Automatic Structure Segmentation for Radiotherapy Planning Challenge 2020: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Automatic Structure Segmentation for Radiotherapy Planning Challenge 2020

Challenge acronym

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

StructSeg 2020

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.

We propose to hold the challenge of automatic structure segmentation for radiotherapy planning 2020 in conjunct with MICCAI 2020. We will provide data of two types of cancers, nasopharyngeal cancer and lung cancer. Four challenge tasks will be organized, including Organ-at-risk segmentation from head and neck CT scans, Organ-at-risk segmentation from chest CT scans, Gross Target Volume segmentation of nasopharynx cancer, Gross Target Volume segmentation of lung cancer. A total of 120 CT scans with more than 1520 organ or tumor annotations will be provided in the chanllenge.

Radiation therapy is one type of important cancer treatment for killing cancer cells with external beam radiation. Treatment planning is vital for the treatment, which sets up the radiation dose distribution for tumor and ordinary organs. The goal of planning is to ensure the cancer cells receiving enough radiation and to prevent normal cells in organs-at-risk (OAR) from being damaged too much. Organs-at-risk are usually the organs that are sensitive to radiation. For instance, optical nerves and chiasma in the head cannot receive too much radiation otherwise the patient risks losing his/her vision. Gross Target Volume (GTV) is the position and extent of gross tumor imaged by CT scanes, i.e. what can be seen. One important step in radiotherapy treatment planning is therefore to delineate the boundaries of tens of OARs and GTV in every slice of a patient's CT scans, which is tedious and occupies much of oncologists' time. Automatic OAR & GTV delineation would substantially reduce the treatment planning time and therefore reduces the overall cost for radiotherapy.

This is a re-holding of the successful StructSeg 2019 challenge with 20% more training data on GTV segmentation provided to the research community. In StructSeg 2019, 718 teams have registered and 34 teams submitted their final models for evaluation. The challenge talks and ceremony have attracted a large number of audience to learn about the top-ranking methods during MICCAI 2019.

By releasing the new training data, continuely running the online evaluation server, holding the challenge talks, the StructSeg 2020 is expected to continually attract much attention from the research community and advance the research on OAR and GTV segmentation significantly.

Challenge keywords

List the primary keywords that characterize the challenge.

Organ-at-risk Segmentation, Gross Target Volume segmentation, nasopharynx cancer, lung cancer

Year

The challenge will take place in ...

2020

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?

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.

1000 teams are expected to register and 50 teams are expected to submit their final models for evluation this year. The numbers are based on the participating teams in StructSeg 2019, where 718 teams registered and 34 teams submitted their final models.

The teams already expressed their willingness to parcipate into the challenge include Shanghai Jiao Tong University, Zhejiang University, University of Eletronic Science and Technology of China, etc.

Publication and future plans

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

There will be no coordination on publication. Each team can submit papers on their own.

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.

Only talks will be given at the challenge ceremony. Projectors, computers, monitors, loud speakers, microphones will be needed

TASK: Organ-at-risk segmentation from head and neck CT scans

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.

22 OARs of 60 nasopharynx cancer patients will be annotated and released to public as the training data. There will be a total of 1,100 annoated structures/organs (22 X 60 = 1320) in the training set. Each of the annotated CT scan is marked by one experienced oncologist and verified by another experienced one. Another 10 patients' CT scans will be used as the test data.

Keywords

List the primary keywords that characterize the task.

Organ-at-risk segmentation, head & neck CT

ORGANIZATION

Organizers

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

Hongsheng Li, Assistant Professor, The Chinese University of Hong Kong;

Ming Chen, Deputy Chief Director, Cancer Hospital of the University of Chinese Academy of Sciences.

b) Provide information on the primary contact person.

Hongsheng Li, The Chinese University of Hong Kong, hsli@ee.cuhk.edu.hk.

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.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event one time.

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 for publicity and own online submission site for final model submission.

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

None at this moment.

Participation policies

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

Fully automatic.

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

No additional data allowed.

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

May not participate.

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

The top-3 teams will be invited to present talks and winning certificates will be provided.

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-10 performing results will be made public.

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

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

We will not publish a challenge paper with submission methods. Each team can submit their paper on their own.

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 on webpage (same as StructSeg 2019). We will let the participants to submit their Docker submissions to our evaluation server. Any valid submission would be notified as "Valid submission with >1% accuracy". Any invalid submission would be informed with an error log file, telling participants why the submission fails to be evaluated.

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 are allowed. However, we will choose the highest submission from the latest three submissions of each team. Since we don't show the actual testing accuracy of each submission, but only show "Valid submission with >1% accuracy". The participants have no way to hack the test set.

Challenge schedule

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

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

Release of training data: Jun. 10th, 2020;
Submission deadline for results: Sept. 10th, 2020;
Announcement of final results: Sept. 15th, 2020.

Ethics approval

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

Ethics approval on data from StructSeg 2019 have been obtained. We only need to re-new the ethics approval for StructSeg 2020.

Data usage agreement

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

Examples:

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

CC BY NC SA.

Code availability

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

Evaluation code will be made public before the system is open for submission.

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

No 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.

None.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Assistance, Treatment planning.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

Cohorts

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

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

Actual cancer patients.

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

Actual cancer patients.

Imaging modality(ies)

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

CT.

Context information

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

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

Pixel-level segmentation annotations.

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.

Head and neck CT scans.

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.

Organ-at-risks in head and neck CT scans.

Assessment aim(s)

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

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

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

Precision, Accuracy.

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 data are collected by Philips Brilliance Big Bore CT scanners at Cancer Hospital of University of Chinese Academy of Sciences.

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

CT.

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.

Cancer Hospital of University of Chinese Academy of Sciences.

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).

None.

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.

22 OARs from 1 patient's head and neck CT scan.

The OARs are left eye, right eye, left lens, right lens, left optical nerve, right optical nerve, optical chiasma, pituitary, brain stem, left temporal lobes, righttemporallobes, spinalcord, leftparotidgland, right parotid gland, left inner ear, right inner ear, left middle ear, right middle ear, left temporomandibular joint, right temporomandibular joint, left mandible, right mandible.

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

Training data: 22 OARs of 60 nasopharynx cancer patients;

Test data: 22 OARs of 10 nasopharynx cancer patients.

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

86% training data and 14% test data.

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

All cases are from actual treatment plan for cancer patients.

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.

Pixel-level segmentation of the organs-at-risk of interest.

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

All annotations are from actual radiation therapy plans following clinical practice.

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.

Experienced oncologists who have performed hundreds of radiation therapy plans.

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

None.

Data pre-processing method(s)

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

Patient-identity information is removed from data the protect patient privacy.

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.

~2-3mm.

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)

1. Dice Similarity Coefficient (DSC);
2. 95% Hausdorff Distance.

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

The two metrics are widely used to measure segmentation accuracy.

Ranking method(s)

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

For each participant pi and each test case cj, 1) first calculate their average 1) DSC and 2) 95\% Hausdorff Distance, Then calculate the average values for each of the two measures over all patients. Rank each participant's two average measures separately. Finally calculate the average rank of the two measures' ranks.

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

No missing result allowed.

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

The ranking method has been successfully tested in StructSeg 2019.

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 propose new analysis to measure the stability and robustness of each algorithm over test cases. For each algorithm on each case, we will calculate its average rank of the two metrics of all tested algorithms. The mean and standard deviation of each algorithm over all cases would then be calculated. The statistics can tell us much about individual algorithms' performance and stability.

Algorithm overall performance: the overall performance of the algorithm can be measured by the mean of the average ranks over all cases.

Algorithm stability: the standard deviation of the average ranks of each algorithm can measure the stability of each algorithm. If an algorithm has high mean average rank, but its standard deviation of the average ranks is large. It

shows that the algorithm's performance is unstable on different test cases.

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

Algorhtim stability is also an important aspect of each method, which might not be shown by their final overall ranking.

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.

- Common problems/biases of the submitted methods;

- Inter-algorithm variability.

TASK: Organ at risk segmentation from chest CT scans

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.

6 OARs of 60 lung cancer patients will be annotated and released to public as the training data. There will be a total of 360 annoated structures/organs (6 X 60 = 360) in the training set. Each of the annotated CT scan is marked by one experienced oncologist and verified by another experienced one. Another 10 patients' CT scans will be used as the test data.

Keywords

List the primary keywords that characterize the task.

Organ-at-risk segmentation, chest CT

ORGANIZATION

Organizers

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

Hongsheng Li, Assistant Professor, The Chinese University of Hong Kong;

Ming Chen, Deputy Chief Director, Cancer Hospital of the University of Chinese Academy of Sciences.

b) Provide information on the primary contact person.

Hongsheng Li, The Chinese University of Hong Kong, hsli@ee.cuhk.edu.hk.

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.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event one time.

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 for publicity and own online submission site for final model submission.

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

None at this moment.

Participation policies

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

Fully automatic.

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

No additional data allowed.

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

May not participate.

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

The top-3 teams will be invited to present talks and winning certificates will be provided.

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-10 performing results will be made public.

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

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

We will not publish a challenge paper with submission methods. Each team can submit their paper on their own.

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 on webpage (same as StructSeg 2019). We will let the participants to submit their Docker submissions to our evaluation server. Any valid submission would be notified as "Valid submission with >1% accuracy". Any invalid submission would be informed with an error log file, telling participants why the submission fails to be evaluated.

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 are allowed. However, we will choose the highest submission from the latest three submissions of each team. Since we don't show the actual testing accuracy of each submission, but only show "Valid submission with >1% accuracy". The participants have no way to hack the test set.

Challenge schedule

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

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

Release of training data: Jun. 10th, 2020;
Submission deadline for results: Sept. 10th, 2020;
Announcement of final results: Sept. 15th, 2020.

Ethics approval

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

Ethics approval on data from StructSeg 2019 have been obtained. We only need to re-new the ethics approval for StructSeg 2020.

Data usage agreement

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

Examples:

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

CC BY NC SA.

Code availability

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

Evaluation code will be made public before the system is open for submission.

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

No 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.

None.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Assistance, Treatment planning.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

Cohorts

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

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

Actual cancer patients.

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

Actual cancer patients.

Imaging modality(ies)

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

CT.

Context information

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

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

Pixel-level segmentation annotations.

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.

Chest CT scans.

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.

Organ-at-risks in chest CT scans.

Assessment aim(s)

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

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

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

Precision, Accuracy.

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 data are collected by Philips Brilliance Big Bore CT scanners at Cancer Hospital of University of Chinese Academy of Sciences.

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

CT.

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.

Cancer Hospital of University of Chinese Academy of Sciences.

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).

None.

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.

6 OARs from 1 patient's chest CT scan.

The OARs are left lung, right lung, spinal cord, esophagus, heart, trachea.

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

Training data: 6 OARs of 60 lung cancer patients;

Test data: 6 OARs of 10 lung cancer patients.

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

86% training data and 14% test data.

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

All cases are from actual treatment plan for cancer patients.

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.

Pixel-level segmentation of the organs-at-risk of interest.

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

All annotations are from actual radiation therapy plans following clinical practice.

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.

Experienced oncologists who have performed hundreds of radiation therapy plans.

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

None.

Data pre-processing method(s)

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

Patient-identity information is removed from data the protect patient privacy.

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.

~2-3mm.

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)

1. Dice Similarity Coefficient (DSC);
2. 95% Hausdorff Distance.

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

The two metrics are widely used to measure segmentation accuracy.

Ranking method(s)

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

For each participant pi and each test case cj, 1) first calculate their average 1) DSC and 2) 95\% Hausdorff Distance, Then calculate the average values for each of the two measures over all patients. Rank each participant's two average measures separately. Finally calculate the average rank of the two measures' ranks.

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

No missing result allowed.

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

The ranking method has been successfully tested in StructSeg 2019.

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 propose new analysis to measure the stability and robustness of each algorithm over test cases. For each algorithm on each case, we will calculate its average rank of the two metrics of all tested algorithms. The mean and standard deviation of each algorithm over all cases would then be calculated. The statistics can tell us much about individual algorithms' performance and stability.

Algorithm overall performance: the overall performance of the algorithm can be measured by the mean of the average ranks over all cases.

Algorithm stability: the standard deviation of the average ranks of each algorithm can measure the stability of each algorithm. If an algorithm has high mean average rank, but its standard deviation of the average ranks is large. It shows that the algorithm's performance is unstable on different test cases.

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

Algorhtim stability is also an important aspect of each method, which might not be shown by their final overall ranking.

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.

- Common problems/biases of the submitted methods;

- Inter-algorithm variability.

TASK: Gross target volume segmentation of nasopharyngeal cancer

SUMMARY

Abstract

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

The 60 GTV annotations of the same 60 nasopharyngeal cancer patients' CT scans will be provided as the training data and another 10 patients' GTV will be used as the test data. Each CT scan is annotated by one experienced oncologist and verified by another one.

Keywords

List the primary keywords that characterize the task.

Gross Target Volume segmentation, head & neck CT

ORGANIZATION

Organizers

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

Hongsheng Li, Assistant Professor, The Chinese University of Hong Kong;

Ming Chen, Deputy Chief Director, Cancer Hospital of the University of Chinese Academy of Sciences.

b) Provide information on the primary contact person.

Hongsheng Li, The Chinese University of Hong Kong, hsli@ee.cuhk.edu.hk.

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.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event one time.

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 for publicity and own online submission site for final model submission.

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

None at this moment.

Participation policies

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

Fully automatic.

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

No additional data allowed.

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

May not participate.

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

The top-3 teams will be invited to present talks and winning certificates will be provided.

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-10 performing results will be made public.

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

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

We will not publish a challenge paper with submission methods. Each team can submit their paper on their own.

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 on webpage (same as StructSeg 2019). We will let the participants to submit their Docker submissions to our evaluation server. Any valid submission would be notified as "Valid submission with >1% accuracy". Any invalid submission would be informed with an error log file, telling participants why the submission fails to be evaluated.

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 are allowed. However, we will choose the highest submission from the latest three submissions of each team. Since we don't show the actual testing accuracy of each submission, but only show "Valid submission with >1% accuracy". The participants have no way to hack the test set.

Challenge schedule

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

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

Release of training data: Jun. 10th, 2020;
Submission deadline for results: Sept. 10th, 2020;
Announcement of final results: Sept. 15th, 2020.

Ethics approval

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

Ethics approval on data from StructSeg 2019 have been obtained. We only need to re-new the ethics approval for StructSeg 2020.

Data usage agreement

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

Examples:

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

CC BY NC SA.

Code availability

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

Evaluation code will be made public before the system is open for submission.

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

No 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.

None.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Assistance, Treatment planning.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

Cohorts

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

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

Actual cancer patients.

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

Actual cancer patients.

Imaging modality(ies)

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

CT.

Context information

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

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

Pixel-level segmentation annotations.

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.

Head and neck CT scans.

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.

Gross Target Volume of nasopharynx cancer.

Assessment aim(s)

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

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

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

Precision, Accuracy.

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 data are collected by Philips Brilliance Big Bore CT scanners at Cancer Hospital of University of Chinese Academy of Sciences.

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

CT.

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.

Cancer Hospital of University of Chinese Academy of Sciences.

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).

None.

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.

GTV from 1 patient's head and neck CT scan.

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

Training data: GTV of 60 nasopharynx cancer patients;

Test data: GTV of 10 nasopharynx cancer patients.

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

86% training data and 14% test data.

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

All cases are from actual treatment plan for cancer patients.

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.

Pixel-level segmentation of the GTV.

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

All annotations are from actual radiation therapy plans following clinical practice.

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.

Experienced oncologists who have performed hundreds of radiation therapy plans.

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

None.

Data pre-processing method(s)

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

Patient-identity information is removed from data the protect patient privacy.

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.

~2-3mm.

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)

1. Dice Similarity Coefficient (DSC);
2. 95% Hausdorff Distance.

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

The two metrics are widely used to measure segmentation accuracy.

Ranking method(s)

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

For each participant pi and each test case cj, 1) first calculate their average 1) DSC and 2) 95\% Hausdorff Distance, Then calculate the average values for each of the two measures over all patients. Rank each participant's two average measures separately. Finally calculate the average rank of the two measures' ranks.

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

No missing result allowed.

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

The ranking method has been successfully tested in StructSeg 2019.

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 propose new analysis to measure the stability and robustness of each algorithm over test cases. For each algorithm on each case, we will calculate its average rank of the two metrics of all tested algorithms. The mean and standard deviation of each algorithm over all cases would then be calculated. The statistics can tell us much about individual algorithms' performance and stability.

Algorithm overall performance: the overall performance of the algorithm can be measured by the mean of the average ranks over all cases.

Algorithm stability: the standard deviation of the average ranks of each algorithm can measure the stability of each algorithm. If an algorithm has high mean average rank, but its standard deviation of the average ranks is large. It shows that the algorithm's performance is unstable on different test cases.

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

Algorhtim stability is also an important aspect of each method, which might not be shown by their final overall ranking.

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.

- Common problems/biases of the submitted methods;

- Inter-algorithm variability.

TASK: Gross target volume segmentation of lung cancer

SUMMARY

Abstract

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

The 60 GTV annotations of the same 60 lung cancer patients' CT scans will be provided as the training data and another 10 patients' GTV will be used as the test data. Each CT scan is annotated by one experienced oncologist and verified by
another one.

Keywords

List the primary keywords that characterize the task.

Gross Target Volume segmentation, chest CT

ORGANIZATION

Organizers

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

Hongsheng Li, Assistant Professor, The Chinese University of Hong Kong;

Ming Chen, Deputy Chief Director, Cancer Hospital of the University of Chinese Academy of Sciences.

b) Provide information on the primary contact person.

Hongsheng Li, The Chinese University of Hong Kong, hsli@ee.cuhk.edu.hk.

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.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event one time.

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 for publicity and own online submission site for final model submission.

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

None at this moment.

Participation policies

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

Fully automatic.

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

No additional data allowed.

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

May not participate.

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

The top-3 teams will be invited to present talks and winning certificates will be provided.

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-10 performing results will be made public.

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

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

We will not publish a challenge paper with submission methods. Each team can submit their paper on their own.

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 on webpage (same as StructSeg 2019). We will let the participants to submit their Docker submissions to our evaluation server. Any valid submission would be notified as "Valid submission with >1% accuracy". Any invalid submission would be informed with an error log file, telling participants why the submission fails to be evaluated.

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 are allowed. However, we will choose the highest submission from the latest three submissions of each team. Since we don't show the actual testing accuracy of each submission, but only show "Valid submission with >1% accuracy". The participants have no way to hack the test set.

Challenge schedule

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

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

Release of training data: Jun. 10th, 2020;
Submission deadline for results: Sept. 10th, 2020;
Announcement of final results: Sept. 15th, 2020.

Ethics approval

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

Ethics approval on data from StructSeg 2019 have been obtained. We only need to re-new the ethics approval for StructSeg 2020.

Data usage agreement

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

Examples:

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

CC BY NC SA.

Code availability

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

Evaluation code will be made public before the system is open for submission.

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

No 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.

None.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Assistance, Treatment planning.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

Cohorts

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

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

Actual cancer patients.

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

Actual cancer patients.

Imaging modality(ies)

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

CT.

Context information

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

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

Pixel-level segmentation annotations.

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.

Chest CT scans.

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.

Gross Target Volume of lung cancer.

Assessment aim(s)

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

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

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

Precision, Accuracy.

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 data are collected by Philips Brilliance Big Bore CT scanners at Cancer Hospital of University of Chinese Academy of Sciences.

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

CT.

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.

Cancer Hospital of University of Chinese Academy of Sciences.

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).

None.

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.

GTV from 1 patient's chest CT scan.

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

Training data: GTV of 60 lung cancer patients;

Test data: GTV of 10 lung cancer patients.

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

86% training data and 14% test data.

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

All cases are from actual treatment plan for cancer patients.

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.

Pixel-level segmentation of the GTV.

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

All annotations are from actual radiation therapy plans following clinical practice.

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.

Experienced oncologists who have performed hundreds of radiation therapy plans.

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

None.

Data pre-processing method(s)

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

Patient-identity information is removed from data the protect patient privacy.

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.

~2-3mm.

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)

1. Dice Similarity Coefficient (DSC);
2. 95% Hausdorff Distance.

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

The two metrics are widely used to measure segmentation accuracy.

Ranking method(s)

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

For each participant pi and each test case cj, 1) first calculate their average 1) DSC and 2) 95\% Hausdorff Distance, Then calculate the average values for each of the two measures over all patients. Rank each participant's two average measures separately. Finally calculate the average rank of the two measures' ranks.

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

No missing result allowed.

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

The ranking method has been successfully tested in StructSeg 2019.

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 propose new analysis to measure the stability and robustness of each algorithm over test cases. For each algorithm on each case, we will calculate its average rank of the two metrics of all tested algorithms. The mean and standard deviation of each algorithm over all cases would then be calculated. The statistics can tell us much about individual algorithms' performance and stability.

Algorithm overall performance: the overall performance of the algorithm can be measured by the mean of the average ranks over all cases.

Algorithm stability: the standard deviation of the average ranks of each algorithm can measure the stability of each algorithm. If an algorithm has high mean average rank, but its standard deviation of the average ranks is large. It shows that the algorithm's performance is unstable on different test cases.

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

Algorhtim stability is also an important aspect of each method, which might not be shown by their final overall ranking.

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.

- Common problems/biases of the submitted methods;

- Inter-algorithm variability.

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.

StructSeg 2019: https://structseg2019.grand-challenge.org/