Endoscopic Vision Challenge 2021: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Endoscopic Vision Challenge 2021

Challenge acronym

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

EndoVis

Challenge abstract

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

Minimally invasive surgery using cameras to observe the internal anatomy is the preferred approach to many surgical procedures. Furthermore, other surgical disciplines rely on microscopic images or use flexible endoscopes for diagnostic purposes. As a result, endoscopic
and microscopic image processing as well as surgical vision are evolving as techniques needed to facilitate computer assisted interventions (CAI). Algorithms that have been reported for such images include 3D surface reconstruction, salient feature motion tracking, instrument detection or activity recognition. However, what is missing so far are common datasets for consistent evaluation and benchmarking of algorithms against each other. As a vision CAI challenge at MICCAI, our aim is to provide a formal framework for evaluating the current state of the art, gather researchers in the field and provide high quality data with protocols for validating endoscopic vision algorithms. EndoVis serves as an umbrella for different kinds of sub-challenges that tackle specific problems and applications in endoscopic/microsopic vision.

Challenge keywords

List the primary keywords that characterize the challenge.

Surgical Vision, Endoscopy, Classification, Segmentation, Detection

Year

The challenge will take place in ...

2021

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

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

none

Duration

How long does the challenge take?

Full day.

Expected number of participants

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

50 (based on numbers from previous EndoVis challenges)

Publication and future plans

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

The joint publications 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.

platform depends on the specific sub-challenges (e.g. grand-challenge.org, DREAM/synapse platform) , no on-site challenges, so far no specific technical equipment is required

TASK: HeiChole Surgical Workflow Analysis and Full Scene Segmentation (HeiSurf)

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.

Understanding the surgical scene via full scene segmentation is a prerequisite for many assistance functions in computer-assisted interventions (CAI), such as robot-assisted surgery and surgical workflow analysis. These are key technologies for the development and seemingless integration of CAI systems
into the workflow in the operating room. These systems may increase the safety and efficiency of the operation through early context-sensitive warnings, OR management and procedure time prediction, continuing surgical education and professional development by objective assessment of surgical skill and competency as well as semi-autonomous assistance. Furthermore, to enable future robotic assisted surgery systems, they will have to understand the surgical scene and workflow and learn from the surgeons skills. Thus, surgical scene understanding and workflow analysis are a prerequisite for the next generation of surgical robotics.
Most segmentation methods in CAI are generally task-specific, e.g.
are limited to certain organs or instruments, or have only been evaluated on datasets that make it difficult to judge if the methods would generalize to an actual laparoscopic setting, because either the datasets are small or not based on human anatomy.
In this challenge we aim to benchmark how well current segmentation methods perform on a representative dataset in terms of accuracy measured via DICE-coefficient and Hausdorff distance. Furthermore, we want to determine if and how segmentation and workflow information influence and benefit from each other in a multi-task setting. For this, we will provide a large dataset, comprising 33 laparoscopic cholecystectomies collected from 3 centers, with full scene frame segmentations every 4 minutes with overall 14 different classes as well as surgical phases and image-wise instrument labels for every frame.

Keywords

List the primary keywords that characterize the task.

Laparoscopic cholecystectomy, full scene segmentation, minimally-invasive surgery

ORGANIZATION

Organizers

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

Heidelberg University Hospital:
Martin Wagner, Jonathan Chen, Benjamin Müller, Beat Müller

National Center for Tumor Diseases (NCT): Dresden

Sebastian Bodenstedt, Stefanie Speidel

Karlsruhe Institute of Technology (KIT):
Paul Scheikl, Franziska Mathis-Ullrich

German Cancer Research Center (DKFZ):
Lena Maier-Hein

b) Provide information on the primary contact person.

Sebastian Bodenstedt: sebastian.bodenstedt@nct-dresden.de
Deputy: Martin Wagner: martin.wagner@med.uni-heidelberg.de

Life cycle type

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

Examples:

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

synapse.org

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

Part of https://endovis.grand-challenge.org/, sub-challenge site TBD

Participation policies

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

Fully automatic.

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

Publicly available data is allowed.

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

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

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

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.

The results of all participating teams will be reported

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).

A challenge paper is planned. From each team at least two members will be listed as co-authors.

Submission method

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

Examples:

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

Docker container on the Synapse platform. Submission instructions will be similar to our challenge from 2019: https://www.synapse.org/#!Synapse:syn18824884/wiki/592580

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 no results on submissions to the teams. We will evaluate submissions on parts of the training data and return these results to the team to allow for sanity checks. The last submission will be the one that is 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

May 1st: Release of first part of the training data
August 1st: Release of second part of the training data
September 10th: Registration deadline

September 15th: Final submission date of docker containers
September 27th or Oct 1st: Presentation at EndoVis

Ethics approval

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

As the data used consists of anonymous surgical video data, no ethics approval is required

Data usage agreement

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

Examples:

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

CC BY NC ND.

Additional comments: During the challenge, the data will be made available upon request and teams will be forbidden to share the data with anyone. After the challenge paper has been published or has been made available on a preprint server the data will be made available under the CC BY-NC-ND license.

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 code for evaluation submissions will be made available with the release of the training data

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

Participants are encouraged,but not required to make their code available as open source.

Conflicts of interest

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

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

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Surgery, Assistance, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation, workflow analysis

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 that underwent laparoscopic gallbladder removal

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

Patients that underwent laparoscopic gallbladder removal

Imaging modality(ies)

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

Laparoscopic video

Context information

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

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

Segmentation masks, frame-wise labels on surgical phases and instrument usage, an ID differentiating between the 3 centers in the dataset

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

none

Target entity(ies)

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

Videos from laparoscopic gallbladder removals, i.e. endoscopic video from the abdominal cavity, mostly showing the liver, gallbladder and surrounding

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 surgical instruments and anatomical structures in the laparoscopic video
- Surgical phase segmentation, i.e. determining which video frame belongs to which surgical phase
- Instrument presence detection, i.e. which instruments are currently visible in a video frame

Assessment aim(s)

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

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

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

Accuracy.

Additional points: - Find image segmentation algorithms that can segment different objects and structures in

laparoscopic scenes with a high amount of accuracy measured by the average DICE coefficient and Hausdorff distance over all classes
- Find algorithms that can accurately identify surgical phases and instrument usage in regards to the average DICE coefficient over all classes
- Investigate whether a multi-task approach to semantic scene segmentation and surgical phase recognition yields a benefit compared to the separate tackling of the individual tasks in regards to accuracy

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).

Video collected from varying types of laparoscopic cameras

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

Videos collected during routine laparoscopic surgeries at participating surgical centers

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

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

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

not available

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 training case is a video of a laparoscopic gallbladder removal. For frames at fixed intervals (every 4 minutes) a segmentation mask will be provided. For each frame a phase label and instrument usage label is provided

A test case is identical, but no annotations will be provided.

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

The dataset will contain 24 training cases (ca. 290 segmentation masks) and 9 test cases (ca. 110 segmentation masks). The training cases will be equally selected from 2 surgical centers. The test cases will contain 3 cases from each of the 3 different centers.

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

One center was selected to only provide testing data, so it is possible to determine if the algorithms can generalize and be used successfully in new centers. Disregarding the "surprise" center, a 80%/20% split is common practice.

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.

No further characteristics.

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 annotation team consists of six medical students, of whom 3 will be annotating each image to ensure maximum annotational accuracy. Data will be validated by a board certified surgeon.
For the workflow labels, the phase segmentations and tool usage from the EndoVis 19 Surgical Workflow challenge will be used. For more information, we refer to the annotation protocol from the Workflow challenge: https://docs.google.com/document/d/1PehU09Q49fUF9HDGN0i8Kr3OOEBm17ZYDRmObgdZIXY/edit?usp=sharing

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 instructions as well as class definitions are determined in the annotation protocol:
General rules:
- Every Pixel must be assigned to one of the given object classes.
- One Pixel cannot have two classes
- Unless it is a subclass as for the instrument class
- The annotator shall always try to annotate the first object in the line of sight
- If the contour cannot be clearly defined, the annotator shall evaluate the Frame and make an educated guess
- When annotating and using Polygon points the Annotator shall always follow the Rule "use as little as possible and as much as necessary".

In addition, for every object class (14) and instrument class (18) detailed annotation rules are provided (due to

space constraints not listed here)

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

The Annotators are medical students who have studied medicine for at least a year. All students underwent annotational Training and are supervised by a trained surgeon.

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 will be merged via majority voting in case of small discrepancies. For larger discrepancies, a surgical expert will manually correct the annotations.
For this, we will separately examine each object
class in an image, if the DICE score between each binary segmentation is above a certain threshold (probably around 0.95, but this will be fine-tuned), majority voting will be applied. If the value is lower, or if disagreement on the occurrence of an object class occurs, the image will be flagged and examined by a surgical expert.

Data pre-processing method(s)

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

The training and test data consists of videos recorded from the endoscopic video feed during surgery and compressed using MPEG-4. Each video was preprocessed to remove frames taken from outside the abdominal cavity, in order to ensure the anonymity of patient and surgical staff. For the semantic segmentation annotation, images were automatically extracted at fixed intervals from the video

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 each image will be examined by multiple persons, errors could only be due to poor lighted areas, motion artefacts or image quality, which make it difficult to provide absolute certain annotations for parts of an image.

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

all described above

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)

The DICE similarity coefficient and Hausdorff distance will be used for ranking the scene segmentation task. The F1 score will be used for ranking purposes for all other tasks. Other common metrics, e.g. accuracy, precision, recall, will also be reported.

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

The DICE coefficient (and the positive correlated IoU) is a metric used commonly in the segmentation and the surgical workflow community and is generally an accepted metric. For segmentation tasks the Hausdorff distance complements the DICE coefficient as it examines the contour while the DICE coefficient examines the volume overlap.

Ranking method(s)

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

For image segmentation: The DICE coefficient and Hausdorf distance will be computed for every class in each selected image in a video and then averaged over each video and then over all videos. We will then compute two ranks, one for each metric. The rank of each team will then be determined through its average rank.
For surgical phases and instrument presence: F1-score will be computed for every different class label in a video, we will then average over all classes and videos
For multi-task: the separate scores will be computed for each category as above and for each score, we will perform a ranking. A team score is then the average rank. This way, we aim to discourage over-optimization of some categories while neglecting other categories.

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

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

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

For the categories: We decided to first average over each video to make sure that each case is weighted equally, even though the videos vary in length.
For multi-task: As results from different categories will have to be merged, we decided to use a ranking scheme

Statistical analyses

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

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

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

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

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

Further analyses

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

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

NA

TASK: Gastrointestinal ImAge NAlysis (GIANA)

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.

GIANA challenge has the objective to compare performances of state-of-the-art Computer Vision algorithms in the context of the analysis of videocoloscopy images and video (polyp detection, localization, segmentation, lesion classification)
After three very successful iterations of the GIANA sub-challenge in the framework of the Endoscopic Vision Challenge (MICCAI 2015, 2017, 2018) here we are back again to offer participants the possibility to propose state-of-the-art approaches in Computer Vision for the content analysis of videocoloscopy videos and images.. Like last occurences, we are not focusing only on polyp detection in colonoscopy images: we also cover polyp segmentation in both colonoscopy images (SD and HD). This year, images from two different manufacturers will be provided. So, this year again, we offer to you the chance to test your detection, segmentation or classification methods in fully publicly available databases. We also enrich the GIANA competition with a new task: histology prediction. In situ determination of the histological type of the lesion is of primary interest for physicians as part of the newly adpoted resect or discard strategy.

Keywords

List the primary keywords that characterize the task.

Videocoloscopy, Colorectal Polyps, Detection, Segmentation, Histology Prediction

ORGANIZATION

Organizers

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

Dr. Jorge Bernal, Associate Professor, UAB, CVC, Barcelona, Spain ;
Prof. Aymeric Histace, Full Professor, ENSEA, ETIS, Cergy, France ; Prof.,- MD. Gloria Fernandes-Esparrach, Clinic Hospital Barcelona, Spain ;
Prof. MD. Xavier Dray, Saint-Antoine Hospital, APHP, Paris, France

b) Provide information on the primary contact person.

Jorge Bernal : jorge.bernal@uab.cat

Life cycle type

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

Examples:

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

grand-challenge.org

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

https://giana.grand-challenge.org/

Participation policies

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

Semi automatic, Fully automatic.

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

For clinical data, only the use of the provided data set are allowed. Pretraining on generic data set are nevertheless allowed but have to indicated.

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 not participate.

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

We may be able to provide a prize for each proposed tasks of the challenge. Additionaly, specific prizes related to the methods proposed can be decided (Prize of the jury for instance) which aim is to reward a specific originality in the processing scheme proposed

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.

Per task, top three performing methods will be announced publicly. For specific prize, only one team will be awarded.

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).

We propose to each participating team to send a 2-pages synthesis of their methods. We also suggest that this year a proceedings is published with an extended version of this two-page abstract. In any case, this year, participating teams will be authorized to publish their own results after the challenge as long as they commit to cite proper references from the organizers

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.

Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail. : https://giana.grand-challenge.org/Rules/

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 plan to provide participants an online evaluation platform in which they can test the performance of their methods in a limited set of examples from the testing set. This would allow participants to progressively test the new iterations of their methodologies without having to wait until challenge day to see how they actually perform.

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

With respect to the timeline, the following dates will be used
Registration date: from March 15th to September 20th
Release of training data: March 15th 2021
Release of testing data: May 15th 2021
Submission date: September 20th
Release date of the results: challenge day

Ethics approval

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

All the data were acquired after an ethical approval of both clinical institutions. Proposed data are full

anonymized and their use is restricted to research and/or teaching 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 ND.

Additional comments: -

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.

Matlab and Python code for evaluation of the expected outputs will be available. The ranking will be based on the performance results through a attribution-of-point system.

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

Due to the formatting of past GIANA events, codes were not required from participating teams. If the format of the challenge is adapted, we will consider to organize an on-site demo for each of the participating teams .

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.

Challenge may be funded by Société Française d'ENdoscopie Digestive. Aymeric HIstace and Xavier Dray are co-funder of Augmented-Endoscopy start up (focusing on Wirelecc capsule image analysis).

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

CAD, Training, Research, Screening, Education.

Task category(ies)

State the task category(ies).

Examples:

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

Classification, Detection, Localization, Segmentation, 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 presenting with risks of colorectal cancer. Included sequences cover lesions in early stage as well as some in a more advanced stage

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

Patients presenting with appearance of one or multiple polyp lesions within the colon. Various stage of lesion developement.

Imaging modality(ies)

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

Videocolonoscopy images and videos. White light imaging. Two manufacturers will be considered (Pentax and Olympus). Datasets : https://giana.grand-challenge.org/Tasks/ + the ETIS Larib dataset https://polyp.grand-challenge.org/EtisLarib/

Context information

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

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

Statistical distribution of the type of polyps, polyp size, location in the colon and morphology according to Paris classification

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.

Colorectal images and videos taken from classic videocoloscopy examinations

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.

Polyp detection, localization, segmentation, histology prediction (prevention of colorectal cancer risks).

Assessment aim(s)

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

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

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

Accuracy.

Additional points: For the case of the polyp detection tasks, we will compare the methods using ROC curves and Area under ROC curve. Only teams able to detect all polyps in all sequences will be considered for the final ranking. With respect to polyp segmentation task, we will rank methods according to IoU score and Hausdorff distance. A separate ranking will be made for each of the two metrics, awarding 3,2 and 1 points to the best three teams. Aggregate of both rankings will be used to determine the final winner. Only methods able to segment all

polyps will be considered for the final ranking. Finally, with respect to polyp classification we will use the confusion matrix as the main metric, calculating classification scores per class and overall accuracy.

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).

Pentax and Olympus videocolonoscopes in a usual set-up.

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

https://giana.grand-challenge.org/PolypDetection/ ; https://giana.grand-challenge.org/PolypSegmentation/ ; https://polyp.grand

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.

Hospital Clinic of Barcelona ; Hospital Saint Antoine, APHP, Paris.

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).

Images and video were acquired and annotated by experienced physicians in both centers.

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 represent a short video or images extracted from videocolonoscopy examinations. For each, a related ground truth is provided taking the form of binary masks. An ellipse for each image of the videos, a pixel-wize binary mask for still images of CVC and ETIS-Larib
Three classes: Adenoma/ Serrated Sessile Adenoma /Hyperplastic with the possibility (due to lesion representation) to group them into adenoma/non-adenoma

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

Polyp detection and localization: 18 videos for training, 18 for testing Polyp segmentation: ETIS Larib : 96 images will be considered for the training, 100 for testing CVC HD segment: 50 images for training, 100 for testing

Histology prediction Around 1000 images for training, 300 for testing as well as negative examples

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

https://giana.grand-challenge.org/PolypDetection/; https://giana.grand-challenge.org/PolypSegmentation/; https://polyp.grand-challenge.org/EtisLarib/

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.

https://giana.grand-challenge.org/PolypDetection/; https://giana.grand-challenge.org/PolypSegmentation/; https://polyp.grand-challenge.org/EtisLarib/

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.

4 different annotators (2 from Hospital Clinic and 2 from Lariboisiere Hospital)

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.

With respect to GIANA challenge, the annotation workflow is as follows. Images and videos are acquired in the exploration rooms of Hospital Clinic, Barcelona. For the case of video recordings for the polyp detection task, the following metadata was added: estimation of polyp size (in mm), morphology (according to Paris classification) and histological class (adenoma vs. no adenoma). All frames from the videos were manually labelled by clinicians using GTCreator annotation tool (publicly available at http://www.cvc.uab.es/CVC-Colon/index.php/gtcreator/) ; in this case, due to the high amount of frames, the image mask consisted of an ellipse covering the polyp. With respect to still frames used in the polyp segmentation and classification task, images were selected by clinicians with the criteria of having at least one shot per polyp (if there are more than one, the polyp appearance should be clearly different from the previous one). With respect to metadata, the following items were provided: polyp size (in mm), morphology (according to Paris classification), location within the colon (rectum, sigma, cecum, etc..) and histological class (adenoma vs. no adenoma). Pixel-wise masks representing the polyp region were created by clinicians using the GTCreator tool. For each of the cases, clinicians are shown how to do the annotations via video examples and hands-on demonstration. ospital Clinic, Barcelona. For the case of video recordings for the polyp detection task, the following metadata was added: estimation of polyp size (in mm), morphology (according to Paris classification) and histological class (adenoma vs. no adenoma). All frames from the videos were manually labelled by clinicians using GTCreator annotation tool; in this case, due to the high amount of frames, the image mask consisted of an ellipse covering the polyp. With respect to still frames used in the polyp segmentation and classification task, images were selected by clinicians with the criteria of having at least one shot per polyp (if there are more than one, the polyp appearance should be clearly different from the previous one). With respect to metadata, the following items were provided: polyp size (in mm), morphology (according to Paris classification), location within the colon (rectum, sigma, cecum, etc..) and histological class (adenoma vs. no adenoma). Pixel-wise masks representing the polyp region were created by clinicians using the GTCreator tool. For each of the cases, clinicians are shown how to do the annotations via video examples and hands-on demonstration.

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.

NA

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.

NA

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.

NA

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)

Detection : Accuracy, F2, Segmentation : Dice ; Histology prediction : Confusion Matrices and related per class performance metrics

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

Accuracy and frame-based metrics are the classic ones used in Computer Vision but that does not give a good view on the clinical usability of the proposed method. For this, F2 score is added to the metrics for detection as well as we propose the use of clinical-based metrics such as reaction time or Average PDR. Regarding other tasks, classic metrics as DIce and Confusion matrices are used since they represent well known and adapted metrics by the community. Additionally, in the case of segmentation, standard deviation will be considered to give a quantitative estimation of the robustness of the proposed methods.

Ranking method(s)

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

For each task and each related performance metrics, a ranking is provided and points are attributed to the first three in decreasing rder (3, 2, and 1 point). Thus we are able for each task to have a final podium but also a global podium if a team takes part to the all tasks proposed.

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

If missing results for a specific task appear, for this specific task, participants results will be computed but won't be able to be taken into account for the final ranking.

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

Main risk we noted all along previous GIANA challenges is that two or more teams have very similar performance for a given task. The way we consider the different metrics with associated point depending on the ranking for a specific metric help to solve in a fair way possible the issue.

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.

NA yet

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

Regarding GIANA, both in Jorge Bernal et al. "Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge" IEEE Transactions on Medical Imaging, 2017, 36 (6), pp.1231 - 1249. and a Springer book to be out in February 2021 (about GIANA 2017 and 2018) we provided a full scheme for the analysis of obtained results using mainly combining algorithms via ensembling, inter-algorithm variability, common problems/biases.

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.

Regarding GIANA, both in Jorge Bernal, Nima Tajkbaksh, F Sánchez, Bogdan Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, Konstantin Pogorelov, Sungbin Choi, Quentin Debard, Lena Maier-Hein, Stefanie Speidel, Danail Stoyanov, Patrick Brandao, Henry Córdova, Cristina Sánchez-Montes, Suryakanth Gurudu, Gloria Fernández-Esparrach, Xavier Dray, Jianming Liang, Aymeric Histace

Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge IEEE Transactions on Medical Imaging, Institute of Electrical and Electronics Engineers, 2017, 36 (6), pp.1231 - 1249. 10.1109/TMI.2017.2664042 DOI : 10.1109/TMI.2017.2664042 and a Springer book to be out in February 2021 (about GIANA 2017 and 2018) we provided a full scheme for the analysis of obtained results using mainly combining algorithms via ensembling, inter-algorithm variability, common problems/biases.

TASK: Surgical Action Triplet Recognition (CholecTriplet2021)

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.

With the development of context-aware decision support in the operating room, it has become imperative to analyze surgical workflow activities at a fine-grained level to foster safety and efficiency. Most of the existing works recognize surgical actions at a coarse-grained level (such as phases, stages, single verb, etc.) leaving out some detailed information needed to analyze surgical workflow at par with the current pace of deep learning and artificial intelligence on activity recognition. Hence, we aim to recognize surgical actions as a triplet of instrument, verb and target.
To this effect, we introduce a new and unique endoscopic dataset, CholecT50, in which every frame has been annotated with labels from the triplet classes. The dataset is very useful for the development and evaluation of algorithms targeting the recognition of instrument-tissue interaction in laparoscopic cholecystectomies.
This sub-challenge focuses on exploiting machine learning methods for the online automatic recognition of surgical actions as a series of triplets. Participants will develop and compete with algorithms to recognize action triplets directly from the provided surgical videos. This novel challenge investigates the state-of-the-art on surgical fine-grained activity recognition and will establish a new promising research direction in computer-assisted surgery.

Keywords

List the primary keywords that characterize the task.

Surgical activity recognition, action triplet, tool-tissue interaction, CholecT50, deep learning, laparoscopic video analysis.

ORGANIZATION

Organizers

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

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

b) Provide information on the primary contact person.

Chinedu Nwoye: nwoye@unistra.fr

Life cycle type

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

Examples:

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

grand-challenge.org

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

https://actiontriplet.grand-challenge.org/cholect50-challenge/

Participation policies

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

Fully automatic.

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

Publicly available data is allowed.

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

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

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

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

e) Define the policy for result announcement.

Examples:

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

The results of all teams will be first presented during the Endoscopic Vision Challenge meeting at MICCAI 2021. Afterwards, the information will be made available to all participating teams. The results will be made publicly available in the form of a joint publication.

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

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

Every participating team can submit at most 2 qualifying authors. The sub-challenge organizers determine the order of the authors in a joint challenge paper. Participants are allowed to publish their own results separately only after a publication of a joint challenge paper (expected by end of 2022).

Submission method

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

Examples:

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

The participating team will submit a docker container with their codes that will be benchmarked internally by the organizers on the unseen test data. Submission instructions and a template docker in the required submission format with specific input/output protocol will be provided.

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

Evaluation metrics will be provided for participants to evaluate their algorithm on their chosen validation data split. During the validation phase, participants make a submission which will be evaluated on the validation data for sanity check. During the final submission stage, only the last submission for each team before the deadline will be evaluated for the challenge ranking.

Challenge schedule

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

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

Mar 15th: Release of first training dataset and the script for the evaluation metrics
Mar 29th: Release of second training dataset
May 31st: Launch of slack interaction forum
Jun 20th: Release of docker submission template
Aug 10th: Validation phase: evaluation server opens for testing containers
Sep 15th: Submission deadline (11:59pm GMT)
Sep 27th or Oct 1: Challenge Day

Ethics approval

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

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

ethics approval is required.

Data usage agreement

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

Examples:

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

CC BY NC.

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

Code availability

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

The script(s) for computing evaluation metrics will be made available to the participating teams.

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. A docker image template will not be shared by the organizers. Each team can choose to provide their source code, though they are not required to. Only a paper describing their method is required.

Conflicts of interest

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

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

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Decision support, Surgery, Assistance, Research.

Additional points: Surgical Workflow Analysis
Surgical action recognition
Research

Task category(ies)

State the task category(ies).

Examples:

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

Classification: surgical action triplet recognition

Cohorts

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

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

Patients undergoing laparoscopic cholecystectomy.

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

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

Imaging modality(ies)

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

Laparoscopic video stream

Context information

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

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

Action triplet annotation (128 classes, each frame has a 1D binary vector, indicating if the corresponding action is being performed.
Additionally, annotations for: Instruments, verbs and targets

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

none

Target entity(ies)

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

Videos from laparoscopic cholecystectomies

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

Recognizing surgical tool-tissue interactions in laparoscopic videos

Assessment aim(s)

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

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

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

Accuracy.

Additional points: • Action-triplet recognition: algorithm with high average precision.
From the predicted action triplet labels, the evaluation software will be able to extract and assess also the average precisions for the correct triplet's components (such as instrument, verb, target, instrument-verb, instrument-target, etc.) .

DATA SETS

Data source(s)

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

Recordings from laparoscopes at University Hospital of Strasbourg, France.

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

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

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

University Hospital of Strasbourg, France

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

Surgeons

Training and test case characteristics

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

Examples:

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

One case is a single laparoscopic intervention. A video of the entire operation will be provided for each case. Each case has an action triplet annotation (128 classes, each frame has a 1D binary vector, each entry indicating if the corresponding action is being performed (1) or not (0)).
Additionally, each case is also provided with supplementary annotations for: Instruments, verbs and targets . All the supplementary annotations follow the same format as the triplet annotation case. Both training and testing cases are annotated with the same parameter.

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

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

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

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

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

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

Annotation characteristics

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

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

Two surgeons annotated the data. The first surgeon annotated 40 videos and the second surgeon annotated 10 videos. Where there is ambiguity, label mediation is provided by a third clinician.

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

Surgeons were given a list of items to annotate and were also educated on the use of the annotation software.

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

Surgical experts involved in both clinical practice and research carried out the annotation.

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.

Disagreements between annotators are solved through mediation

Data pre-processing method(s)

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

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

Sources of error

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

Possible source error:

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

- Disagreement in the beginning and termination of action (est. 0.08 – 1.00 secs)
- Disagreement in action similarity (estimated 1-5 frames per video)
- Possibility of unrepresented class in training/testing data (max. 3/128)

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

none

ASSESSMENT METHODS

Metric(s)

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

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

Action triplet recognition will be assessed by mean average precision (mAP).
The mAP is aggregated as a mean of all APs over all frames in all the test videos.

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

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

Ranking method(s)

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

The performance metrics (triplet mAP) produces a scalar value which will be used to rank the model performance specifically in descending order.
In case of a tile, we evaluate also the mAP of the components of the triplets namely: instrument-verb, instrument-target, instrument, verb, and target, in that order.

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

Only full submissions for the task will be considered

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

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

Statistical analyses

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

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

The level of significance in differences in metrics and model ranking stability will be analyzed using Wilcoxon signed-rank test.
We will also examine the performance of each model on each test video.

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

Wilcoxon signed-rank test is a non-parametric statistical hypothesis test to determine whether the median difference between two sets of observation is significant especially if the differences between pairs of data are non-normally distributed.
Also, analyzing the algorithm's performance on each test case and then on the whole will help to understand how the data label skewness affects the task learning and reveals the strength of each algorithm given a peculiar data distribution.

Further analyses

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

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

For the joint challenge paper, combining algorithms strengths and/or model ensemble may be evaluated.
Furthermore future direction of further works will be proposed following the algorithms strengths and weaknesses.

TASK: Placental Vessel Segmentation and Registration in Fetoscopy (FetReg)

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.

Fetoscopy Laser Photocoagulation (FLP) is a widely used procedure for the treatment of Twin-to-Twin Transfusion Syndrome (TTTS). TTTS occurs in approximately 10-15% of monochronic twins pregnancies. In TTTS, the flow of blood between the two fetuses becomes uneven as a result the donor experiences slow growth while the recipient is at risk of heart failure due to the excess of blood it takes. During FLP, the abnormal vascular anastomoses are identified, and laser ablated to regulate the flow of blood. The procedure is particularly challenging due to the limited field –of –view, poor manoeuvrability of the fetoscopy, poor visibility due to fluid turbidity and variability in light source, and unusual position of the placenta. This may lead to increased procedural time and incomplete ablation, resulting in persistent TTTS. Computer-assisted intervention can help overcome these challenges by expanding the fetoscopic field of –view and providing better visualization of the vessel map. This in turn can guide the surgeons in better localizing abnormal anastomoses.
We propose a challenge for placental vessel segmentation and registration for mosaicking in clinical fetoscopy. This challenge aims to provide a benchmark multi-centre dataset for placental vessel segmentation and registration. We aim to access consistent mosaics for fetoscopy video clips of sufficiently long duration. Participants will be provided with a large placenta vessel segmentation dataset with ground-truth annotations and fetoscopy video clips without any annotations for training and validation. The evaluation on unseen test data will be performed through docker submissions. The test data will not be released to the participants. This dataset, acquired across three different fetal medicine institutions, will be the first large scale multicentred data that will be made publicly available along with the challenge.
The two main sub-tasks of the FetReg challenge are: Sub-task 1: Placental vessel segmentation
Sub-task 2: Placental vessel registration and RGB frame registration for mosaicking

Keywords

List the primary keywords that characterize the task.

Fetoscopy; Segmentation; Registration; Video Mosaicking; Surgical Data Science

ORGANIZATION

Organizers

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

Sophia Bano1, Alessandro Casella2,3, Francisco Vasconcelos1, Sara Moccia3,4, Danail Stoyanov1
1 Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and Department of Computer Science, University College London, London, UK
2 Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy
3 Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy
4 Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy

b) Provide information on the primary contact person.

Sophia Bano: sophia.bano@ucl.ac.uk

Life cycle type

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

Examples:

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

grand-challenge.org

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

Part of https://endovis.grand-challenge.org/, sub-challenge site TBD

Participation policies

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

Fully automatic.

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

Publicly available data is allowed.

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

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

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

We will offer certificates and cash awards to the top three teams per task subject. 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.

All docker submissions will be evaluated on an unseen test dataset. The test dataset will not be made available to the participants. This data will be acquired using the same procedure as the training dataset with a different patient ID. The submitted results will be announced on the day of the challenge sorted in descending performance metric values.

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. A joint journal is intended to be published within 8 months of the challenge. Only the first author and the last author of the submitted paper of the proceeding will be considered and invited for a co-authorship. 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.

The submission would be made by docker containers. As mentioned by the MICCAI 2021 organisers, NVIDIA is offering to make cloud-based GPUs available, and we would like to use them for the online evaluation of the dockers submitted.
For the segmentation task, the algorithm should produce masks according to the format provided in the ground-truth and store them in a predefined directory.
For the registration task, the algorithm should return the homography estimation between all consecutive pairs of frames. This will be used to compute a stitched image and the consistency score calculated on 5 frames for each frame as presented in [Bano et al. MICCAI2020].
We will provide the script for computing the evaluation metrics.

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

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

Challenge schedule

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

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

Challenge website: 1st April 2021
Training data release - I (frames - vessel segmentation): 15th April 2021
Training data release – II (sequences – registration): 1st May 2021
Team registration open: 30th August 2021
Submission of docker images: 6th September 2021
Docker submission deadline: 17th September 2021
Methodology report submission: 17th September 2021 (only docker submissions accompanied by methodology report will be tested and included on the challenge day)
Spotlight presentation: Challenge day in MICCAI2021
Decision of FetReg challenge winners: Challenge day in MICCAI2021
Joint journal submission: March 2022

Ethics approval

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

All the data used in the training and testing of the challenge will be published for research and educational purpose. The data will be made openly available through a challenge website after the challenge ends.
We already have the ethics in place for releasing fully anonymised data in this domain for research purposes only. The dataset has been acquired from three different centres across Europe.

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 SA.

Additional comments: -

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.

Data preparation and evaluation codes will be made available via GitHub.

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

Participants are encouraged to make their code publicly available, but this is not a mandatory requirement.

Conflicts of interest

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

Only the organizers will have access to the test data and labels.
There is no conflict of interest.
We are currently looking for sponsors and we will update the organizers of MICCAI/EndoVis once it has been finalized.

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, Surgery, Assistance, Research, Education.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation, Registration, Video Mosaicking

Cohorts

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

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

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

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

Clinically acquired in vivo fetoscopy videos.

Imaging modality(ies)

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

Fetoscopy TTTS procedure videos captured from fiber-optic fetoscope(s).

Context information

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

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

No further information other than image data will be provided.
• The annotations provided for segmentation sub-task will include binary masks displaying the placental vessels.
• No ground-truth will be provided for the registration task.

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.

Fetoscopic laser photocoagulation (FLP) procedure videos that captured the fetal side of the placenta through a fetoscopic camera. The procedures were performed for the treatment of Twin-to-Twin-Transfusion Syndrome (TTTS).

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.

Same as 19 (a).

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)).

Segmentation sub-task: Mean Intersection over union (mIOU).
Mosaicking and registration task: 5-frame consistency score from [Bano et al., MICCAI2020]

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).

Karl Storz straight and 30 degrees curved fetoscope(s).

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

Multi-centre data captured at three European sites reflecting real clinical practices. The dataset captures the environment variability present among different patients. The dataset also captures the variability introduced by different light source settings and video compression.

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.

• Fetal Medicine Unit, University College London Hospital, London, UK (UCLH)
• Department of Fetal and Perinatal Medicine, Istituto Giannina Gaslini, Genoa, Italy (IGG)
• Department of Development and Regeneration, University Hospital Leuven, Leuven, Belgium (UZLeuven)
UCLH is contributing 6-9 recordings.
IGG is contributing 6-9 recordings.
UZLeuven is contributing 3-4 recordings.
Training data will contain at least 3 videos from each centre. Testing data will contain at least 1 video from each centre.
A part of our dataset release will include the dataset that we previously released in Bano et al. [MICCAI2020] paper. This dataset is available publicly: https://www.ucl.ac.uk/interventional-surgical-sciences/fetoscopy-placenta-data

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

Expert Obstetricians and Fetal Surgeons were involved in acquiring all the data since it is a rare and delicate procedure. The dataset was acquired during surgical procedures and was not specifically designed for this challenge.

Training and test case characteristics

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

Examples:

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

Both training and test cases represent fetoscopy videos from several fetoscopy laser photocoagulation procedures, each performed on a different patient. The data will be from approximately 20 fetoscopy videos.
Training cases for the segmentation (sub-task 1) will be annotated for vessel presence. Each training frame will have a respective ground-truth binary vessel mask. Fetoscopy video clips will be provided for the unsupervised registration and mosaicking (sub-task 2) during training. No ground-truth will be provided for sub-task 2.
For generalization, test cases will be from totally unseen fetoscopy videos (not included in training case).

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

Sub-task 1: Vessel segmentation
Training data: 1500 frames from 15 procedures with ground-truth masks.
Validation: Participants can choose train-validation split from the training data.
Testing data: 500 frames from 5 procedures with ground-truth masks. Test data will not be released but will be evaluated using the submitted docker containers.
Sub-task 2: Registration and mosaicking
Training data: 30 clips from 15 procedures without any ground-truth annotations.
Testing data: 10 video clips from 5 procedures without any ground-truth annotations. The 5-frame consistency [Bano et al. MICCAI2020] will be used as a metric for the unsupervised evaluation. Test data will not be released but will be evaluated using the submitted docker containers.

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

These numbers are chosen to balance a good trade-off in annotation effort while introducing sufficient visual diversity. Moreover, this procedure is not prevalent unlike other endoscopic procedures and only limited data is available. Moreover, not all recordings are suitable for annotation and algorithm development due to some having extremely poor visibility, low resolution and heavy compression.

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

none

Annotation characteristics

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

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

Fetoscopy videos acquired are decomposed into frames and excess black background is cropped to obtain squared images which are then resized to 448 x 448-pixel resolution. A subset of approx. 100 non-overlapping frames is selected from each procedure frames. These images are loaded in the pixel annotation tool (https://github.com/abreheret/PixelAnnotationTool) for vessel annotations. The manual annotation using this tool gives binary vessel maps. Four academic researchers and staff members are involved in the vessel annotations. Two fetal medicine specialists are involved in verifying the correctness of the annotations on a subset of data.

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

Annotation protocol
1. The annotators are instructed to use the same JSON that include the vessel label. This JSON is loaded in the annotation tool at the start of the annotation procedure.
2. The annotators are asked to colour over the visible vessels using the 'paint option' available in the tool.
3. The annotator can zoom in or out or change the paint head size to accommodate different vessel sizes.
4. All visible vessels in a frame are labelled by the annotators.

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.

Clinical experts: More than 10 years
Academic researcher and staff: more than 3 years of 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.

Not required. All annotators will agree on the annotation protocol prior to starting the annotations. Each image will be annotated by only one annotator. A sub-set of annotated images will be verified by the clinical expert in the initial phase of the annotation process. From our previous experience of annotating ~500 placental vessels [Bano et al. MICCAI2020] where we followed the same strategy of verification as mentioned above, we observe strong agreement between the annotator and the expert.

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.

Fetoscopy videos acquired are decomposed into frames and excess black background is cropped to obtain square images which are then resized to 448 x 448-pixel resolution.

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.

Not applicable. All annotators and experts thoroughly review the annotations before agreeing on the final annotation.

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

none

ASSESSMENT METHODS

Metric(s)

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

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

• Sub-task 1 - Segmentation: Mean Intersection over union (mIOU).
Semantic segmentation already has well-defined evaluation metrics that have been used in the relevant literature. mIOU is the most effective one and will be used for this task.
Suppl 4: Structured description of a challenge design
• Sub-task 2 – Registration and Video Mosaicking: 5-frame consistency score from [Bano et al. MICCAI2020] which measure the structural similarity index (SSIM) for vessel maps and Photometric Error (PE) for RGB images.
The major challenge with registration in intraoperative surgical images is the unavailability of the groundtruth. For the specific case of fetoscopy, a monocular camera is used which limits having any additional information that could be relevant for groundtruth generation. Recent work [Bano et al. MICCAI2020] has proposed a metric for evaluating the consistency of generated mosaics quantifying the accumulated drift error which commonly occurs during mosaicking. We plan to use the 5-frame consistency score as proposed by [Bano et al. MICCAI2020] for evaluating the obtained homographies. It is worth noticing here that obtaining manual groundtruth homographies is nearly impossible in the in vivo fetoscopy videos due to poor visibility, texture paucity and low resolution of the images.

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

• The selected metrics are the standard ones widely used for segmentation evaluation.
• The 5-frame consistency metric is the only known metric for evaluating the video frame registration error in the absence of ground-truth.

Ranking method(s)

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

• Sub-task 1- Segmentation: The highest value of the mIOU on the overall test dataset will be used to declare the winner.
• Sub-task 2 – Registration and Mosaicking: Highest 5-frame consistency score with minimum standard deviation, which will be the equally weighted sum of SIMM and PE metrics, will be declared the winner.

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

not allowed

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

The segmentation metrics are used to capture the overlap between the ground-truth and predicted vessel maps. Quantitative evaluation of registration and mosaicking is difficult due to the absence of the groundtruth homographies. The use of 5-frame consistency metric allows for capturing the drifting error in nearby frames. This quantifies the consistency of the generated mosaics.

Statistical analyses

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

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

Intended post challenge

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

Intended post challenge

Further analyses

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

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

Intended post challenge

TASK: PEg TRAnsfer Workflow recognition by different modalities (PETRAW)

SUMMARY

Abstract

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

Automatic and online recognition of surgical workflows is mandatory to bring computer-assisted surgery (CAS) applications inside the operating room. According to the type of surgery, different modalities could be used for workflow recognition. In the case of robotic-assisted surgeries and virtual reality training sessions, video and kinematic data are easily available. However, even if these modalities are available, numerous methods on state of art only focus on one of them. Some methods only used video-based method [1], [2], others only kinematic data [3], [4]. Last year we submitted the MISAW sub-challenge as part of EndoVis at MICCAI2020 (challenge description and results available at https://www.synapse.org/MISAW, paper under writing), offering to participant to combine both modalities for the workflow recognition. To the best of our knowledge, there are very few studies about the added value to combine multiple modalities. Whereas, some study as demonstrate that the addition of external information improve the recognition, as the presence of tools [5].

Segmentation of the surgical scene is very important for surgical understanding and an active area of research. For example, in 5 editions of EndoVis has proposed 6 sub-challenges based on this topic. However, to the best of our knowledge, semantic segmentation was never used for surgical workflow recognition.

The "PEg TRAnsfer Workflow recognition by different modalities" (PETRAW) sub-challenge provides a unique dataset for online automatic recognition of surgical workflow of a peg transfer training session. The objective of peg transfer session is to transfer 6 blocks from the left to the right and back. Each block must be extracted from a peg with one hand, transferred to the other hand, and inserted in a peg at the other side of the board. The dataset contains video, kinematic, and segmentation data for at least 100 sequences. Participants are challenged to recognize all levels of granularity of the surgical workflow (phases, steps, and action verb) with different modalities configurations. Participants can submit results for uni-modality-based models (video-based model, kinematic-based model or semantic segmentation-based model) and multi-modality-based models (video + kinematic based model or video + kinematic + semantic segmentation-based model). Each model has to be able to recognize all granularities in the same model. In the case of models using semantic segmentation information, and to reflect the fact that this modality is rarely available, the participants are challenged to use the output of a segmentation method as input for PETRAW.

Keywords

List the primary keywords that characterize the task.

Surgical Process Model, Workflow recognition, Multi-modality, Multi-granularity, Semantic segmentation

ORGANIZATION

Organizers

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

Arnaud Huaulmé, Univ Rennes, INSERM, LTSI - UMR 1099, F35000, Rennes, France
Kanako Harada, Department of Mechanical Engineering, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
Pierre Jannin, Univ Rennes, INSERM, LTSI - UMR 1099, F35000, Rennes, France

b) Provide information on the primary contact person.

Arnaud Huaulmé: arnaud.huaulme@univ-rennes1.fr

Life cycle type

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

Examples:

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

Main website: www.grand-challenge.org
Docker submission: www.synapse.org

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

Part of https://endovis.grand-challenge.org/, sub-challenge site www.synapse.org/PETRAW (not created yet)

Participation policies

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

Fully automatic.

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

Publicly available data is allowed.

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

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

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

The award policy will be dependent on the ranking on each task (see item 27). Challenges prizes depending on the

availability of sponsorship.

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.

According to the number of participants, all or the top five performing methods will be announced publicly during the challenge day. The remaining teams could decide whether or not their identity should be publicly revealed (e.g. in the 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 participating teams that reveal their identity can nominate two members of their team as co-authors for the challenge publication.
The method description submitted by the participant will be used in challenge publication. Personal data of the participant will include their names, affiliation, and contact addresses. References used in the method's description may be published in the challenge results as well.
Participating teams may publish their own results separately with an explicit allowance from the challenge organizers once the challenge publication has been accepted for publication.

Submission method

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

Examples:

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

Participants have to create a synapse project named "PETRAW_YourTeamName", with YourTeamName as the name of the team. To be valid, a submission must contain the following elements: A docker image, a method write-up. The docker image will take two local folders as input and provide an output folder where the algorithm outputs will be stored. Algorithm output must provide for each test case and each timestamp the recognized value separated by a tabulation. Results must provide phase, step, left-hand, and right-hand action verbs for each timestamp. In the case of tasks based on segmentation modality, the algorithm must also provide as output the segmentation of the test cases.
Please note, that according to MICCAI 2021 policy, additional elements could be asked to have a valid submission. For example, for MICCAI 2020, a pre-record talk was also asked. Participants will be informed as soon as the information will be known by the organizers.

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.

Submission of multiple docker images is possible, however, we will not provide results or leaderboard to participants before the challenge day. Only the last run is officially counted to compute challenge 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

- May 31st: Opening of the registration and release of the training cases.
- August 1st: Opening of the submission.
- September 12th (23:59 PST): Registration & Submission deadline.
- September 27th or October 1st: challenge day.
- After the challenge day: release of the results

Ethics approval

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

Not applicable. Data does not include patient information.

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.

Additional comments: Publicly available for non-commercial use after the challenge and the challenge paper submission.

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.

Organizers' evaluation scripts will be publicly available (open access) after the challenge.

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.

This work was partially funded by ImPACT Program of Council for Science, Technology and Innovation, Cabinet Office, Government of Japan.
All challenge organizers and some members of their institute had access to training and test cases. Therefore, there may participate but are not eligible for awards.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Training, Decision support, Surgery, Assistance, Research, Education.

Task category(ies)

State the task category(ies).

Examples:

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

Five workflow recognition tasks divided into two categories:
Uni-modality surgical workflow recognition (video only, kinematic only, semantic segmentation only).
Multi-modality surgical workflow recognition (video + kinematic, video + kinematic + semantic segmentation).
All models have to recognize all granularities: phases, steps, action verb of the left hand, and action verb of the right hand.

Please note that for models using semantic segmentation modality, the workflow recognition model must use the output of a segmentation method. The results of this segmentation will be assessed but not be rank. This challenge focus on workflow recognition not on 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.

Training sessions on peg transfer task performed on virtual reality simulator or real platform.

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

Training sessions on peg transfer task performed on virtual reality simulator.

Imaging modality(ies)

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

2D videos

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 context information is given along with the images

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.

On the final application, data would be acquired on any type of peg transfer training 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.

The algorithm target is the automatic workflow recognition of several granularity levels (phases, steps, and action verbs) on peg transfer training tasks.

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)).

Robustness, Accuracy.

Additional points: Robustness, Accuracy.
Additional points: The algorithms must be applicable for online applications. The recognition has to be accurate, robust, and reproducible.
Please note that the semantic segmentation used to train workflow recognition model will be assessed but not ranked for this challenge, only the workflow recognition will be.

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 VR simulator used for the data acquisition was developed at the department of mechanical engineering at the University of Tokyo [6] and is composed of: a core laptop (i7-700HQ, 16Go RAM, GTX 1070), a 3D rendering setup (3D screen (24 inches, 144Hz) and 3D glasses), Two user interfaces.
The Kinematic and scene information were acquired synchronously acquired at 30 Hz during the simulation of the task. The 2D video and semantic segmentation were generated after the session thanks to scene information. Workflow annotation was automatically computed thanks to scene information by ASURA. All information about the automatic annotation is described in [7]. Since this publication, the information provided by the VR simulator has been enhanced to be able to differentiate correctly the action verb “catch” and “touch”. Annotations were manually check by organizers to ensure that no issue exists.

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

Kinematic data represent the position, rotation quaternion, Forceps aperture angle(degrees), Linear velocity (obtained from simulation, not derived from position), and Angular velocity (obtained from simulation, not derived from orientation) for left and right tool. Position and linear velocity are in centimeters, angle and angular velocity in degrees.
Video and semantic segmentation data were generated at posterior thanks to the scene information at a resolution of 1920x1080.
Workflow annotation details were provided at item 23.

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

All data was acquired/computed by the MediCis team, of LTSI Laboratory from the University of Rennes, thanks to a virtual reality simulator developed by the department of mechanical engineering of the University of Tokyo.

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).

All cases were acquired by a non-medical person expert in the VR simulator manipulation. The data acquisition was divided by sessions of 5 cases performed. A minimum of 5 hours rest time was observed between to session to limit the fatigue of the subject.
This choice was made to have the same level of expertise for all cases. Moreover, the COVID-19 crisis does not allow us to recruit multiple participants with enough availability to have a comparable number of cases by subjects.

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 are composed of:
Kinematic data of left arm (px_l, py_l, pz_l, q1_l, q2_l, q3_l, q4_l, ape_angle_l, lin_velo_x_l, lin_velo_y_l, lin_velo_z_l, ang_velo_x_l, ang_velo_y_l, ang_velo_z_l at 30hz). Same for right arm noted px_r, ...
2D video data (1920x1080, 30Hz)
Semantic segmentation (1920x1080, 30Hz). Six classes are presented: background (Hexadecimal code: 0000FF), base (FFFFFF), left tool (FF0000), right tool (00FF00), pegs (000000), blocks (FF00FF).
Workflow annotation, containing for each timestamp (30Hz) the label for phase, step, left-hand and right-hand action verbs.

The test cases will not be provided to the participants. Participants have to generate the semantic segmentation of test cases for the tasks based on this modality. The segmentation is provided for training cases in order to train segmentation model.

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

The minimum number of cases for training will be 60, and 40 for testing. This number is subject to increase depending on the feasibility of acquiring more data. Currently (12/07/2020), the COVID19 crisis only allows us to record 80 cases.
No validation cases are provided. It is up to the participants to split the training dataset into training and validation data.

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

The minimum number of cases was chosen to represent the generic way to realize the peg transfer task, but also to include some errors in the process (e.g. drop a peg on the board). The repartition was choosing to have 60% of data for training 40% for testing. This repartition will be kept in case of an increase of the cases number.

To limit the effect of immediate learning or fatigue inside a single session (5 cases recorded by session), 3 cases of one session were randomly chosen for training and the 2 others for testing.

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.

Participants have to generate the semantic segmentation of the test case to reflect the difficulty to have this information on a real application. Participants could choose the number of classes they need, e.g. only one class for the 6 blocks.

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.

Workflow annotation was performed automatically thank to ASURA [7] and manually checked to ensure that no issue exists.

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.

NA

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.

NA

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.

Workflows annotation was modified to switch from continuous sequence to a discrete sequence at 30Hz (synchronize to kinematic and video data). For each timestamp, the annotation would like as follow: timestamp_number, phase_value, step_value, action_verb_Left_Hand_value, action_verb_Hand_value. With a tabulation as separator

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 automatic annotation is depending of the scene information provided by the simulator. Despite the precautions taken (improvement has done since the publication of the method), it is possible that this information may not be sufficiently discriminating for ASURA to properly differentiate certain actions. However, due to the replicability of the method, if one of this case appears, it will be always interpreted in the same way by ASURA.

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)

The assessment for the workflow recognition will used the following metrics.
Frame by frame scores: balanced accuracy, precision, recall, f1 for all class.
Application-dependent scores: balanced accuracy, precision, recall, f1 for all class.

Even if the segmentation will not be rank, we will use the following metric to asses the algorithm used: Mean Intersection-Over-Union.

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

Balanced accuracy: global result of the recognition by taking into account unbalanced class.
Precision, Recall, f1: Local results of the recognition to take into account under-represented classes.
Frame by Frame scores: classic metric used for accuracy precision recall and f1.
Application-dependent (AD) scores: The frame by frame scores is not representative of the time precision needed for an application. AD-scores re-estimate classic scores by using acceptable delay thresholds for a transitional window [5].
Mean Intersection-Over-Union: classic metric used for semantic segmentation assessment.

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.

To perform the ranking, we will use a metric-based aggregation on the balanced Application-Dependent accuracy. For one participant, we will aggregate metric values over all test cases, and aggregate overall metrics to obtain a final score.

The tasks consist of recognizing 4 different elements (phase, steps, left-harm and right-harm action verb) the score s for a recognition algorithm (a_i) will be the mean of the score for each element s_elem:
Linear equation: s (a_i)= (s_phase(a_i)+s_step(a_i)+s_left_verb(a_i) +s_right_verb(a_i))/4
Scientific equation:

$$s_{multi}(a_i) = \frac{s_{phase}(a_i) + s_{step}(a_i) + s_{left\_verb}(a_i) + s_{right\_verb}(a_i)}{4}$$

With s_elem for a recognition algorithm (a_i) :
Linear equation: s_elem (a_i)=sum(balanced_AD_accuracy_trial_t)/T ; with sum from t=0 until T.
Scientific equation: s_elem (a_i)=(_(t=0)^Tbalance_AD_accuracy_trial_t)/T

$$s_{elem}(a_i) = \frac{\sum_{t=0}^{T} balance\_AD\_accuracy\_trial\_t}{T}$$

One ranking will be made by task.
The semantic segmentation of the test cases will not be rank.

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

The tasks consisting of recognize 4 different elements. In the case of missing results for one element, we will consider results as good as a total random recognition. For example, if an element is a 3 classes problem, the missing result will be set to 1/3. For a 12 classes problem, to 1/12.

For tasks based on semantic segmentation granularity, if participants are not able to provide segmentation of the test cases, their participation in these tasks will be canceled.

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

We decided to use a metric-based aggregation according to the conclusion of [8] reporting this type of aggregation as one of the most robust.

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.

Ranking stability will be investigated via bootstrapping and by comparing ranking method. The statical analysis will be performed ChallengeR package provided by [9].

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

Bootstrapping was identified by [8] as an appropriate approach to investigate ranking 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: Objective Surgical Skill Assessment in VR Simulation (SimSurgSkill)

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.

Timely and effective feedback within surgical training plays a critical role for surgeons in development of essential skills required to perform safe and efficient surgery. Due to the typical busy schedules of expert surgeons, mentor feedback can be challenging to receive. There is an opportunity for automated feedback on technical skills through objective metrics during surgical training. Most related work focuses on developing algorithms to predict subjective assessment scores like OSATS or GEARS. This challenge is aimed at developing automated skills assessment algorithms using virtual reality (VR) based surgical tasks and objective metrics that are provided directly from the virtual environment (e.g., needle drops, excessive force, etc.). The ability to design algorithms to predict these objective metrics might enable meaningful automated surgical score reports.

Keywords

List the primary keywords that characterize the task.

surgical training; skill assessment; virtual reality

ORGANIZATION

Organizers

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

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

b) Provide information on the primary contact person.

Aneeq Zia: aneeq.zia@intusurg.com

Life cycle type

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

Examples:

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

One time event with fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

synapse.org

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

Part of https://endovis.grand-challenge.org/, sub-challenge site TBD

Participation policies

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

Fully automatic.

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

Publicly available data is allowed.

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

May not participate.

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 to TBD

e) Define the policy for result announcement.

Examples:

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

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

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

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

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

Submission method

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

Examples:

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

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

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

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

Challenge schedule

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

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

Release of training cases in June 2021; registration ongoing; submission date September 2021; release of results at MICCAI 2021

Ethics approval

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

An existing Western IRB will be used.

Data usage agreement

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

Examples:

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

CC BY NC ND.

Additional comments: -

Code availability

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

Open source on challenge site

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

Open and private code submission will be accepted

Conflicts of interest

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

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

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Training, Surgery, Research.

Task category(ies)

State the task category(ies).

Examples:

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

Classification, Detection, 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.

Basic surgical tasks during surgical training; subjects will be surgeons of varying experiences

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

Virtual reality exercise of the same activities as the target cohort; subjects will be surgeons and nonsurgeons with varying experience operating the robot

Imaging modality(ies)

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

one channel of endscopic 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).

Each video clip will come with ground truth objective surgical skills based metrics

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 VR using the da Vinci simulator on various basic training and procedural steps.

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 objective metrics relating to surgical skills for tasks being performed in VR. There will be total of 4-5 metrics to predict.

Assessment aim(s)

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

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

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

Accuracy.

Additional points: The assessment will be based off of closeness in ground truth objective metrics and predicted objective metrics which will be measured through correlation coefficient and mean absolute error.

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 surgical simulator.

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 and potentially hospitals

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 range from beginners (early in their learning curve) to experts (practicing 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.

Our dataset will consist of 3-4 surgical tasks in VR that are available in the da Vinci simulator. Training and testing cases will all be in VR. One case will consist of a video clip captured from the da Vinci simulator and will be accompanied by objective metrics available through the report presented to the participant at the end on the console.

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

The total dataset will comprise of 150 or more VR videos of different tasks. The tasks being performed will range from basic ones like suturing/knot tying to procedural steps like anastomosis in prostatectomy. From the 150 videos, 100 will be kept as training with 25 for validation and testing set each.

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 metric values within the training/validation/testing set

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.

Ground truth will be extracted from the performance report displayed after performing any task in the da Vinci simulator. No human annotation will be needed in this dataset.

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.

NA

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.

NA

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

NA

Data pre-processing method(s)

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

Raw video frames will not be altered

Sources of error

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

no image annotation to be performed, hence not applicable

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)

correlation coefficient and mean absolute error between ground truth and predicted metrics will be used for assessment

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

These metrics have been extensively used to report results for regression problems in literature

Ranking method(s)

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

The performance rank will be based on the rank of average of correlation coefficient and mean absolute error

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

Missing results will be penalized and given a 0 score for those cases

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

Using correlation coefficient/absolute errors for ranking seems most reasonable for such a regression problem.

Statistical analyses

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

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

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

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

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

NA

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.

Task 3 (CholecTriplet2021):
Nwoye, Chinedu Innocent, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. "Recognition of instrument-tissue interactions in endoscopic videos via action triplets."
In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 364-374. Springer, Cham, 2020.

Task 5 (PETRAW):
[1] D. Sarikaya and P. Jannin, "Surgical Gesture Recognition with Optical Flow only," arXiv, Apr. 2019, Accessed: Nov. 19, 2020. [Online]. Available: http://arxiv.org/abs/1904.01143.
[2] I. Funke, S. Bodenstedt, F. Oehme, F. von Bechtolsheim, J. Weitz, and S. Speidel, "Using 3D Convolutional Neural Networks to Learn Spatiotemporal Features for Automatic Surgical Gesture Recognition in Video," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, vol. 11768 LNCS, pp. 467–475, doi: 10.1007/978-3-030-32254-0_52.
[3] F. Despinoy et al., "Unsupervised trajectory segmentation for surgical gesture recognition in robotic training," IEEE Trans. Biomed. Eng., vol. 63, no. 6, pp. 1280–1291, Aug. 2015, doi: 10.1109/TBME.2015.2493100ï.
[4] R. DiPietro and G. D. Hager, Automated Surgical Activity Recognition with One Labeled Sequence, vol. 11768 LNCS. Springer, 2019, pp. 458–466.
[5] O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, and P. Jannin, "Automatic data-driven real-time segmentation and recognition of surgical workflow," Int. J. Comput. Assist. Radiol. Surg., Oct. 2016, doi: 10.1007/s11548-016-1371-x.
[6] S. . Heredia Perez, K. Harada, and M. Mitsuishi, "Haptic Assistance for Robotic Surgical Simulation," 27th Annu. Congr. Japan Soc. Comput. Aided Surg., vol. 20(4), pp. 232–233, Nov. 2018.
[7] A. Huaulmé, F. Despinoy, S. A. Heredia Perez, K. Harada, M. Mitsuishi, and P. Jannin, "Automatic annotation of surgical activities using virtual reality environments," Int. J. Comput. Assist. Radiol. Surg., vol. 14, no. 10, pp. 1663–1671, Jul. 2019, doi: 10.1007/s11548-019-02008-x.
[8] L. Maier-Hein et al., "Why rankings of biomedical image analysis competitions should be interpreted with care," Nat. Commun., vol. 9, no. 1, p. 5217, Dec. 2018, doi: 10.1038/s41467-018-07619-7.
[9] M. Wiesenfarth, A. Reinke, B. A. Landman, M. J. Cardoso, L. Maier-Hein, and A. Kopp-Schneider, "Methods and open-source toolkit for analyzing and visualizing challenge results," arXiv, Oct. 2019, Accessed: Nov. 18, 2020. [Online]. Available: http://arxiv.org/abs/1910.05121.

Further comments

Further comments from the organizers.

EndoVis serves as an umbrella for different kinds of sub-challenges that tackle specific problems and applications in endoscopic/microsopic vision, this year we have 6 sub-challenges described as tasks in this proposal (HeiSurf, GIANA, CholecTriplet2021, FetReg, PETRAW, SimSurgSkill).