

SELMA3D 2025: 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.

SELMA3D 2025: Self-supervised learning for 3D light-sheet microscopy image segmentation

Challenge acronym

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

SELMA3D 2025

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 modern biological research, the ability to visualize and analyze complex structures within tissues and organisms is crucial. Traditional imaging techniques often struggle to provide a cellular-resolution, 3D view of bio-samples while preserving their structural integrity. The combination of tissue clearing and light-sheet microscopy (LSM) overcomes these limitations, serving as a powerful method for high-contrast, ultra-high-resolution imaging. This approach enabled detailed visualization of a wide range of biological structures, including cellular and subcellular structures, organelles and processes, across diverse samples [1].

Tissue clearing techniques render inherently opaque biological samples transparent, allowing light to penetrate deeply into the tissue [2] and imaging reagents (e.g., fluorophores or antibodies), while preserving their structural integrity and molecular content. Various fluorophores or antibodies can be employed to selectively stain specific biological structures within samples and enhance their contrast under microscopy [3]. After staining and tissue clearing, LSM provides rapid 3D imaging of intricate biological structures with high spatial resolution, offering valuable insights into various biomedical fields, such as neuroscience [4], immunology [5], oncology [6] and cardiology [7].

Automated image analysis approaches enable scientists to extract structural and functional, cellular and subcellular information from LSM images of various biosamples at an accelerated pace. To analyze LSM images, segmentation plays a pivotal and essential role in identifying and distinguishing different biological structures [8]. For large LSM images, such as those of whole organs or organisms, manual segmentation is time-intensive, with individual images containing up to 10000^3 voxels. As a result, there is a growing demand for automatic segmentation methods. Recent strides in deep learning-based segmentation models offer promising solutions for automating the segmentation of LSM images [9-10]. While these models achieve performance comparable to

expert human annotators, their success largely relies on supervised learning, which requires extensive, high-quality manual annotations. These models are usually task-specific, designed for particular structures, with limited generalizability across different applications [11]. Therefore, the widespread adaptation of deep learning-based segmentation models is constrained, as the annotation for every specific LSM image segmentation task requires experts with domain knowledge, making the process impractical for many scenarios.

It is crucial to develop generalizable models capable of serving multiple LSM image segmentation tasks. Self-supervised learning offers significant advantages in this regard, as it allows deep learning models to pretrain on large-scale, unannotated datasets, thereby learning useful and generalizable representations of LSM image data. Subsequently, the model can be fine-tuned on a smaller labeled dataset for specific segmentation tasks [12]. 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 properties of LSM images e.g. the high signal-to-noise ratio, makes them particularly well-suited for self-supervised learning.

The SELMA3D 2024 challenge represents a significant advancement in self-supervised learning research within the field of 3D LSM images. To the best of our knowledge, it is the first attempt to benchmark self-supervised learning for LSM image segmentation tasks. Self-supervised learning offers potential benefits for models across a range of downstream tasks. The SELMA3D challenge focuses on segmentation, a crucial task in LSM image analysis. In the 2024 edition, we categorized biological structures frequently studied in LSM images into two main types based on morphology: tree-like structures, including vessels and microglia, and spot-like structures, including cell nuclei, c-Fos⁺ cells and amyloid-beta plaques. Participants were asked to develop a universal self-supervised learning approach for 3D LSM semantic segmentation, one that can benefit segmentation of both types of structures. The top performing teams achieved Dice scores exceeding 70% for both structure types. However, the results obtained by different participants suggested a single self-supervised learning strategy struggles to consistently enhance feature learning for both types simultaneously.

Based on the results and experience from 2024, for the second edition of the SELMA3D challenge hold in 2025, we propose to expand and improve the challenge in the following aspects:

Firstly, we redefine the classification of biological structures into two categories: isolated structures and contiguous structures. Isolated structures refer to distinct, spatially separate components without physical connections, for example cell nuclei, c-Fos⁺ cell, amyloid-beta plaques and microglia. In contrast, contiguous structures highlight the physical continuity between parts of a structure that are connected without interruption, for example vessels and nerves. Based on this classification, the SELMA3D 2025 challenge will be divided into two tasks: 1) self-supervised segmentation of isolated structures in 3D light-sheet microscopy images, 2) self-supervised segmentation of contiguous structures in 3D light-sheet microscopy images. This division reduces the time required for participants to download and preprocess training datasets for each task. Moreover, it allows participants to develop self-supervised learning strategies tailored to the morphological characteristics of specific structures. Participants can select a task based on their research interests and available time, providing greater flexibility and focus.

Secondly, we aim to vastly enhance the dataset by expanding both the number and diversity of samples. In SELMA3D 2024, our dataset comprised 35 whole-brain or brain-subregion 3D images, with 9 images of contiguous structures and 26 of isolated structures. This year, through collaborations with other laboratories and research

institutes, e.g. Alain Chédotal's lab [13-15], we increase the total to 58 3D images, including 28 images of contiguous structures and 30 of isolated structures. Additionally, we have expanded the quantity of annotated patches, providing a richer foundation for model training. Furthermore, while last year's images were limited to brain regions, this year we have incorporated images from various organs and regions across the body to enhance data diversity. We have also broadened the range of biological structures included. The SELMA3D 2024 dataset contained c-Fos+ cells, cell nuclei, amyloid-beta plaques, blood vessels, and microglia. This year, we have expanded it to include neurons, lymphatic cells, and fluorescent proteins for isolated structure segmentation, as well as peripheral nerves, gut nerves, and lymphatic vessels for contiguous structure segmentation. This expansion enhances the foundation for model training by diversifying the dataset and providing participants with a broader range of examples to develop and evaluate their approaches. Ultimately, the increased dataset size and diversity will enable a more in-depth exploration of how self-supervised learning can improve performance and generalize across various tasks.

Thirdly, we plan to expand the quantitative evaluations for the segmentation tasks. Building on the insights gained from SELMA3D 2024, we will incorporate a more comprehensive set of evaluation metrics tailored to the morphological characteristics of the targeted structures. These enhanced segmentation evaluations are designed to provide deeper insights into model performance, addressing both accuracy and morphological consistency. By offering detailed and structure-specific evaluations, we aim to better guide the development of advanced self-supervised learning strategies and improve the interpretability of results.

Given the above aspects, we aim to optimize the challenge setting, establishing a more comprehensive benchmark for self-supervised learning in 3D LSM image segmentation. We look forward to organizing the second edition of the SELMA3D challenge and welcoming submissions.

References:

- [1] E.H.K. Stelzer, F. Strobl, B. Chang, et al. Light sheet fluorescence microscopy. *Nature Reviews Methods Primers* 1(1): 73, 2021 Nov.
- [2] H.R. Ueda, A. Ertürk, K. Chung, et al. Tissue clearing and its applications in neuroscience. *Nature Reviews Neuroscience* 21(2): 61-79, 2020, Jan.
- [3] P.K. Poola, M.I. Afzal, Y. Yoo, et al. Light sheet microscopy for histopathology applications. *Biomedical engineering letters* 9: 279-291, 2019 July.
- [4] H.R. Ueda, H.U. Dodt, P. Osten, et al. 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. Krimtza, et al. 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, et al. Tissue clearing to examine tumour complexity in three dimensions. *Nature Reviews Cancer*, 21(11): 718-730, 2021 July.
- [7] P. Fei, J. Lee, R.R.S. Packard, 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, et al. Efficient processing and analysis of large-scale light-sheet microscopy data. *Nature protocols* 10: 2015: 1679-1696, 2015 Oct.
- [9] N. Kumar, P. Hrobar, M. Vagenknecht, et al. 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, J.C. Paetzold, O. Schoppe, et al. Machine learning analysis of whole mouse brain vasculature. *Nature Methods* 17: 442-449, 2020 Mar.
- [11] Y. Zhou, M.A. Chia, S.K. Wagner, et al. A foundation model for generalizable disease detection from retinal images. *Nature* 622: 156-163, 2023 Sept.
- [12] R. Krishnan, P. Rajpurkar, E.J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering* 6: 1346-1352, 2022 Aug.
- [13] M. Belle, D. Godefroy, C. Dominici, et al. A simple method for 3D analysis of immunolabeled axonal tracts in a transparent nervous system. *Cell reports* 9(4): 1191-1201, 2014 Nov.
- [14] M. Belle, D. Godefroy, G. Couly, et al. Tridimensional visualization and analysis of early human development. *Cell* 169(1): 161-173, 2017 Mar.
- [15] R. Blain, G. Couly, E. Shotar, et al. A tridimensional atlas of the developing human head. *Cell* 186(26): 5910-5924, 2023 Dec.

Challenge keywords

List the primary keywords that characterize the challenge.

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

Year

2025

Novelty of the challenge

Briefly describe the novelty of the challenge.

SELMA3D challenge is the first challenge to benchmark self-supervised learning specifically for 3D LSM image segmentation. The challenge emphasizes self-supervised learning, which reduces the reliance on extensive manual annotations by pretraining on large unannotated datasets and fine-tuning on smaller labeled sets. Based on morphological characteristics, common biological structures in LSM are classified into two types: isolated and contiguous structures. Participants are asked to develop tailored self-supervised strategies for each structure type.

Task description and application scenarios

Briefly describe the application scenarios for the tasks in the challenge.

While the downstream task for the self-supervised learning of 3D light-sheet microscopy images in this challenge is segmentation, the pretrained model or self-supervised learning strategy can be leveraged for a variety of other downstream tasks, such as object detection, anomaly detection, image denoising and so on.

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?

2 Hours

In case you selected half or full day, please explain why you need a long slot for your challenge.

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.

During the SELMA3D challenge held in 2024, we had 84 registered participants. After the challenge concluded, we received additional registrations and expressions of interest from potential participants. Moreover, during our promotion efforts last August, some individuals expressed interest in participating but noted that the large-scale dataset posed a significant challenge, making it difficult to complete their work before the deadline. This feedback suggests considerable potential for increased participation. With the growing attention on SSL pre-training, particularly with the rise of large vision models, we anticipate even greater participation this year. To further boost engagement, we will launch a coordinated social media campaign across LinkedIn, Bluesky, and X to promote the challenge. For SELMA3D 2025, we expect at least 120 participants to register.

In 2024, 5 teams submitted algorithms. Most teams reported that limited time hindered their ability to develop more advanced solutions. To encourage more submission, we adjust the challenge settings in several aspects to make the workload more manageable. First, we split the SELMA3D 2024 task into two separate tasks, dividing the dataset into images of isolated structures and images of contiguous structures. This allows participants to choose a task that aligns with their research interests and available time, thereby reducing both data preprocessing efforts and model training time.

Second, we will provide preprocessing code to assist participants in handling the dataset, specifically for cropping sequential slices into 3D patches. This code, originally developed in our previous work, will be adapted for this challenge. For reference, the preprocessing code can be found here:

https://github.com/erturklab/SCP-Nano/blob/main/2_image_cropping.py.

Additionally, we will begin publicizing and promoting the challenge earlier. Last year, the late promotion limited the time available for teams to develop solutions despite their interest. This time, we will start outreach sooner and release the training dataset earlier, ensuring participants have ample time to develop and refine their algorithms. These improvements aim to make the challenge more accessible, encourage broader participation, and support the development of more advanced solutions.

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 from both 2024 and 2025 in a journal article.

The SELMA3D 2024 challenge results have been summarized in a pre-print on arXiv[1].

[1] Y. Chen, R. Al-Maskari, I. Horvath et al. SELMA3D challenge: Self-supervised learning for 3D light-sheet microscopy image segmentation. arXiv:2501.03880, 2025.

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. Participants are expected to utilize their own computing resources for algorithm development. Organizers will employ grand-challenge.org for evaluation during the testing phases.

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

TASK 1: Self-supervised segmentation of isolated structures in 3D light-sheet microscopy images

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 cellular resolution while minimizing photobleaching and phototoxicity. This makes LSM a powerful tool for visualizing a variety of biological structures and studying their functions and developmental processes.

Task 1 is dedicated to developing self-supervised learning strategies for segmenting isolated structures in 3D light-sheet microscopy images. Isolated structures is one of the primary types of biological structures commonly studied in microscopy. These structures are characterized by spatially distinct components that lack physical continuity or direct connections. Examples include cell nuclei, specific cell types such as neurons or immune cells, and pathological formations like plaques (e.g., amyloid-beta plaques in Alzheimer's disease).

In this task, participants will be provided with a training dataset consisting of two subsets:

1. Unannotated subset: A large collection of 3D LSM images of isolated structures derived from both mouse and human samples. This subset includes more than 30 high-resolution images, each exceeding 4×10^{10} voxels, amounting to over 70,000 patches of size $256 \times 256 \times 256$ voxels. These unannotated images are designed to support model pretraining using self-supervised learning techniques, enabling the model to learn generalizable representations of isolated structures.
2. Annotated subset: A curated selection of 3D LSM image patches representing the same isolated structures in the unannotated subset, but with precise manual annotations. This subset enables participants to fine-tune their models for the segmentation of isolated structures, leveraging the pre-training phase.

Keywords

List the primary keywords that characterize the task.

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

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, Abid Abrar

[Cornell University, New York, USA]

Johannes C. Paetzold

[Institut de la Vision, Paris, France]

Yorick Gitton, Alain Chedotal

[Ludwig Maximilian University of Munich, Germany]

Ying Chen

[University of Zurich, Switzerland]

Bjoern Menze, Kaiyuan Yang

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

Repeated event as open call challenge

Challenge venue and platform

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

MICCAI 2025

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).

SELMA3D 2024 website: <https://selma3d.grand-challenge.org/>; SELMA3D 2025 website: will appear on the SELMA3D 2024 website

Participation policies

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

Fully automatic

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

No additional data allowed

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

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

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

The challenge webpage will publicly acknowledge the top 3 teams, and they will receive a Jellycat Selma SLOTH 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 after submission, 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 prior to the announcement. The top 5 teams will be invited to prepare a 5-10 minute presentation for the challenge session.

Participants who wish to retract their submissions can opt for their performance to be reported anonymously online and in the publication, or they can request complete removal 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 SELMA3D 2024 and SELMA3D 2025 challenges. 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.

Submissions for evaluation on both the preliminary and the final test sets will be done through submitted docker containers, i.e. type 2 submissions on Grand Challenge. Submitted containers will be evaluated by organizers on grand-challenge.org. Participants will be provided with Docker templates and instructions to assist with their submissions. For reference, please refer to the SELMA3D 2024 Docker templates and instructions:

https://github.com/YingChen7/SELMA3D_challenge-submission-example. This year, we will enhance support by offering step-by-step tutorials with images to guide participants through the process. Additionally, we will organize live Q&A sessions and webinars on containerization and deployment to assist teams that may lack expertise in this area. These initiatives aim to make the submission process smoother and more accessible for all participants.

In addition to the docker containers, each participating team is required to submit a 2-3 page summary outlining their methods and approaches along with the Docker container for the final test set submission. This summary is a mandatory part of the final test phase submission and is also necessary for co-authorship in the final challenge journal paper. Furthermore, based on our experience summarizing the SELMA3D 2024 submissions in the archive paper, we will provide more detailed guidelines on what participants should include in their summaries. For the self-supervised learning stage, the summary should cover the self-supervised learning strategy, data preprocessing steps, and network architecture. For the fine-tuning stage, participants should describe the fine-tuning strategy, data preprocessing steps, and the final segmentation network architecture. These guidelines will ensure clarity and consistency across submissions.

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 some examples from the preliminary and final test sets without annotations. These examples can be used to test and validate their algorithms locally before official submission for evaluation. This allows participants to assess the performance of their methods and make any necessary adjustments prior to submitting them to Grand Challenge.

Each team is allowed a maximum of 15 submissions for evaluation on the preliminary test set. The results from the preliminary test set will contribute to the public leaderboard, providing an initial, unofficial ranking. Teams can use the evaluation results to refine their algorithms and improve performance..

It is crucial to note that the final ranking will be determined exclusively by the performance on the final test set. For the final test set, each team is provided with 2 opportunities to upload their Docker containers, allowing participants to compare the model's performance with and without the self-supervised learning stage. In case of technical issues, participants may re-submit their containers. However, only the results from the last submission will be considered when calculating the official 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 1st, 2025
- Training set release: April 10th, 2025
- Release of preliminary test set samples (without annotation): June 1st, 2025
- Opening of submission and leaderboard for the preliminary test set: June 20th, 2025
- Release of final test set samples (without annotation): June 30st, 2025
- Opening of submission and leaderboard for the final test set: July 10th, 2025
- Contacting top-performing teams and planning for the in-person session: Aug 10th - September 10th, 2025
(teams requiring visas will be contacted earlier, and we will coordinate with MICCAI to send invitation letters for visa clearance.)
- In-person challenge event: September 23th, 2025

Additional point:

The performance results will be released public immediately after submission and evaluation on the Grand Challenge platform.

Compared to SELMA3D 2024, the 2025 edition will release the training set earlier, giving participants at least an extra month to develop their methods. Besides, we plan to conduct a series of both online and in-person seminars to promote the challenge and address any questions participants may have.

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 task 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 [1-6]. 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, M.I. Todorov, R. Cai, et al. Cellular and molecular probing of intact human organs. *Cell* 180(4): 796-812, 2020 Feb.
- [2] H. Mai, Z. Rong, S. Zhao, et al. Scalable tissue labeling and clearing of intact human organs. *Nature Protocols* 17: 2188-2215, 2022 July.
- [3] H.S. Bhatia, A. Brunner, F. Öztürk, et al. Spatial proteomics in three-dimensional intact specimens. *Cell* 185(26): 5040-5058, 2022 Dec.
- [4] D. Kaltenecker, R. Al-Maskari, M. Negwer, et al. Virtual reality empowered deep learning analysis of brain activity. *Nature Methods*(21): 1306–1315, 2024 April.
- [5] M. Belle, D. Godefroy, C. Dominici, et al. A simple method for 3D analysis of immunolabeled axonal tracts in a

transparent nervous system. Cell reports 9(4): 1191-1201, 2014 Nov.

[6] M. Belle, D. Godefroy, G. Couly, et al. Tridimensional visualization and analysis of early human development. Cell 169(1): 161-173, 2017 Mar.

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 (Attribution-NonCommercial)

Code availability

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

The evaluation code for the SELMA3D 2024 challenge is publicly available on GitHub:
https://github.com/YingChen7/SELMA3D_challenge-evaluation/.

The evaluation code for SELMA3D 2025 will be accessible to the public on GitHub as well.

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 the preliminary and final test sets with annotations 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 isolated structures from diverse bio-samples. A self-supervised learning strategy for LSM segmentation of isolated structures holds potential benefits across various segmentation applications for whole organ or whole body, including but not limited to cell or cell nuclei 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 isolated structures derived from mouse and human samples. These images were produced by the Institute for Tissue Engineering and Regenerative Medicine (iTERM), the Institute for Stroke and Dementia Research and the Institut de la Vision between 2012 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 samples obtained from both mice and humans following a tissue clearing protocol[1-4].

Individual specimen information will not be made available.

References:

[1] S. Zhao, M.I. Todorov, R. Cai, et al. Cellular and molecular probing of intact human organs. *Cell* 180(4): 796-812, 2020 Feb.

[2] A. Ertürk, K. Becker, N. Jähring, et al. Three-dimensional imaging of solvent-cleared organs using 3DISCO. *Nature Protocols* 7: 1983-1995, 2012 Nov.

[3] N. Renier, Z. Wu, D.J. Simon, et al. iDISCO: a simple, rapid method to immunolabel large tissue samples for

volume imaging. Cell 159(4): 896-910, 2014 Nov.

[4] N Renier, EL Adams, C Kirst et al. Mapping of brain activity by automated volume analysis of immediate early genes. Cell 165(7): 1789-1802, 2016 June.

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

This task focuses on a self-supervised learning strategy to develop a generalized model for semantic segmentation of isolated structures in 3D light-sheet microscopy (LSM) images.

To prevent participants from solely relying on the annotated training set for model training—thereby achieving favorable segmentation results without effectively utilizing the self-supervised learning stage—variations will be introduced between the annotated training set, preliminary test set, and final test set. The objective is to encourage the development of self-supervised learning strategies capable of pretraining models with robust generalization capabilities.

To be specific, the annotated training set will encompass cell nuclei, c-Fos+ cells, and amyloid-beta plaque from mouse and human samples. The preliminary and final test sets will not only contain patches of the biological structures present in the training set but will also introduce patches from two distinct biological structures: fluorescent protein (EGFP) and microglia.

Assessment aim(s)

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

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

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

Accuracy

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 from Ali Euturk's lab was performed using objective lens 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 (Miltényi Biotec, chamber size of 250 × 90 × 70 mm, for large samples) coupled to a white light laser module (NKT SuperK Extreme EXW-12).

LSM imaging from Alain Chédotal's lab was performed with an ultramicroscope (LaVision BioTec) using InspectorPro software (LaVision BioTec) or a Blaze light-sheet microscope (Miltényi Biotec) equipped with sCMOS camera 5.5MP (2560×2160 pixels) controlled by Inspector Pro 7.5.3 acquisition software (Miltényi Biotec)

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 fluorophores or antibodies were used to selectively bind to specific isolated structures in the sample, enhancing their visibility in the images.

Different tissue clearing methods were used following our previous work [1-4].

[1] S. Zhao, M.I. Todorov, R. Cai, et al. Cellular and molecular probing of intact human organs. *Cell* 180(4): 796-812, 2020 Feb.

[2] A. Ertürk, K. Becker, N. Jähring, et al. Three-dimensional imaging of solvent-cleared organs using 3DISCO. *Nature Protocols* 7: 1983-1995, 2012 Nov.

[3] N. Renier, Z. Wu, D.J. Simon, et al. iDISCO: a simple, rapid method to immunolabel large tissue samples for volume imaging. *Cell* 159(4): 896-910, 2014 Nov.

[4] N Renier, EL Adams, C Kirst et al. Mapping of brain activity by automated volume analysis of immediate early genes. *Cell* 165(7): 1789-1802, 2016 June.

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, the Institute for Stroke and Dementia Research and the Institut de la Vision.

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 standardized imaging workflow, which includes structure staining, tissue clearing, and LSM imaging, was performed by experienced graduates and technicians who had undergone at least six months of training to master the process.

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 large 3D LSM image capturing either the full organ or part of an organ from a mouse or human. In the annotated subset of the training set, as well as in the preliminary and final test sets, each case consists of a 3D patch image cropped from the large 3D LSM image, accompanied by a corresponding voxel-wise annotation of the stained biological structures in the form of a 3D binary image. All raw LSM images and image patches are single-channel.

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

Training Set:

1) Training subset with no annotations: (A total of 30 3D images, each exceeding 4×10^{10} voxels, amounting to over 70,000 patches of size $256 \times 256 \times 256$ voxels.)

- 18 3D images of c-Fos+ cells [1]
- 4 3D images of cell nuclei [2]
- 4 3D images of amyloid-beta plaques [3]
- 2 3D images of neuron [4]
- 2 3D images of lymphatic cells [5]

2) Training subset with annotations:

- 9 3D patches of c-Fos+ cells
- 7 3D patches of cell nuclei
- 20 3D patches of amyloid-beta plaques

Preliminary test set:

- 2 3D patches of c-Fos+ cells
- 1 3D patches of cell nuclei
- 3 3D patches of amyloid-beta plaques
- 3 3D patches of microglia [6]
- 3 3D patches of fluorescent protein (EGFP)

Final test set:

- 4 3D patches of c-Fos+ cells
- 2 3D patches of cell nuclei
- 11 3D patches of amyloid-beta plaques
- 17 3D patches of microglia
- 17 3D patches of fluorescent protein (EGFP)

References:

- [1] D. Kaltenecker, R. Al-Maskari, M. Negwer, et al. Virtual reality empowered deep learning analysis of brain activity. *Nature Methods*(21): 1306–1315, 2024 April.
- [2] S. Zhao, M.I. Todorov, R. Cai, et al. Cellular and molecular probing of intact human organs. *Cell* 180(4): 796-812, 2020 Feb.
- [3] H.S. Bhatia, A. Brunner, F. Öztürk, et al. Spatial proteomics in three-dimensional intact specimens. *Cell* 185(26):

5040-5058, 2022 Dec.

[4] M. Belle, D. Godefroy, C. Dominici, et al. A simple method for 3D analysis of immunolabeled axonal tracts in a transparent nervous system. *Cell reports* 9(4): 1191-1201, 2014 Nov.

[5] M. Belle, D. Godefroy, G. Couly, et al. Tridimensional visualization and analysis of early human development. *Cell* 169(1): 161-173, 2017 Mar.

[6] H. Mai, Z. Rong, S. Zhao, et al. Scalable tissue labeling and clearing of intact human organs. *Nature Protocols* 17: 2188-2215, 2022 July.

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 sourced from our previously published works. A thorough inspection and verification process is meticulously carried out on all images to ensure their quality. The provided images encompass a wide range of isolated structures within LSM images, capturing significant variability.

In the preliminary and final test sets, less than half of the patches correspond to biological structures also present in the training set, including c-Fos+ cells, cell nuclei and amyloid-beta plaques. The remainder consists of new biological structures, including microglia and fluorescent protein. Performance on these test sets will provide insights into the model's ability to generalize to unseen structures.

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.

None

e) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

In the training set, the unannotated subset consists of large 3D LSM images showcasing a diverse range of isolated structures from both mouse and human bio-samples. This unannotated data is designed for pretraining models using self-supervised learning techniques. Additionally, we provide a smaller annotated subset, comprising image patches with precise voxel-wise labels of these structures, to enable participants to fine-tune their models for segmentation tasks.

In the preliminary test set or validation set, few patches (6 patches) present the same isolated structures present in the training set, while the other few patches (6 patches) present newly introduced isolated structures. This composition allows participants to assess their algorithms' performance on both seen structures and unseen structures. By using this setup, participants gain insights into how well their models generalize to unseen data while also validating their performance on previously seen structures.

In the final test set, the proportion of unseen isolated structures is increased to 2/3 (34 patches), compared to 50% in the preliminary test set, surpassing the number of seen structures present in the training data. This intentional imbalance ensures a robust evaluation of the model's ability to generalize to completely new data, a critical goal of self-supervised learning. By focusing on unseen isolated structures, the final test set provides a more rigorous test of the participants' algorithms and emphasizes the core of the challenge, that is to develop self-supervised learning methods for robust and generalizable segmentation.

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, which are manually performed by an expert annotator with experience in LSM imaging. Four expert annotators participated in this stage, each handling different cases. After the initial annotation, an expert with extensive LSM imaging experience verifies and refines the annotations. Additionally, two leaders from our organizing team conduct a thorough review of all annotations to ensure their accuracy and reliability.

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 were crafted from scratch using virtual reality (VR) by annotators who underwent a minimum of three months of specialized training. For each image patch, which contains a single kind of biological structure, the annotators were asked to voxel-wisely label that biological structure with a value of 1, while assigning a value of 0 to the rest background. After the initial annotation process, additional manual refinements were carried out in 3D using VR, incorporating feedback from LSM imaging experts and team leaders to ensure accuracy and consistency.

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 subset of the training set for self-supervised learning, given the considerable size of a full 3D LSM image, each 2D plane within the 3D image is stored as a 16-bit signed TIFF image file.

For the annotated subset of the training set, as well as the preliminary and final test sets, small patches extracted from the large 3D image, accompanied by their corresponding annotations, are stored in MHA (MetaImage) format with 16-bit signed precision.

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. Due to the significant time and effort required to annotate a single case, we do not have multiple annotations for the same case from different annotators, and as a result, we are unable to provide inter- and intra-annotator variability. Notably, an expert with extensive LSM imaging experience, along with two leaders from our organizing team, collaboratively review and either approve or revise the annotations. As these three individuals work together to assess and finalize the annotations, our goal is to ensure their accuracy and reliability.

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 focus of Task 1 is to develop a self-supervised learning strategy that enhances the performance of semantic segmentation models for isolated structures.

For assessing segmentation results of isolated structures, two metrics will be employed:

1. volumetric Dice similarity coefficients (DSC)
2. Betti matching error in dimension 0 [1]

[1] N. Stucki, J.C. Paetzold, S. Shit, et al. Topologically faithful image segmentation via induced matching of persistence barcodes. In *International Conference on Machine Learning*, 2023, pp. 32698-32727.

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

The selected metrics align with the recommendations of the Metrics Reloaded toolkit [1]. The volumetric Dice similarity coefficient is a widely used metric to assess the overall overlap between the ground truth and the predicted segmentation. It calculates the similarity of two sets by comparing the shared voxel volume relative to the total voxel volume, providing a global measure of segmentation accuracy.

In the segmentation tasks of isolated structures, the primary objective is to detect each individual component accurately. The Betti matching error in dimension 0 quantifies the discrepancies in the number and spatial

arrangement of individual segments. This metric provides valuable insights into how well the algorithm captures distinct components, helping to evaluate its performance in separating and identifying individual structures.

[1] L. Maier-Hein, A. Reinke, P. Godau, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods* 21: 195–212, 2024 Feb.

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 of a submitted algorithm is determined through the following process:

- Compute the metric scores for each test case;
- Calculate the average of the metric scores across all test cases for each individual metric;
- Rank the averaged scores for each metric independently based on its specific optimization trend (e.g., higher is better or lower is better);
- Determine the overall ranking of the submitted algorithm by calculating the mean rank across all metrics;
- If two or more algorithms have equal final ranks, the prize will be shared equally among them.

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

For volumetric Dice similarity coefficients, if the submitted method fails to produce a result for a test case, the metric for that test case will be assigned the most penalizing value of 0.

For Betti matching error in dimension 0, if the submitted method fails to produce a result for a test case, the metric will default to the absolute Betti error for that case.

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

There is no interdependence between test cases. While multiple test cases may originate from the same 3D LSM image, there is no overlap between any two test cases. To evaluate algorithm performance, various metrics are calculated independently for each test case.

We adhere to the recommendations of L. Maier-Hein et al. [1] for determining the final ranking of algorithms. In general, there are two primary ranking schemes for aggregating metric scores across test cases. The first scheme is metric-based aggregation or aggregate-then-rank, which means metric scores are first aggregated across all test cases using statistical measures such as the mean or median. Algorithms are then ranked based on the aggregated score. The second scheme is case-based aggregation or rank-then-aggregate. In this scheme, each test case is ranked individually, and the final rank of an algorithm is determined by aggregating the ranks across all test cases.

According to L. Maier-Hein et al [1], single-metric rankings exhibit greater statistical robustness when metric-based aggregation is employed and when the mean is used instead of the median for aggregation. This scheme ensures more reliable comparisons between algorithms.

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

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.

We will employ bootstrapping and leave-one-out analyses to assess the robustness and stability of the rankings, as described by L. Maier-Hein et al [1]. The results will be presented during the in-person challenge event.

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

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

Bootstrapping or sampling with replacement, involves creating multiple resampled datasets by randomly selecting test cases from the original test set, with some cases potentially appearing more than once and others not at all. This method enables an assessment of ranking stability across different subsets of the test set, offering insights into how rankings might fluctuate under varying sampling conditions.

Leave-one-out analysis, on the other hand, systematically removes one case at a time from the test set and applies the ranking scheme to the reduced subset. This approach evaluates the influence of excluding individual test cases, providing a detailed understanding of how specific cases impact the overall rankings.

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.

The preliminary and final test sets include newly introduced structures not present in the annotated training set. To evaluate the generalization capability of the semantic segmentation models, we will separately assess segmentation performance for both seen and unseen structures. This analysis will highlight any variations in performance on the newly introduced structures compared to those included in the training set. By designing the challenge this way, we aim to motivate participants to prioritize self-supervised learning approaches that enhance model generalization.

TASK 2: self-supervised segmentation of contiguous structures in 3D light-sheet microscopy images

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 cellular resolution while minimizing photobleaching and phototoxicity. This makes LSM a powerful tool for visualizing a variety of biological structures and studying their functions and developmental processes.

Task 2 is dedicated to developing self-supervised learning strategies for segmenting contiguous structures in 3D light-sheet microscopy images. Contiguous structures represent another key type of biological structures commonly studied in microscopy. These structures are characterized by the physical continuity of their components, where different parts are connected without interruption, forming a continuous entity. Examples include blood vessels, lymphatic vessels, and nerves.

In this task, participants will be provided with a training dataset consisting of two subsets:

1. Unannotated subset: A large collection of 3D LSM images of contiguous structures derived from both mouse and human samples. This subset includes more than 28 high-resolution images, each exceeding 3×10^{10} voxels, amounting to over 50,000 patches of size $256 \times 256 \times 256$ voxels. These unannotated images are designed to support model pretraining using self-supervised learning techniques, enabling the model to learn generalizable representations of contiguous structures.
2. Annotated subset: A curated selection of 3D LSM image patches representing the same contiguous structures in the unannotated subset, but with precise manual annotations. This subset enables participants to fine-tune their models for the segmentation of contiguous structures, leveraging the pre-training phase.

Keywords

List the primary keywords that characterize the task.

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

ORGANIZATION

Organizers

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

Same as Task 1

- b) Provide information on the primary contact person.

Same as Task 1

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

Repeated event as open call challenge

Challenge venue and platform

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

Same as Task 1

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

Same as Task 1

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

Same as Task 1

Participation policies

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

Fully automatic

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

No additional data allowed

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

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

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

Same as Task 1

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.

Same as Task 1

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).

Same as Task 1

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.

Same as Task 1

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.

Same as Task 1

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

Same as Task 1

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 task 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 [1-5]. 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] M.I. Todorov, J.C. Paetzold, O. Schoppe, et al. Machine learning analysis of whole mouse brain vasculature. *Nature Methods* 17: 442-449, 2020 Mar.
- [2] H. Mai, J. Luo, L. Hoeher, et al. Whole-body cellular mapping in mouse using standard IgG antibodies. *Nature Biotechnology* 42: 617-627, 2023 July.
- [3] M. Belle, D. Godefroy, C. Dominici, et al. A simple method for 3D analysis of immunolabeled axonal tracts in a transparent nervous system. *Cell reports* 9(4): 1191-1201, 2014 Nov.
- [4] M. Belle, D. Godefroy, G. Couly, et al. Tridimensional visualization and analysis of early human development. *Cell* 169(1): 161-173, 2017 Mar.
- [5] R. Blain, G. Couly, E. Shotar, et al. A tridimensional atlas of the developing human head. *Cell* 186(26): 5910-5924, 2023 Dec.

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 (Attribution-NonCommercial)

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.

Same as Task 1

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

Same as Task 1

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.

Same as Task 1

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 contiguous structures from diverse bio-samples. A self-supervised learning strategy for LSM segmentation of contiguous structures holds potential benefits across various segmentation applications for whole organ or whole body, including but not limited to vascular or nerve 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 contiguous structures derived from mouse and human samples. These images were produced by the Institute for Tissue Engineering and Regenerative Medicine (iTERM), the Institute for Stroke and Dementia Research and the Institut de la Vision between 2012 and 2023.

Imaging modality(ies)

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

Same as Task 1

Context information

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

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

Same as Task 1

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

Same as Task 1

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 samples obtained from both mice and humans following a tissue clearing protocol [1-5].

Individual specimen information will not be made available.

References:

[1] S. Zhao, M.I. Todorov, R. Cai, et al. Cellular and molecular probing of intact human organs. *Cell* 180(4): 796-812, 2020 Feb.

[2] A. Ertürk, K. Becker, N. Jährling, et al. Three-dimensional imaging of solvent-cleared organs using 3DISCO. *Nature Protocols* 7: 1983-1995, 2012 Nov.

[3] H. Mai, J. Luo, L. Hoeher, et al. Whole-body cellular mapping in mouse using standard IgG antibodies. *Nature*

Biotechnology 42: 617–627, 2023 July.

[4] N. Renier, Z. Wu, D.J. Simon, et al. iDISCO: a simple, rapid method to immunolabel large tissue samples for volume imaging. *Cell* 159(4): 896-910, 2014 Nov.

[5] N Renier, EL Adams, C Kirst et al. Mapping of brain activity by automated volume analysis of immediate early genes. *Cell* 165(7): 1789-1802, 2016 June.

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

This task focuses on a self-supervised learning strategy to develop a generalized model for semantic segmentation of contiguous structures in 3D light-sheet microscopy (LSM) images.

To prevent participants from solely relying on the annotated training set for model training—thereby achieving favorable segmentation results without effectively utilizing the self-supervised learning stage—variations will be introduced between the annotated training set, preliminary test set, and final test set. The objective is to encourage the development of self-supervised learning strategies capable of pretraining models with robust generalization capabilities.

To be specific, the annotated training set will encompass blood vessels and peripheral nervous data from mouse and human samples. The preliminary and final test sets will not only contain patches of the biological structures present in the training set but will also introduce patches from a distinct biological structure: lymphatic vessels.

Assessment aim(s)

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

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

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

Accuracy

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).

Same as Task 1

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 fluorophores or antibodies were used to selectively bind to specific contiguous structures in the sample, enhancing their visibility in the images. Different tissue clearing methods were used following our previous work [1-5].

[1] S. Zhao, M.I. Todorov, R. Cai, et al. Cellular and molecular probing of intact human organs. *Cell* 180(4): 796-812, 2020 Feb.

[2] A. Ertürk, K. Becker, N. Jährling, et al. Three-dimensional imaging of solvent-cleared organs using 3DISCO. *Nature Protocols* 7: 1983-1995, 2012 Nov.

[3] H. Mai, J. Luo, L. Hoeher, et al. Whole-body cellular mapping in mouse using standard IgG antibodies. *Nature Biotechnology* 42: 617-627, 2023 July.

[4] N. Renier, Z. Wu, D.J. Simon, et al. iDISCO: a simple, rapid method to immunolabel large tissue samples for volume imaging. *Cell* 159(4): 896-910, 2014 Nov.

[5] N Renier, EL Adams, C Kirst et al. Mapping of brain activity by automated volume analysis of immediate early genes. *Cell* 165(7): 1789-1802, 2016 June.

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.

Same as Task 1

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).

Same as Task 1

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.

Same as Task 1

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

Training Set:

- 1) Training subset with no annotations: (A total of 28 3D images, each exceeding 3×10^{10} voxels, amounting to over 50,000 patches of size $256 \times 256 \times 256$ voxels.)
 - 14 3D images of blood vessels [1,]

- 14 3D images of peripheral nerves [2-6]

2) Training subset with annotations:

- 16 3D patches of blood vessels
- 11 3D patches of peripheral nerves

Preliminary test set:

- 3 3D patches of blood vessels
- 2 3D patches of peripheral nerves
- 2 3D patches of gut nerves
- 2 3D patches of lymphatic vessels

Final test set:

- 5 3D patches of blood vessels
- 5 3D patches of peripheral nerves
- 6 3D patches of gut nerves
- 6 3D patches of lymphatic vessels

References:

- [1] M.I. Todorov, J.C. Paetzold, O. Schoppe, et al. Machine learning analysis of whole mouse brain vasculature. *Nature Methods* 17: 442-449, 2020 Mar.
- [2] H. Mai, J. Luo, L. Hoeher, et al. Whole-body cellular mapping in mouse using standard IgG antibodies. *Nature Biotechnology* 42: 617-627, 2023 July.
- [3] R. Cai, C. Pan, A. Ghasemigharagoz, et al. Panoptic imaging of transparent mice reveals whole-body neuronal projections and skull-meninges connections. *Nature Neuroscience* 22: 317-327, 2019 Dec.
- [4] M. Belle, D. Godefroy, C. Dominici, et al. A simple method for 3D analysis of immunolabeled axonal tracts in a transparent nervous system. *Cell reports* 9(4): 1191-1201, 2014 Nov.
- [5] M. Belle, D. Godefroy, G. Couly, et al. Tridimensional visualization and analysis of early human development. *Cell* 169(1): 161-173, 2017 Mar.
- [6] R. Blain, G. Couly, E. Shotar, et al. A tridimensional atlas of the developing human head. *Cell* 186(26): 5910-5924, 2023 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 derived from our previously published works. A thorough inspection and verification process is meticulously carried out on all images to ensure their quality. The provided images encompass a wide range of contiguous structures within LSM images, capturing significant variability.

In the preliminary and final test sets, part of the patches present biological structures also shown in the training set, including blood vessels and peripheral nerves. The other part consists of gut-specific nerves and a new biological structure, lymphatic vessels. Performance on these test sets will provide insights into the model's ability to generalize to unseen structures.

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.

None

e) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

In the training set, the unannotated subset consists of large 3D LSM images showcasing a diverse range of contiguous structures from both mouse and human bio-samples. This unannotated data is designed for pretraining models using self-supervised learning techniques. Additionally, we provide a smaller annotated subset, comprising image patches with precise voxel-wise labels of these structures, to enable participants to fine-tune their models for segmentation tasks.

In the preliminary test set or validation set, few patches (5 patches) present the same contiguous structures present in the training set, while the other few patches (4 patches) present newly introduced contiguous structures. This composition allows participants to assess their algorithms' performance on both seen structures and unseen structures. By using this setup, participants gain insights into how well their models generalize to unseen data while also validating their performance on previously seen structures.

In the final test set, the proportion of unseen contiguous structures is increased to 55% (12 patches), compared to 44% in the preliminary test set, surpassing the number of seen structures present in the training data. This intentional imbalance ensures a robust evaluation of the model's ability to generalize to completely new data, a critical goal of self-supervised learning. By focusing on unseen contiguous structures, the final test set provides a more rigorous test of the participants' algorithms and emphasizes the core of the challenge, that is to develop self-supervised learning methods for robust and generalizable segmentation.

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.

Same as Task 1

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.

Same as Task 1

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.

Same as Task 1

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.

Same as Task 1

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.

Same as Task 1

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

Same as Task 1

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 focus of Task 2 is to develop a self-supervised learning strategy that enhances the performance of semantic segmentation models for contiguous structures.

For assessing segmentation results of contiguous structures, four metrics will be employed:

1. volumetric Dice similarity coefficients (DSC)
2. Centerline-Dice similarity coefficients (cIDice) [1]
3. Hausdorff Distance 95% Percentile (HD95)
4. Betti matching error in dimension 0 [2]

[1] S. Shit, J.C. Paetzold, A. Sekuboyina, et al. cIDice - a novel topology-preserving loss function for tubular structure segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16555-16564.

[2] N. Stucki, J.C. Paetzold, S. Shit, et al. Topologically faithful image segmentation via induced matching of persistence barcodes. In International Conference on Machine Learning, 2023, pp. 32698-32727.

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

The selected metrics align with the recommendations of the Metrics Reloaded toolkit [1]. The volumetric Dice similarity coefficient is a widely used metric to assess the overall overlap between the ground truth and the predicted segmentation. It calculates the similarity of two sets by comparing the shared voxel volume relative to the total voxel volume, providing a global measure of segmentation accuracy.

The Hausdorff Distance 95% Percentile measures the maximum distance between two sets of points. This metric

focuses on the most significant discrepancies between the predicted and ground truth segmentations, helping to identify how well the boundaries of the segmented structure align. This makes it especially useful for evaluating segmentation accuracy in contiguous structures with complex or irregular boundaries.

In the segmentation tasks of contiguous structures, the goal is not only to capture the correct boundaries but also to preserve the topology of the underlying anatomical structures. The Betti matching error is a specialized metric used to evaluate how well the predicted segmentation matches the topological features of the ground truth. Betti matching error helps to identify whether the segmentation algorithm preserves the connectedness and branching patterns of contiguous structures.

Centerline-Dice evaluates the voxel-wise overlap of the central axis of contiguous structures. cDice helps assess how well the centerline of the predicted structure matches the centerline of the ground truth, offering a clear measure of how much of the contiguous structure is correctly identified and captured.

[1] L. Maier-Hein, A. Reinke, P. Godau, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods* 21: 195–212, 2024 Feb.

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.

Same as Task 1

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

For volumetric Dice similarity coefficients and centerline-Dice similarity coefficients, if the submitted method fails to produce a result for a test case, the metric for that test case will be assigned the most penalizing value of 0.

For Hausdorff Distance 95% Percentile, if the submitted method fails to produce a result for a test case, the metric will be set to the diagonal distance of the cuboid defined by the patch dimensions.

For Betti matching error in dimension 0, if the submitted method fails to produce a result for a test case, the metric will default to the absolute Betti error for that case.

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

Same as Task 1

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.

Same as Task 1

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

Same as Task 1

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.

Same as Task 1

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