) and the original training data
(1,500) in a fully supervised manner, using the labels unknown to participants. This simulates a
scenario where participants request 500 additional images to be labelled, out of 2,000, to improve their model.
The test data is the same as task 1.

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

We consider that it provides a good balance of training, validation and available samples fit for the tasks.

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.

Specific to the tool segmentation task, there will be three types of tools used: resection device, forceps and balloon. In terms of colon folds, the distinctive edges will be annotated. There will be frame sets that will include both tools and folds annotations. We estimate a 40% distribution of tools images out of the entire dataset (training, unlabelled, test), while the rest will contain just folds.

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.

Semi-automatic annotations provided by a specialised company for all cases.

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.

Semi-automatic annotations provided by a specialised company for all cases.

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 data cases have been annotated by non-medical labelling experts. The experts belong to a popular annotation company in AI with over 10-year experience.

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.

The annotations of all sets will consist of tool and folds segmentation maps where pixel values will be attributed to either fold edge, tool part or background.

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.

For obtaining instrument frames, the videos have been pre-sampled based on timestamped colonoscopy audio transcripts (meta-data from EndoMapper). For obtaining non-instrument frames, a frame selection algorithm(*) was first used for initial selection of clear frames, followed by further manual filtering.

(*) Barbed et al. Semantic analysis of real endoscopies with unsupervised learned descriptors, MIDL 2022

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.

Instruments have complex shapes, with thin wires and small holes that could be missed by annotators. We estimate that these would amount, on average, to less than 1% of all instrument pixels.
For the most part, folds edges are visually very clear curves, however, there may be ambiguities when annotating their extremities. Additionally, some folds can be missed in parts of the image that are either very dark or contain significant light reflections. We estimate that these would amount, on average, to less than 5% of all fold pixels.

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 both tasks: mAP for both instrument class and fold edges.

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

For Instrument segmentation, F1-score is one of the most well-established metrics in the literature.
On the other hand, fold edges are extremely thin structures, and threshold-dependent metrics (F1-score, Jaccard, etc) are not adequate in this context. Out of popular threshold-independent metrics for edges (OIS, ODS, AP) and metrics designed for thin structures (centerline Dice score), we select AP as this is also an established metric for semantic segmentation of other shapes and can be applied effectively to both edges and instruments. Therefore, we will average both classes with mAP. Other metrics (F1-score, centerline Dice score, Jaccard) will be computed for reporting and metric evaluation purposes.

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.

Rank stability analysis will be performed and reported with the challenge results. For the final joint challenge paper, we will assess the results of the participating teams for statistical significance. If the underlying test assumptions are fulfilled (depending on the properties of the submitted data which we can only assess after the challenge deadline) we will use a Students-T-test, otherwise a Wilcoxon signed-rank test could be more appropriate.

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.

Sub-task 1 segmentation: The highest mAP.
Sub-task 2 active learning: The highest mAP on the selected subset.

Statistical analyses

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

- description of the missing data handling,
- details about the assessment of variability of rankings,

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

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

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.

NA

TASK 8: SurgVU: Surgical Visual Understanding

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.

Machine learning models that can detect and track surgical context from endoscopic video will enable transformational interventions. For example, the ability to automatically categorize surgical progress (i.e. Phase, Step, Task, or Action) and the instruments used, will allow for improved assessments of surgical performance, efficiency, and tool choreography, as well as new analyses of OR resource planning. Indeed, the theoretical and practical applications are broad and far-reaching. Obtaining the data needed to train these models, however, is resource intensive and time consuming. Clinical videos need to be annotated, frame by frame, segmenting the surgical categories and identifying the bounding boxes and/or key points around surgical instruments, under a broad variety of conditions. Moreover, ongoing annotator training is needed to stay up to date with surgical methods and instrument innovation. Importantly, in robot-assisted surgery, instrument installation/uninstallation information can be programmatically harvested from the system, providing proxy annotations for tool presence in the video feed. Additionally, standard surgical ontologies are being developed with objective definitions, easing the burden on human annotators. In this challenge we invite the surgical data science community to take part in two categories. The first category requires participants to train a model to localize tools and their corresponding key-points in videos, using tool presence data as weak labels. The second category invites participants to segment videos into different surgical steps being performed. Winning solutions to either category would significantly reduce the annotation loads required for training models, and avail themselves to a wealth of clinical applications.

Keywords

List the primary keywords that characterize the task.

Weak learning; Object detection; Object localization; Activity recognition

ORGANIZATION

Organizers

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

Aneeq Zia, Max Berniker, Conor Perreault, Rogerio Nespolo, Ziheng Wang, Anthony Jarc, Intuitive Surgical

b) Provide information on the primary contact person.

Aneeq Zia
aneeq.zia2@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).

N/A

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).

TBA

Participation policies

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

Only automatic methods are allowed.

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.

Any publicly available dataset will be allowed for pre-training. Any private dataset used will need to be released publicly at the time of submission.

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.

organizers institute may participate in the challenge but will not be eligible for awards

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.

Results of all teams will be first presented at the EndoVis challenge at MICCAI 2024. Complete leaderboard will be announced publicly on the challenge website.

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

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

The organizers will publish a challenge paper within six months after the challenge. Following which, the participating teams can publish their own results from the challenge citing the challenge paper. Possibility of a combined publication amongst the participating teams/organization team will also be discussed after the challenge

Submission method

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

Examples:

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

Submission instructions will be posted to the website and sent via email. Results will be submitted via a docker container through 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 2024
Registration: until Aug 2024
Submission deadline end of September, 2024;
Challenge Day: Day of Endovis 2024

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. Informed consent was obtained from all surgeons in the study.

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 (Attribution-NonCommercial-NoDerivs)

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.

All teams will be required to upload their code repos to github and make it public as part of the final submission.

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; 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.

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 of various skill levels during 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 training videos will come with ground truth tool presence labels along with surgical step labels while the testing data will have ground truth tool bounding box labels along with surgical step labels.

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 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 surgical tool bounding boxes and surgical steps.

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)).

For bounding box detection, the assessment will be done using mean average precision over multiple intersection-over-union (IOU) values - this metric is standard for COCO dataset. For surgical step classification task, average f1-score across all steps will be used for evaluation.

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 training being performed on a porcine model. These videos will be long and variable in length where multiple surgical training steps will be present in each video along with different surgical tools.

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

We will have approx. 200 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 has an average length of approx 4 hrs.

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 and surgical steps 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) while the surgical step distribution will be much more balanced across training and testing 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.

We will use a crowd (5+ annotaters) to annotate tool bounding boxes for our testing set. The annotations will not be redundant as bounding box annotations are not that subjective. The surgical step annotations will be done by a team of domain knowledge experts for training and testing sets.

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

For 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. For surgical steps, the annotators will be provided by clear step starting and ending times for annotations.

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 and surgical steps.

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 videos in the dataset will not be pre-processed and will be available in raw form.

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 dont expect it to be more than 5% For surgical steps, there can be some sources of error as this type of annoation is temporal in nature. However, as the robotic surgical training steps are very well defined (as opposed to steps in clinical procedures), we do not expect there to be any significant error in annotations (<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. The step labels would not have any other relevant source of error.

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. For surgical step recognition, we will use the average of f1-scores across all classes.

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. The surgical step category is a standard classification problem where average f1-score can accurately measure the performance of the models.

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.

Using the standard metrics being used within the object detection and surgical step recognition research seems like the right way to rank teams. The spatial detection metric tests the algorithms for detection of objects of different sizes (for tool detection category) which will be useful in differentiating high and low performing teams. Similarly, mean f1-score will reward and penalize the predictions for correct/incorrect predictions accordingly.

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.

The performance rank will be based on the rank of the evaluation metrics (e.g mAP IoU 0.5:0.05:0.95 for bounding box detection and mean f1-scores for surgical step recognition) - the higher the value of this metric, the higher the ranking of that team will be

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. Bootstrapping will also be used for uncertainty analysis.

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.

Additional analysis will be performed on the results to check for ranking variability. On top of the mAP across multiple IoU values for the first category, we will test ranking of the teams when using individual IoU values (e.g 0.5, 0.6, etc) instead of averaging over all. In addition to this, we will evaluate per class metrics for both challenge categories as well.

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.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., ... & Kopp-Schneider, A. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. Nature communications, 9(1), 5217.

Further comments

Further comments from the organizers.

NA