Deep Generative Model Challenge for Domain Adaptation in Surgery 2021: Structured description of the challenge design

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

CHALLENGE ORGANIZATION

Title

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

Deep Generative Model Challenge for Domain Adaptation in Surgery 2021

Challenge acronym

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

AdaptOR 2021

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.

Mitral regurgitation (MR) is the second most frequent indication for valve surgery in Europe and may occur for organic or functional causes [1]. Mitral valve repair, although considerably more difficult, is prefered over mitral valve replacement, since the native tissue of the valve is preserved. It is a complex on-pump heart surgery, often conducted only by a handful of surgeons in high-volume centers. Minimally invasive procedures, which are performed with endoscopic video recordings, became more and more popular in recent years. However, data availability and data privacy concerns are still an issue for the development of automatic scene analysis algorithms. The AdaptOR challenge aims to address these issues by formulating a domain adaptation problem „from simulation to surgery“: We provide a smaller number of datasets from real surgeries, and a larger number of annotated recordings of training and planning sessions from a physical mitral valve simulator. The goal is to reduce the considerable domain gap between simulation and intraoperative cases, e.g. by incorporating generative models, as in [2,3].
The task associated to the domain adaptation itself is to detect a varying number of 2D landmarks per frame [4] in the target domain. The landmarks are defined by the placement of sutures during mitral annuloplasty (entry and exit points into the tissue), which renders useful for surgical skill assessment and detailed intraoperative documentation. The evaluation metrics of this challenge will be related to how well these points could be identified in unseen intraoperative scenes, therefore it is also possible to only come up with a solution to a landmark detection problem in a single domain. More complex methods, however, would leverage data from both domains and adapt them on input-, output-, and/or feature level.
Due to the specific clinical motivation of improving the realism of surgical simulation [2,3], the AdaptOR challenge especially aims to provide a framework for comparison of the performance of different image-to-image

translation approaches. Such approaches need to learn how to sucessfully transform the images into an intraoperative appearance, thereby not altering already realistic entities of the image (surgical instruments, sutures, needles etc.). While this can be merely assessed visually, and we will show example results during the workshop, we hypothesize that the success of landmark detection may be an indicator for the quality of the transfer with respect to the consistency of sutures in both domains.

Challenge keywords

List the primary keywords that characterize the challenge.

Generative Models, Generative Adversarial Networks, Domain Adaptation, Domain Transfer, Landmark Detection, Mitral Valve, Heart Valve, Endoscopy, Surgical Training, Surgical Simulation

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.

Deep Generative Models for Medical Image Computing and Computer Assisted Intervention 2021 (DGM4MICCAI)

Duration

How long does the challenge take?

Half 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.

We expect about 10 teams to participate.
Our estimation is based on number of participants in other related challenges:
•M&M 2020, which was a cardiac challenge on domain adaptation (14 teams)
•Past EndoVis 18, 19, 20, which are challenges related to endoscopy (about 5-10 teams)

Based on our previous publications on this use-case [2,3] and related use-cases [5,6], we have received several requests from researchers internationally to use our code and/or dataset. Therefore we assume that there will be enough interest by the community.

Publication and future plans

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

Participants will submit an 8-page paper on their methods adhering to the same schedule of the workshop. The paper will be published in the LNCS proceedings. After the challenge took place, a journal paper (preferable TMI or MedIA) will be submitted to summarize the challenge results. Two authors of each team will be invited as co-

authors.

Our envisioned goal is to extend the dataset with additional cases and potentially establish a recurring AdaptOR event to support progress in this application field. In 2022, we would like to focus on stereo tasks and method generalization.

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.

The challenge will not be on-site. The online challenge will use the synapse platform.

TASK: Domain Adaptation for Landmark Detection

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.

described above

Keywords

List the primary keywords that characterize the task.

Domain Adaptation, Generative Models, Landmark Detection, Deep Learning, Machine Learning

ORGANIZATION

Organizers

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

Jun. Prof. Dr. Sandy Engelhardt, University Hospital Heidelberg
Dr. Anirban Mukhopadhyay, Technical University Darmstadt
Prof. Dr. Raffaele De Simone, University Hospital Heidelberg
Lalith Sharan, University Hospital Heidelberg
Antonia Stern, University Hospital Heidelberg
Julian Brand, University Hospital Heidelberg
Henry Krumb, Technical University Darmstadt

b) Provide information on the primary contact person.

Sandy Engelhardt, University Hospital Heidelberg
Email: sandy.engelhardt@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.

Docker submission: https://docs.synapse.org/
The challenge will be linked on grand-challenge.org (will be done once the proposal is accepted).

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

https://adaptor2021.github.io/ (site under construction)

Participation policies

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

Fully automatic.

Additional points: Only fully automatic approaches are allowed

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

No additional data and no models pre-trained on other datasets are 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.

Members of the organizers' institutes may participate in the challenge but are not eligible for awards.

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

Certificates will be provided for the top 3 performing teams. Upon acceptance of the challenge, we will seek for sponsorship of the winner(s) of the challenge from industry partners.

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 the results will be made available publicly. All teams will be invited to the half-day challenge event at DGM4MICCAI workshop to present their work in more detail.

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).

Challenge submission should be accompanied by an 8-page LNCS format paper, describing individual methods by the participants in detail. The paper will be published in the „Deep Generative Model" workshop proceedings after the workshop. There are no restrictions on the number of authors.

After the challenge, the challenge organizers will publish one challenge journal paper together with two participants of each challenge team summarizing the results. Each team should nominate two authors (typically the first and last author). An embargo period until the availability of this journal paper will be put in place.

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 challenge cohort splits into two sets:
1) data acquired during simulating mitral valve repair on a surgical simulator ("Sim-Domain"),
2) intraoperative endoscopic data from mitral valve repair ("Intraop-Domain").

For the purpose of result verification and to encourage reproducibility and transparency, all entries must submit the following:

• Docker container on the Synapse platform. More information will be provided on the website.
- (mandatory:) When running a pre-defined command on novel input data from the Intraop-Domain, the model should output a JSON file, which should include the input file name and a list of the x- and y-coordinates of the detected landmarks.
- (optional:) When running a pre-defined command on a Sim-Domain image, the model should output an image, which was transformed into the Intraop-Domain and vice versa. These results do not play a role in the final rankings, but should provide insights on the quality of image-to-image transformation. Example images will be shown during the workshop event and in the joint publication. Depending on the number of submissions, an additional user-study with domain experts might be conducted at a later stage.
• We encourage participants to provide their code open source. The URL should be added in the LNCS submission.

• Participants agree that the challenge organizers are allowed to use their submitted docker containers to run further meta-analysis.

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 challenge will be split into three phases: Training phase, Platform testing phase, Testing phase.

During training phase, the participating teams will be able to independently validate their results using cross-validation on the training data.
During the platform testing phase, they are allowed to use the official submission platform to resolve potential technical issues. We will use dummy datasets for sanity checks, e.g. to ensure the submission is in the correct format.

During the test phase, participants are allowed to make in total three submissions. The best result out of these three is selected as final result.

Furthermore, violation to the following rules will lead to disqualification:

- Each team is only allowed to register once and all submissions must be done from the same account.
- A single participant is only allowed to be part of one 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 date of the training cases: 01/04/2021
• registration date: until 30/05/2021
• platform testing: 1/06/2021-15/06/2021
• submission of docker container: 15/06/2021-07/07/2021
• LNCS paper submission date: 15/07/2021
• associated workshop days: either 27/09/2021 or 1/10/2021
• release date of the results: date of the workshop

(subject to change depending on the MICCAI 2021 deadlines)

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).

We have received approval from the Local Ethics Commitee from University Hospital Heidelberg to use the anonymized data. The registration numbers are S-658/2016 and S-777/2019.

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: By registering in the challenge, each team agrees (1) to use the data provided only in the scope of the challenge and (2) to neither pass it on to a third party nor to use it for any additional publication or for commercial use. After the challenge, the data will be made publicy available for non-commercial use.

Additional comments: By registering in the challenge, each team agrees (1) to use the data provided only in the

scope of the challenge and (2) to neither pass it on to a third party nor to use it for any additional publication or for commercial use. After the challenge, the data will be made publicy available for non-commercial use.

Code availability

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

We will make the code available on the synapse platfrom that will be used to compute the metrics for ranking.

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

Teams are encouraged to provide their code open source and to add the URL in the LNCS paper.

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.

No conflict of interest. Only challenge organizing team will have access to test case labels during the challenge.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Surgery.

Additional points: Intraoperative Support

Task category(ies)

State the task category(ies).

Examples:

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

Detection.

Cohorts

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

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

Patients with mitral insufficiency, which undergo minimally-invasive mitral valve repair.

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

The challenge cohort splits into two sets:
1) data acquired during simulating mitral valve repair on a surgical simulator ("Sim-Domain“),
2) intraoperative endoscopic data from mitral valve repair ("Intraop-Domain").

Imaging modality(ies)

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

Endoscopy

Context information

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

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

• Image coordinates of the landmarks to be detected.
• To which (anonymized) patient and domain the frame belongs to

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

No further information.

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.

Mitral valve shown in endoscopy

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.

Sutures placed in the mitral valve annulus

Assessment aim(s)

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

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

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

Precision, Sensitivity.

Additional points: Find sutures with high sensitivity (TPR) and high precision (PPV) in endoscopic frames.

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).

Sim-Domain:

Image1S 3D 30 degree optics (Karl Storz SE & CO KG).
• Imaging systems used:
Image1 Connect TC200, with resolution of 1080x1920, 25 fps
... for a stereo-pair saved in top-down format.
• Recorders used:
-Karl Storz AIDA
-DVI2PCIe capture card with Epiphan video capture software

Intraop-domain:

Image1S 3D 30 degree optics (Karl Storz SE & CO KG).
• Imaging systems used:

-Image1 Connect TC200, with resolution of 1080x1920, 25 fps
-Image1S Connect TC200, 2160x3840, 25 fps
... for a stereo-pair saved in top-down format.
• Recorders used:
-Karl Storz AIDA

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

Sim-Domain: Data was acquired on the minimally invasive training simulator (modified housing from (MICS MVR surgical simulator, Fehling Instruments GmbH & Co. KG, Karlstein, Germany). Camera angle is mainly from the upper left side. Light intensities 85-100%.

Intraop-Domain: Data was acquired during minimally invasive surgery (rightlateral thoracotomy). The distance and the camera orientation with respect to the mitral valve depends on the anatomical conditions of the patient. Light intensities varied between 85-100%.

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

The data was acquired at University Hospital Heidelberg (in MIC training lab and ORs).

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

The data sets are anonymized. No further characteristics are 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.

One case refers to one frame extracted from the simulation sessions and the intraoperative videos.
Stereo-frames were saved in top-down format (left image top, right image bottom) and split and treated completely independent.
Relevant scenes of mitral annuloplasty were identified before extracting the frames from the entire video recordings. In scenes with rapid changes, every 10th frame was extracted, in scenes with only few changes, every 240th frame was extracted and in every other scene, every 120th frame was extracted.

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

Training Sim-Domain:

2708 mono frames from 10 simulations (192-374 frames each) with approx. 33500 annotated landmarks.

Training Intraop-Domain:

2376 mono frames from 4 patients (372-794 frames each) with almost 24000 annotated landmarks.

Testing Intraop-Domain:

500 mono frames from 5 patients (100 frames from each). Patients are different from the training set.

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

To reflect a real-world scenario, the split must be conducted on the level of the patients or simulations, respectively. The training-test ratio is 14:5, which means that 26% are used for testing. According to [5], the median ratio of training cases to test cases in past challenges is 0.75.

The idea behind the challenge is to keep the number of intraoperative patients low to force participants to incorporate the frames from Sim-Domain in the training process to achieve better generalization performance.

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 tried to relatively balance the number of frames per simulation/patient in the training data set. We completely balanced the number of frames in the test data per patient. Therefore, each patient in the test set has a similar influence on the final score.

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 ground truth was produced by two students with basic knowledge of the surgical steps. They both followed a pre-defined labeling strategy.

Training set: Annotations by student2 were additionally checked by student1 and vice versa.

Test set: Annotations were made by both students independently and the mean was computed to determine the final landmark position.

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

The annotation was conducted in temporal order and simultaneously on the left and right image of the stereo pair. Annotation was conducted with the program „label me“.

Besides the coordinates, the rater annotated whether he/she found it easy/medium/hard to annotate the frame. More details on the label strategy can be found on the website: https://adaptor2021.github.io/

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.

Students with technical background, who received a briefing about mitral annuloplasty.

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.

-

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.

Stereo-frames were saved in top-down format (left image top, right image bottom) and split and treated completely independent. Image format varied and was reduced to 512 x 288 to reduce computational costs.

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 suture when entering or exiting the tissue is usually not just a single point, but a small region. Therefore, annotations by different annotators may have a small variation. This is accounted for in the ground truth metric: The landmark counts as detected if it is in a certain radius around the annotated ground truth pixel.
A small interobserver study on 100 frames with 5 observers revealed that 84.2% of the observers were able to correctly identify and label a landmark. Note that this includes beside identification also labeling (i.e., providing a name), which is not a task in this challenge.

In some situations, the points of interests are slightly occluded by other sutures or instruments. In this situation, assessment of temporal information helped.

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

We have discarded intraoperative recordings with fogging scenes.

ASSESSMENT METHODS

Metric(s)

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

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

We will report true positives, false positives and false negatives.

A landmark is counted as true positive, if it lies within a radius of 6 pixels around the manually labeled point, same as in [4]. This accounts for the fact that the region, where the suture enters or exits the tissue, is usually a small region and not just a single pixel. Finally, we report Sensitivity (TPR) and Precision (PPV).

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

Both ratios of correctly detected landmarks against the true number of sutures and the predicted amount of sutures are equally important. It will implicitly provide information on true positives (prediction matches ground truth), false positives (algorithm detects wrong points), and false negatives (algorithm misses landmarks). All cases are relevant with respect to the potential application (e.g. automatic surgical skill 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.

Sensitivity and precision are computed over all landmarks in the test sets. It is not differentiated whether the prediction is particulary well for certain frames/patients/simulations and worse for others.
The balanced F-score (F1 score) presents the harmonic mean of precision and recall and will be used to determine the ranking (the higher the better).
We exclude false negative rate (FNR) in the ranking, since it is related to TPR by TPR = 1-FNR. In the case where all metrics are tied, we will accept to have multiple teams with the same ranking.

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

Not providing coordinates for a frame will increase the count for false negatives, therefore the sensitivity will decrease.

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

We regard sensitivity and precision as equally important, therefore the F_beta-score with beta=1 is used, which represents the harmonic mean.

Statistical analyses

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

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

We will assess the ranking variability with Kandall's tau analysis. In particular, we will investigate whether a region of slightly different size around the ground truth point will lead to different challenge rankings.

We will provide more insights into the variability of ranking; e.g. considering the arithmetic average of precision and recall and the more general F-score (F_beta) , with beta = 0.5, 2 etc.
The results during the challenge event will be reported with transparency.

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

Kendall's tau may quantify differences between rankings (1: identical ranking; -1: inverse ranking). However, even for high values of Kendall's tau, critical changes in the ranking may occur [7]. Therefore, we will additionally provide the complete alternative ranking lists.

Further analyses

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

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

In our surgical training scenario, there is a clinical need for transforming not-so-realistic phantom data into more realistic surgical images [2,3]. Therefore, we encourage the participants to use image-to-image translation approaches, however, this is not mandatory.

In general, we think the underlying detection task could be solved differently:
1) Training of a landmark detection approach only on the Intrap-Domain
2) Using the Sim-Domain for Pre-Training/Dataset Fusion
3) Incorporating the Sim-Domain by using a combination of more advanced input-, output-, feature-level domain adaptation approaches, possibly in an end-to-end training manner
4) Others.

The authors should detail on their approaches in their submitted LNCS papers.

In case an image-to-image translation task was solved, we will provide visual examples of the generative model's output for visual comparison. These results are qualitative and will not be considered in the ranking scheme. We hypothesize that the quantitative assessment for landmark detection may be an indicator for the quality of the domain transfer with respect to the consistency of sutures in both domains.

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.

[1] https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Volume-16/Mitral-valve-incompetence-epidemiology-and-causes

[2] Engelhardt S., De Simone R., Full P.M., Karck M., Wolf I. (2018) Improving Surgical Training Phantoms by Hyperrealism: Deep Unpaired Image-to-Image Translation from Real Surgeries. In: Frangi A., Schnabel J., Davatzikos C., Alberola-López C., Fichtinger G. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science, vol 11070. Springer, Cham, doi: 10.1007/978-3-030-00928-1_84

[3] Engelhardt, S., Sharan, L., Karck, M., De Simone, R., Wolf, I. (2019), Cross-Domain Conditional Generative Adversarial Networks for Stereoscopic Hyperrealism in Surgical Training. In: Shen D. et al. (eds) Medical Image

Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11768. Springer, Cham, pp 155-163, doi: https://doi.org/10.1007/978-3-030-32254-0_18

[4] Stern, A., Sharan, L., Romano, G., Koehler, S., Karck, M., De Simone, R., Wolf, I., Engelhardt, S., Heatmap-based 2D Landmark Detection with a Varying Number of Landmarks, Bildverarbeitung für die Medizin 2021 (accepted), arXiv:2101.02737

[5] Pfeiffer, M., Funke, I., Robu, R. M., Bodenstedt, S., Strenger, L., Engelhardt, S., Roß, T., Clarkson, M.J., Gurusamy, K., Davidson, B.R., Maier-Hein, L., Riediger, C., Welsch, T., Weitz, J., Speidel, S., Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. Shen D. et al. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11768. Springer, Cham, pp 119-127, doi: https://doi.org/10.1007/978-3-030-32254-0_14

[6] Sahu, M., Strömsdörfer, R., Mukhopadhyay, A., Zachow, S. (2020): Endo-Sim2Real: Consistency Learning-Based Domain Adaptation for Instrument Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, S. 784-794, Springer, 23rd International Conference on Medical Image Computing and Computer Assisted Intervention, virtual Conference, 04.-08.10., doi: 10.1007/978-3-030-59716-0_75

[7] Maier-Hein, L., Eisenmann, M., Reinke, A. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat Commun 9, 5217 (2018). https://doi.org/10.1038/s41467-018-07619-7

Further comments

Further comments from the organizers.