Body Maps: Towards 3D Atlas of Human Body: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Body Maps: Towards 3D Atlas of Human Body

Challenge acronym

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

BodyMaps

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.

Variations in organ sizes and shapes can indicate a range of medical conditions, from benign anomalies to life-threatening diseases. Precise organ volume measurement is fundamental for effective patient care, but manual organ contouring is extremely time-consuming and exhibits considerable variability among expert radiologists. Artificial Intelligence (AI) holds the promise of improving volume measurement accuracy and reducing manual contouring efforts. We formulate our challenge as a semantic segmentation task, which automatically identifies and delineates the boundary of various anatomical structures essential for numerous downstream applications such as disease diagnosis,
prognosis, and surgical planning. Our primary goal is to promote the development of AI algorithms and to benchmark the state of the art in this field. The BodyMaps challenge particularly focuses on assessing and improving the generalizability and efficiency of AI algorithms in medical segmentation across diverse clinical settings and patient demographics. In light of this, the innovation of our BodyMaps challenge includes the use of (1) large-scale, diverse datasets for both training and evaluating AI algorithms, (2) novel evaluation metrics that emphasize the accuracy of hard-to-segment anatomical structures, and (3) penalties for algorithms with extended inference times. Specifically, this
challenge involves two unique datasets. First, AbdomenAtlas, the largest annotated dataset [Qu et al., 2023, Li et al., 2023], contains a total of 10,142 three-dimensional computed tomography (CT) volumes. In each CT volume, 25 anatomical structures are annotated by voxel. AbdomenAtlas is a multi-domain dataset of pre, portal, arterial, and delayed phase CT volumes collected from 88 global hospitals in 9 countries, diversified in age, pathological conditions, body parts, and race background. The AbdomenAtlas dataset will be made available to the public progressively during the challenge period and the participants are encouraged to use any other public/private datasets for training AI algorithms. Second, W-1K is a proprietary collection of 1,000 CT volumes, where 15 anatomical structures are annotated by voxel. The CT volumes and annotations of W-1K will be reserved for external validation of AI algorithms. The final score will be calculated on the W-1K dataset, measuring both segmentation performance and inference speed of the AI algorithms. Note that the segmentation performance

will not only be limited to the average segmentation performance but also prioritize the performance of hard-to-segment structures. We hope our BodyMaps challenge can set the stage for larger-scale clinical trials and offer exceptional opportunities to practitioners in the medical imaging community.

References

Qu, C., T. Zhang, H. Qiao, J. Liu, Y. Tang, A. L. Yuille, Z. Zhou. "AbdomenAtlas-8K: Annotating 8,000 CT Volumes for Multi-Organ Segmentation in Three Weeks." In NeurIPS, 2023.
Li, W., A. L. Yuille, Z. Zhou. "How Well Do Supervised 3D Models Transfer to Medical Imaging Tasks?" 2023.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Anatomical Structure Segmentation, Organ Volume Measurement, Domain Adaptation, Domain Generalization, Computed Tomography

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?

N/A

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.

N/A

Publication and future plans

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

N/A

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.

N/A

TASK 1: BodyMaps

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.

Variations in organ sizes and shapes can indicate a range of medical conditions, from benign anomalies to life-threatening diseases. Precise organ volume measurement is fundamental for effective patient care, but manual organ contouring is extremely time-consuming and exhibits considerable variability among expert radiologists. Artificial Intelligence (AI) holds the promise of improving volume measurement accuracy and reducing manual contouring efforts. We formulate our challenge as a semantic segmentation task, which automatically identifies and delineates the boundary of various anatomical structures essential for numerous downstream applications such as disease diagnosis,
prognosis, and surgical planning. Our primary goal is to promote the development of AI algorithms and to benchmark the state of the art in this field. The BodyMaps challenge particularly focuses on assessing and improving the generalizability and efficiency of AI algorithms in medical segmentation across diverse clinical settings and patient demographics. In light of this, the innovation of our BodyMaps challenge includes the use of (1) large-scale, diverse datasets for both training and evaluating AI algorithms, (2) novel evaluation metrics that emphasize the accuracy of hard-to-segment anatomical structures, and (3) penalties for algorithms with extended inference times. Specifically, this
challenge involves two unique datasets. First, AbdomenAtlas, the largest annotated dataset [Qu et al., 2023, Li et al., 2023], contains a total of 10,142 three-dimensional computed tomography (CT) volumes. In each CT volume, 25 anatomical structures are annotated by voxel. AbdomenAtlas is a multi-domain dataset of pre, portal, arterial, and delayed phase CT volumes collected from 88 global hospitals in 9 countries, diversified in age, pathological conditions, body parts, and race background. The AbdomenAtlas dataset will be made available to the public progressively during the challenge period and the participants are encouraged to use any other public/private datasets for training AI algorithms. Second, W-1K is a proprietary collection of 1,000 CT volumes, where 15 anatomical structures are annotated by voxel. The CT volumes and annotations of W-1K will be reserved for external validation of AI algorithms. The final score will be calculated on the W-1K dataset, measuring both segmentation performance and inference speed of the AI algorithms. Note that the segmentation performance will not only be limited to the average segmentation performance but also prioritize the performance of hard-to-segment structures. We hope our BodyMaps challenge can set the stage for larger-scale clinical trials and offer exceptional opportunities to practitioners in the medical imaging community.

References

Qu, C., T. Zhang, H. Qiao, J. Liu, Y. Tang, A. L. Yuille, Z. Zhou. "AbdomenAtlas-8K: Annotating 8,000 CT Volumes for Multi-Organ Segmentation in Three Weeks." In NeurIPS, 2023.
Li, W., A. L. Yuille, Z. Zhou. "How Well Do Supervised 3D Models Transfer to Medical Imaging Tasks?" 2023.

Keywords

List the primary keywords that characterize the task.

Anatomical Structure Segmentation, Organ Volume Measurement, Domain Adaptation, Domain Generalization, Computed Tomography

ORGANIZATION

Organizers

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

Organizers:
Johns Hopkins University: Wenxuan Li, Yu-Cheng Chou, Jieneng Chen, Chongyu Qu, Alan Yuille, Zongwei Zhou
University of Science and Technology of China: Qi Chen

Technical Support:

Johns Hopkins University: Yaoyao Liu, Angtian Wang, Junfei Xiao (Johns Hopkins University)
NVIDIA: Yucheng Tang

Annotation Team:

Experts:
Shanghai Jiao Tong University: Xiaoxi Chen
The First Affiliated Hospital, Zhejiang University School of Medicine: Jincheng Wang

Trainees:
The First Hospital of China Medical University: Huimin Xue
Johns Hopkins University: Yixiong Chen
Shengjing Hospital of China Medical University: Yujiu Ma
Southeast University: Yuxiang Lai
Rutgers University: Hualin Qiao
China Medical University: Yining Cao, Haoqi Han, Meihua Li, Xiaorui Lin, Yutong Tang, Jinghui Xu

b) Provide information on the primary contact person.

Zongwei Zhou, PhD
Computer Science, Johns Hopkins University
248 Malone Hall, 3400 N Charles St, Baltimore, MD 21218
E-mail: zzhou82@jh.edu
Phone: +1 (480)738-2575

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

- There will be repeated events with an annual fixed submission deadline.
- The challenge will be hosted on the CodaLab platform.
- Eligibility for awards will be limited to submissions received before the deadline of each event.
- The challenge submission portal will remain open, and the dataset will continue to be accessible even after the event deadline.
- In the future, we may supplement the challenge with additional data and/or annotations.

Challenge venue and platform

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

ISBI (2024) coordinated with the MICCAI Special Interest Group for Biomedical Image Analysis Challenges (BIAS)

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

CodaLab (https://codalab.lisn.upsaclay.fr/competitions/16919)

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

https://ccvl.jhu.edu/bodymaps

Participation policies

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

semi-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.

Additional public/private data and pre-trained models are allowed for AI training.

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.

Participants who have the access to the W-1K dataset are not eligible for awards and will not be listed on the leaderboard.

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

(1) We will provide cash prizes (alternative way: equal value Amazon gift card) for the top 5 teams:
- First prize: 500 USD
- Second prize: 300 USD
- Third prize: 200 USD
- 4rd-5th: 100 USD

This will be funded by the Lustgarten Foundation for Pancreatic Cancer Research and the Patrick J. McGovern Foundation Award.
(2) A certification will be awarded to the top 10 performing teams at the ISBI 2024 conference.

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 submission results will be reported on the leaderboard.
- All participating teams have the opportunity to publish their results on the ISBI 2024 and other vision conference proceedings.
- The top 10 performing methods (teams) will be announced publicly and invited to give oral presentations during the main conference.
- We will also host corresponding workshops at premier computer vision conferences (e.g., CVPR, ECCV, MICCAI), where participants with the most successful and innovative entries are invited to present their methods and results.

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 should register for this challenge with their real names, affiliations (including department, full name of university/institute/company, country), and affiliation E-mails.
- Incomplete and redundant registrations will be removed without notice. Each team can have at most ten people.
- Participants are not allowed to register multiple teams and accounts. Participants from the same research group are also not allowed to register multiple teams. BodyMaps organizers keep the right to disqualify such participants.
- All participants must submit a complete solution to this challenge for testing. A complete solution includes a Docker container (tar file) and a qualified methodology paper (at least 2 pages, LNCS format).
- All participants should agree that the submitted short papers can be publicly available to the community on the challenge website, and organizers can use the information provided by the participants, including scores, predicted labels, and papers.
- All participating teams have the opportunity to publish their own results/findings on the ISBI 2024 or other vision conference proceedings. Every member of the participating teams may qualify as an author.
- We invite researchers from diverse backgrounds, including those from underrepresented regions in medical AI research, to participate in the BodyMaps Challenge. This diversity will enrich the challenge with a wide range of perspectives and approaches, enhancing the quality and applicability of the developed solutions.

Submission method

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

Examples:

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

The challenge submission is based on the Docker container. Once participants join the challenge, they must submit a complete solution to organizers via email (to: bodymaps.official@gmail.com) for testing. A complete solution refers to a Docker container (tar file). The challenge organizers are responsible for providing an evaluation server and evaluating the performance of Docker containers on the W-1K dataset. More submission details will be published at the time point of the challenge announcement.

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 leverage TotalSegmentator and DAP Atlas as the validation datasets. Participants are encouraged to test their algorithms using the external validation datasets before making their final submissions. A unique feature of our challenge lies in the fact that superior performance on the TotalSegmentator and DAP Atlas does not guarantee similar results on our official test set. This feature highlights the challenge of algorithm generalizability, a key focus of our competition. Our test set is constructed from a proprietary W-1K dataset of 1,000 CT volumes, where 15 anatomical structures are annotated by voxel. CT volumes and annotations of W-1K will not be disclosed to the public at any time and are exclusively reserved for external validation of AI algorithms. To avoid overfitting the test set, we only offer successful submission opportunities twice a month.
- All teams can make docker submissions via email (the detailed submission format for validation will be provided when the challenge starts).
- If the Docker container does not work, we will return back the error information to the participants. Participants with technical failure are allowed to resubmit their algorithms with one extra time.
- When the evaluation is finished, we will return back the evaluation metrics via email.
- All valid submission results will be reported on the leaderboard.

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 running and registration open: 01/08/2024
- Release of the starter code: 01/10/2024
- Release of the dataset: 01/20/2024
- Paper submission deadline: 04/06/2024
- Evaluation submission deadline: 04/15/2024
- Release of final results (paper submission and evaluation submission decision): 04/20/2024
- Challenge days (ISBI main conference): 05/27/2024 - 05/30/2024

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).

For AbdomenAtlas, we will only disseminate the annotations of the CT volumes separately, and participants will retrieve the original CT volumes, if needed, from the original sources (websites). Everything we intend to create and license-out will be in separate files and no modifications are necessary to the original CT volumes. We have consulted with the lawyers at Johns Hopkins University, confirming the permissions of distributing the annotations based on the license of each dataset. We will further include detailed download instructions on our challenge webpage.

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 (Attribution-NonCommercial-NoDerivs)

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.

Code for algorithm evaluation will be accessible for all participants on GitHub when the challenge starts (https://github.com/johnson111788/BodyMaps).

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

Publication of algorithm code will be a prerequisite for award eligibility. To this end, the code will need to be published on a publicly accessible repository of the teams' choice within three days (before 04/18 11:59 PM EST) after the submission deadline.

Conflicts of interest

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

This challenge will be supported by the Lustgarten Foundation for Pancreatic Cancer Research and the Patrick J. McGovern Foundation Award. Participants who have access to the W-1K dataset will be allowed to participate but not eligible for awards and not listed on the leaderboard.

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

Research, Intervention assistance, Treatment planning, Organ volume measurement

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 final biomedical application targets patients requiring diagnostic or therapeutic interventions involving computed tomography (CT) imaging, particularly those undergoing abdominal imaging for various conditions such as tumors, trauma, or vascular diseases. This cohort includes a diverse population in terms of age, gender, and ethnicity.

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

The challenge data is acquired from patients represented in the AbdomenAtlas [Qu et al., 2023, Li et al., 2023] and W-1K datasets, encompassing a broad spectrum of pathological conditions, age groups, and demographic backgrounds. This ensures that the challenge reflects a real-world, diverse patient population. Detailed statistics can be found in the corresponding publications.
We will provide 75K masks and 1.2M annotated images that are taken from 68 hospitals worldwide, spanning four distinct phases: pre, portal, arterial, and delayed.

Reference
Qu, C., T. Zhang, H. Qiao, J. Liu, Y. Tang, A. L. Yuille, Z. Zhou. "AbdomenAtlas-8K: Annotating 8,000 CT Volumes for Multi-Organ Segmentation in Three Weeks." In NeurIPS, 2023.
Li, W., A. L. Yuille, Z. Zhou. "How Well Do Supervised 3D Models Transfer to Medical Imaging Tasks?" 2023.

Imaging modality(ies)

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

AbdomenAtlas provides computed tomography (CT) volumes with four different enhancement phases:
pre (16%),arterial (37%), portal (44%), and delayed (3%). W-1K provides contrast CT volumes in both venous and arterial phases.

Context information

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

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

No additional information will be provided regarding the image data.

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

No additional information will be provided regarding the patient in general.

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 challenge data consists of chest and abdominal computed tomography (CT) scans, with the scan range typically extending from the thoracic inlet to the pelvic floor. If clinically relevant, scans can be extended to cover the entire body of the target cohort. This imaging protocol is representative of standard clinical practice for comprehensive examinations.
For the challenge cohort, the dataset includes a diverse array of cases from multiple global institutions, reflecting a broad spectrum of patient demographics and pathologies. For the target cohort in the final biomedical application, the scope could potentially extend to include full-body scans, encompassing a more extensive array of anatomical structures, to cater to varied clinical needs. We aim to map the anatomy labels to the ontology in the Radlex or UMLS database, aiming to establish a more comprehensive integrated database of medical informatics and imaging for longer-term application.

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 challenge structures that the participating algorithms have been designed to focus on multi-organ segmentation within the chest and abdomen regions, while the target ones will focus on a more extensive range of anatomical structures, including those found in the cardiac region, vertebrae, and muscles. Our preliminary findings indicate that AI algorithms trained on the AbdomenAtlas dataset demonstrate promising transfer learning abilities. These abilities extend to other anatomical structures beyond the abdomen, suggesting that the algorithms can adapt and perform effectively on various body regions, including the chest, head, and extremities. This adaptability is crucial for the algorithms to be practical and versatile in diverse medical imaging scenarios.

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)).

Our primary goal is to promote the development of advanced AI algorithms and to benchmark the state of the art in this field. Our BodyMaps challenge particularly focuses on assessing and improving the generalizability and efficiency of AI algorithms in medical segmentation across diverse clinical settings and patient demographics. Therefore, three properties of the algorithms must be optimized to perform well in this challenge.
1. Segmenting anatomical structures with high accuracy, especially those hard-to-segment structures, including but not limited to those with low contrast in soft tissues resulting in
indistinct boundaries and organs with complex morphologies like tubular structures.
2. Performing robustness when analyzing CT volumes sourced from a large variety of scanner type, contrast enhancement, patient demographics, and variations in organ appearances
caused by different imaging protocols or patient positioning (e.g., rotations along the vertical axis between 30 and

60 degrees).
3. Optimizing computational cost and reducing inference time. We encourage the AI algorithms to efficiently segment 15 anatomical structures within three-dimensional CT volumes, achieving
this within a timeframe of 90 seconds and a GPU memory consumption of 24GB during testing, on average per case.

These criteria are intended to push the boundaries of current medical image segmentation technology and ensure that the resulting AI models can be effectively translated into clinical practice, improving patient outcomes through better diagnostics and treatment planning.
We will evaluate the segmentation performance of 15 anatomical structures using the W-1K dataset. Their class indexes are provided as follows.

Spleen: 1
Right kidney: 2
Left kidney: 3
Gallbladder: 4
Liver: 5
Stomach: 6
Aorta: 7
Inferior Vena Cava: 8
Pancreas: 9
Right adrenal gland: 10
Left adrenal gland: 11
Duodenum: 12
Colon: 13
Intestine: 14
Celiac Trunk: 15

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).

Challenge data contain CT scans taken by a variety of vendors, e.g. Philips, Siemens, GE, and Toshiba. The CT volumes span four contrast-enhancement phases: pre, arterial, portal and delayed.

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

All the released 3K CT volumes were taken from publicly available datasets. Different from a naive combination of these public datasets (only containing 13K masks), our released AbdomenAtlas
provides 75K annotated organ masks for these CT volumes, substantially increasing the number of masks by 5.7 times.
Hereby, in the following, we differentiate our extensively annotated dataset from its original public counterpart by

referring to it as AbdomenAtlas-XXX.
- AbdomenAtlas-LiTS: The in-plane image resolution ranges from 0.56 mm to 1.0 mm, and 0.45 mm to 6.0 mm in slice thickness. Also, the number of axial slices ranges from 42 to 1026.
- AbdomenAtlas-KiTS: The CT scan for each patient is the most recent preoperative CT study containing at least one series in late arterial contrast phase that depicts the entirety of the abdomen.
- AbdomenAtlas-CHAOS: The images include a single CT scan of the upper abdomen area. The CT volumes were acquired at the portal venous phase after contrast agent injection, average axial slice number is 160, slice thickness 2.0-3.2 mm.
- AbdomenAtlas-BTCV: The 50 scans were captured during portal venous contrast phase with variable volume sizes (512 x 512 x 85 - 512 x 512 x 198) and field of views (approx. 280 x
280 x 280 mm3 - 500 x 500 x 650 mm3). The in-plane resolution varies from 0.54 x 0.54 mm2 to 0.98 x 0.98 mm2, while the slice thickness ranges from 2.5 mm to 5.0 mm.
- AbdomenAtlas-MSD: The pancreas and liver dataset include 420 portal venous phase CT scans were obtained with the following reconstruction and acquisition parameters: pitch/table speed 0.984-1.375/39.37-27.50 mm; automatic tube current modulation range, 220-380 mA; noise index, 12.5-14; 120 kVp; tube rotation speed, 0.7-0.8 ms; scan delay, 80-85 s; and axial slices reconstructed at 2.5 mm intervals.

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.

- AbdomenAtlas-LiTS: Acquired from seven clinical sites all over the world, including Rechts der Isar Hospital, the Technical University of Munich in Germany; Radboud University Medical Center, the Netherlands; Polytechnique Montreal and CHUM Research Center in Canada; Sheba Medical Center in Israel; the Hebrew University of Jerusalem in Israel; Hadassah University Medical Center in Israel; IRCAD in France. We will provide all the download links for the original CT volumes to participants when the challenge starts.

- AbdomenAtlas-KiTS: University of Minnesota Medical Center.

- AbdomenAtlas-CHAOS: Collected from the Department of Radiology, Dokuz Eylul University Hospital, Izmir, Turkey.

- AbdomenAtlas-MSD: Data contributing institutions included: 1) Center for Biomedical Image Computing and Analytics (CBICA), University of Penn- sylvania, PA, USA, 2) University of Alabama at Birmingham, AL, USA, 3) Heidelberg University, Germany, 4) University Hospital of Bern, Switzerland, 5) University of Debrecen, Hungary, 6) Henry Ford Hospital, MI, USA, 7) University of California, CA, USA, 8) MD Anderson Cancer Center, TX, USA, 9) Emory University, GA, USA, 10) Mayo Clinic, MN, USA, 11) Thomas Jeffer- son University, PA, USA, 12) Duke University School of Medicine, NC, USA, 13) Saint Joseph Hospital and Medical Center, AZ, USA, 14) Case Western Reserve University, OH, USA, 15) University of North Carolina, NC, USA, 16) Fondazione IRCCS Istituto Neurologico Carlo Besta, Italy, 17) Washington University School of Medicine in St. Louis, MO, USA, and 18) Tata Memorial Centre, Mumbai, India. The heart dataset was provided by King's College London (London,
United Kingdom). This liver dataset is from several clinical sites, including Ludwig Maximilian University of Munich (Germany), Radboud University Medical Center of Nijmegen (The Netherlands), Polytechnique and CHUM Research Center Montreal (Canada), Tel Aviv University (Israel), Sheba Medical Center (Israel), IR- CAD Institute Strasbourg (France), and Hebrew University of Jerusalem (Israel). The dataset for Hippocampus is taken from the Psychiatric Genotype/Phenotype Project data repository at Vanderbilt University Medical Center (Nashville, TN,

USA).

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).

Creating 75K high-quality organ masks for 3K CT volumes requires extensive medical knowledge and annotation cost (much more difficult than annotating natural images). Based on our experience and those reported in [Park et al., 2020], trained radiologists annotate abdominal organs at a rate of 30-60 minutes per organ per three-dimensional CT volume. This translates to 3000K human hours for completing released AbdomenAtlas. We employed a highly efficient annotation method, combining AI with the expertise of four senior radiologists with 8-15 years of clinical experience and six junior radiologists with 3-5 years of experience. All radiologists have received training in utilizing annotation standards and annotation software to guarantee the quality and consistency of annotations. We use active learning [Qu et al., 2023] to overcome this challenge and produce the largest annotated dataset to date.

Reference

Park, S., L. C. Chu, E. K. Fishman, A. L. Yuille, B. Vogelstein, K. W. Kinzler, K. M. Horton et al. "Annotated normal CT data of the abdomen for deep learning: Challenges and strategies for implementation." Diagnostic and interventional imaging, 2020.
Qu, C., T. Zhang, H. Qiao, J. Liu, Y. Tang, A. L. Yuille, Z. Zhou. "AbdomenAtlas-8K: Annotating 8,000 CT Volumes for Multi-Organ Segmentation in Three Weeks." In NeurIPS, 2023.

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.

In the BodyMaps Challenge, a 'case' refers to a 3D CT volume accompanied by expert annotations of 25 anatomical structures. The goal is to segment these structures accurately. For training purposes, we plan to release 3,000 annotated CT volumes and using any other public/private datasets is encouraged. Also, we recommend using external datasets (comprising 1,737 CT volumes) to validate before submitting for testing. The use of the private dataset (1,000 CT volumes in total) for testing ensures a comprehensive representation of diverse medical scenarios, enhancing the generalizability of developed algorithms and mirroring the variability encountered in clinical practice.

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

The AbdomenAtlas will be released to the public by stage (1000 annotated CT volumes per stage). At the end of this time challenge, it will provide 3,000 CT volumes with per-voxel annotations for 25 structures. The dataset will be used as the training set to develop their AI algorithms. The participants are encouraged to use any other public/private datasets for training AI algorithms
For external validation, TotalSegmentator and DAP Atlas are two recommended publicly available datasets, providing 1,204 and 533 CT volumes, respectively.
For official testing and benchmarking, we have 1,000 CT volumes from the proprietary W-1K dataset, which will not be shared publicly to preserve the integrity of the evaluation process.

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

The distribution of cases (3,000 for training) has been carefully chosen to reflect the variability and breadth of clinical conditions encountered in medical practice. By recommending additional 1,737 cases for external validation, we provide participants with a fair and controlled comparison among each other, promoting methodological diversity and innovation.

The proprietary test set size (1,000) was selected to provide a comprehensive but manageable dataset for evaluating generalizability and performance of the AI models.

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.

The training and validation datasets are curated to include a wide range of scanners, demographic backgrounds, and imaging protocols, ensuring that the algorithms developed are robust and generalizable. The test set, being from a proprietary source, introduces a real-world challenge of domain adaptation, testing the algorithms' ability to perform accurately on previously unseen data with different distributions. This aligns with the challenge's goal of advancing AI algorithms that can be readily deployed in clinical settings.

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 involve 10 expert annotators, consisting of 4 senior annotators and 6 junior annotators.

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.

Annotation Standard:
Our annotation focused on 25 organs and structures, including 16 abdominal organs (esophagus, stomach, duodenum, intestine, colon, rectum, liver, gallbladder, spleen, pancreas, right kidney, left kidney, right adrenal gland, left adrenal gland, bladder, prostate), 2 thorax organs (right lung and left lung), 5 vascular structures (aorta, celiac trunk, postcava, portal vein and splenic vein, hepatic vessel), and 2 skeletal structures (left and right proximal femur).

1. The esophagus contour should include the entire esophageal wall and lumen along with any esophageal lesions, while adjacent structures such as the trachea, aorta, and surrounding fat and muscle should be excluded. The esophagus and stomach are bounded by the cardia. The stomach contour should encompass the entire stomach wall and lumen including the fundus, body, antrum, and pylorus, as well as any gastric lesions, while adjacent structures, organs, and surrounding fat should be excluded. The stomach and duodenum are bounded by the pylorus. The duodenum contour should include the entire duodenal wall and lumen 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. The duodenum and jejunum are bounded by the Treitz ligament. The intestine contour should include the jejunum and ileum wall and lumen from the ligament of Treitz to the ileocecal valve, along with any intestinal lesions. It should exclude surrounding fat, mesentery and mesenteric vessels. The ileum and colon are bounded by the ileocecal valve. The colon contour should include the entire wall and lumen of cecum, appendix, ascending colon, transverse colon, descending colon and sigmoid colon, as well as any colon lesions, while adjacent structures, surrounding fat, mesentery and omentum should be excluded. The sigmoid take-off is taken as an imaging landmark for the rectosigmoid junction. The rectum contour should include the entire rectal wall, lumen and any lesions, while adjacent structures, surrounding fat and muscle should be excluded.

2. The liver contour should include all the liver parenchyma and any lesions. The intrahepatic vessels and intrahepatic bile ducts need to be covered, while excluding surrounding fat, adjacent structures and organs. The gallbladder contour should encompass the entire gallbladder wall and lumen, including the fundus, body and neck, as well as any gallstones or polyps. The cystic duct, the surrounding liver parenchyma and fat should be excluded. The pancreatic contour should encompass all pancreatic parenchyma including the head, body and tail, as well as any pancreatic lesions and pancreatic duct. The surrounding vessels and fat should be excluded. The spleen contour should include all splenic parenchyma and any splenic lesions. It should exclude adjacent structures and extra splenic vessels. The adrenal gland contour should include the entire adrenal gland and any adrenal lesions. It should exclude adjacent structures and surrounding fat. The kidney contour should include the renal parenchyma, while excluding renal pelvis, ureter, extrarenal blood vessels, surrounding fat and any adjacent structures. The bladder contour should include the entire adrenal wall, lumen and any bladder lesions. It should exclude adjacent structures and surrounding fat. The prostate contour should include the whole prostate parenchyma, prostatic urethra and any prostate lesions. It should exclude adjacent structures, surround fat, prostatic venous plexus and bilateral seminal vesicles.

3. The lung contour should include the entire lung parenchyma, pulmonary bronchovascular bundle, visceral pleura and any pulmonary lesions. It should exclude pleural effusion, pneumothorax, parietal pleura, mediastinal structures, and chest wall.

4. The aortic, celiac trunk contour should include the entire lumen of the arteries. The artery wall and calcification, ulcers, thrombosis, dissection should also be included. The inferior vena cava, portal vein and splenic vein contour should include the entire lumen and cover the walls, as well as intraluminal thrombus and tumor thrombus. The hepatic vessel contour should include all intrahepatic vessel wall and lumen, as well as intraluminal thrombus and tumor thrombus.

5. The proximal femur contour should include the femoral head, neck and the region 5 cm distal to the lesser trochanter, with both cortical bone and spongy bone, as well as any lesions. It should exclude surrounding muscles, vessels.

Annotation software operation guide:
1. We use 3D-Slicer to do segmentation. The Segmentations module in 3D-Slicer manages segmentations. Each segmentation can contain multiple segments, which correspond to one structure or ROI. Each segment can contain multiple data representations for the same structure

2. Import an AI predicted segmentation. 3D volumes in NRRD (.nrrd or .nhdr) and Nifti (.nii or .nii.gz) file formats can be directly loaded as segmentation: Drag-and-drop the volume file to the application window (or use menu: File / Add Data, then select the file). In the Description column choose Segmentation. Click OK.

3. Export segmentation file. Open Export to files section in the Segmentations module (or in Segment editor module: choose Export to files, in the drop-down menu of Segmentations button). Choose destination folder, file format, etc. Click Export.

4. Edit your segmentation file. Segment editor is a module for specifying segments in images. Some of the tools have a painting interface, offering editing of overlapping segments, displaying in both 2D and 3D views.

5. Paint on your segment. Pick the radius (in millimeters) of the brush to apply. Left click to apply a single circle. Left click and drag to fill a region. A trace of circles is left which are applied when the mouse button is released. Sphere mode applies the radius to slices above and below the current slice.

6. Draw on your segment. Left click to lay individual points of an outline. Left drag to lay down a continuous line of points. Left double-click to add a point and fill the contour. Alternatively, right click to fill the current contour without adding any more points.

7. Erase from your segment. Same as the Paint effect, but the highlighted regions are removed from the selected segment instead of added. If the Masking / Editable area is set to a specific segment then the highlighted region is removed from the selected segment and added to the masking segment. This is useful when a part of a segment has to be separated into another segment.

Our annotation pipeline involved an interactive segmentation approach, an integration of AI algorithms and human expertise, which premises to improve the efficiency while upholding high-quality annotations. Six junior radiologists revised the annotations predicted by our AI algorithms under the supervision of senior radiologists, and in turn, the AI algorithms improved their predictions by learning from these revised annotations. This interactive process continued to enhance the quality of annotations until no major revision was needed, as confirmed by these radiologists. "No major revision" includes small pixel-level variations at organ edges, whereas macroscopic labeling errors such as mislabeling or missing areas of organs are not permissible. Subsequently, four senior radiologists examine the final visualizations for accuracy for all the annotations before releasing. The junior radiologists were responsible for reviewing the correctness of the annotations and marking

the patient ID for any major discrepancies. Those wrong cases are then reviewed by the senior radiologists. Our uniform annotation standards, largely following those in [Ma et al., 2023], require trained radiologists to spend approximately 5-60 minutes annotating each organ in a three-dimensional CT volume, depending on the size of the organ and the complexity of the surrounding tissue.

Reference

Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z. and Zhang, F., 2023. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. arXiv preprint arXiv:2308.05862.

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.

Four senior radiologists have 8-15 years clinical experience, six junior radiologists have 3-5 years clinical experience.

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

N/A

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.

The training data for the BodyMaps Challenge, derived from various publicly available datasets, will undergo a standardized pre-processing procedure to ensure uniformity.
- This involves converting all original imaging files (for example, some data in the DICOM files) to the NIfTI format, suitable for widespread use in the medical imaging research community.
- Additionally, to maintain consistency across datasets where annotations may vary-for instance, some datasets might label 'kidney' as a single class while others separate it into 'Left kidney' and 'right kidney'-we will standardize these classifications.

This ensures that all training data is harmonized, thereby reducing potential bias or confusion when training AI algorithms. The pre-processing pipeline will be meticulously documented and made
available to challenge participants to facilitate reproducibility and transparency in the development and evaluation of their algorithms.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

The primary source of error we anticipate in image annotation stems from the inherent variability in human interpretation, which can lead to inconsistencies in organ boundary identifications. To mitigate this, we have established a rigorous annotation protocol and provided segmentation tools application training, overseen by senior radiologists with extensive clinical experience in CT image interpretation.

We estimate the inter-annotator variability to be within the range of 5-10%, a value derived from preliminary studies that measured annotation differences among our team. This variability is an essential consideration as it represents the realistic uncertainty present in clinical settings and is thus a critical factor in evaluating algorithm performance.

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

Besides annotation variability, other potential sources of error are scanning equipment and patient condition, such as beam hardening, partial volume averaging, quantum mottle and inconsistency between different devices from scanning equipment, motion artifacts, contrast medium injection flow rate, body fluid distribution from patient condition. These errors can affect image quality and thus algorithm performance. We will document the prevalence of such artifacts within our dataset and provide this information to participants, ensuring they can develop algorithms robust to these real-world imaging challenges.

ASSESSMENT METHODS

Metric(s)

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

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

The segmentation accuracy metric:
1. Weighted Dice Similarity Coefficient (wDSC). This metric evaluates the overlap between algorithm output and ground truth, with a weighting factor that reflects the segmentation difficulty for each structure. The weight for each structure's DSC is estimated based on the per-class segmentation performance reported in our preliminary experiments. We average the performance of UNet[Ronneberger et al., 2015], Swin UNETR[Hatamizadeh et al., 2021], and SegResNet[Myronenko et al., 2019] and subtract the performance from one to measure the difficulty of different structures. Then we apply Softmax on the result to get the weight that sums to one. Some structures are inherently more difficult to segment than others due to blurry boundaries, small in size, and tubular structures. Our weighted metric is novel compared to the common practice in segmentation challenges, where only the average DSC is calculated uniformly across all classes.
2. Weighted Normalized Surface Distance (wNSD): The wNSD emphasizes the accuracy of the boundary delineation between the predicted segmentation and the ground truth. This is particularly important for precise organ volume measurement and subsequent surgical planning.

The segmentation efficiency metric:
3. Standardized Running Time: Taking into account the time required for docker startup, data reading, and saving segmentation results, we recommend that the total time consumption per case should, on average, be within 90 seconds. The standardized running time for each case is considered a critical efficiency metric. It is calculated by dividing the actual time consumption by 90 seconds. Furthermore, to accommodate the workload during the testing stage, a dynamic time limit is established for each case, determined by the spacing and sizes of each test case. During the inference process, if the inference time exceeds this time limit for more than 20% of the cases,

the inference will be terminated, and the submission will be classified as a failure submission.
4. Standardized Area Under GPU Memory-Time Curve (MB) [Ma et al. FLARE 2023]: The memory efficiency of the algorithm is evaluated over time, taking into account the computational resources utilized, as indicated by the GPU memory-time curve. It is recommended that the GPU memory consumption be kept below 24GB, aligning with the affordability and availability of such GPUs in most medical centers. The standardized Area Under GPU Memory-Time Curve for each case is considered another critical efficiency metric. It is calculated by dividing the Area Under GPU Memory-Time Curve by 24*1024*90.

Reference

Ma, J., B. Wang. MICCAI FLARE23: Fast, Low-resource, and Accurate oRgan and Pan-cancer sEgmentation in Abdomen CT. https://codalab.lisn.upsaclay.fr/competitions/12239
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.
Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2021, September). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In International MICCAI Brainlesion Workshop (pp. 272-284). Cham: Springer International Publishing.
Myronenko, A. (2019). 3D MRI brain tumor segmentation using autoencoder regularization. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4 (pp. 311-320). Springer International Publishing.

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

We have chosen the Weighted Dice Similarity Coefficient (wDSC) and Weighted Normalized Surface Distance (wNSD) for their precision in reflecting clinical needs. These metrics are tailored to emphasize the segmentation accuracy of clinically significant structures and the boundary delineation's precision, crucial for treatment planning and diagnostic accuracy in medical procedures.

1. For both DSC and NSD metrics, greater weights will be placed on structures that are challenging to segment. This prioritization is based on several factors. Firstly, structures with low contrast in soft tissues, often presenting as indistinct boundaries. Secondly, organs that exhibit complex morphologies, such as tubular structures. Finally, variations in organ appearances due to differing protocols, for instance, when subjects are rotated along the vertical axis between 30 and 60 degrees, will also be considered.

2. DSC is particularly valuable in biomedical contexts where certain anatomical structures are more critical to segment accurately due to their diagnostic or therapeutic importance. By assigning higher weights to these structures, the wDSC ensures that algorithms are fine-tuned to perform exceptionally well where it matters most, reflecting real-world clinical priorities.

3. NSD refines the assessment by emphasizing the precision of segmentation at the boundaries of the structures. In many medical procedures, such as surgical planning or radiotherapy, the exact delineation of the boundary is crucial. Accurate boundary segmentation can significantly impact treatment outcomes, making the wNSD a highly relevant metric for evaluating algorithms against clinical needs.

Efficiency metrics like Running Time and the Area Under GPU Memory-Time Curve are chosen to ensure the practical deployment of segmentation algorithms in a clinical setting.

4. Running time: In many medical scenarios, time is a critical factor, and algorithms must be able to process images swiftly without sacrificing accuracy. The Running Time metric, therefore, serves as a benchmark for performance in time-sensitive clinical environments.

5. Area Under GPU Memory-Time Curve addresses the resource constraints commonly found in clinical settings. Medical centers require algorithms that are not only accurate and fast but also efficient in terms of computational resources. The memory-time curve provides a measure of an algorithm's memory usage over time, rewarding algorithms that maintain low memory profiles throughout their running time, which is essential for deployment in clinical settings with limited computational resources (e.g., ordinary laptop/PC).

This combination of metrics ensures that the challenge outcomes will not only contribute to scientific advancement but also to the practical application of AI in medicine. The reason is that the evaluation metrics mirror the multifaceted demands of clinical application: accuracy in critical areas, precision at organ boundaries, swift processing times, and computational efficiency. This performance assessment ensures that the algorithms developed through the BodyMaps Challenge are not only technically advanced but also clinically viable.

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.

This ranking method ensures a balanced evaluation, taking into account both the quality of the segmentation and the computational efficiency of each algorithm, which is crucial for practical clinical deployment.

Step 1: Metric Calculation for Each Case
- Calculate the Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) for each case in the test dataset (W-1K).
- Record the Running Time for each testing case.
- Compute the Area Under the GPU Memory-Time Curve for each case in the test dataset.

Step 2: Aggregate Metric Scores
- Calculate weighted DSC (wDSC) and weighted NSD (wNSD) for all the cases by assigning higher weights to structures that are hard to segment.
- Calculate the mean for the Running Time scores, and the Area Under the GPU Memory-Time Curve scores across all the cases.

Step 3: Individual Metric Ranking
- Rank all algorithms based on their wDSC score, wNSD score, mean Running Time score (mT), and mean Area Under the GPU Memory-Time Curve score (mAUC), separately.

Step 4: Final Ranking
- For the final ranking, combine the weighted segmentation accuracy metrics (wDSC, wNSD) with the efficiency metrics (mean Standardized Running Time, mean Standardized AUC) using
a composite score. Calculate the composite score by summing the wDSC and wNSD and then dividing by the product of mean Standardized Running Time (mST) and mean Standardized AUC (mSAUC).

The formula for the composite score is given by: Composite Score = (wDSC + wNSD)/(mST x mSAUC)

- Rank the algorithms in descending order of their composite scores to establish the final leaderboard.

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

- Upon submission, if an algorithm fails to produce a result for a case within the allotted time frame (90 seconds) or exceeds the memory constraints (24GB for GPU memory consumption), the case will be designated as a warning case for that algorithm. This will adversely impact the algorithm's composite score.
- Algorithms that have warning cases for more than a threshold percentage - 20% - of the test cases may be disqualified from the final ranking to maintain the integrity and competitiveness of the challenge.

This method for handling submissions with missing (i.e. warning) results ensures that all submissions are treated fairly and that the final ranking reflects the performance of the algorithms under the same conditions, providing an equitable basis for comparison. It also encourages participants to ensure their algorithms are robust and can handle the maximum range of test cases in the challenge.

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

This ranking scheme was used because it aligns with the primary goals of the BodyMaps Challenge: to develop AI algorithms that are both accurate in segmentation and efficient in processing.

- Multidimensional Accuracy: By using the Weighted Dice Similarity Coefficient (wDSC) and Weighted Normalized Surface Distance (wNSD), the ranking scheme accounts for both the volumetric accuracy of segmentations and the precision of their boundaries. These accuracy metrics are weighted to emphasize the clinical importance of certain structures, which may require greater segmentation accuracy due to their diagnostic or therapeutic significance.

- Computational Efficiency: Efficiency in clinical settings is also important. The incorporation of the Running Time and Area Under the GPU Memory-Time Curve as efficiency metrics ensures that algorithms are practical for real-world use.

- Composite Scoring: The final composite score calculation, which combines accuracy and efficiency metrics, is designed to favor algorithms that achieve a balance between high-quality segmentation results and efficient use of computational resources. This is crucial in clinical practice where both accurate and rapid analysis can significantly impact patient outcomes.

- Fair and Comprehensive Ranking: The step-by-step process, from individual metric calculation to the final

composite score, ensures that each aspect of performance is considered. This methodological rigor guarantees that algorithms are ranked fairly, based on their performance across all evaluated criteria.

- Clinical Relevance: The chosen ranking scheme mirrors the multifaceted criteria that clinical tools must meet - they must not only provide precise and reliable results but also integrate seamlessly into the fast-paced, resource-sensitive clinical environment. The scheme encourages the development of algorithms that can be effectively and efficiently used in clinical decision-making processes.

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.

- Handling Missing Data: In cases where algorithm submissions do not produce results within the allocated time or exceed memory constraints, those instances are treated as warning data. Warning data will negatively affect the algorithm's composite score. Furthermore, if the warning data exceed 20% of the total test cases, the submission will be deemed a failure.

- Assessment of Variability in Rankings: The mean Running Time and the Area Under the GPU Memory-Time Curve scores may introduce variability in the rankings due to hardware optimization. Algorithms may be optimized differently for the computational resources provided, such as CPU cores and GPU types, which can affect their performance efficiency. To assess this variability, we will compute measures of central tendency (mean, median) and dispersion (standard deviation, interquartile range) for the efficiency metrics to understand the distribution of algorithm performances on the provided hardware setup.

- Software for Data Analysis: All data analyses, encompassing the calculation of metrics, rankings, and statistical tests, are conducted using both proprietary and third-party Python scripts, as well as Excel software. These software tools are instrumental in performing the analyses with precision and efficiency. The tools utilized for analysis will be detailed in the challenge documentation and open-sourced code to ensure transparency and reproducibility.

By using these methods, we ensure that the statistical analysis of the BodyMaps Challenge is robust, allowing for a fair comparison of algorithms and providing valuable insights into their performance.

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

- Handling Missing Data: Giving a limit of tolerance for missing results due to timeouts or memory constraints is a standard approach in challenges. This policy is straightforward and ensures that all participants pay attention to the penalties for non-compliance with the efficiency requirements. It reflects the real-world clinical environment where complete and timely results are essential, and therefore, it maintains the relevance and applicability of the challenge outcomes.

- Variability of Rankings: The inclusion of statistical measures of central tendency and dispersion for the efficiency metrics is crucial for a comprehensive understanding of each algorithm's performance across the provided hardware setup. By analyzing the spread and central values of running times and memory usage, we can identify patterns and outliers in the data. This approach ensures that the ranking process accounts for variability due to hardware differences, which is a common scenario in clinical deployments.

- Choice of Software: The use of established statistical software packages and related code is justified by the need for accuracy, reproducibility, and transparency in the analysis process. These packages are equipped with the necessary algorithms and functions to perform complex calculations and provide a standard by which results can be verified by third parties if there is a need.

These methods are integral to the challenge as they ensure that the ranking of algorithms is not only indicative of their segmentation performance but also reflects their efficiency, stability, and generalizability which are primary in clinical settings where algorithms must perform reliably under various conditions. These methods uphold the scientific rigor and integrity of the BodyMaps Challenge.

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.

Computing p-values for Pairwise Comparisons: In a challenge with numerous submissions, it is highly likely to encounter ranking order instability, particularly around the median submission whose performance tends to cluster. We will employ the significance map [Wiesenfarth et al. 2021] to graphically visualize the p-value for performance comparisons between any two teams. This map allows for the interpretation of each pixel's value as the statistical significance that the team represented on the x-axis is truly performing better/worse than the team on the y-axis (e.g., at the p = 0.05 level).

Reference

Wiesenfarth, M., A. Reinke, B. A. Landman, M. Eisenmann, L. A. Saiz, M. J. Cardoso, L. Maier-Hein, and A. Kopp-Schneider. Methods and open-source toolkit for analyzing and visualizing challenge results, 2021.

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