TriALS: Triphasic-aided Liver Lesion Segmentation in Non-contrast CT: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

TriALS: Triphasic-aided Liver Lesion Segmentation in Non-contrast CT

Challenge acronym

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

TriALS-NCCT

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.

According to the latest WHO Globocan in 2020 and the most updated national cancer registry in Egypt in 2014, liver cancer ranks first as the most common cancer representing 20-35% of all cancer prevalence. Due to its high burden and cumulative risk to the population, various campaigns for the treatment of hepatitis virus were started and sealed to mitigate its effect. In screening and diagnosis for liver lesions, contrast agents are mostly used to make abnormalities more visible due to the low contrast differentiation between lesions and liver tissue. Contrast agents supply was heavily affected by COVID-19 global supply chain disruption and local economic instability afterward. Our goal is to establish a benchmark for liver tumor segmentation on non-contrast CT scans and show the potential of utilizing multi-phase data to enhance the training process, thereby enhancing lesion detection accuracy in NC imaging when access to contrast agents is restricted. Our challenge distinguishes itself with multi-phase CT (non-contrast and contrast-enhanced arterial (ART), portal venous (PV), and delayed phase cuts) intending to increase the upper bound for liver lesion segmentation.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

CT, segmentation, liver, lesion, hepatocellular carcinoma (HCC)

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 expect a large number of participants (20 to 30 teams). Compared to previous challenges, Liver Tumor Segmentation Benchmark (LiTS) (doi.org/10.1016/j.media.2022.102680), had 43 participating teams divided into ISBI 2017 (17 teams) and MICCAI 2017 (26 teams). Previous challenges highlighted the great interest of research teams in the segmentation of liver tumors. Lastly, challenges utilizing contrast-enhanced MRI data (atlas-challenge.u-bourgogne.fr/leaderboard) received more than 10 submissions in MICCAI 2024.
Unlike single-phase CT datasets, such as the LiTS dataset that provides only PV-phase CT images, we distinguish the challenge with multi-phase (non-contrast-enhanced (NC), arterial (ART), portal venous (PV), and delay phase signals) in the aim to increase the upper bound for liver lesion segmentation. To the best of our knowledge, this is the first challenge to utilize lesion segmentation in mult-phase CT imaging, where there is a lack of this type of data in the literature.
Moreover, our goal is to benchmark non-contrast CT performance and demonstrate how multi-phase data can refine training, thereby enhancing lesion detection accuracy in NC imaging with limited access to contrast agents.

Publication and future plans

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

Each team will be allowed to nominate two authors for the joint publication, detailing the challenge outcomes. Before the challenge starts, we will submit a paper on the challenge data and a proposed baseline method and will provide a baseline code to all participants. If not published at the beginning of the challenge, a pre-print arXiv version will be uploaded with all the necessary code for segmentation and evaluation. Notably, only teams whose methods outperform this baseline will be included in the joint publication.
All teams will be invited to provide a poster describing their methods. A select number of teams, chosen for their diversity and the novelty of their solutions, will also be invited to give presentations. It's important to note that until the joint challenge paper is published, participants must refrain from publishing any related results independently.

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.

synapse.org. Synapse will be used for communicating and uploading docker images where we will run the docker images on our local server. Hardware Restrictions during inference will be provided to ensure fairness and practicality. Our team has already succeeded in participating in and winning previous biomedical challenges and therefore is familiar with the challenges' pitfalls and potential improvements.

TASK 1: Automatic Lesion Segmentation in Multiphase CT Imaging

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.

Segmentation of liver lesions in multi-phase CT with a particular focus on non-contrast CT inference when multi-phase CT is unavailable due to the resources and limited access to contrast agents.

Keywords

List the primary keywords that characterize the task.

CT, segmentation, multi-phase, liver, tumor, hepatocellular carcinoma (HCC)

ORGANIZATION

Organizers

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

Technical Team:
[The Hong Kong University of Science and Technology, Hong Kong SAR]
Marawan Elbatel, Xiaomeng Li.
Clinical Team:
[1*AI Center of Excellence, Ain Shams University, Cairo, Egypt]
[2*Department of Radiology, Ain Shams University, Cairo, Egypt]
Mohamed Ghonim 1*2*, Mohanad Ghonim 1*2*, Amr Muhammad Abdo Salem 1*2*, Nouran Elghitany 1*2*, Noha Elghitany 1*2*, Amira Adel 2*, Susan Adil Ali 2*, Aya Yassin 1*2*

b) Provide information on the primary contact person.

Marawan Elbatel (mkfmelbatel@connect.ust.hk)

Life cycle type

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

Examples:

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

One-time event with a fixed submission deadline.

Challenge venue and platform

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

MICCAI.

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

synapse.org

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

N/A

Participation policies

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

Fully automatic.

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

Publicly available data is allowed, and publicly available pre-trained models. Participants are allowed to use public data only if they are publicly available at the opening, when the challenge starts, including pre-trained models. Institutions can utilize private data to benchmark their algorithm, submissions utilizing private data would not be eligible for awards.

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 are not eligible for awards and are not listed in the final leaderboard.

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

Currently, the challenge prize will be honorary and symbolic, but we will try to find a sponsor until the challenge.

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

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

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

Each team will be allowed to nominate two authors for the joint publication, detailing the challenge outcomes. Before the challenge starts, we will submit a paper on the challenge data and a proposed baseline method and

will provide a baseline code to all participants. If not published at the beginning of the challenge, a pre-print arXiv version will be uploaded with all the necessary code for segmentation and evaluation. Notably, only teams whose methods outperform this baseline will be included in the joint publication.
All teams will be invited to submit a report describing their methods. A select number of teams, chosen for their diversity and the novelty of their solutions, will be invited to give in-person presentations. It's important to note that until the joint challenge paper is published, participants must refrain from publishing any related results independently.

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.

A sample docker submission with a baseline method as well as a complete step-by step submission instructions will be released. Moreover, code for evaluation and testing will be provided to participants. Teams will submit the docker containers via synapse. At the final testing stage, reporting of the methodology would be required.

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.

Participants will be allowed to submit three submissions for evaluating their algorithms on a validation sample to ensure the correctness of the docker container as well as provide preliminary results. For the final evaluation of the test dataset, only the last submitted container will be used.

Challenge schedule

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

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

Challenge website and challenge registration opens: 17th April 2024
Training data release: 17th April 2024
Team registration open: 17th April 2024
Validation Phase Open: 5th August 2024
Team registration closes: 20th August 2024
Docker submission deadline: 11th September 2024
Methodology submission: 11th September 2024

Release of Results and Challenge Day: (6th or 10th of October 2024)

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 dataset is collected and acquired at Ain Shams University Hospitals, Cairo, Egypt. Since the collected data are fully anonymized to be completely untraceable (no personal fields) and the study is retrospective, it is not necessary to go through the process to get an ethics approval number, yet the project file is being finalized and will be submitted in January for the official response from the ethics committee and approval if required.

Data usage agreement

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

Examples:

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

CC BY-NC-ND,
Additional comments: Teams should sign a data usage agreement before downloading data.

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 evaluation script to produce the ranking as well as a docker baseline with detailed instructions will be available for participants to use.

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

Participating teams are not required yet encouraged to release their code as open-access.

Conflicts of interest

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

No conflict of interest. We will update MICCAI if a sponsor is to be included.

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

Screening, Diagnosis, Decision support, Intervention planning, 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 is the future patients admitted at Ain Shams University Hospitals, Cairo, Egypt with any liver lesion detected on CT. Deep learning models trained on the African population may not generalize well on different populations (e.g different vendors, pathologies). However, interested researchers will have an overview of the state-of-the-art methods concluded and summarized by the challenge. Moreover, methods trained on cross-domain datasets can be evaluated for domain generalization purposes.

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

A retrospective and prospective cohort of Adult male and female patients > 18 years with contrast and non-contrast enhanced abdominal CT studies, scans were acquired from the Radiology Department at Ain Shams University Hospitals, Cairo, Egypt.

Imaging modality(ies)

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

CT without contrast and with contrast in different phases:
Time at which the acquisition is performed after iodinated contrast* agent injection:
- Arterial time: Upon bolus tracking in descending aorta by smart prep (usually about 35-45 sec post-injection);
- Portovenous time: 20 seconds post arterial (usually about 65 to 70 sec post-injection);
- Delayed time: 3 minutes post porto venous (Usually up to five minutes post-injection).
*Contrast agents used are variable due to changes in supply chains and limited import. We sometimes use non-ionic contrast agents ("Omnipaque"), when unavailable, we use ionic contrast agents ("Telebrix")

Context information

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

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

None

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

No clinical information on patients.

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.

Liver visible as a whole in contrast and non-contrast enhanced CTs of the abdomen, pelvis-abdomen, and chest.

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

The algorithm aims to automatically segment the associated lesion(s) within the liver.

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)).

Find highly accurate liver segmentation algorithm for CT images, importantly on non-contrast CT without requiring multi-phase contrast information during inference.

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).

- GE Healthcare Optima 128 slice
- Toshiba Prime Aquilion 64 slice
- GE BrightSpeed CT 16 slice

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

Data acquisition includes CT in different phases (without contrast, arterial, portal venous, delayed). Acquisition parameters vary in a normal range, and the time of the acquisition after the contrast agent (arterial, portal, delay) varies due to the vascularity diversity of patients.

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.

Images were acquired at Ain Shams University Hospitals, Cairo, Egypt.

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).

Radiology residents, as well as, specialists and consultants of 3+ and 10+ years of experience.

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 Training, validation, and test cases represents four 3D CT images (non-contrast, arterial, portal, delayed) and has been labeled to include all lesions within the liver region.

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

Sixty cases for training, ten for validation, and thirty for testing.

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

With the lack of publicly available multi-phase liver datasets, a 60:30 train/test split will be used for significant assessment of the results.

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.

There is a strong imbalance between the semantic classes: lesion(s), and background.

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.

We use a hybrid Human-Algorithm labeling scheme. Firstly, we deploy our algorithm for the preliminary segmentation of all scans by ensuring that the phases are well registered. We propagate the lesion from the most visible phase and add to it in other phases. Our algorithm has been trained on publicly available liver tumor datasets (LiTS) and a few-shot in-house data manually delineated by a medical doctor resident. Secondly, all segmentations in all phases are corrected when necessary and modified by a medical doctor resident. Finally, all cases are reviewed by another medical doctor resident for inter-rater-variability. To address variability, both residents sit together to resolve conflicts. In case no agreement is reached, experienced radiologists intervene to resolve.

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.

Lesion(s) segmentation: all lesion volumes were considered (if more than one was present).
We provide a dedicated plan and training material to explain the instructions to annotators, including an introduction to ITK-SNAP and tips on how to use it. We also share lesions that we think might be conflicting to reach conjoint decisions with our consultants on how to proceed evidence-based with such lesions further on.

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.

Two radiologist residents medically trained with experience in delineating previous international challenges are responsible for annotating the cases to calculate inter-rater-variability as a human baseline.

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

None

Data pre-processing method(s)

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

All images will be in their native CT modality space, mirroring the 'raw' format commonly encountered in clinical settings. We challenge participants to tackle data pre-processing in novel and creative ways (e.g. normalization, windowing). While our baseline code release will encompass a data preprocessing scheme, we encourage participants to explore and apply their unique 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.

There exist uncertainties in the image annotations. For that purpose, we use a hybrid Human-Algorithm labeling scheme and would provide inter-reader variability, with cases reviewed by two medical residents, if no consensus between them, the scan will be reviewed by additional radiologists(more than three years of experience). In all cases, non-contrast CT scans are acquired with variable slice thicknesses of 1.25/5/10 mm (according to the CT it is performed on usually) compared to contrast-enhanced CT where the slice thickness is usually 1.25/5 mm. Another expected issue that would be faced is the trial to actually find some lesions in the non-contrast phase and therefore annotate them. This might not be always avoidable and we might accept not being able to annotate some lesions in the non-contrast phase alone. When faced with such an issue of having a lesion on contrast-enhanced phases yet not being able to locate them on non-contrast images, we might resort to label propagation from the phase that the lesion is most visible in or the possibility of excluding such studies.

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

No other source of error.

ASSESSMENT METHODS

Metric(s)

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

- Example 1: Dice Similarity Coefficient (DSC)

- Example 2: Area under curve (AUC)

We will follow established challenges, LiTS (doi.org/10.1016/j.media.2022.102680) and ATLAS (https://atlas-challenge.u-bourgogne.fr/) to compute the metrics for our challenge, focusing on per-case voxel-wise segmentation metrics:
-Dice Similarity Coefficient.
-Average Symmetric Surface Distance (ASD).
-5 mm Surface Dice (SD).
-Maximum Symmetric Surface Distance (MSD):

Acknowledging the distinctive nature of our challenge, which involves the segmentation of lesions in a multi-phase setting, we will adapt the application of these metrics in two distinct scenarios:
1) In the presence of multi-phase imaging data during inference.
2) In scenarios relying solely on non-contrast CT data, without access to multi-phase imaging.

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

The selected metrics, emphasizing clinical relevance, aim to enhance liver lesion(s) segmentation by leveraging multi-phase data.
Addressing the shortage of contrast in our target cohort for the final biomedical application, we prioritize algorithms advancing tumor detection even without contrast-enhanced phases.

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 "rank then aggregate" approach is utilized to favor algorithms that perform consistently well across all metrics. Notably, when evaluating diverse metrics that may not lend themselves to direct normalization or comparison, "rank then aggregate" promotes fairness and transparency.

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

We will assign the worst score on all metrics for missing predictions.

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

Rank then aggregate. Specifically, we will follow the following steps:
1) Individual Metric Ranking: We will compare the performance of the algorithms on each metric for all teams individually.
2) Aggregated Ranking for Final Evaluation: The mean rank of each algorithm is computed across all metrics providing the final rank for the leaderboard.

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 Wilcoxon signed-rank test will be employed following confirmation of non-normal distribution through the Shapiro-Wilk test.

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

Being robust against outliers and non-normal data, we selected the Wilcoxon signed rank due to its lack of reliance on the assumption of normal distribution.

Further analyses

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

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

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.

N/A

Further comments

Further comments from the organizers.

N/A