Self-supervised learning for 3D light-sheet microscopy image segmentation: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Self-supervised learning for 3D light-sheet microscopy image segmentation

Challenge acronym

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

SELMA3D

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.

In the realm of modern biological research, the ability to visualize and understand complex structures within tissues and organisms is crucial. Traditional imaging methods often face challenges in providing detailed, 3D views without compromising sample integrity. Light-sheet microscopy (LSM) after tissue clearing and specific structure staining overcomes these limitations, making it an efficient, high contrast, and ultra-high resolution method for visualizing a wide array of biological structures in diverse samples, such as cellular and subcellular structures, organelles and processes[1].

In the structure staining step, various dyes, fluorophores, or antibodies can be employed to selectively label specific biological structures within samples and enhance their contrast under microscopy[2]. In the tissue clearing step, while preserving sample integrity and fluorescence of labeled structures, inherently opaque biological samples are rendered transparent, allowing light to penetrate deeper into the tissue[3]. Integrating with structure staining and tissue clearing steps, LSM provides researchers with unprecedented capabilities to visualize intricate biological structures with high spatial resolution, offering new insights into various biomedical research fields such as neuroscience[4], immunology[5], oncology[6] and cardiology[7].

To analyze LSM images in different biomedical research fields, segmentation plays a pivotal and essential role in identifying and distinguishing different biological structures[8]. For very small-scale LSM images, image segmentation can be done manually. However, in whole-organ or body LSM cases, manual segmentation is time-intensive, single images can have 10000^3 voxel, hence automatic segmentation methods are highly demanded. Recent strides in deep learning-based segmentation methods offer promising solutions for automated segmentation of LSM images[9-10]. Although these methods reached segmentation performances comparable to expert human annotators, their success largely relies on supervised learning from extensive training sets of manually annotated images which are specific to one kind of structure staining. However, large-scale annotation for diverse LSM image segmentation tasks poses a great challenge.

Self-supervised learning proves advantageous in this context, as it allows deep learning models to pretrain on large-scale, unannotated datasets, learning useful and general representations of LSM image data. Subsequently, the model can be fine-tuned on a smaller labeled dataset for specific segmentation tasks[11]. Notably, self-supervised learning has not been extensively explored within the LSM field, despite the presence of vast sets of LSM data of different biological structures. Some of the properties of LSM images e.g. the high signal-to-noise-ratio makes the data specifically well suited for self-supervised learning.

In this challenge, we aim to host an inaugural MICCAI challenge on self-supervised learning for 3D LSM image segmentation, encouraging the community to develop self-supervised learning methods for general segmentation of various structures in 3D LSM images. With an effective self-supervised learning method, extensive 3D LSM images with no annotations can be leveraged to pretrain segmentation models. This encourages models to capture high-level representations that are generalizable across different biological structures. Subsequently, the pretrained models can be finetuned on substantially smaller annotated datasets, thereby significantly minimizing the annotation efforts in various 3D LSM segmentation applications.

References:

[1] E.H.K. Stelzer, F. Strobl, B. Chang, F. Preusser, S. Preibisch, K. McDole and R. Fiolka. Light sheet fluorescence microscopy. Nature Reviews Methods Primers 1(1): 73, 2021 Nov.

[2] P.K. Poola, M.I. Afzal, Y. Yoo, K.H. Kim and E. Chung. Light sheet microscopy for histopathology applications. Biomedical engineering letters 9: 279-291, 2019 July.

[3] H.R. Ueda, A. Ertürk, K. Chung, V. Gradinaru, A. Chédotal, P. Tomancak and P.J. Keller. Tissue clearing and its applications in neuroscience. Nature Reviews Neuroscience 21(2): 61-79, 2020, Jan.

[4] H.R. Ueda, H.U. Dodt, P. Osten, M.N. Economo, J. Chandrashekar and P.J. Keller. Whole-brain profiling of cells and circuits in mammals by tissue clearing and light-sheet microscopy. Neuron, 106(3): 369-387, 2020 May.

[5] D. Zhang, A.H. Cleveland, E. Krimitza, K. Han, C. Yi, A.L. Stout, W. Zou, J.F. Dorsey, Y. Gong and Y. Fan. Spatial analysis of tissue immunity and vascularity by light sheet fluorescence microscopy. Nature Protocols: 1-30, 2024 Jan.

[6] J. Almagro, H.A. Messal, M.Z. Thin, J. Rheenen and A. Behrens. Tissue clearing to examine tumour complexity in three dimensions. Nature Reviews Cancer, 21(11): 718-730, 2021 July.

[7] P. Fei et al. Cardiac light-sheet fluorescent microscopy for multi-scale and rapid imaging of architecture and function. Scientific Reports 6: 22489, 2016 Mar.

[8] F. Amat, B. Höckendorf, Y. Wan, W.C. Lemon, K. McDole, P.J. Keller. Efficient processing and analysis of large-scale light-sheet microscopy data. Nature protocols 10. 2015: 1679-1696.

[9] N. Kumar, P. Hrobar, M. Vagenknecht, J. Soukup, N. Patterson, P. Bloomingdale, T. Freshwater, S. Bardehle, R. Peter, R. Mangadu. A Light sheet fluorescence microscopy and machine learning-based approach to investigate drug and biomarker distribution in whole organs and tumors. bioRxiv 2023.09.16.558068.

[10] M.I. Todorov et al. Machine learning analysis of whole mouse brain vasculature. Nature Methods 17: 442-449, 2020 Mar.

[11] R. Krishnan, P. Rajpurkar, E.J. Topol. Self-supervised learning in medicine and healthcare. Nature Biomedical Engineering 6: 1346-1352, 2022 Aug.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Light-sheet microscopy, 3D image, deep learning, self-supervised learning, pretraining, image segmentation

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.

No associated workshop.

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.

Drawing upon recent publications and the growing research interest in self-supervised learning and 3D LSM analysis as well as the uniqueness of the dataset provided, we anticipate the involvement of a minimum of 25 teams.

Our strategy involves organizing a series of seminars, encompassing both online and in-person formats, to effectively publicize and promote our challenge. This initiative aims to not only attract a broader pool of participants but also to address any queries or concerns they may have during the process.

Publication and future plans

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

Yes, we plan to summarize and present the challenge results in a journal article.

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.

The challenge will be hosted on grand-challenge.org online. Participants are expected to utilize their individual computing resources for algorithm development. Organizers will employ grand-challenge.org for prediction evaluation on the validation set and docker evaluation on the testing set.

Regarding the in-person event, we require projectors, microphones, loudspeakers, and cameras to facilitate hybrid participation.

TASK 1: Self Supervised semantic segmentation of brain structures in 3D microscopy

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.

When combined with tissue clearing and specific structure staining, LSM has unique advantages for imaging large and intact samples with high resolution while minimizing photobleaching and phototoxicity. This makes LSM a powerful tool for studying the structure, function, and dynamics of the brain, advancing the understanding of normal brain physiology and neurological disorders[1-2].

For the analysis of brain LSM images in different applications, segmentation is a fundamental step to delineate targeted biological structures, such as blood vessels and neuron cells. However, manual segmentation of these biological structures in a whole-brian image or large-scale annotation for deep-learning segmentation model development is time-consuming and laborious. To address this issue, we propose the task of developing self-supervised learning methods for 3D brain LSM image, contributing to semantic segmentation models of various biological structures in brain LSM images.

In this task, each participant will receive a training dataset comprising two sets. The first set includes a large (> 6x10^11 voxels, equivalent to > 35000 images of 256x256x256 voxels) of whole-brain 3D LSM images of both mice and human samples without annotations, this set should facilitate model pretraining through self-supervised learning. This dataset will be one of the largest datasets ever provided to a MICCAI challenge. The second set consists of cropped patches from whole-brain 3D LSM images with precise annotations, enabling the fine-tuning of the model for semantic segmentation tasks.

[1] Yang, Xiao, et al. Laminin-coated electronic scaffolds with vascular topography for tracking and promoting the migration of brain cells after injury. Nature Biomedical Engineering 7: 1282-1292, 2023 Oct.
[2]Pesce, Luca, et al. 3D molecular phenotyping of cleared human brain tissues with light-sheet fluorescence microscopy. Communications Biology 5: 447, 2022 May.

Keywords

List the primary keywords that characterize the task.

brain LSM image, 3D image, deep learning, self-supervised learning, pretraining, semantic segmentation

ORGANIZATION

Organizers

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

[Helmholtz Munich, Germany]
Ali Erturk, Luciano Höher, Rami Al-Maskari, Izabela Horvath, Mayar Ali

[Imperial College London, UK]
Johannes C. Paetzold

[Ludwig Maximilian University of Munich, Germany]
Ying Chen

[University of Zurich, Switzerland]
Bjoern Menze, Kayuan Yang

[Technical University of Munich]
Daniel Rückert, Martin Menten, Alexander Berger, Laurin Lux

b) Provide information on the primary contact person.

Johannes C. Paetzold (j.paetzold@ic.ac.uk);
Ying Chen (Ying.Chen@campus.lmu.de)

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

The challenge will adhere to a predetermined submission deadline for showcasing results and presenting awards at the in-person event during the MICCAI 2024 conference.

Depending on the reception and participation in this challenge, we plan to replicate and expand it for MICCAI 2025 and beyond. This expansion involves not only increasing the number of samples for LSM images but also enhancing sample diversity by incorporating various organs or whole body samples from different species. We already possess hundreds of terabyte-sized 3D LSM images. We can cooperate with other labs leading in LSM worldwide to realize further expansion to the dataset. Additionally, we will broaden the diversity of labeled structures in LSM.

Challenge venue and platform

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

N/A

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

grand-challenge.org

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

We are creating a competition on grantchallenge.org in accordance with the details of this proposal.

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.

Participants are constrained to utilize the training data provided exclusively by this challenge, which comprises two subsets. The first subset is designated for self-supervised learning and encompasses whole-brain 3D LSM images without annotations. In the second subset we provide a few patches from whole-brain images with annotations for participants to evaluate their performance on. Importantly, the annotations are on structures different to the validation and test set structures. This is to make sure that participants develop self-supervised algorithms.

During the validation phase, participants should not manually annotate the images to plainly train a supervised model for the prediction submission.

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 from the organizers' groups are welcome to join, and their results can be featured in publications and on the leaderboard. However, they are not eligible for awards. Individuals and teams from external labs or departments, not affiliated with the organizers' groups, are encouraged to participate and are eligible for both awards and inclusion on the leaderboard.

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

The challenge webpage will publicly acknowledge the top three teams, and they will receive a Jellycat Brezel as a souvenir during the in-person challenge event. However, no monetary awards will be granted.

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 performance results will be disclosed publicly, and outstanding submissions will be acknowledged during the in-person challenge event. However, participating teams have the option to decide whether to make their results public any time before the day of the announcement. The top 5 teams will be invited to prepare a 10 minute presentation for the challenge session to showcase and discuss their methods. Following the public announcement, a detailed analysis of the submitted results will be available upon request.

If a participant wishes to retract their submission after the announcement, their performance will either be reported in an anonymized manner both online and in the publication, or it will be removed from the leaderboard and publication.

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

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

We plan to collaborate closely with participants to produce a comprehensive journal article summarizing the key results and analyses derived from this challenge. All participating teams that submit work contributing to the algorithm development are welcome to contribute to our challenge publication. Up to three authors from each participating team will be acknowledged as authors of the article. Any additional authors from the submissions may be included upon request according to the ICMJE authorship guidelines.

Furthermore, we encourage all participating teams to independently submit their results without imposing any publication embargo.

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.

1)Submission for validation set:
we will make the images of the validation set public while keeping the annotations private. Participants can submit their predictions, i.e. binary segmentation images stored in NIfTI format, for our validation set on the challenge website for automated evaluation. This set serves as a means for participants to refine their methods for unseen data and assess their performance relative to others.

To prevent participants from submitting manually labeled annotations on the validation images to get a final high ranking, we introduce a completely private test set, consisting of private test data and annotations. The validation set is not utilized for the final evaluation. Besides, it is explicitly prohibited to manually annotate images during validation to train a supervised model for the test submission.

2) Submission for test set:
Submissions for evaluation on the test set will be facilitated through submitted docker containers, which will undergo evaluation by organizers on grand-challenge.org. In addition to the docker containers, each participating team is encouraged to submit a 2-3 page summary detailing their methods and approaches within one week after the docker submission deadline.

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 have access to the validation set without annotations, and their predictions on this set can be submitted to grand-challenge.org for automated evaluation. Each team is limited to two submissions per day for evaluation on the validation set. The validation set results contribute to the public leaderboard, offering an initial unofficial ranking. It is crucial to note that the final ranking will be determined based solely on the performance on the test set.

For the final test set, each team is provided with three opportunities to upload their containers. In the event of technical issues, participants are allowed to retry their docker submissions. However, only the results from the last run will be officially considered in computing the challenge results.

Challenge schedule

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

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

Preliminary Schedule:

- Challenge website launch: April 10th, 2024
- Training set release: June 18th, 2024
- Validation set (without annotation) release: June 25th, 2024
- Opening of submission system and leaderboard for the validation set: July 7th, 2024
- Opening of submission system for the final test set: July 24th, 2024
- Contacting top performing teams and planning for the in-person session: Aug 10th - October 1st, 2024 (teams requiring visas will be contacted earlier, and we will coordinate with MICCAI to send invitation letters for visa clearance.)
- In-person challenge event: October 6th, 2024

Additional point:
we plan to conduct a series of both online and in-person seminars to publicize and broadcast our challenge, addressing any questions from participants.

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 data utilized in this challenge is a set of research data which consists of previously published data and data acquired by the challenge organizers within the ethics approval of specific studies [1-4]. As a result, the data has received approval from the pertinent ethical committee, and no further ethics approval is necessary. It's important to note that the data has been anonymized, with sample information removed.

References:
[1] S. Zhao et al. Cellular and molecular probing of intact human organs. Cell 180(4): 796-812, 2020 Feb.
[2] M.I. Todorov et al. Machine learning analysis of whole mouse brain vasculature. Nature Methods 17: 442-449, 2020 Mar.
[3] H. Mai, Z. Rong, S. Zhao, R. Cai, H. Steinke, I. Bechmann, A. Ertürk. Scalable tissue labeling and clearing of intact human organs. Nature Protocols 17: 2188-2215, 2022 July.
[4] H.S. Bhatia et al. Spatial proteomics in three-dimensional intact specimens. Cell 185(26): 5040-5058, 2022 Dec.
[5] D. Kaltenecker et al. Virtual reality empowered deep learning analysis of brain activity. Accepted by Nature Methods.

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

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 organizers' evaluation code will be accessible to the public on both GitHub and the challenge website.

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

The submitted docker containers from all participants will be publicly accessible unless participants disagree. Participants who do not agree to make docker public will not be eligible for awards, and will not be listed in the leaderboard.

Additionally, we strongly encourage participants to share their code with the public.

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.

Monetary awards will not be provided.

Access to validation annotations and the private test set will be restricted to the main organizers and their annotation team exclusively.

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, Education, Training.

Task category(ies)

State the task category(ies)

Examples:

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

- Tracking

Semantic segmentation (self/semi-supervised)

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 3D LSM images capturing a broad spectrum of biological structures in samples from diverse species. A universal self-supervised learning approach for 3D LSM semantic segmentation holds potential benefits across various segmentation applications for whole organ or whole body, including but not limited to vessels and different cell segmentation.

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

The challenge data cohort is 3D LSM images of mice and human brains produced by the Institute for Tissue Engineering and Regenerative Medicine (iTERM) and the Institute for Stroke and Dementia Research between 2019 and 2023.

Imaging modality(ies)

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

Light-sheet microscopy (LSM)

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 contextual information will be made available.

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

No clinical information about the human samples will be made available.

Target entity(ies)

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

The dataset encompasses 3D LSM images of brain samples obtained from both mice and humans following tissue clearing protocols[1-2].

Individual specimen information will not be made available.

References:
[1] S. Zhao et al. Cellular and molecular probing of intact human organs. Cell 180(4): 796-812, 2020 Feb.
[2] A. Ertürk, K. Becker, N. Jährling, C.P. Mauch, C.D. Hojer, J.G. Egen, F. Hellal, F. Bradke, M. Sheng, H.U. Dodt. Three-dimensional imaging of solvent-cleared organs using 3DISCO. Nature Protocols 7: 1983-1995, 2012 Nov.

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.

A self-supervised learning method aimed at achieving a generalized model for semantic segmentation of diverse labeled biological structures in brain 3D LSM images, encompassing tree-like structures like vessels and spot-like structures like cells.

To prevent participants from solely relying on the annotated training set for training models and achieving promising segmentation results without engaging in the self-supervised learning stage, we will introduce variations between the annotated training set, validation set, and test set. The objective is to lead participants to develop self-supervised learning methods that can pretrain models with a high degree of generalization.

To be specific, the annotated training set will encompass brain blood vessel data, c-Fos labeled brain cell data involved in neural activity, cell nucleus data, and Alzheimer's disease plaque data from mouse samples. In the validation and testing sets, data from different mouse and human samples is used. This will be c-Fos labeled data from other samples. Additionally, a different biological structure data, Microglial data, will be added.

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)).

The principal aim of this challenge is to develop a self-supervised learning method that enhances the generalization capabilities of semantic segmentation models in high-resolution microscopic images. To evaluate the model's generalization performance, assessments will be conducted on 3D brain LSM images of various labeled structures. These structures can generally be categorized into two types: spot-like structures, such as different types of cells, and tree-like structures, such as vessels. For assessing segmentation results of spot-like structures, two metrics are used, volumetric Dice similarity coefficients and Betti matching error [1]. For assessing segmentation results of tree-like structures, volumetric Dice similarity coefficients and Betti matching error together with another metric, centerline-Dice similarity coefficients (clDice) [2], are utilized.

In the validation phase, participants submit their predictions on our validation set to the challenge website, and the predictions will be automatically evaluated by the above metrics.

In the testing phase, participants submit their docker containers to the challenge website, which will undergo evaluation by organizers on grand-challenge.org by the above metrics. Please note, the participants are supposed to singley submit the segmentation model after finetuning, which excels at segmenting different brain structures.

[1] N. Stucki, J. C. Paetzold, S. Shit, B. Menze, U. Bauer. Topologically faithful image segmentation via induced matching of persistence barcodes. In International Conference on Machine Learning, 2023, pp. 32698-32727.
[2] S. Shit et al. clDice - a novel topology-preserving loss function for tubular structure segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16555-16564.

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).

LSM imaging was performed using a 4× objective lens (Olympus XLFLUOR 340) equipped with an immersion corrected dipping cap, mounted either on an UltraMicroscope II (LaVision BioTec, chamber size of 72 × 74 × 35 mm, for small samples) or a prototype UltraMicroscope (Miltenyi Biotec, chamber size of 250 × 90 × 70 mm, for large samples) coupled to a white light laser module (NKT SuperK Extreme EXW-12).

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

The data was obtained through the following procedure: structure staining, tissue clearing, LSM imaging.

Different dyes and stains were used to selectively bind to specific structures in the sample, enhancing their visibility in the images.

Different tissue clearing methods were used following our previous work [1-2].

[1] S. Zhao et al. Cellular and molecular probing of intact human organs. Cell 180(4): 796-812, 2020 Feb.
[2] A. Ertürk, K. Becker, N. Jährling, C.P. Mauch, C.D. Hojer, J.G. Egen, F. Hellal, F. Bradke, M. Sheng, H.U. Dodt. Three-dimensional imaging of solvent-cleared organs using 3DISCO. Nature Protocols 7: 1983-1995, 2012 Nov.

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.

All raw image data were acquired from the Institute for Tissue Engineering and Regenerative Medicine and the Institute for Stroke and Dementia Research. The mice samples were purchased from Charles River (CRL, Brussels); human samples were sourced from donation organizations, such as the International Institute for the Advancement of Medicine.

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).

The data acquisition followed a routine: structure staining, tissue clearing, and LSM imaging. Depending on the targeted structure for study, various dyes or stains were employed to selectively bind to specific structures in the sample, enhancing their visibility in contrast to the rest of the sample. Two tissue clearing methods were implemented, as detailed in our prior works [1-2]. All images were captured using a 4× objective lens (Olympus XLFLUOR 340). And these images specifically showcase brain regions.

[1] S. Zhao et al. Cellular and molecular probing of intact human organs. Cell 180(4): 796-812, 2020 Feb.
[2] A. Ertürk, K. Becker, N. Jährling, C.P. Mauch, C.D. Hojer, J.G. Egen, F. Hellal, F. Bradke, M. Sheng, H.U. Dodt. Three-dimensional imaging of solvent-cleared organs using 3DISCO. Nature Protocols 7: 1983-1995, 2012 Nov.

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 unannotated subset of the training set designed for self-supervised learning, each case comprises a sizable 3D LSM image capturing the entire brain of a mouse or parts of a human brain. In the annotated subset of the training set, validation set, and testing set, each case consists of a 3D patch image cropped from a whole-brain image, and the corresponding pixel-wise annotation of labeled structures which is a 3D binary image. All raw LSM images and image patches are of two channels, both structure channel and autofluorescent channel.

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

Training Set:
1)Training subset with no annotations: (A total of 20 samples, each with an image size in the order of 5000 x 6000 x 1000, which is equivalent to > 35000 images of 256x256x256 voxels)
-6 brains with cells labeled by neural activity marker[1],
-11 brains with blood vessel marker[2],
-1 brain with cell nucleus marker[3],
-2 brains with Alzheimer's disease plaque marker[4].
2) Training subset with annotations:
-36 patches from the set above with annotations, each with a patch size of 128×128×128. Half of these patches are of blood vessels, the other half patches are of the rest structure markers.

Validation set:
-5 patches of unseen brains with microglia marker[4] with annotations
-5 patches of unseen brains with neural activity marker with annotations

Testing set:
-10 patches of unseen brains with microglia marker with annotations
-10 patches of unseen brains with neural activity marker with annotations

References:

[1] D. Kaltenecker et al. Virtual reality empowered deep learning analysis of brain activity. Accepted by Nature Methods.
[2] M.I. Todorov et al. Machine learning analysis of whole mouse brain vasculature. Nature Methods 17: 442-449, 2020 Mar.
[3] S. Zhao et al. Cellular and molecular probing of intact human organs. Cell 180(4): 796-812, 2020 Feb.
[4] H.S. Bhatia et al. Spatial proteomics in three-dimensional intact specimens. Cell 185(26): 5040-5058, 2022 Dec.

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

The images in the training set are from our previous published works. Rigorous inspection and verification processes are meticulously conducted on all images to guarantee their quality. The provided images sufficiently cover variability of biological structures within LSM images. Besides, around half of the training set, both the unannotated and annotated subsets, is tree-like structure data (vessels). The rest is spot-like structure data(cells involved in neural activity, cell nucleus, Alzheimer's disease plaque)

In the validation and testing set, some data is of biological structure also in the previous training set (cells involved in neural activity). The rest is of a new biological structure, microglia cell, which is composed of branches like vessels and a small cellular body. Image numbers of the two biological structures will be equal. Performance on such validation and testing sets can reflect the model generalization.

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.

In the training phase, the training subset with no annotations encompasses 3D brain LSM images representing various biological structures from both mice and human brain samples. We provide a significantly smaller training subset with annotations, which contains brain image patches of these biological structures from mice samples, for participants to finetune models.

In the validation phase, half of the validation set are brain patches of the same biological structure as the training set. Another half are brain patches of a new biological structure. All patches in the validation set come from mice brains that are distinct from those in the training set.

In the testing phase, still half of the testing set are brain patches of biological structure which is also shown in the previous training set, and another half are brain patches of a new biological structure. But here we have more patches compared to the validation set from both mice and human brains. A larger testing set helps to enhance the confidence in the model's generalization to unseen data.

Annotation characteristics

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

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

The manual annotation and verification processes are conducted in 3D using virtual reality (VR) for visualization efficiency. Each case undergoes a hierarchical annotation process, beginning with initial semantic segmentation annotations performed manually by 4 expert annotators experienced in LSM imaging. These initial annotations are subsequently verified and fine-tuned by an expert with extensive LSM imaging experience. Finally, two leaders within our organizing team conduct a comprehensive review of all annotations, either approving or revising them.

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

The initial 3D manual annotations are crafted from scratch using virtual reality (VR). Following feedback from the LSM imaging expert and leaders, further manual revisions are conducted in 3D using virtual reality (VR).

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

The initial manual annotations are conducted by expert annotators with in-depth biological and anatomy training, ensuring a comprehensive understanding of LSM. Subsequently, an expert with three years of professional experience in LSM reviews and refines the initial annotations. The final annotations are then determined and approved by two leaders with five or more years of professional experience in LSM.

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.

In the unannotated part of the training set for self-supervised learning, given the considerable size of a whole-brain LSM image, each 2D plane within the 3D image data is preserved as a 16-bit signed TIFF image file.

In the annotated part of the training set for finetuning, validation set and testing set, small patches extracted from the entire brain images, accompanied by their corresponding annotations, are stored in NIfTI format with 16-bit signed precision and in LPS+ orientation.

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.

Our annotation process follows a hierarchical structure with multiple levels of verification and approval. Notably, during the conclusive phase, two leaders collaborate to thoroughly review and either approve or revise the annotations. Given that these two experts collectively assess and finalize the annotations, we anticipate minimal intra-annotator variations.

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

We aim to minimize the number of artifacts in our images. However, stripes are a common type of artifact found in LSM images[1]. These artifacts can lead to over- or under-segmentation errors.

[1] J. Mayer, A. Robert-Moreno, J. Sharpe, J. Swoger. Attenuation artifacts in light sheet fluorescence microscopy corrected by OPTiSPIM. Light: Science & Applications 7: 70, 2018 Oct.

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 labeled structures to be segmented in this challenge can be categorized into two kinds. The first kind comprises spot-like structures, such as different types of cells, while the second kind includes tree-like structures, like vessels.

For the first kind of structures, we will evaluate the segmentation results using two metrics:
1) volumetric Dice similarity coefficient;
2) Betti matching error in dimension 0.

For the second kind of structures, the assessment of segmentation results will involve three metrics:
1) volumetric Dice similarity coefficient;
2) Betti matching error in dimension 0 and 1;
3) centerline Dice similarity coefficient.

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

The volumetric Dice similarity coefficient quantifies the voxel overlap between the ground truth and the segmentation prediction.

In the segmentation tasks of spot-like structures, the primary objective is to detect each individual spot component. In the segmentation tasks of tree-like structures, the aim is to ensure that the segmented structures preserve the topology of the underlying anatomy. The Betti matching error quantifies the spatial matching between topological features of ground truth and segmentation prediction. Betti matching error is not only sensitive to the number of components but also the location of every component.

Centerline-Dice is suitable for evaluating voxel-wise overlap for tubular and curvilinear structures. clDice can measure how much of the tree-like structures are covered in the prediction.

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 ranking for each metric will be determined through the Wilcoxon signed-rank test, with the appropriate hypothesis ('greater' or 'lesser') depending on the metric, conducted on the test-set. To enhance robustness, bootstrapping (sampling with replacement) will be employed to obtain reliable rankings.

For both 'spot-like' and 'tree-like' structure segmentations, we will employ an average rank of these evaluation metrics for the final leaderboard.

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

For volumetric Dice, if the submitted method fails to produce a result on a test case, the metric for that test case will be set to its most penalizing value, i.e. 0.

For Betti matching error, if the submitted method fails to produce a result on a test case, the metric for that test case will be calculated as absolute Betti error.

For clDice, if the submitted method fails to produce a result on a test case, the metric for that test case will be set to its most penalizing value, i.e. 0.

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

The Wilcoxon signed-rank test, with the appropriate hypothesis ('greater' or 'lesser'), is a recommended ranking scheme for metrics according to the literature [1]. This statistical test assesses the statistical significance of differences in the test data between two teams being compared. It offers advantages such as the ability to analyze ordinal data and mitigate the influence of outliers. Similar ranking schemes have been employed in other medical challenges, such as the BraTS challenge (http://braintumorsegmentation.org/).

[1] L. Maier-Hein et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nature communications 9: 1-13, 2018.

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.

Each team will undergo comparison with other teams using the Wilcoxon signed-rank test to ascertain whether there exists a statistically significant difference between the two teams being compared.

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

The Wilcoxon signed-rank test has advantages such as the ability to analyze ordinal data and mitigate the influence of outliers [1]. Similar statistical methods were used in other challenges such as BraTS challenge (http://braintumorsegmentation.org/) with positive feedback from the participants.

[1] P. Mishra, C.M. Pandey, U. Singh, A. Keshri and M. Sabaretnam. Selection of appropriate statistical methods for data analysis. Annals of cardiac anaesthesia 22(3): 297, 2019.

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.

Compared to the annotated training set, the validation set and the final test set include new labeled structures, microglial cells. To showcase the generalizable capacity of semantic segmentation models, we will assess the segmentation results separately for previous structures and the newly introduced structure, analyzing any changes in segmentation performance on the new types of labeled structure. We hope that the design of our challenge encourages participants to focus on self/semi supervised learning.

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.

References are inserted in-place for the relevant text-fields.

Further comments

Further comments from the organizers.

None