

MICCAI Abdominal Multi-Organ Segmentation Challenge 2022: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

MICCAI Abdominal Multi-Organ Segmentation Challenge 2022

Challenge acronym

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

AMOS

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.

Abdominal multi-organ segmentation is one of the most attractive topics in the field of medical image analysis, which plays an important role in supporting clinical workflows such as disease diagnosis and treatment planning. The recent success of deep learning methods applied for abdominal multi-organ segmentation expose the lack of large-scale comprehensive benchmarks for developing and comparing such methods. While several benchmark datasets [1-4] for abdominal organ segmentation are available, the limited number of organs of interest and training samples still limits the power of modern deep models and makes it difficult to provide a fully comprehensive and fair estimate of different methods.

To address the above drawbacks and further promote the development of medical image segmentation technology, we present AMOS, a large-scale, clinical and diverse abdominal multi-organ segmentation benchmark. It provides 500 CT and 100 MRI scans collected from multi-center, multi-vendor, multi-modality, multi-phase, multi-disease patients, each with voxel-level annotations of 15 abdominal organs. It is the most comprehensive benchmark of its kind to date.

Specifically, the AMOS 2022 challenge contains two tasks in which the participating teams can take place and submit their result(s):

a) Task 1 - Segmentation of abdominal organs (CT only): as a mostly regular task, Task 1 aims to comprehensively evaluate the performance of different segmentation methods across large-scale and great diversity CT scans, a total of 500 cases with annotations of 15 organs (spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus) are presented.

b) Task 2 - Segmentation of abdominal organs (CT & MRI): this task extends the image modality target of Task 1 to the MRI modality. Under such a “Cross Modality” setting, a single algorithm is required to segment abdominal organs from both CT and MRI. Specifically, additional 100 MRI scans with the same type of annotation will be provided.

Totally, the AMOS 2022 challenge focuses on the comprehensive evaluation of state-of-the-art methods for the segmentation of abdominal multi-organ in both clinical CT and MRI scans. With this challenge, we hope to bring together researchers and practitioners at the interplay of medical image analysis, computer vision, and machine learning to contribute their effort to the development of the corresponding techniques. We will also provide workshops to encourage participants to discuss open problems in the related areas.

- [1] Bilic, P., Christ, P. F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., ... & Menze, B. H. (2019). The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056.
- [2] Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., ... & Weight, C. (2021). The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis, 67, 101821.
- [3] Kavur, A. E., Gezer, N. S., Bar, M., Aslan, S., Conze, P. H., Groza, V., ... & Selver, M. A. (2021). CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis, 69, 101950.
- [4] Ma, J. (2021). Flare21 - Grand Challenge. grand. Retrieved December 2, 2021, from <https://flare.grand-challenge.org/FLARE21/>.

Challenge keywords

List the primary keywords that characterize the challenge.

Abdominal Multi-Organ Segmentation, Large-Scale Dataset, CT, MRI

Year

The challenge will take place in ...

2022

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.

We conservatively estimate the final participation from (round) 50 teams, considering that:

- 1) The number of teams registering in organ segmentation competitions has continued to rise over the past years. (Lits2019: 27 teams, Kits2019: 106 teams, Chaos2020: >1500 teams, Flare2021: 380 teams). Considering that only 35 teams participated in the final testing phase of Flare2021, we expect the final number of participants to be around 50 teams throughout.
- 2) We have received 30 explicit requests to download AMOS 2022 training data when it becomes available, as well as emails from 8 international groups interested in participating in the AMOS 2022 challenge.

- 3) We will advertise the event in related international mailing lists, as well as the media platform (e.g., WeChat media platform, academic accounts with 100K active followers).
- 4) The challenge organizers are affiliated with the National Health Commission (NHC) of the People's Republic of China, so we will also actively use this NHC platform to promote the competition.

Publication and future plans

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

We plan to coordinate the publication of a top journal following the Challenge that describes the design, data, clinical relevance, and summarizes the results of the AMOS Challenge. All teams that submit by the deadline and present their results at MICCAI 2022 will be included in the paper. Each team can have a maximum of two co-authors on this paper.

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.

AMOS 2022 is an off-site challenge and algorithms are run using the participants' computing infrastructure. On the day of the challenge, a projector, a computer, and two microphones will be needed for the presentations of the top methods.

TASK: Segmentation of Abdominal organs in 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.

Task 1 aims to comprehensively evaluate the performance of different segmentation methods across large-scale and diverse CT scans. A total of 500 cases with annotations of 15 organs are presented.

Keywords

List the primary keywords that characterize the task.

Abdominal Multi-Organ Segmentation, Medical Image Analysis, Large Scale Dataset, CT

ORGANIZATION

Organizers

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

Ruimao Zhang, The Chinese University of Hong Kong, Shenzhen, China

Zhen Li, The Chinese University of Hong Kong, Shenzhen, China

Xiang Wan, Shenzhen Research Institute of Big Data, Shenzhen, China

Guanbin Li, Shenzhen Research Institute of Big Data, Shenzhen, China

Haotian Bai, The Chinese University of Hong Kong, Shenzhen, China

Jie Yang, The Chinese University of Hong Kong, Shenzhen, China

Lingyan Zhang, Department of Medical Imaging, Longgang Central Hospital of Shenzhen, China

Wanling Ma, Department of Radiology, Longgang District People's Hospital of Shenzhen City & The Third Affiliated Hospital (Provisional) of The Chinese University of Hong Kong (Shenzhen), China

Ping Luo, The University of Hong Kong, Hongkong, China

Yuanfeng Ji, The University of Hong Kong, Hongkong, China

S. Kevin Zhou, The University of Science and Technology of China, Suzhou, China &

The Chinese University of Hong Kong, Shenzhen, China

Shuguang Cui, The Chinese University of Hong Kong, Shenzhen, China

b) Provide information on the primary contact person.

Ruimao Zhang(zhangruimao@cuhk.edu.cn), Zhen Li (lizhen@cuhk.edu.cn), The Chinese University of Hong Kong, Shenzhen, China and Ping Luo (pluo@cs.hku.hk), The University of Hong Kong, Hongkong, China.

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

Open call challenge.

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

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

www.amos.sribd.cn will become available after approval

Participation policies

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

Fully automatic.

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

No additional data allowed.

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

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

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

We will provide prizes (cash and certificates) to the Top 5 teams, the exact amount of the cash prize is not yet known. Sponsors for the awards are currently being communicated. Details will be announced on the Challenge homepage later.

e) Define the policy for result announcement.

Examples:

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

The Top 10 performing methods will be announced publicly.

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 plan to solicit short papers in the length of 4 pages from the teams that describe their approach in detail.

Please note that final phase participants are required to submit short papers or will not be eligible for awards and inclusion in the rankings. In addition, for the content of the short paper, original or non-original methods are fine, but they should describe their solution for the AMOS challenge in a factual way. Furthermore, based on the collected paper proceedings, we intend to coordinate a journal manuscript to the related top journal of medical image analysis, which aims to describe the challenge designing, the participation solution, and summarize the results of the MICCAI AMOS 2022 Challenge.

Submission method

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

Examples:

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

There are validation and test submission phases, and the processes are different.

Validation phases: the participants are required to send the output of their algorithms as a single compressed zip file via the grand-challenge.org submission system.

Testing phases: the participants will be requested to submit their algorithm in the form of a docker container to our docker repository, which will be available after approval; we believe this will enable more thorough confirmation of reproducibility. After the competition, the docker models will be released with the consent of the corresponding participants.

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

We intend to release a validation set in May, allowing participants to tune their methods. The validation data ground truth will not be provided to the participants, but multiple submissions to the online evaluation platform will be allowed. For the testing phase, only the last run submitted docker container is officially counted to compute results.

Challenge schedule

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

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

Registration dates: From challenge's approval until the submission deadline of short papers reporting method and preliminary results (see below)

Expected Website Opens: March 28, 2022

Expected Training Set Released: March 30, 2022

Expected Validation Set Released & First Stage Submission: May 1, 2022

Expected First stage Deadline: July 1, 2022

Expected Submission of short papers, reporting method & preliminary results: July 15, 2022

Expected release of second stage data & evaluation within 72 hours (only for participants with submitted papers)

July 20, 2022 - July 23, 2022

Expected Announcement of the final Top 10 ranked teams: Aug 15, 2022

Ethics approval

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

All data contributions to this study have been reviewed and approved by the Research Ethics Committee of Longgang District People's Hospital (reference number: 2021077) and the Research Ethics Committee of Longgang District Central Hospital (reference number: 2021ECJ012). The approved documents can be found in <https://drive.google.com/drive/folders/1UNHjEgau85rit-DBAKg9kGv6REkiHU6Z?usp=sharing>

Data usage agreement

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

Examples:

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

CC BY NC SA.

Code availability

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

The metric calculation code will be released at the same time as the training data, and the ranking code will be available after the end of the challenge.

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

Participants of the test phase of the competition will be asked to submit their models in the form of docker during the test phase. Detailed instructions for creating docker containers will be provided.

Conflicts of interest

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

This challenge is supported by the Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001), NSFC under the project "The Essential Algorithms and Technologies for Standardized Analytics of Clinical Texts" (12026610) and the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, by the Open Research Fund from Shenzhen Research Institute of Big Data No. 2019ORF01005, by NSFC-Youth 61902335, by Key Area R&D Program of Guangdong Province with grant No.2018B030338001.

Yuanfeng Ji, Haotian Bai, and clinical evaluators will have access to the test case labels.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

CAD, Decision support, Treatment planning, Diagnosis, Assistance, Surgery, Intervention planning, Screening.

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.

The multi-institutional cohort of patients diagnosed with abdominal tumors/abnormalities was clinically scanned with CT acquisition protocols.

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

The multi-institutional cohort of patients diagnosed with abdominal tumors/abnormalities was clinically scanned with CT acquisition protocols.

Imaging modality(ies)

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

Multi-Phase CT Scan

Context information

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

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

N/A

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

Age, Gender, Report Finding, and Impression. [Note that all provided information will be subject to availability.]

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.

Abdominal CT scan.

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.

15 organs of interest, including the spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus.

Assessment aim(s)

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

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

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

Accuracy, Precision.

Additional points: Find a highly precise abdominal organ segmentation algorithm for CT images, that works

robustly in diverse clinical scenarios.

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 CT scans are collected from various brands of CT scanners from two medical centers, including Aquilion ONE scanner from Toshiba, Brilliance16 scanner from Philips, Revolution ACT, Optima CT660/Optima CT540 scanners from GE, uCT 530 scanner from United Imaging, SOMATOM Force scanner from Siemens. No additional performance assessment devices during the acquisitions.

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

Acquisition details are different for each different institution, as these scans we use are representative of real clinical protocols. For example, 50 CT scans collected from the same scanner are obtained following criteria: 120kVp; data collection diameter of 500 mm; exposure time of 500-800 ms, Xray tube current 50-400 mA. Images were reconstructed at a section thickness of 2.5-5 mm with a standard FC08 convolutional kernel and with a reconstruction diameter range of 400-500 mm. We intend to release all the available details in the future.

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 provided data describe CT scans, acquired with different clinical protocols and various scanners from:

1. Longgang District Central Hospital (SZ, CHINA)
2. Longgang District People's Hospital (SZ, CHINA)

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).

Radiographers with at least 10 years of experience acquired the data using clinically defined protocols.

Training and test case characteristics

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

Examples:

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

A case describes a CT scan for a single abdominal tumors/abnormalities patient at a single time point, and each case has an annotated label map corresponding to 15 abdominal organs.

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

There is a total of 500 abdominal CT cases, each scan comes from a different patient. We made a 60%/10%/30% split to get 300 training cases, 50 validation cases, and 150 testing cases.

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

The total number of cases is determined by the cost of collection and annotation that we can afford to in the current stage.

We chose the above 60%/10%/30% split due to:

(a) The training, validate/test data respectively represent 60%, 40%, which is a common practice in machine learning.

(b) Compared to the previous challenge which usually provides 100 scans totally, our challenge provides 300 training cases, 50 validations cases, and 150 testing cases, which is believed to provide enough data variability for model training, model selection, and model evaluation.

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.

Data were collected on patients with abdominal tumors (majority) or other abnormalities. Moreover, among the 500 CT scans collected, males and females are 314 and 186, respectively. For the age distribution, the patients' minimum, maximum, median, and mean ages are 14, 94, 54, and 53.64 years, respectively. The distribution of these factors will be consistent between the training/validate/test splits.

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.

Considering the massive cost of "annotate from scratch" of all 3d data, we decide to first train a segmentation model to collaborate with human annotators. Specifically, 50 cases were manually annotated from scratch and used to train a model for pre-labeling subsequent cases. Next, 5 well-trained junior radiologists check and revisit the segmentation results, each case is assigned to one annotator. To reduce the sources of errors associated with image annotation, we further invited three senior radiologists with more than 10 years of clinical experience to participate in the final validation. They review the annotations, try to revise or send feedback to the corresponding annotator regarding the quality. Such progress is iterated several times and eventually, well-labeled annotations with consensus will be included in our dataset.

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

The radiologists were asked to use ITK-SNAP 3.6 toolbox for checking and revision.

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

The Pre-labeling model was trained with the nnUNet package. 5 junior radiologists with more than 5 years of

clinical work experience. 3 senior radiologists each have more than 10 years of clinical experience.

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

No aggregation was applied.

Data pre-processing method(s)

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

Data are not pre-processed except for information required for identification. The data will be released in the NIfTI format.

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.

Sources of error in the image segmentation annotation depend mostly on the annotators. In the final verify stage, 5 junior radiologists are involved to systematically validate all the annotations to ensure accurate annotations. Besides, following the strategy used in MICCAI Kits, MICCAI Flare, we will establish a github repository to solicit possible errors from data users.

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

N/A

ASSESSMENT METHODS

Metric(s)

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

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

The classical medical segmentation metrics: Dice Similarity Coefficient (DSC), and normalized surface dice (NSD), will be used to assess different aspects of the performance of the segmentation methods.

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

We chose the (DSC, NSD) as validation metrics because of their simplicity, popularity, rank stability, and ability to assess the accuracy of the predictions.

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.

All two metrics will be calculated for each label within each of the predicted label maps of the cases in the testing

set. The mean and standard deviation of each label will be calculated. For both two metrics, the participating algorithms will be ranked from low to high, where the highest score receives the highest-scoring rank. The metrics will be averaged over each label, and the participating teams will be ranked based on both average metrics. The final leaderboard place will be determined by averaging the leaderboard places between the two rankings, called "rank-then-aggregate."

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

Regarding the validation set, all predictions must be present in the zip file of the submission. Incomplete submissions will not be evaluated and counted as validation submissions. In the final testing phase, all predictions for the test set will be made locally using the team's docker containers. Therefore, no missing results are expected on the test set.

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

This ranking scheme is standard and has been adopted in previous challenges with satisfactory results, and we believe the design takes into account transparency and fairness to the participants.

Statistical analyses

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

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

The variability of the rankings will be characterized using bootstrap methods, which will be performed in Python.

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

The bootstrap method has been successfully applied in many previous challenges, meanwhile, bootstrap is a simple non-parametric approach that relies on minimal assumptions.

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.

N/A

TASK: Segmentation of Abdominal organs in both CT and MRI 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.

Task 2 extends the image modality target of Task1 to the MRI modality, under such a "Cross Modality" setting, a single algorithm is required to segment abdominal organs from both CT and MRI. This task aims to comprehensively evaluate the performance of different segmentation methods across large-scale and diverse CT and MRI scans. Additional 100 MRI scans with the same annotation will be provided.

Keywords

List the primary keywords that characterize the task.

Abdominal Multi-Organ Segmentation, Medical Image Analysis, Large Scale Dataset, CT, MRI

ORGANIZATION

Organizers

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

Ruimao Zhang, The Chinese University of Hong Kong, Shenzhen, China

Zhen Li, The Chinese University of Hong Kong, Shenzhen, China

Xiang Wan, Shenzhen Research Institute of Big Data, Shenzhen, China

Guanbin Li, Shenzhen Research Institute of Big Data, Shenzhen, China

Haotian Bai, The Chinese University of Hong Kong, Shenzhen, China

Jie Yang, The Chinese University of Hong Kong, Shenzhen, China

Lingyan Zhang, Department of Medical Imaging, Longgang Central Hospital of Shenzhen, China

Wanling Ma, Department of Radiology, Longgang District People's Hospital of Shenzhen City & The Third Affiliated Hospital (Provisional) of The Chinese University of Hong Kong (Shenzhen), China

Ping Luo, The University of Hong Kong, Hongkong, China

Yuanfeng Ji, The University of Hong Kong, Hongkong, China

S. Kevin Zhou, The University of Science and Technology of China, Suzhou, China & The Chinese University of Hong Kong, Shenzhen, China

Shuguang Cui, The Chinese University of Hong Kong, Shenzhen, China

b) Provide information on the primary contact person.

Ruimao Zhang(zhangruimao@cuhk.edu.cn), Zhen Li (lizhen@cuhk.edu.cn), The Chinese University of Hong Kong, Shenzhen, China and Ping Luo (pluo@cs.hku.hk), The University of Hong Kong, Hongkong, China.

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

Open call challenge.

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

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

www.amos.sribd.cn will become available after approval

Participation policies

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

Fully automatic.

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

No additional data allowed.

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

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

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

We will provide prizes (cash and certificates) to the Top 5 teams, the exact amount of the cash prize is not yet known. Sponsors for the awards are currently being communicated. Details will be announced on the Challenge homepage later.

e) Define the policy for result announcement.

Examples:

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

The Top 10 performing methods will be announced publicly.

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 plan to solicit short papers in the length of 4 pages from the teams that describe their approach in detail.

Please note that final phase participants are required to submit short papers or will not be eligible for awards and inclusion in the rankings. In addition, for the content of the short paper, original or non-original methods are fine, but they should describe their solution for the AMOS challenge in a factual way. Furthermore, based on the collected paper proceedings, we intend to coordinate a journal manuscript to the related top journal of medical image analysis, which aims to describe the challenge designing, the participation solution, and summarize the results of the MICCAI AMOS 2022 Challenge.

Submission method

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

Examples:

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

There are validation and test submission phases, and the processes are different.

Validation phases: the participants are required to send the output of their algorithms as a single compressed zip file via the grand-challenge.org submission system.

Testing phases: the participants will be requested to submit their algorithm in the form of a docker container to our docker repository, which will be available after approval; we believe this will enable more thorough confirmation of reproducibility. After the competition, the docker models will be released with the consent of the corresponding participants.

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

We intend to release a validation set in May, allowing participants to tune their methods. The validation data ground truth will not be provided to the participants, but multiple submissions to the online evaluation platform will be allowed. For the testing phase, only the last run submitted docker container is officially counted to compute results.

Challenge schedule

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

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

Registration dates: From challenge's approval until the submission deadline of short papers reporting method and preliminary results (see below)

Expected Website Opens: March 28, 2022

Expected Training Set Released: March 30, 2022

Expected Validation Set Released & First Stage Submission: May 1, 2022

Expected First stage Deadline: July 1, 2022

Expected Submission of short papers, reporting method & preliminary results: July 15, 2022

Expected release of second stage data & evaluation within 72 hours (only for participants with submitted papers)

July 20, 2022 - July 23, 2022

Expected Announcement of the final Top 10 ranked teams: Aug 15, 2022

Ethics approval

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

All data contributions to this study have been reviewed and approved by the Research Ethics Committee of Longgang District People's Hospital (reference number: 2021077) and the Research Ethics Committee of Longgang District Central Hospital (reference number: 2021ECJ012). The approved documents can be found in <https://drive.google.com/drive/folders/1UNHjEgau85rit-DBAKg9kGv6REkiHU6Z?usp=sharing>

Data usage agreement

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

Examples:

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

CC BY NC SA.

Code availability

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

The metric calculation code will be released at the same time as the training data, and the ranking code will be available after the end of the challenge.

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

Participants of the test phase of the competition will be asked to submit their models in the form of docker during the test phase. Detailed instructions for creating docker containers will be provided.

Conflicts of interest

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

This challenge is supported by the Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001), NSFC under the project "The Essential Algorithms and Technologies for Standardized Analytics of Clinical Texts" (12026610) and the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, by the Open Research Fund from Shenzhen Research Institute of Big Data No. 2019ORF01005, by NSFC-Youth 61902335, by Key Area R&D Program of Guangdong Province with grant No.2018B030338001.

Yuanfeng Ji, Haotian Bai, and clinical evaluators will have access to the test case labels.

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

CAD, Decision support, Treatment planning, Diagnosis, Assistance, Surgery, Intervention planning, Screening.

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.

The multi-institutional cohort of patients diagnosed with abdominal tumors/abnormalities was clinically scanned with CT and MRI acquisition protocols.

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

The multi-institutional cohort of patients diagnosed with abdominal tumors/abnormalities was clinically scanned with CT and MRI acquisition protocols.

Imaging modality(ies)

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

Multi-Phase CT and MRI Scan

Context information

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

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

N/A

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

Age, Gender, Report Finding, and Impression. [Note that all provided information will be subject to availability.]

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.

Abdominal CT and MRI scan.

- 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.

15 organs of interest, including the spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus.

Assessment aim(s)

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

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

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

Accuracy, Precision.

Additional points: Find a highly precise abdominal organ segmentation algorithm for CT and MRI images, that works robustly in diverse clinical scenarios.

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 CT scans are collected from various brands of CT scanners from two medical centers, including Aquilion ONE scanner from Toshiba, Brilliance16 scanner from Philips, Revolution ACT, Optima CT660/Optima CT540 scanners from GE, uCT 530 scanner from United Imaging, SOMATOM Force scanner from Siemens. No additional performance assessment devices during the acquisitions.

The MRI scans are collected from various brands of MRI scanners from two medical centers, including Achieva 1.5T scanner from Philips, Prisma 3.0T scanner from Siemens, Signa hde 1.5T scanner from GE.

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

Acquisition details are different for each different institution, as these scans we use are representative of real clinical protocols. For example, 50 CT scans collected from the same scanner are obtained following criteria: 120kVP; data collection diameter of 500 mm; exposure time of 500-800 ms, Xray tube current 50-400 mA. Images were reconstructed at a section thickness of 2.5-5 mm with a standard FC08 convolutional kernel and with a reconstruction diameter range of 400-500 mm.

The same is applicable to the acquisition of MRI scans, 30 Contrast-enhanced T1-weighted imaging was performed with the Prisma scanner from Longgang District People's Hospital, with an in-plane resolution of 1 mm × 1 mm, flip angle: 15°, in-plane matrix of 400×400, and slice thickness of 1.0 to 2 mm.

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 provided data describe CT and MRI scans, acquired with different clinical protocols and various scanners from:

1. Longgang District Central Hospital (SZ, CHINA)
2. Longgang District People's Hospital (SZ, CHINA)

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).

Radiographers with at least 10 years of experience acquired the data using clinically defined protocols.

Training and test case characteristics

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

Examples:

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

A case describes a CT or MRI scan for a single abdominal tumors/abnormalities patient at a single time point, and each case has an annotated label map corresponding to 15 abdominal organs.

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

There is a total of 500 abdominal CT cases and 100 abdominal MRI cases, each scan comes from a different patient. We made a 60%/10%/30% split to get 300 CT + 60 MRI training cases, 50 CT + 10 MRI validation cases, and 150 CT + 30 MRI testing cases.

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

The total number of cases is determined by the cost of collection and annotation that we can afford to in the current stage.

We chose the above 60%/10%/30% split due to:

(a) The training, validate/test data respectively represent 60%, 40%, which is a common practice in machine learning.

(b) Compared to the previous challenge which usually provides 100 scans totally, our challenge provides 300 CT + 60 MRI training cases, 50 CT+ 10 MRI validations cases, and 150 CT + 30 MRI testing cases, which is believed to provide enough data variability for model training, model selection, and model evaluation.

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.

Data were collected on patients with abdominal tumors (majority) or other abnormalities. Moreover, among the 500 CT scans collected, males and females are 314 and 186, respectively. For the age distribution, the patients' minimum, maximum, median, and mean ages are 14, 94, 54, and 53.64 years, respectively. For the 100 MRI scans, males and females are 55 and 45, and the patients' minimum, maximum, median, and mean ages are 22, 85, 50, and 48.71 years, respectively. The ratio of the number of patients diagnosed with tumors to the number of patients with other abnormalities is 3:2. The distribution of these factors will be consistent between the training/validate/test splits.

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.

Considering the massive cost of “annotate from scratch” of all 3d data, we decide to first train a segmentation model to collaborate with human annotators. Specifically, 50 CT and 20 MRI cases were manually annotated from scratch and used to train a model for pre-labeling subsequent cases. Next, 5 well-trained junior radiologists check and revisit the segmentation results, each case is assigned to one annotator. To reduce the sources of errors associated with image annotation, we further invited three senior radiologists with more than 10 years of clinical experience to participate in the final validation. They review the annotations, try to revise or send feedback to the corresponding annotator regarding the quality. Such progress is iterated several times and eventually, well-labeled annotations with consensus will be included in our dataset.

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

The radiologists were asked to use ITK-SNAP 3.6 toolbox for checking and revision.

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

The Pre-labeling model was trained with the nnUNet package. 5 junior radiologists with more than 5 years of clinical work experience. 3 senior radiologists each have more than 10 years of clinical experience.

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

No aggregation was applied.

Data pre-processing method(s)

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

Data are not pre-processed except for information required for identification. The data will be released in the NIfTI format.

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.

Sources of error in the image segmentation annotation depend mostly on the annotators. In the final verify stage, 5 junior radiologists are involved to systematically validate all the annotations to ensure accurate annotations. Besides, following the strategy used in MICCAI Kits, MICCAI Flare, we will establish a github repository to solicit possible errors from data users.

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

N/A

ASSESSMENT METHODS

Metric(s)

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

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

The classical medical segmentation metrics: Dice Similarity Coefficient (DSC), and normalized surface dice (NSD), will be used to assess different aspects of the performance of the segmentation methods.

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

We chose the (DSC, NSD) as validation metrics because of their simplicity, popularity, rank stability, and ability to assess the accuracy of the predictions.

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.

All two metrics will be calculated for each label within each of the predicted label maps of the cases in the testing set. The mean and standard deviation of each label will be calculated. For both two metrics, the participating algorithms will be ranked from low to high, where the highest score receives the highest-scoring rank. The metrics will be averaged over each label, and the participating teams will be ranked based on both average metrics. The final leaderboard place will be determined by averaging the leaderboard places between the two rankings, called "rank-then-aggregate."

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

Regarding the validation set, all predictions must be present in the zip file of the submission. Incomplete submissions will not be evaluated and counted as validation submissions. In the final testing phase, all predictions for the test set will be made locally using the team's docker containers. Therefore, no missing results are expected on the test set.

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

This ranking scheme is standard and has been adopted in previous challenges with satisfactory results, and we believe the design takes into account transparency and fairness to the participants.

Statistical analyses

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

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

The variability of the rankings will be characterized using bootstrap methods, which will be performed in Python.

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

The bootstrap method has been successfully applied in many previous challenges, meanwhile, bootstrap is a simple non-parametric approach that relies on minimal assumptions.

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.

N/A

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] Bilic, P., Christ, P. F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., ... & Menze, B. H. (2019). The liver tumor segmentation benchmark (lits). arXiv preprint arXiv:1901.04056.
- [2] Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., ... & Weight, C. (2021). The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis, 67, 101821.
- [3] Kavur, A. E., Gezer, N. S., Bar, M., Aslan, S., Conze, P. H., Groza, V., ... & Selver, M. A. (2021). CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis, 69, 101950.
- [4] Ma, J. (2021). Flare21 - Grand Challenge. grand. Retrieved December 2, 2021, from <https://flare.grand-challenge.org/FLARE21/>.

Further comments

Further comments from the organizers.

N/A