# Tooth Fairy: A Cone-Beam Computed Tomography Segmentation Challenge: Structured description of the challenge design

## Version 2.0 (after MICCAI2023 reviews)

## CHALLENGE ORGANIZATION

### Title

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

Tooth Fairy: A Cone-Beam Computed Tomography Segmentation Challenge

### Challenge acronym

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

ToothFairy

### 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.

Many recent works in dentistry and maxillofacial imagery focused on the Inferior Alveolar Nerve (IAN) canal detection, also thanks to the recently introduced Cone Beam Computerized Tomography (CBCT) that guarantees advantages w.r.t conventional CTs (e.g., lower radiation dose, lower costs and improved spatial resolution). The three-dimensional information acquired with CBCT can be crucial to plan a wide number of surgical interventions with the aim of preserving noble anatomical structures like the inferior alveolar canal, an osseous structure of the mandible which contains the homonymous nerve, artery and vein. Identifying the canal ensures its preservation in cases of impacted third molar extraction, implant positioning or removal of cystic lesions by preventing damages to dental or neural structures that would significantly reduce the quality of life. Artificial intelligence in general, and deep learning models in particular, can support medical personnel in surgical planning procedures by providing a voxel-level segmentation of the IAN, which is more accurate than bi-dimensional annotation commonly used in daily clinical practice.
Unfortunately, the small extent of available 3D maxillofacial datasets has strongly limited the performance of deep learning-based techniques. On the other hand, a huge amount of sparsely 2D-annotated data is produced daily in the maxillofacial practice, becoming the *de facto* standard in radiology medical centers for dentistry and maxillofacial purposes.
Although the amount of sparsely labeled images is significant, the adoption of those data still raises an open problem. The incomplete detection of nerve positioning is often sufficient to facilitate a positive outcome of surgical intervention, but it is not an accurate anatomical representation. Nevertheless, 2D annotations fail to identify a considerable amount of inner information about the IAN position and the bone structure. Additionally, deep learning approaches frame the presence of dense 3D annotations as a crucial factor, but the availability of such annotations is strongly limited by the exceptionally large amount of time required.

The challenge we propose aims at pushing the development of deep learning frameworks to segment the inferior alveolar nerve by incrementally extending the amount of publicly available 3D-annotated CBCT scans. Moreover, specific task for the segmentation of bones and teeth might be addressed in further editions of the challenge.

### Challenge keywords

List the primary keywords that characterize the challenge.

Inferior Alveolar Canal, Segmentation, 3D Volumes,

### 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.

Not Yet Defined/Available

### 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.

20

### Publication and future plans

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

We plan to coordinate a challenge results publication and open a special issue.

### 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 platform used to run the platform challenge will be grand-challenge.org. For the on-site event we will need a projector and microphones.

# TASK: IAC 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.

See challenge description.

### Keywords

List the primary keywords that characterize the task.

See challenge keywords.

## ORGANIZATION

### Organizers

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

Federico Bolelli (1), Luca Lumetti (1), Mattia Di Bartolomeo (1), Shankeeth Vinayahalingam (2), Alexandre Anesi (1), Bram van Ginneken (2), Costantino Grana (1)

(1) Università degli Studi di Modena e Reggio Emilia, Italy
(2) Radboud University Medical Center, The Netherlands

b) Provide information on the primary contact person.

Federico Bolelli, Università degli Studi di Modena e Reggio Emilia (federico.bolelli@unimore.it)

### 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 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) used to run the challenge.

grand-challenge.org

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

www.toothfairychallenge.eu

## Participation policies

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

Fully automatic.

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

Publicly available data is allowed.

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.

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

The challenge will award first three classified institutions/research group. For each institution/research group only the last submission will be considered in the final ranking.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

The top three performing methods will be announced publicly and they must share their code on a public repository (e.g. GitHub). If the submission authors will not publicly share the code with appropriate README for replicating the experiments, the prize(s) will shift to next classified institution(s).

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

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

All the members of the team qualify as authors of the challenge submission. A paper resuming challenge results and including description of each significant proposal will be published after the challenge: at most two authors for each team may be eligible as authors of this paper.

All the participants are allowed to submit their own results in any venue (conferences, workshops, etc) with embargo restriction: 6 months after the MICCAI2023 event.

We are planning to organize a special issue open (not only) to challenge participants for publishing and presenting their results.

## 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.

Since the test dataset will not be released, all the participants must submit a docker container or a GitHub repository, following the template that we provided at https://github.com/AImageLab-zip/ToothFairy. The submission instructions are detailed at https://toothfairy.grand-challenge.org/how-to-submit/.

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.

Teams can have multiple submissions on the preliminary test set (a subset of the entire test set composed by 3 volumes) to verify the quality of their methods and the compliance with submission policies. However, the maximum number of submissions per team in the preliminary test phase (1st July 2023 to 31st July 2023) is 15. If a team (from different users) makes more that 15 submissions in the preliminary test phase, it will be disqualified.

## 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

The challenge schedule is defined as follows:

- Release of training cases: 30th March 2023;
- Release of test cases: test cases will not be publicly released;
- Submission date(s) (Preliminary Phase): from 1st July 2023 to 31st July 2023;
- Submission date(s) (Final Phase): from 25th July 2023 to 31st July 2023.
- Associated workshop days: to be defined;
- Release date of the results: 1st September 2023.

Submissions performed between the 1st and the 31st of July 2023 will be evaluated on a subset of the test (3 volumes). This way, all the participants will have a feedback on the correctness of the submitted docker/code and access to a temporary leaderboard. Submissions performed in the Final Phase will produce the final ranking.

## 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).

Data composing the training set have already received the approval of the ethical committee and they can be

shared for research purposes only. More specifically, the ethical approval for this study was obtained from Comitato Etico dell'Area Vasta Emilia Nord (Approval Number 1374/2020/OSS/ESTMO SIRER ID 1275 - NAI-CBCT-D).

### 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-SA

Any publication using our data must explicitly reference to this challenge and cite [1, 2].

### 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 code used to evaluate the submissions is publicly available on GitHub at https://github.com/AImageLab-zip/ToothFairy

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

As previously mentioned, the top three performing methods must share their code on a public repository (e.g., GitHub). For all the others, the code publication is warmly suggested.

### 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.

We are in contact with different companies and institutions that are eligible to become sponsor of the challenge, including NVIDIA, Affidea, NewTom, and others. Member of the companies that will be sponsor of the challenge can participate with submission, but will not be eligible for prizes. As mentioned, the test set will not be publicly release, but we plan to release GT masks after the challenge.

## 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

Medical data management, Decision support, Surgery, Research, Treatment planning, Intervention planning.

## 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.

Final biomedical applications will acquire data using standard CBCT scanning protocol.

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

Target cohort includes patients older than 12 years who performed a CBCT for monitoring or surgery planning. Only CBCTs that were unreadable or where there was a massive alteration of the normal anatomy (mainly due to previous surgical resective procedures) have been excluded, thus allowing a wide variability close to a real clinical setting.

### Imaging modality(ies)

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

Any kind of image technique that can be used to automatically segment 3D volumes is accepted.

### Context information

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

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

Along with the images, the inferior alveolar canal segmentation ground truth will be provided in the form of image, i.e., a 3D volume having the same size of the image data with 0 values corresponding to background (non-IAN voxels) and 1s corresponding to IAN-voxels .

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

No patient clinical data is provided.

### 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.

Lower jawbone of the patients, acquired by means of Cone Beam Computer Tomography.

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

The algorithm must target the Inferior Alveolar Canal present in the lower jawbone, by providing a binary volume where the voxels that are part of the IAN are identified with 1s and the others are marked with 0s.

### Assessment aim(s)

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

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

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

The aim is to provide a reliable tool that can be integrated in the daily clinical practice, in particular in surgical planning and execution. A correct and three-dimensional detection of the IAC is mandatory in mandibular surgeries and will significantly improve the bi-dimensional annotation performed by radiology technicians so far.

## 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).

All the data of the training set has been obtained using a NewTom/NTVGiMK4, 3 mA, 110 kV, 0.3 mm cubic voxels. Instead, test data are obtained using a standard CBCT scanning protocol (i-CAT, 3D Imaging System, Imaging Sciences International Inc, Hatfield, PA, USA) in "Extended Field" modus (FOV: 16 cm diameter/22 cm height; scan time: 2 × 20s; voxel size: 0.4 mm)

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

Pixel spacing and intra-slice distance is always 0.3 millimeters. Data volumes are already converted to the Hounsfield Unit (HU) and their values range between -1000 and 5264. Volume shapes range from (148, 272, 334) to (171, 423, 462) for the Z, Y and X axes respectively.

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.

The 3D CBCT volumes composing training dataset have been acquired by the Affidea center located in Modena, Italy. Affidea is a leading pan-European healthcare group specialized in the provision of advanced diagnostics, specialist outpatient services, laboratory analyses, physiotherapy and rehabilitation, cancer diagnosis and treatment. It counts 312 different centers in 15 different countries, with about 11 000 professionals. The test set is from the department of oral and maxillofacial surgery in Radboud University Nijmegen Medical Centre.

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).

Both data acquisition and data annotation tasks were performed by medical experts with many (>5) years of experience.

## Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Both training and test cases are CBCT volumes of the lower jawbone. Each case is associated with a 3D "sparse" annotation where the canal path is identified by a 1-voxel-thin line. Additional, a subset of these CBCTs is accompanied with a dense annotations, i.e., binary volumes where all the voxels that are part of the IAN are identified with 1s and the others are marked with 0s.

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

The dataset is composed by 443 volumes of training and 50 volumes reserved as test set. All the 443 volumes will be provided with the sparse annotation. Among them, only 153 also have a "complete" dense annotation. Each challenge participant can decide how to split training data to create the internal validation set.

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

The test set comprises a number of cases that is (more than) enough to evaluate the performance of algorithms under different perspectives. Having more (and more) data as training set is always helpful, but expensive and time consuming (extremely so). The current amount of training data selected for the challenge represents the biggest training set available in literature at the moment of writing this proposal.

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.

As mentioned, the test will not be released for ethical issues. We chose to use set of data coming from different institutions to keep the training set as big as possible. When selecting test cases we ensured that the clinical distributions (i.e., age, sex, etc) and the annotators are aligned with those of the training set.

### 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.

All the annotations have been obtained by means of a specifically developed software that allows experts to annotated the inferior alveolar canal in a 3D manner. Everything has been described in [1] and [3] and the software is open-source. Annotators are doctors with more that 5 years of experience. The total number of annotators involved is 5.

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 annotators are medical experts and the annotation tool has been developed following their instructions/requirements thus maximizing the annotation accuracy. For this reason, no instructions are needed for the annotators.

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.

Every case has been annotated by medical experts with years (>5) of experience in the maxillofacial field.

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 multiple annotations are available in our scenario, meaning that all CBCT volumes are annotated by a single expert only.

### Data pre-processing method(s)

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

The data pre-processing pipeline is described in [1]. In short, the ground truth is generated from medical experts annotations by applying the Delaunay triangulation followed by computing the alpha-shape to smooth the obtained mesh.

### 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.

Volumes have been manually annotated by medical experts, but no multiple annotations are available for the same CBCT.

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

No other relevant source of error has been identified.

## 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 metric used to rank the submitted proposals is the Dice Similarity Coefficient (DSC). The 95th percentile Hausdorff Distance (HD95) will be also included. Please refer to the ranking method(s) section for more details.

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

Together with IoU (Intersection over Union), DSC represents the goal standard for the evaluation of segmentation algorithms. Since their meaning is practically the same, there is no reason to preserve both the metrics.

### 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 schema involves the following steps:

1) Calculate the Dice score (average on all volumes), the HD95 (average on all volumes), the maximum used memory (Mem), and the total execution time (Time) for all cases;
2) Rank the Dice, HD95, maximum used memory, and running time, independently;
3) Average the rankings obtained at point 2 for Dice and HD95 to produce the final rank.
4) If two or more final ranks obtained at point 3 are equal, compute the average of the rankings obtained at point 2 for Mem and Time to break ties.
5) If two or more ranks are still equal, it's a tie: the prize will be evenly split.

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

If an algorithm does not produce any result for a specific CBCT volume we will consider the predicted segmentation as a volume of 0s, meaning that the Dice score for that image will be equal to 0.

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

There is no interdependence between cases. All of CBCTs (both in train and test) belong to a different patient and have been annotated by only one expert. The selected metrics will provide homogeneous numbers on different patients that can be later averaged to provide the final rank. Different metrics (i.e. DICE and Hausdorff distance) calculated separately are later aggregated to produce the final ranking.

### 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.

No statistical analyses methods have been used in the scope of this challenge.

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.

Challenge participants are allowed to use ensemble methods.

## 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.

[1] Cipriano, M., Allegretti, S., Bolelli, F., Di Bartolomeo, M., Pollastri, F., Pellacani,A., Minafra, P., Anesi, A., Grana, C.: Deep Segmentation of the Mandibular Canal: a New 3D Annotated Dataset of CBCT Volumes. IEEE Access10, 11500–11510(2022). https://doi.org/10.1109/ACCESS.2022.31448402.

[2] Cipriano, M., Allegretti, S., Bolelli, F., Pollastri, F., Grana, C.: Improving Segmentation of the Inferior Alveolar Nerve through Deep Label Propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 21137–21146. IEEE (Jun 2022)

[3] Mercadante, C., Cipriano, M., Bolelli, F., Pollastri, F., Anesi, A., Grana, C., et al.: A cone beam computed tomography annotation tool for automatic detection of the inferior alveolar nerve canal. In: 16th International Conference on Computer Vision Theory and Applications-VISAPP 2021. vol. 4, pp. 724–731. SciTePress (2021)