Fast, Low-resource, Accurace, and Robust Organ and Pan-cancer Segmentation: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Fast, Low-resource, Accurace, and Robust Organ and Pan-cancer Segmentation

Challenge acronym

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

FLARE

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.

Organ and cancer segmentation in medical images, especially from 3D CT and MR scans, is fundamentally important for accurate diagnosis, treatment planning, and monitoring the progression of diseases. Precise segmentation results of organs and pathological lesions can aid clinicians in formulating personalized treatment strategies, which are essential for optimal patient outcomes. Deep learning-based methods have significantly revolutionized these tasks, achieving unprecedented levels of accuracy and automation compared to traditional model-based methods. However, a notable limitation is that most deep learning models are tailored for specific types of cancer, such as brain cancer, lung cancer, liver cancer, and so on. As a result, the generalizability of these algorithms across various cancer types remains a challenge. Another main barrier that hinders the real-world deployment of existing methods is the algorithm's efficiency because deep learning models usually require considerable computing for running, such as GPU, CPU, and RAM.

During the past three years, we have organized three challenges to address these limitations with community efforts.

- FLARE 2021: four abdominal organs segmentation in CT scans; Data: 511 CT scans
- FLARE 2022: 13 abdominal organs segmentation in CT scans; Data: 2300 CT scans
- FLARE 2023: 13 abdominal organs and pan-cancer segmentation in CT scans; Data: 4500 CT scans

Nowadays, the winning solution can simultaneously segment 13 organs and various abdominal lesions within 10 seconds for a 3D CT scan with over 1,000,000 voxels, which significantly improved the segmentation accuracy, efficiency, and generalization ability. In FLARE 2024, we aim to further promote the development of pan-cancer segmentation and model deployment on low-resource settings (e.g., no GPU available) by extending the challenge

to the following three tasks:

- Subtask 1: Pan-cancer segmentation in CT scans
- Subtask 2: Abdominal CT organ segmentation on laptop
- Subtask 3: Unsupervised domain adaptation for abdominal organ segmentation in MRI Scans

We will provide a comprehensive and large-scale dataset with more than 10,000 CT scans and 1,000 abdominal MR scans, which is a multi-racial, multi-center, multi-disease, multi-phase, multi-manufacturer, and multi-modality dataset. This challenge would mark a significant stride for universal cancer segmentation models and applicable toolsets for CT and MR image analysis.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Segmentation, Pan-cancer, Organs, Efficient, Cross-modality

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

n/a

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 had 48 and 37 successful docker submissions in FLARE 2022 and 2023, respectively. We expect 35-50 docker submissions in FLARE 2024.

Publication and future plans

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

All participants have the opportunity to publish their methods on LNCS proceedings. Top 5 teams in each task will be invited to write a challenge summary 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.

We plan to organize an on-site challenge and only need basic equipment: projectors, computers, monitors, loud speakers, microphones

TASK 1: Pan-cancer segmentation 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.

Deep learning has been the most popular method for cancer segmentation in CT scans, but most of the deep learning models are tailored for specific cancer types. With the increasing prevalence of various cancers and the growing demand for personalized medicine, there is a critical need for a universally applicable, resource-efficient, and highly accurate pan-cancer segmentation method. In MICCAL FLARE 2023, we focused on the pan-cancer segmentation in the abdomen and the winning solution achieved very promising results, where the segmentation performance of the pan-cancer model was comparable and even better than specialist cancer segmentation models.

These encouraging results motivate us to further scale up the dataset and extend the task to pan-cancer segmentation in whole-body CT scans. Our challenge task has three main features:

- Task: whole-body cancer segmentation rather than just abdominal cancer
- Dataset: 10,000 CT scans, covering whole-body cancer types
- Evaluation Metrics: segmentation accuracy (DSC, NSD) and segmentation efficiency (running time and GPU consumption with tolerance)

We also noticed that a recent lesion segmentation challenge (ULS, https://uls23.grand-challenge.org/) shares a similar goal. However, the ULS mainly focuses on CT volume of interest rather than the complete CT scans. Moreover, our challenge provides a significantly larger dataset and the evaluation metrics consider bother segmentation accuracy and efficiency.

Keywords

List the primary keywords that characterize the task.

Pan-cancer, Lesion, Segmentation, Human, Universal Model

ORGANIZATION

Organizers

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

Jun Ma (University of Toronto, University Health Network, Vector Institute)
Bo Wang (University of Toronto, University Health Network, Vector Institute)

b) Provide information on the primary contact person.

junma.ma@mail.utoronto.ca

Life cycle type

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

Examples:

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

Open call (challenge opens for new submissions after conference deadline)

Challenge venue and platform

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

N/A

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

CodaLab (Similar to the MICCAI FLARE 2023 Challenge)
Each task will have an independent challenge platform.

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

N/A

Participation policies

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

Any user interactions are not allowed.

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

No additional data is allowed.

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

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

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

We will provide certificates and souvenirs for the top-5 teams.

e) Define the policy for result announcement.

Examples:

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

All results 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).

All participants have the opportunity to publish their methods on LNCS proceedings. Top 5 teams in each task will be invited to write a challenge summary paper.

Submission method

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

Examples:

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

Docker containers

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

We will provide a validation set for participants to pre-evaluate their algorithms. The testing set has the same format as the validation set. To avoid overfitting the testing set, we only offer one submission opportunity on the testing 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

15 March 2023 (12:00 AM EST): Launch challenge registration and release training and validation data. Docker submission of validation set opening.

15 April 2023 (12:00 AM EST): Deadline for the first validation submission.

15 May 2023 (12:00 AM EST): Deadline for the second validation submission.

15 June 2023 (12:00 AM EST): Deadline for the third validation submission.

15 July 2023 (12:00 AM EST): Deadline for the third validation submission. Docker and short paper submission of the testing set opening.

15 August 2023 (12:00 AM EST): Deadline for the testing submission.

10 September 2023 (12:00 AM EST): Invite top teams to prepare presentations and participate in the MICCAI2023 Satellite Event.

6/10 October 2023: Announce final result

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).

Our dataset was curated from public domains and global data contributors with license permits. There is no patient information included. Thus, we do not need additional ethics approval.

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 available with the training data.

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

Top 5 teams should make their code publicly available. The remaining teams are encouraged to make their code publicly available.

Conflicts of interest

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

Only the organizers have access to the test case labels.
We are currently looking into which companies are willing to sponsor the challenge in the form of awards and computational resources

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Intervention assistance
Intervention follow-up
Intervention planning
Prognosis
Research

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 target cohort will be patients with various cancer types where the CT scans were acquired in different centers with different manufacturers.

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

The challenge cohort is composed of patients with over 15 cancer types, such as head-neck cancer, lung cancer, liver cancer, kidney cancer, pancreas cancer, and so on, which were curated from 50+ different medical centers.

Imaging modality(ies)

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

Computed Tomography

Context information

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

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

The image data will be released as compressed Nifti1Image objects stored in .nii.gz files. These files will include an "affine matrix" defining the voxel size of each image.

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

No additional information regarding each patient will be released.

Target entity(ies)

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

The dataset covers whole-body regions, including head, chest, abdomen, and pelvis.

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.

This is a binary segmentation task where the segmentation target is the lesion.
Participants have to submit a Docker image that can segment all measurable abdominal lesions based on Response Evaluation Criteria In Solid Tumors (RECIST) criteria.

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)).

Hardware requirements, Runtime, Accuracy, Robustness, Sensitivity, Specificity.

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 dataset contains CT scans from various manufacturers, such as Siemens, GE, Philips, Toshiba, Imatron, Vital, and PHMS.

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

Our dataset is adapted from TCIA, MSD, KiTS, and data contributors from local hospitals under the license permission. Thus, the acquisition protocols are very diverse. Detailed Information can be found in the following references.
[1] Simpson A, et al. A Large Annotated Medical Image Dataset for the Development and Evaluation of Segmentation Algorithms. ArXiv Preprint ArXiv:1902.09063.
[2] Heller N, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019).

[3] Clark K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of Digital Imaging 26, no. 6 (2013): 1045-1057.
[4] Ma J, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[5] Grossberg A, et al. Imaging and Clinical Data Archive for Head and Neck Squamous Cell Carcinoma Patients Treated with Radiotherapy. Scientific Data 5:180173 (2018)
[6] Fedorov A, et al. DICOM re encoding of volumetrically annotated Lung Imaging Database Consortium (LIDC) nodules. Medical Physics (2020)

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.

Alberta Health Services
Barretos Cancer Hospital, Barretos, Sao Paulo, Brazil
Beaumont Health System, Royal Oak, MI
BioPartners, CA
Boston Medical Center, Boston, MA
Brigham and Women's Hospital, Boston, MA
Cleveland Clinic
Cureline, Inc. team and clinical network, Brisbane, CA
Hebrew University of Jerusalem
International Institute for Molecular Oncology, Poznan
IRCAD Institute Strasbourg
Lahey Hospital & Medical Center, Burlington, MA
Ludwig Maxmilian University of Munich
M Health Fairview
M.D. Anderson Cancer Center, Houston TX
Mayo Clinic, Rochester, MN
Memorial Sloan Kettering Cancer Center (New York, NY, USA)
National Cancer Institute, Bethesda, MD
National Institutes of Health, Bethesda MD
Polytechnique & CHUM Research Center Montral
Radboud University Medical Center of Nijmegen
Roswell Park Cancer Institute, Buffalo NY
Sheba Medical Center
St. Joseph's Hospital and Medical Center, Phoenix, AZ
Tel Aviv University
The National Institutes of Health Clinical Center(NIH)
University of California, San Francisco, CA
University of Chicago
University of North Carolina, Chapel Hill, NC
University of Pittsburgh/UPMC, Pittsburgh, PA
University of Sheffield
University of Southern California
Washington University School of Medicine, St. Louis, MO
Cancer Center West China Hospital
Erasmus MC Cancer Institute, Rotterdam
Fudan University Shanghai Cancer Center

Longgang Central Hospital of Shenzhen, China
Longgang District People Hospital
University Hospital Basel

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).

Data annotation is overseen by the challenge clinical chair Dr. Jian He. He is an experienced radiologist who specializes in abdomen CT.

Training and test case characteristics

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

Examples:

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

One case, in this challenge, is a single CT scan associated with/without a binary segmentation mask.

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

The total number of cases is 10,600. There are 10,000/100/500 cases for training, validation, and testing, respectively.
The training set includes 7000 cases with partial labels (only the primary cancer is labeled) and 3000 unlabeled cases.

- head-neck cancer: 2000 labeled cases
- lung cancer: 2000 labeled cases and 1000 unlabeled cases
- Abdominal cancer (e.g., liver cancer, kidney cancer): 3000 labeled cases and 2000 unlabeled cases

It should be noted that none of the labels in the validation set and testing set is publicly available, and all the testing cases are from new centers that did not appear in the training set.

Moreover, we are discussing potential collaboration with new data contributors. The testing set could be expanded based on the availability of new datasets.

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

The data is provided based on the current availability. The challenge uses a partial-label learning task setting where only the primary tumor is labeled in each CT scan. This setting is in line with real practice because each medical department mainly focuses on one specific cancer type. In FLARE 2022-2023, the top teams have shown

significant performance improvements by using pseudo-label learning with unlabeled data. Thus, we continue to provide unlabeled data in the training set.

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

All the testing cases will be hidden from participants and the Docker container will be used for submission. Importantly, none of the labels is publicly available in the testing set.

Annotation characteristics

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

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

The annotations were generated with the strategy of semi-automatic segmentation by MedSAM and manual correction.

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.

Annotators and radiologists were asked to use 3D Slicer software to annotate the measure lesions based on RECIST.

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 annotation process involved a team of five junior radiologists with one to five years of experience and two senior radiologists with more than 10 years of experience.

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

To make all the labels consistent across the datasets from different medical centers, an experienced radiologist finally checked and revised all the annotations.

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.

We converted the dicom images to NifiTy format. The cases have various orientations. Participants should develop their own pre-processing methods.

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.

Error is unavoidable in medical image segmentation. We will have a forum to collect feedback from participants about possible errors that they discover in the dataset.

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)

Segmentation accuracy metrics: Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD).
Segmentation efficiency metrics: Running time and GPU memory consumption (area under GPU memory-time curve).

All metrics will be used to compute the ranking. We give GPU memory consumption a 4GB tolerance, GPU Memory = max(0, Real GPU Memory-4096MB) because all the previous top teams (i.e., FLARE21-23) can achieve state-of-the-art segmentation accuracy within this tolerance. We also give running time a tolerance (the running time of the winning solution in FLARE 2023) because it satisfies the typical clinical requirements and the code has been publicly available.

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

The metrics are complementary. Specifically, DSC and NSD are used to measure the region error and boundary error, respectively.
Running time and area under GPU memory-time curve are used to measure the inference speed and GPU consumption, respectively.

Ranking method(s)

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

The main benefit of this ranking scheme is that it can aggregate the metrics with different dimensions. A similar ranking scheme was also employed in the MICCAI BraTS Challenge 2017-2023 and FLARE 2021-2023.

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

Missing results on test cases will result in the ranks being set to the maximum (the number of teams). Specifically, missing results will get a zero value for DSC, NSD and the equivalent worst value for running time, and area under GPU memory-time curve.

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

Following the recommendation in ChallengeR, the ranking scheme includes the following three steps:
Step 1. For each testing case, we compute the four metrics;

Step 2. Rank participants for each of the 500 testing cases and each metric; Thus, each participant will have 2000 (500x4) rankings.

Step 3. Average all these rankings.

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 testing cases with missing results will be assigned the worst performance. Thus, there is no missing data for the results analysis. Ranking variability will be characterized using the bootstrap.

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

The Bootstrap is a simple nonparametric method 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.

We will verify the performance of the ensembles of top teams.

TASK 2: Abdominal CT Organ Segmentation on Laptop

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 past MICCAI FLARE challenges have demonstrated the possibility to achieve a good trade-off between segmentation accuracy and efficiency. For example, in the FLARE 2022 challenge, the winning solution can accurately segment a large CT scan, comprising over 1,000,000 voxels, in less than 10 seconds, all while consuming less than 2GB of GPU memory. Despite these advancements, a critical consideration arises in low-resource environments, such as laptops or hospital imaging edge devices, where GPU resources may not be available.

This context leads us to the question: Is it feasible to adapt state-of-the-art abdominal segmentation models for deployment in non-GPU environments without compromising on segmentation accuracy? To explore this possibility, we set up this task to benchmark **CPU-based abdominal segmentation algorithms.** The the best of our knowledge, this is the first challenge for benchmarking CPU-based segmentation models for 3D CT scans.

In an effort to ensure comparability with the winning solution of the FLARE 2022 Challenge, this task will utilize the same dataset, which includes 2050 cases for model training. The old validation set and testing set are merged as a new validation set with 250 cases. We provide a new testing set with 300 cases for this challenge. The evaluation metrics include DSC, NSD, and running time, providing a holistic assessment of the algorithm performance

Keywords

List the primary keywords that characterize the task.

Abdomen, Segmentaiton, Organ, Efficiency

ORGANIZATION

Organizers

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

Jun Ma (University of Toronto, University Health Network, Vector Institute)
Bo Wang (University of Toronto, University Health Network, Vector Institute)

b) Provide information on the primary contact person.

junma.ma@mail.utoronto.ca

Life cycle type

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

modifications (repeated event).

Examples:

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

Open call (challenge opens for new submissions after conference deadline)

Challenge venue and platform

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

N/A

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

CodaLab (Similar to the MICCAI FLARE 2023 Challenge)
Each task will have an independent challenge platform.

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

N/A

Participation policies

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

Any user interactions are not allowed.

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

No additional data is allowed.

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

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

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

We will provide certificates and souvenirs for the top-5 teams.

e) Define the policy for result announcement.

Examples:

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

All results 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).

All participants have the opportunity to publish their methods on LNCS proceedings. Top 5 teams in each task will be invited to write a challenge summary paper.

Submission method

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

Examples:

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

Docker containers

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

We will provide a validation set for participants to pre-evaluate their algorithms. The testing set has the same format as the validation set. To avoid overfitting the testing set, we only offer one submission opportunity on the testing 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

15 March 2023 (12:00 AM EST): Launch challenge registration and release training and validation data. Docker submission of validation set opening.

15 April 2023 (12:00 AM EST): Deadline for the first validation submission.

15 May 2023 (12:00 AM EST): Deadline for the second validation submission.

15 June 2023 (12:00 AM EST): Deadline for the third validation submission.

15 July 2023 (12:00 AM EST): Deadline for the third validation submission. Docker and short paper submission of the testing set opening.

15 August 2023 (12:00 AM EST): Deadline for the testing submission.

10 September 2023 (12:00 AM EST): Invite top teams to prepare presentations and participate in the MICCAI2023 Satellite Event.

6/10 October 2023: Announce final result

Ethics approval

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

The challenge dataset is the same as the MICCAI FLARE 2022 dataset. Thus, we do not need additional ethics approval.

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 available with the training data.

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

Top 5 teams should make their code publicly available. The remaining teams are encouraged to make their code publicly available.

Conflicts of interest

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

Only the organizers have access to the test case labels.
We are currently looking into which companies are willing to sponsor the challenge in the form of awards and computational resources

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Intervention assistance
Intervention follow-up
Intervention planning
Prognosis
Research

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 target cohort will be patients with abdominal CT scans that were acquired in different centers with different manufacturers.

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

The challenge cohort is composed of patients with various abdominal cancers as well as healthy subjects, which were curated from 30+ different medical centers.

Imaging modality(ies)

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

Computed Tomography

Context information

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

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

The image data will be released as compressed Nifti1Image objects stored in .nii.gz files. These files will include an "affine matrix" defining the voxel size of each image.

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

No additional information regarding each patient will be released.

Target entity(ies)

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

The provided CT scans include the abdomen. Occasionally the chest and/or pelvis as well

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.

Participants have to submit a Docker image that can segment the liver, spleen, pancreas, right kidney, left kidney, stomach, gallbladder, esophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum from CT scans. The segmentation process is **not allowed to use GPU**.

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)).

Hardware requirements, Runtime, Accuracy, Robustness, Sensitivity, Specificity.

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 dataset contains CT scans from various manufacturers, such as Siemens, GE, Philips, Toshiba, Imatron, Vital, and PHMS.

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

The training set is the same as the MICCAI FLARE 2022 challenge. Details are available at
Ma J. et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. arXiv preprint arXiv:2308.05862.

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.

Alberta Health Services
Barretos Cancer Hospital, Barretos, Sao Paulo, Brazil
Beaumont Health System, Royal Oak, MI
BioPartners, CA
Boston Medical Center, Boston, MA
Brigham and Women's Hospital, Boston, MA
Cleveland Clinic
Cureline, Inc. team and clinical network, Brisbane, CA
Hebrew University of Jerusalem
International Institute for Molecular Oncology, Poznan

IRCAD Institute Strasbourg
Lahey Hospital & Medical Center, Burlington, MA
Ludwig Maxmilian University of Munich
M Health Fairview
M.D. Anderson Cancer Center, Houston TX
Mayo Clinic, Rochester, MN
Memorial Sloan Kettering Cancer Center (New York, NY, USA)
National Cancer Institute, Bethesda, MD
National Institutes of Health, Bethesda MD
Polytechnique & CHUM Research Center Montral
Radboud University Medical Center of Nijmegen
Roswell Park Cancer Institute, Buffalo NY
Sheba Medical Center
St. Joseph's Hospital and Medical Center, Phoenix, AZ
Tel Aviv University
The National Institutes of Health Clinical Center(NIH)
University of California, San Francisco, CA
University of Chicago
University of North Carolina, Chapel Hill, NC
University of Pittsburgh/UPMC, Pittsburgh, PA
University of Sheffield
University of Southern California
Washington University School of Medicine, St. Louis, MO
Cancer Center West China Hospital
Erasmus MC Cancer Institute, Rotterdam
Fudan University Shanghai Cancer Center
Longgang Central Hospital of Shenzhen, China
Longgang District People Hospital
University Hospital Basel

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).

Data annotation is overseen by the challenge clinical chair Dr. Jian He. He is an experienced radiologist who specializes in abdomen CT.

Training and test case characteristics

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

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).

- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

One case, in this challenge, is a single CT scan associated with/without a multi-class segmentation mask

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

The total number of cases is 2600. There are 2050/250/300 cases for training, validation, and testing, respectively. In
the training set, 50 cases have 13-organ annotations. For the remaining 2000 unlabeled cases, we will provide pseudo labels that are generated by running the winning solution in FLARE 2022.
It should be noted that we will provide a new testing set (300 cases) for this task and the original testing test in FLARE 2022 will be used as the validation set.

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

In the FLARE 2022 challenge, the winning solution achieved superior performance on internal and external validation sets by using just 50 labeled cases and 2000 unlabeled cases. This task is an extension of the FLARE 2022 challenge, aiming to answer the question: would it be possible to make state-of-the-art segmentation models deployable on a laptop (without GPU).
In order to make an apple-to-apple comparison to the FLARE 2022 top solutions, we used the same training set in this task. Since the original testing set images have been publicly available, we provided a new testing set of 300 cases for this task.

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 the testing cases will be hidden from participants and the Docker container will be used for submission. Importantly, none of the labels is publicly available in the testing set.

Annotation characteristics

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

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

The annotaiton details are available at
Ma J. et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. arXiv preprint arXiv:2308.05862.
Specifically, we applied a hierarchical annotation pipeline that consisted of three stages to ensure accuracy and consistency throughout the annotation process. In the first stage, an annotation consensus was formulated by a senior radiologist, an oncologist, and a surgeon based on radiation therapy oncology group consensus (RTOG) panel guideline and Netter's anatomical atlas. Specifically, the liver contour should include all hepatic parenchyma and all liver lesions. The hepatic vessels inside the liver also need to be covered. If the vessels are located outside the liver (i.e., the entrance of portal hepatitis) based on the coronal view, they should be excluded from the liver contour. The kidney contour should include the renal parenchyma while excluding adjacent structures such as blood vessels and surrounding fat.

The spleen contour should include all splenic parenchyma and any splenic lesions. It should exclude adjacent structures such as the splenic vessels (arteries and veins), particularly those located outside the spleen. The pancreatic contour should encompass all pancreatic parenchyma including the head, body, and tail, as well as any pancreatic lesions. Exocrine, endocrine components, and the pancreatic duct need to be included, but the surrounding vessels and fat should be excluded. The aortic contour should include the entire lumen of the aorta, from the aortic root to the bifurcation. The aortic wall (including the aortic calcification) should also be included. The inferior vena cava contour should include the entire lumen and cover the walls. The adrenal gland contour should include the entire adrenal gland, both cortex and medulla, and any adrenal lesions. The gallbladder contour should encompass the entire gallbladder wall, including the body, fundus, and neck, as well as any gallstones or polyps. The cystic duct and the surrounding liver parenchyma should be excluded. The esophagus contour should include the entire esophageal wall, while adjacent structures such as the trachea, aorta, and surrounding fat and muscle should be excluded. The stomach contour should encompass the entire stomach wall including the fundus, body, antrum, and pylorus, as well as any gastric lesions. The duodenum contour should include the entire duodenal wall from the duodenal bulb to the ligament of Treitz, along with any duodenal lesions. It should exclude surrounding structures such as the head of the pancreas, common bile duct, and surrounding vasculature.
Before the annotation process, all annotators were required to learn and follow the annotation consensus.

The second stage of our annotation pipeline involved a human-in-the-loop approach aimed at enhancing annotation throughput. To facilitate this process, we utilized five 3D U-Net models trained via 5-fold cross-validation on existing abdomen CT datasets. Leveraging the predictions generated by these models, junior annotators performed manual refinements with the assistance of MedSAM on 100 randomly selected predictions, which were then checked and revised by senior radiologists. This process was iterated seven times until all the images were labeled by one of the junior annotators and further refined by the senior radiologists. In the third and final stage, we aimed to identify potential annotation errors. We trained a new set of U-Net models using five-fold cross-validation, with special attention given to images exhibiting low Dice Similarity Coefficient (DSC) scores (<0.75), which were then double-checked by the senior radiologists.

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.

Annotators and radiologists were asked to use 3D Slicer software or ITK-SNAP to annotate the organs.

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 annotation process involved a team of five junior radiologists with one to five years of experience and two senior radiologists with more than 10 years of experience.

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

To make all the labels consistent across the datasets from different medical centers, an experienced radiologist finally checked and revised all the annotations.

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.

We converted the dicom images to NifiTy format. The cases have various orientations. Participants should develop their own pre-processing methods.

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.

Error is unavoidable in medical image segmentation. We will have a forum to collect feedback from participants about possible errors that they discover in the dataset.

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)

Segmentation accuracy metrics: Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD).
Segmentation efficiency metrics: Running time.
We give running time a tolerance (the running time of the winning solution in FLARE 2022) because it satisfies the typical clinical requirements and the code has been publicly available.

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

The metrics are complementary. Specifically, DSC and NSD are used to measure the region error and boundary error, respectively.
Running time is used to measure the inference speed.

Ranking method(s)

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

The main benefit of this ranking scheme is that it can aggregate the metrics with different dimensions. A similar ranking scheme was also employed in the MICCAI BraTS Challenge 2017-2023 and FLARE 2021-2023.

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

Missing results on test cases will result in the ranks being set to the maximum (the number of teams). Specifically, missing results will get a zero value for DSC, NSD, and the equivalent worst value for running time.

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

Following the recommendation in ChallengeR, the ranking scheme includes the following three steps:
Step 1. For each testing case, we compute the three metrics;

Step 2. Rank participants for each of the 300 testing cases and each metric; Thus, each participant will have 900 (300x3) rankings.

Step 3. Average all these rankings.

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 testing cases with missing results will be assigned the worst performance. Thus, there is no missing data for the results analysis. Ranking variability will be characterized using the bootstrap.

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

The Bootstrap is a simple nonparametric method 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.

We will verify the performance of the ensembles of top teams.

TASK 3: Unsupervised domain adaptation for Abdomen organ segmentation in MRI

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.

Abdominal organ segmentation in CT scans has garnered significant attention during the past decade. Many segmentation benchmarks or challenges have been established, such as MICCAI BTCV, MICCAI MSD, MICCAI LiTS, MICCAI KiTS, MICCAI FLARE, and MICCAI AMOS. However, abdominal MRI, another commonly used imaging technology for abdominal disease diagnosis, remains relatively under-explored. One of the main bottlenecks is the scarcity of annotated abdominal MRI scans in the community. For example, the training set only contains 40 labeled MRI scans in the MICCAI AMOS challenge. In contrast, there are many abdominal CT scans with annotations or high-quality pseudo labels (about 90% in DSC) in the community and it is easy to collect many unlabeled MRI scans. This disparity motivates us to ask a question: **How can we build an abdominal MRI organ segmentation model without MRI annotations?**

To address this question, we set up this unsupervised cross-modality domain adaptation task for abdominal organ segmentation in MRI scans. We provide 2300 labeled abdominal CT scans and 1000 unlabeled MRI scans, representing the source and target domains, respectively. The challenge for participants is to use this dataset to develop a multi-organ MRI segmentation model without using any MRI annotations. Additionally, we offer another 100 and 300 MRI scans for validation and testing purposes, respectively. To summarize, this challenge task has three main features:

- Task: the first unsupervised cross-modality domain adaptation for abdominal organ segmentation
- Dataset: a diverse and large-scale abdominal CT (2300) and MRI dataset (1400)
- Evaluation Metrics: segmentation accuracy (DSC, NSD) and segmentation efficiency (running time and GPU consumption with tolerance)

Keywords

List the primary keywords that characterize the task.

Domain Adaptation, Abdomen, Segmentaion, Organ

ORGANIZATION

Organizers

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

Jun Ma (University of Toronto, University Health Network, Vector Institute)
Bo Wang (University of Toronto, University Health Network, Vector Institute)

b) Provide information on the primary contact person.

junma.ma@mail.utoronto.ca

Life cycle type

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

Examples:

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

Open call (challenge opens for new submissions after conference deadline)

Challenge venue and platform

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

N/A

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

CodaLab (Similar to the MICCAI FLARE 2023 Challenge)
Each task will have an independent challenge platform.

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

N/A

Participation policies

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

Any user interactions are not allowed.

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

No additional data is allowed.

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

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

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

We will provide certificates and souvenirs for the top-5 teams.

e) Define the policy for result announcement.

Examples:

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

All results 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).

All participants have the opportunity to publish their methods on LNCS proceedings. Top 5 teams in each task will be invited to write a challenge summary paper.

Submission method

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

Examples:

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

Docker containers

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

We will provide a validation set for participants to pre-evaluate their algorithms. The testing set has the same format as the validation set. To avoid overfitting the testing set, we only offer one submission opportunity on the testing 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

15 March 2023 (12:00 AM EST): Launch challenge registration and release training and validation data. Docker submission of validation set opening.

15 April 2023 (12:00 AM EST): Deadline for the first validation submission.

15 May 2023 (12:00 AM EST): Deadline for the second validation submission.

15 June 2023 (12:00 AM EST): Deadline for the third validation submission.

15 July 2023 (12:00 AM EST): Deadline for the third validation submission. Docker and short paper submission of the testing set opening.

15 August 2023 (12:00 AM EST): Deadline for the testing submission.

10 September 2023 (12:00 AM EST): Invite top teams to prepare presentations and participate in the MICCAI2023 Satellite Event.

6/10 October 2023: Announce final result

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).

Our dataset was curated from public domains and global data contributors with license permits. There is no patient information included. Thus, we do not need additional ethics approval.

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 available with the training data.

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

Top 5 teams should make their code publicly available. The remaining teams are encouraged to make their code publicly available.

Conflicts of interest

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

Only the organizers have access to the test case labels.
We are currently looking into which companies are willing to sponsor the challenge in the form of awards and computational resources

MISSION OF THE CHALLENGE

Field(s) of application

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

Examples:

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

Intervention assistance
Intervention follow-up
Intervention planning
Prognosis
Research

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 target cohort will be patients with abdominal CT and MRI scans that were acquired in different centers with different manufacturers.

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

The challenge cohort is composed of patients with various abdominal cancers as well as healthy subjects, which were curated from 30+ different medical centers.

Imaging modality(ies)

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

Computed Tomography and Magnetic Resonance Imaging

Context information

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

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

The image data will be released as compressed Nifti1Image objects stored in .nii.gz files. These files will include an "affine matrix" defining the voxel size of each image.

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

No additional information regarding each patient will be released.

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in

laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

The provided CT and MRI scans include the abdomen. Occasionally the chest and/or pelvis as well

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.

Participants have to submit a Docker image that can segment the liver, spleen, pancreas, right kidney, left kidney, stomach, gallbladder, esophagus, aorta, inferior vena cava, right adrenal gland, left adrenal gland, and duodenum from MRI 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)).

Hardware requirements, Runtime, Accuracy, Robustness, Sensitivity, Specificity.

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 dataset contains CT and MRI scans from various manufacturers, such as Siemens, GE, Philips, Toshiba, Imatron, Vital, and PHMS.

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

The CT scans in the training set are the same as the MICCAI FLARE 2022 challenge. Details are available at
Ma J. et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. arXiv preprint arXiv:2308.05862.
The MRI scans are from TCAI, AMOS, and data contributors from local hospitals under the license permission. Detailed Information can be found in the following references.
[1] Clark K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. Journal of Digital Imaging 26, no. 6 (2013): 1045-1057.
[2] Ji, Y., et al. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems, 35, (2023):36722-36732.

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.

Alberta Health Services
Barretos Cancer Hospital, Barretos, Sao Paulo, Brazil
Beaumont Health System, Royal Oak, MI
BioPartners, CA
Boston Medical Center, Boston, MA
Brigham and Women's Hospital, Boston, MA
Cleveland Clinic
Cureline, Inc. team and clinical network, Brisbane, CA
Duke University
Hebrew University of Jerusalem
International Institute for Molecular Oncology, Poznan
IRCAD Institute Strasbourg
Lahey Hospital & Medical Center, Burlington, MA
Ludwig Maxmilian University of Munich
M Health Fairview
M.D. Anderson Cancer Center, Houston TX
Mayo Clinic, Rochester, MN
Memorial Sloan Kettering Cancer Center (New York, NY, USA)
National Cancer Institute, Bethesda, MD
National Institutes of Health, Bethesda MD
Polytechnique & CHUM Research Center Montral
Radboud University Medical Center of Nijmegen
Roswell Park Cancer Institute, Buffalo NY
Sheba Medical Center
St. Joseph's Hospital and Medical Center, Phoenix, AZ
Tel Aviv University
The National Institutes of Health Clinical Center(NIH)
University of California, San Francisco, CA
University of Chicago
University of North Carolina, Chapel Hill, NC
University of Pittsburgh/UPMC, Pittsburgh, PA
University of Sheffield
University of Southern California
Washington University School of Medicine, St. Louis, MO
Cancer Center West China Hospital
Erasmus MC Cancer Institute, Rotterdam
Fudan University Shanghai Cancer Center
Longgang Central Hospital of Shenzhen, China
Longgang District People Hospital
University Hospital Basel

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).

Data annotation is overseen by the challenge clinical chair Dr. Jian He. He is an experienced radiologist who specializes in abdomen CT.

Training and test case characteristics

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

Examples:

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

One case, in this challenge, is a single CT or MRI scan associated with/without a multi-class segmentation mask

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

There are 2300 CT scans and 1000 MRI scans for model training. The CT scans are from the FLARE 2022 dataset where 50 cases have ground-truth labels and the remaining cases have pseudo labels (generated by the FLARE 2022 winning solution, around 90% DSC score). The 1000 MRI scans are unlabeled. The validation set and testing set contain 100 and 300 MRI scans, respectively.

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

As an unsupervised domain adaptation task, all the images in the source domain can be used for model training. For the target domain, we provide 100 and 300 new cases for validation and testing, respectively, where the number of cases is enough to draw statistically significant conclusions.

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 the testing cases will be hidden from participants and the Docker container will be used for submission. Importantly, none of the labels is publicly available in the testing set.

Annotation characteristics

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

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

Since this task has the same segmentation targets as Task 3, we applied the same annotation protocols.

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.

Annotators and radiologists were asked to use 3D Slicer software or ITK-SNAP to annotate the organs.

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 annotation process involved a team of five junior radiologists with one to five years of experience and two senior radiologists with more than 10 years of experience.

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

To make all the labels consistent across the datasets from different medical centers, an experienced radiologist finally checked and revised all the annotations.

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.

We converted the dicom images to NifiTy format. The cases have various orientations. Participants should develop their own pre-processing methods.

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.

Error is unavoidable in medical image segmentation. We will have a forum to collect feedback from participants about possible errors that they discover in the dataset.

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)

Segmentation accuracy metrics: Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD). Segmentation efficiency metrics: Running time and GPU memory consumption (area under GPU memory-time curve).

All metrics will be used to compute the ranking. We give GPU memory consumption a 4GB tolerance, GPU Memory = max(0, Real GPU Memory-4096MB) because all the previous top teams (i.e., FLARE21-23) can achieve state-of-the-art segmentation accuracy within this tolerance. We also give running time a tolerance (the running time of the winning solution in FLARE 2022) because it satisfies the typical clinical requirements and the code has been publicly available.

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

The metrics are complementary. Specifically, DSC and NSD are used to measure the region error and boundary error, respectively.
Running time and area under GPU memory-time curve are used to measure the inference speed and GPU consumption, respectively.

Ranking method(s)

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

The main benefit of this ranking scheme is that it can aggregate the metrics with different dimensions. A similar ranking scheme was also employed in the MICCAI BraTS Challenge 2017-2023 and FLARE 2021-2023.

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

Missing results on test cases will result in the ranks being set to the maximum (the number of teams). Specifically, missing results will get a zero value for DSC, NSD and the equivalent worst value for running time, and area under GPU memory-time curve.

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

Following the recommendation in ChallengeR, the ranking scheme includes the following three steps:
Step 1. For each testing case, we compute the four metrics;

Step 2. Rank participants for each of the 300 testing cases and each metric; Thus, each participant will have 1200 (300x4) rankings.

Step 3. Average all these rankings.

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 testing cases with missing results will be assigned the worst performance. Thus, there is no missing data for the results analysis. Ranking variability will be characterized using the bootstrap.

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

The Bootstrap is a simple nonparametric method 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.

We will verify the performance of the ensembles of top teams.

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.

N/A

Further comments

Further comments from the organizers.

N/A