

# Semi-supervised Teeth Segmentation: Structured description of the challenge design

## CHALLENGE ORGANIZATION

### Title

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

Semi-supervised Teeth Segmentation

### Challenge acronym

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

Semi-TeethSeg

### 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.

Computer-aided design (CAD) tools are increasingly popular in modern dental practice, particularly for treatment planning or comprehensive prognosis evaluation. In particular, the 2D panoramic X-ray image is an efficient way for dentists to determine invisible caries, impacted teeth and supernumerary teeth among children. Additionally, the 3D dental cone-beam computed tomography (CBCT) examination has been widely applied in orthodontics and endodontics due to its low ray quantity. To the best of our knowledge, there is no open-access 2D public dataset for children's teeth and no open 3D dental CBCT dataset, which limits the development of deep learning algorithms for segmenting teeth and automatically analyzing diseases.

The Semi-TeethSeg challenge will release more than 2,000 labelled panoramic X-ray images and 5,000 labelled CBCT slices to enable researchers to accurately segment teeth regions using deep-learning approaches. To encourage the study of tooth feature representation based on a large amount of raw dental data, we further release unlabelled 2D and 3D dental images including more than 1,000 panoramic X-ray images and 30,000 CBCT slices. Several robust semi-supervised-based tooth segmentation methods will be proposed via the Semi-TeethSeg challenge to facilitate the development of CAD-based dentistry.

### Challenge keywords

List the primary keywords that characterize the challenge.

Teeth segmentation, Panoramic X-ray image, Cone Beam Computed Tomography volume, Dentistry

### Year

The challenge will take place in ...

2023

## FURTHER INFORMATION FOR MICCAI ORGANIZERS

## **Workshop**

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

None for the moment

## **Duration**

How long does the challenge take?

Half day.

## **Expected number of participants**

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

We expect 100 participants

## **Publication and future plans**

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

We intend to coordinate with excellent participants a high-level journal article summarizing the main results and analysis obtained from this challenge. All participants can be authors of the article.

All participating teams can publish their own results independently after an embargo time of 6 months.

## **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 organized on [grand-challenge.org](http://grand-challenge.org) for publicity and own online submission site for final model submission.

## TASK: Semisupervised Teeth Segmentation

### 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.

CAD tools are increasingly popular in modern dental practice, particularly for treatment planning or comprehensive prognosis evaluation. In particular, the 2D panoramic X-ray image is an efficient way for dentists to determine invisible caries, impacted teeth and supernumerary teeth among children. Additionally, the 3D dental cone-beam computed tomography (CBCT) examination has been widely applied in orthodontics and endodontics due to its low ray quantity. To our knowledge, only three 2D panoramic tooth datasets are open-access in public, indicating that tooth image collection and labeling are not easy. Different from the teeth images of adults, children's teeth images in our dataset are the main innovation, which is not mentioned in other works. Up to now, we have labeled about 200 children's teeth images and 1000 adults' teeth images. The number of labeled children's teeth images will be increased in the following months as we will invite more dentists to mark the children's images.

In this challenge, we provide two modalities including the 2D panoramic tooth images and the 3D CBCT tooth volumes. These two modalities are evaluated separately with different metrics. For the 2D panoramic image dataset, we only rank all results by the 2D Dice metric. For the 3D CBCT volume dataset, we encourage 3D segmentation method rather than 2.5D or 2D slice-based methods.

The development of deep learning algorithms for the segmentation of teeth and automatic disease analysis still faces some challenges

- The current amount of data and annotation in open-source datasets is still insufficient, limiting the algorithm development.
- There is the problem that manual annotation cannot segment the boundaries of teeth, such as cross-over parts of teeth, which may be difficult to define.
- Effective methods for representing tooth features are critical and urgently needed, especially for semi-supervised tooth segmentation.

#### Keywords

List the primary keywords that characterize the task.

Teeth segmentation, Dental Panoramic radiographs, Cone Beam Computed Tomography volume, Dentistry, Semisupervised learning

### ORGANIZATION

#### Organizers

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

Organizing team:

Fan Ye, College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China.

Weiwei Cui, School of Electronic Engineering and Computer Science, Queen Mary University of London, the United Kingdom.

Xingru Huang, School of Electronic Engineering and Computer Science, Queen Mary University of London, United Kingdom.

Yaqi Wang, College of Media Engineering, Communication University of Zhejiang, China.

Shuai Wang, School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China.

Clinical evaluators and annotation approvers:

Yifan Zhang, Hangzhou Geriatric Stomatology Hospital, Hangzhou Dental Hospital Group, Division of Advanced Prosthetic Dentistry, Tohoku University Graduate School of Dentistry, State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, West China Hospital of Stomatology, Sichuan University, China.

Yilong Li, School of Electronic Engineering and Computer Science, Queen Mary University of London, the United Kingdom.

Data contributors:

Yifan Zhang, Hangzhou Geriatric Stomatology Hospital, Hangzhou Dental Hospital Group, Division of Advanced Prosthetic Dentistry, Tohoku University Graduate School of Dentistry, State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, West China Hospital of Stomatology, Sichuan University, China.

Liaoyuan Zeng, School of Information and Communication Engineering, the University of Electronic Science and Technology of China, China.

Sponser:

Yaqi Wang, College of Media Engineering, Communication University of Zhejiang, China.

Shuai Wang, School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China.

b) Provide information on the primary contact person.

Yaqi Wang, College of Media Engineering, Communication University of Zhejiang, China. Email: wangyaqi@cuz.edu.cn.

## Life cycle type

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

Examples:

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

One time event with fixed submission deadline.

## Challenge venue and platform

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

MICCAI.

b) Report the platform (e.g. [grand-challenge.org](http://grand-challenge.org)) used to run the challenge.

[grand-challenge.org](http://grand-challenge.org) for publicity and own on line submission site for final model submission.

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

We are creating a competition on [grantchallenge.org](http://grantchallenge.org) following the details of the modified proposal.

### **Participation policies**

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

**Fully automatic.**

**Additional points:** We intend to coordinate with excellent participants a high-level journal article summarizing the main results and analysis obtained from this challenge. All participants can be authors of the article.

**All participating teams can publish their own results independently after an embargo time of 6 months.**

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 policy defined.**

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.

**Prospective sponsorship from the College of Media Engineering of Communication University of Zhejiang.**

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 made public.**

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 intend to coordinate with excellent participants a high-level journal article summarizing the main results and analysis obtained from this challenge. All participants can be authors of the article.**

**All participating teams can publish their own results independently after an embargo time of 6 months.**

### **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.

**Submission instructions will be on the webpage. We will let the participants submit their Docker submissions to our evaluation server.**

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.

**Multiple submissions are allowed. However, only the last submission will be considered for the challenge results. Also, the number of Docker submissions will be limited to one per day.**

### **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

Registration is open from 1st March to 15th March 2023.

15th March 2023: Expected release of the first batch of training data.

15th April 2023: Expected release of the second batch of training data.

1st July 2023: Expected release of the public test set+ opening of the submission portal for the public test set.

15th July 2023: Submission of testset closes. Docker submission portal opens.

30th July 2023. Docker submission deadline.

Announcement of results at Workshop/MICCAI.

### **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).

**Ethical approval was obtained from Medical Ethics Committee of Sichuan Provincial People's Hospital and the University of Electronic Science and Technology Hospital Research Ethics Committee.**

### **Data usage agreement**

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

Examples:

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

CC BY NC ND.

### Code availability

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

Evaluation code will be made public before the system is open for submission.

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

Not required, however we encourage the participants to publish their code on github or the challenge platform .

### 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.

There are no conflicts of interest.

Yifan Zhang is the principal sponsor of the challenge by collecting and providing clinical data

Only the organisers and members of their immediate team have access to test case labels.

## 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

Diagnosis, Assistance, Screening, Prognosis, CAD, Research, Surgery, Education, Prevention.

Additional points: 2D/3D Segmentation: for dental panoramic X-rays, the overall segmentation masks of the corresponding tooth structure is given ; for each voxel in the Cone Beam Computed Tomography scan, give its label (tooth class or background class).

## Task category(ies)

State the task category(ies).

Examples:

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

**Segmentation.**

## Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

The target cohort is patients who need treatment for common dental diseases (caries, periapical periodontitis, pulpitis), orthodontic and endodontic restorative treatment.(patients requiring orthodontic, and endodontic prosthetic treatment).

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

Panoramic X-rays are used to view overall tooth structure, accurately segment teeth, and pinpoint dental disease foci, and 3D Cone Beam Computed Tomography scan for patient dentition requiring orthodontic, and endodontic prosthetic treatment.

## Imaging modality(ies)

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

Panoramic X-ray/3D Cone Beam Computed Tomography scan

## Context information

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

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

**No additional context information will be given.**

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



No additional context information will be given.

### 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.

Data is acquired for the intraoral cavity. For a given patient, one 2D/3D scans will be acquired covering full/part of the upper and lower jaws with teeth.

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

The target of the participating algorithms is all visible teeth in a given 2D or 3D scan.

### 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, Precision.

## DATA SETS

### Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

The pediatric dental data set was obtained from cases of patients who visited the Hangzhou Xiasha Dental Hospital from March to June 2022, as well as panoramic radiographs from instrument scans at the time of the visit and intraoral 62 scans and CBCT, for a total of 123 scans.

All the 3D CBCT scans were acquired with a OP300, manufactured by Instrumentarium Orthopantomograph®. CBCT slices were acquired in the DICOM format at the University of Electronic Science and Technology of China Hospital. All CBCT slices were scanned before dental operations, with a resolution of  $266 \times 266$  pixels in the axial view. The in-plane resolution is about  $0.25 \times 0.25$ mm<sup>2</sup> and the slice thickness range from 0.25 mm to 0.3 mm.

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

In general, no particular protocol is defined for the panoramic dental X-rays and 3D CBCT scans

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.

Panoramic radiographs were obtained from the Hangzhou Downtown Dental Hospital and international public data sets, and all data do not contain personal private information. 3D scans are originally acquired in partner dental clinics located in China. All acquired clinical data are anonymized.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Data are acquired by orthodontists with more than five years of professional experience. Scans were annotated by fifteen dentists. Twelve junior dentists with at least two years of experience manually marked all teeth regions.

### 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.

A total of 2885 dental panoramic radiographs were annotated by dentists. All 2D scans were divided into a training set as well as a test set according to the ratio of 7:3.

Any visible tooth in 22 CBCT volumes is annotated by medical experts. For all annotated CBCT volumes, we made a data split to get public training set (13 volumes, 60%), public testing set (2 volumes, 10%), and a hidden testing set used for participant ranking (7 volumes, 30%). Additionally, to facilitate 3D tooth feature representation, 100 unlabelled CBCT volumes is released for training.

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

For the 2D panoramic dataset, 2,200 dental panoramic radiographs were used for the training set, 500 for the test set, and 185 for the validation set. More unlabelled 2D scans will be collected and released on this challenge.

For the 3D CBCT images:

- Public training set 3D scans acquired for 100 unlabelled CBCT scans and 13 labelled CBCT scans.
- Public testing set 3D scans acquired for 2 labelled CBCT scans.
- Hidden testing set 3D scans acquired for 5 labelled CBCT scans

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

We collected and screened 193 panoramic dental radiographs of pediatric patients, pooled 2692 images from three international public adult dental datasets, and annotated all panoramic radiographs with the help of Efficient Interactive Segmentation (EISeg), an intelligent and efficient segmentation software.

Additionally, We collect and publish a 3D CBCT dataset called CTooth+. Our CTooth+ fully maintains the three-dimensional characteristics of teeth, and the number of data samples exceeds 30k slices, far exceeding the existing 2D dental datasets. The data set consists of 5504 annotated CBCT images of 22 patients and 25876 unlabeled images of 146 patients.

Semi-supervised learning (SSL) requires less expert annotations for model training, relieving the time and labour burden associated to data annotation. To our knowledge, there is no SSL-based tooth volume segmentation method published mainly due to data limitation. Therefore, we review the 2022 MICCAI TeethSeg challenge and split our dataset according to their proportion on training, validation and test cases.

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.

Data are collected for patient requiring either preoperative examination or prosthetic treatment.

All dental scans are categorized into four classes, including missing teeth with appliance, missing teeth without appliance, teeth with appliance and teeth without appliance. We will try to tend the teeth characteristics with an even distribution among all cases.

### 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.

Scans were annotated by 15 dentists. Twelve junior dentists with at least two years of experience manually marked all teeth regions. Three senior experts with at least ten years of experience were invited to evaluate the tooth annotations.

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.

Twelve junior dentists first delineate tooth regions slice-by-slice in the axial view. Then the annotations were modified according to the coronal view and sagittal view.

The senior dentists experts assessed the annotation quality, and marked a quality level (excellent, good, fail and poor) on each tooth annotation. "Excellent" annotations were stored in the CTooth+ dataset directly. "Good" annotations were fine-tuning according to the experts' feedback. "Fair" and "Poor" annotations and their feedback were put back into the unlabelled data pool and were marked again.

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.

Professional annotation software ITKSNAP (<http://www.itksnap.org/pmwiki/pmwiki.php>) is in charge of performing the annotations for twelve junior dentists.

Professional image editing software Adobe Photoshop (<https://www.adobe.com/products/photoshop.html>) is in charge of fine-tuning the "Good" annotations for senior dentists.

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.

No Aggregation

### **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.

No preprocessing steps are required.

### **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.

Scans were annotated by 15 dentists including twelve junior dentists and three senior dentists. As most of the annotation progress relies on semi-automated labelling software (ITK-SNAP and Adobe Photoshop), boundaries inaccuracy may exist on root regions. Additionally, the aim of task 2 is to predict accurate tooth volume regions with both labelled and unlabelled CBCT data, so we encourage participating teams to focus on how to use a large number of unlabelled data.

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

N/A

## **ASSESSMENT METHODS**

### **Metric(s)**

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

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

The segmentation inference results are evaluated using dice similarity coefficient (DSC), Hausdorff distance(HD), average symmetric surface distance (ASSD), surface overlap (SO) and surface dice (SD).

We will release the code on how to calculate the eight performance metrics with the python3 version.

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

These two modalities are evaluated separately with different metrics. The structure of this challenge is:

└ Task1: Semi-supervised 2D Teeth Segmentation on panoramic X-ray images

└ Ranking method: rank=sorted(2D\_Dice)

└ Task2: Semi-supervised 3D Teeth Segmentation on CBCT volumes

└ Ranking method: rank=sorted(3D\_Dice)+reverse\_sorted(SD).

It is noticeable that the in-plane spacing is  $0.25 \times 0.25\text{mm}^2$  as all the CBCT volumes were scanned from the same CBCT scanner. To calculate the distance-based metric, the participants can call the volume spacing by packages like pydicom.

### 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 global ranking will be based on an averaged sum of the eight metrics, including DSC, HD, ASSD, SO and SD.

To keep the balance of value-based metrics and distance-based metrics, we delete the HD and ASSD metrics from our final evaluation scheme.

For the 3D segmentation task, to consider the value-based and the distance-based metrics simultaneously, we learn from the “consensus ranking” and adjust the final evaluation scheme as:

rank1=sorted(Dice) (1)

rank2=reverse\_sorted(SD) (2)

final\_rank=rank1+rank2 (3)

For the value-based metric, we rank the Dice scores of all submissions with a 3D dice metrics. The larger the Dice score is, the higher the rank1 is. For the distance-based metric, we rank the 3D surface distance (SD) of all submissions. With the opposite to the dice score, the smaller the SD score is, the higher the rank2 is. Then the final rank is added by the rank1 and the rank2. If two submissions have the same final rank, we follow the priority: final rank>rank1>rank2>Dice>SD.

In addition, we will provide intermediate rankings based on each metric in order to highlight the performances of the competing methods in each of eight metrics. Winners will be the top global ranking methods. Although the challenge tasks are not expected to run in real time, we will add statistics on runtime performances for each participant algorithm as estimated on the challenge platform.

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

The evaluation system will return submit application back once the current submission does not include all the results.

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

We do not make a difference between all the evaluation metrics of tooth volume segmentation. Therefore, considering a simple weighting scheme would be easy to interpret and will lead to a fair ranking. In addition, we will provide intermediate rankings based on each metric in order to highlight the performances of the competing methods in each of the evaluation metrics.

### Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

The proposed statistical analysis to rank participants is a weighted mean. Ranking variability will be characterized using the bootstrap method. Missing data is handled as described above.

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

N/A

### 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.

There is no further analysis applicable that is not discussed above.

The ensuing journal article about the challenge will have a detailed analysis on inter-algorithm variability, algorithm-human variability, and an evaluation of the ensemble of algorithms

### 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.

The dataset of 3D dental scans is published in two conference papers[1],[2].

[1] Cui, W., Wang, Y., Zhang, Q., Zhou, H., Song, D., Zuo, X., Jia, G., & Zeng, L. (2022). CTooth: A Fully Annotated 3D Dataset and Benchmark for Tooth Volume Segmentation on Cone Beam Computed Tomography Images. ICIRA.

[2] Cui, W., Wang, Y., Li, Y., Song, D., Zuo, X., Wang, J., Zhang, Y., Zhou, H., Chong, B.S., Zeng, L., & Zhang, Q. (2022). CTooth+: A Large-scale Dental Cone Beam Computed Tomography Dataset and Benchmark for Tooth Volume Segmentation. DALI@MICCAI.