Segmenting the Inferior Alveolar Canal in CBCTs Volumes: The ToothFairy Challenge

Federico Bolelli, *Associate Member, IEEE*, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani, Kevin Marchesini, Niels van Nistelrooij, Pieter van Lierop, Tong Xi, Yusheng Liu, Rui Xin, Tao Yang, Lisheng Wang, Haoshen Wang, Chenfan Xu, Zhiming Cui, Marek Wodzinski, *Member, IEEE*, Henning Müller, *Member, IEEE*, Yannick Kirchhoff, Maximilian R. Rokuss, Klaus Maier-Hein, Jaehwan Han, Wan Kim, Hong-Gi Ahn, Tomasz Szczepański, Michal K. Grzeszczyk, Przemyslaw Korzeniowski, Vicent Caselles-Ballester, Xavier Paolo Burgos-Artizzu, Ferran Prados Carrasco, Stefaan Berge', Bram van Ginneken, Alexandre Anesi, and Costantino Grana, *Member, IEEE*

***Abstract*— In recent years, several algorithms have been developed for the segmentation of the Inferior Alveolar Canal (IAC) in Cone-Beam Computed Tomography (CBCT) scans. However, the availability of public datasets in this domain is limited, resulting in a lack of comparative evaluation studies on a common benchmark. To address this scientific gap and encourage deep learning research in the field, the ToothFairy challenge was organized within the MICCAI 2023 conference. In this context, a public dataset was released to also serve as a benchmark for future research. The dataset comprises 443 CBCT scans, with voxel-level annotations of the IAC available for 153 of them, making it the largest publicly available dataset of its kind. The participants of the challenge were tasked with developing an algorithm to accurately identify the IAC using the 2D and 3D-annotated scans. This paper presents the details of the challenge and the contributions made by the most promising methods proposed by the participants. It represents the first comprehensive comparative evaluation of IAC segmentation methods on a common benchmark dataset, providing insights into the current state-of-the-art algorithms and outlining future research directions. Furthermore, to ensure reproducibility and promote future developments, an open-source repository that collects the implementations of the best submissions was released.**

***Index Terms*— Segmentation, tooth, neural network, X-ray imaging, computed tomography.**

Received 10 September 2024; revised 6 November 2024; accepted 20 December 2024. Date of publication 25 December 2024; date of current version 3 April 2025. This work was supported in part by the University of Modena and Reggio Emilia and Fondazione di Modena, through the Funds of Fondo di Ateneo per la Ricerca 2023 under Grant FAR 2023 and Grant FARD 2023; and in part by the Italian Ministry of Research, under the complementary actions to the National Recovery and Resilience Plan (NRRP) "Fit4MedRob—Fit for Medical Robotics" under Grant PNC0000007. The work of Marek Wodzinski was supported by Poland's High-Performance Computing (HPC) Infrastructure PLGrid (HPC Centers: Academic Computer Centre (ACC) Cyfronet AGH) for providing computer facilities and support within computational under Grant PLG/2023/016239. The work of Shankeeth Vinayahalingam was supported by the Radboud AI for Health collaboration between Radboud University and Radboudumc. The work of Niels van Nistelrooij was supported by the Berlin Institute of Health in Cooperation with Charité-Universitätsmedizin Berlin. The work of Tomasz Szczepański, Michal K. Grzeszczyk, and Przemyslaw Korzeniowski was supported in part by the Minister of Science and Higher Education "Support for the Activity of Centers of Excellence established in Poland under Horizon 2020" under Contract MEiN/2023/DIR/3796; in part by the European Union's Horizon 2020 Research and Innovation Program under Grant 857533; in part by the Sano Project carried out within the International Research Agendas Program of the Foundation for Polish Science, funded by the European Union under the European Regional Development Fund; and in part by the Sano Centre for Computational Medicine, Health Informatic Group Team, Nawojki, Kraków, Poland (https://sano.science/). *(Alexandre Anesi and Costantino Grana contributed equally to this work.)* *(Corresponding author: Federico Bolelli.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Comitato Etico dell'Area Vasta Emilia Nord under Approval No. 1374/2020/OSS/ESTMO SIRER ID 1275-NAICBCT-D.

Please see the Acknowledgment section of this article for the author affiliations.

Digital Object Identifier 10.1109/TMI.2024.3523096

I. Introduction

THE use of Cone Beam Computed Tomography (CBCT) is a standard procedure for the diagnostic assessment of the maxillofacial complex. Compared to an OrthoPantomoGraphy (OPG), a CBCT exposes the patient to slightly higher radiations, but also provides three-dimensional (3D) information that is missing in the two-dimensional (2D) OPG [1]. Other advantages of the CBCT are fast data acquisition, lower radiation exposure compared to conventional CT scans, good resolution of high-density regions (i.e., bony and dental tissues), and finally, the low cost of this technology.

The availability of CBCT scans allows for better identification of some anatomical structures, which is crucial for surgical planning since their preservation increases the post-operative quality of life [2], [3]. Among them, the Inferior Alveolar Canal (IAC) is a bony structure that can be visualized on CBCT scans of the jawbone. It runs along the mandibular bone from the mandibular foramen to the mental foramen. Its content is represented by the inferior alveolar neurovascular bundle, which includes the Inferior Alveolar Nerve (IAN), artery, and vein [4].

The IAN provides sensitivity to the homolateral lower cheek, lip, chin, teeth, gums, and veins [5]. Its preservation is mandatory during various mandibular surgical procedures, such as extraction of impacted teeth, Open Reduction and Internal Fixation (ORIF) of mandibular fractures, orthognathic procedures, placing of dental implants, removal of benign tumors and cysts, reduction of the mandibular height, and

Fig. 1. ToothFairy challenge logo.

preprosthetic procedures [2], [3], [6], [7], [8], [9]. CBCTs can help for a better 3D identification of the mandibular canal, allowing customization of the abovementioned procedures. Nevertheless, a manual and accurate voxel-based segmentation of the IAC is time-consuming, and only a small amount of publicly available data exists. Automatic segmentation of the IAC has the potential to facilitate the work of surgeons, especially when combined with other tools, such as tooth segmentation algorithms and orthognathic surgery planning software.

A. State of the Art Datasets for the Automatic Segmentation of the IAC

The training of neural networks for segmentation usually requires large datasets that must be accurately annotated. In the context of IAC segmentation, the most common type of annotation performed in the daily medical routine is a sparse one (2D), due to its fast execution time. However, this annotation type hide important details about the morphology of the mandibular canal into the bone and IAN position. From a clinical perspective, annotations performed on 2D isolated slices lack depth and spatial context, making it possible to identify the overall flow of the IAC but affecting related measurements, e.g., the exact distance between teeth roots and the canal itself. Moreover, when dealing with 2D images, cases where the boundaries between structures appear blurred or smeared are common, especially when objects lie partially within the imaging slice thickness. Such an effect leads to inaccurate delineation of structures, which can compromise the precision needed for treatment, leading to suboptimal plans or outcomes. 3D annotations would instead allow us to fully exploit the capabilities offered by 3D networks, mitigating the previously outlined limitations.

Although 2D annotations of medical images are easily accessible, cheap to obtain, and regularly produced by specialized centers through daily practice, there is almost no publicly available dataset in the literature. As a result, deep learning studies applied to medical imaging, particularly in the maxillofacial domain, often rely on privately held internal datasets. As an example, the work of Jaskari et al. [10] made use of a dataset with 637 scans sparsely annotated by two medical professionals. Such data is not publicly available, and its use is restricted by Finnish law and the General Data Protection Regulation (EU). Training and test data unavailability applies to many other recent proposals on the field [11], [12] and represents a major flaw in computer science research. Indeed, the practice of adopting only private datasets leads to a major information gap in the research community, preventing the possibility of fairly replicating experiments and effectively comparing different technical proposals.

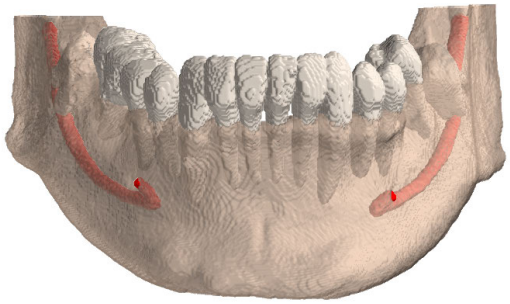


Fig. 2. CBCT of the lower jawbone with IAC marked in red.

B. The ToothFairy Challenge and Dataset

This paper aims at introducing the *ToothFairy* challenge (Fig. 1) along with the dataset that was released as part of it. The challenge, hosted by the MICCAI 2023 conference, is designed to drive advancements in deep learning segmentation networks for IAC detection by setting a common benchmark setup that should be taken as a reference for further proposal. Challenge participants were tasked with developing an algorithm that precisely identifies the IAC within the lower jawbone, using both 2D- and 3D-annotated CBCT scans. The algorithm's objective was to (completely automatically) generate a binary volume where voxels corresponding to the IAC were labeled as 1s (Fig. 2). The final goal of the challenge was to find accurate and comprehensive three-dimensional detection of the IAC to create tools that integrate into daily clinical practice, especially aiding in surgical planning and execution.

To address the ToothFairy challenge, the homonymous dataset was developed: a new public maxillofacial dataset where the IAC was annotated by medical experts at the voxel level. Actually, the dataset improves upon and extends the Maxillo dataset, previously released by Cipriano et al. [13], and it is the largest dataset with 3D annotations of the IAC that is publicly available to the scientific community.

The key contributions of this article to the scientific community can be summarized as follows:

- Detailed description of the *ToothFairy* challenge, highlighting the rationale behind the chosen training and test schema and the adopted evaluation metrics;
- Introduction of a new private dataset with 3D annotations of the inferior alveolar canal. Such a dataset is accessible through Grand Challenge[1] platform and sets a common benchmark to allow for a fair comparison of future proposals;
- Description of the most promising methods for IAC segmentation, proposed by the participants of the ToothFairy challenge, together with an in-depth performance analysis highlighting their strengths and weaknesses, outlining a path for further research in the field;
- An open-source repository that collects the challenge evaluation software and the implementations of the

[1] https://toothfairy.grand-challenge.org/

submitted models for the IAC segmentation is released together with the pre-trained network weights, ensuring reproducibility.[2]

The rest of the paper is organized as follows: Sec. II describes the state-of-the-art algorithms for the segmentation of the inferior alveolar nerve. In Sec. III, the details about the dataset and its annotation process are described. Sec. IV includes a summary of the segmentation methods proposed by the participants of the challenge and a more detailed description of the best-performing methods. In Sec. V the evaluation protocol is detailed; the results are reported in Sec. VI and discussed in Sec. VII.

Finally, in Sec. VIII and Sec. IX, future research directions are delineated and conclusions are drawn.

II. Related Work

Since the introduction of CBCT scan technology in the early 2000s, the scientific community spent much effort on creating automatic systems for the segmentation of the IAC from 3D volumes acquired employing such image modality.

A. Classical Computer Vision Techniques

The first approaches were based on classical computer vision techniques. Kainmueller et al. [16] elaborated a method based on the Statistical Shape Model (SSM) of the nerve and the bone, optimizing the prediction with a Dijkstra-based procedure. Another similar method was proposed by Abdolali et al. [17], who introduced a pre-processing phase before the statistical model, relying on *low-rank decomposition*, and substituting the tracing algorithm with a *fast marching* to determine the optimal path between the mandibular and mental foramen. One last relevant example is the work published by Wang et al. [18], based on the multi-plane and curved surface reconstruction to generate a multi-planar image set. The resulting images were grouped using a K-means clustering algorithm on the texture features, only considering the area of interest. This approach allowed an enhancement of the contrast in the IAN canal image. Then, the edges of the mandibular canal were identified using a 2D line-tracking technique, and these results were refined by applying a fourth-order polynomial to obtain the final segmentation. However, these approaches have limitations because they rely on the annotation of the mandibular bone in the training set, which requires an extra manual effort.

B. Deep Learning Models for IAC Segmentation

Thanks to the spread of Deep Learning in the field of medical diagnosis, segmentation methods using Deep Neural Networks (DNNs) have been proposed in recent years to solve the problem of IAN localization. One of the pioneering approaches is that of Jaskari et al. [10], who leveraged the U-Net architecture [19], trained on a coarsely annotated dataset. Their approach, tested on 15 volumes with voxel-level annotations, achieved better results than traditional computer vision methods. However, it still has limitations as it does not use densely annotated images for training. In the same year, Kwak et al. [11] compared 2D SegNet, 2D U-Net, and 3D U-Net trained on a proprietary dataset of images labeled at a cross-sectional level with annotations provided at intervals of 1 mm. Unfortunately, a direct comparison with their work is unfeasible, as neither the dataset nor the source code are available. Moreover, the authors employed arguable evaluation metrics, making the reported results of minor scientific relevance.

Another approach is the one by Lahoud et al. [12], who trained a 3D U-Net first on a coarsely annotated dataset, obtained by interpolating some control points and imposing a fixed uniform diameter of 2.50 mm, and afterward fine-tuned the model feeding it with 126 voxel-level-annotated CBCTs. Again, both the training and test data, as well as the source code of the proposed algorithmic solution, are privately stored, making any fair comparison impossible.

One important milestone regarding the IAC segmentation was set by the release of the first publicly available CBCT dataset, including both 2D and 3D medical annotations of the mandibular canal [13], [20]. Alongside the release of the dataset, the authors proposed a deep learning model for the 3D segmentation of the IAC, named PosPadUNet3D, which is a modified version of the 3D U-Net. Since the volumes can not be entirely fed into the network during training, they are divided into adjacent sub-volumes, and the position information is exploited by means of a positional embedding attached to the encoder's output. The segmentation pipeline based on the PosPadUNet3D model consists of three different steps. In the first stage, called "deep label expansion," the images annotated with both 2D and 3D labels are used to supervise a deep label propagation neural network trained to generate 3D labels from sparse 2D labels. Second, the aforementioned network is employed to generate synthetic 3D annotations for the volumes provided only with 2D labels. Finally, the synthetic dataset is employed to pre-train the segmentation CNN, which is further fine-tuned through the voxel-level annotations made by medical experts. This process has been shown to substantially improve the final segmentation of the IAC on the CBCT volumes.

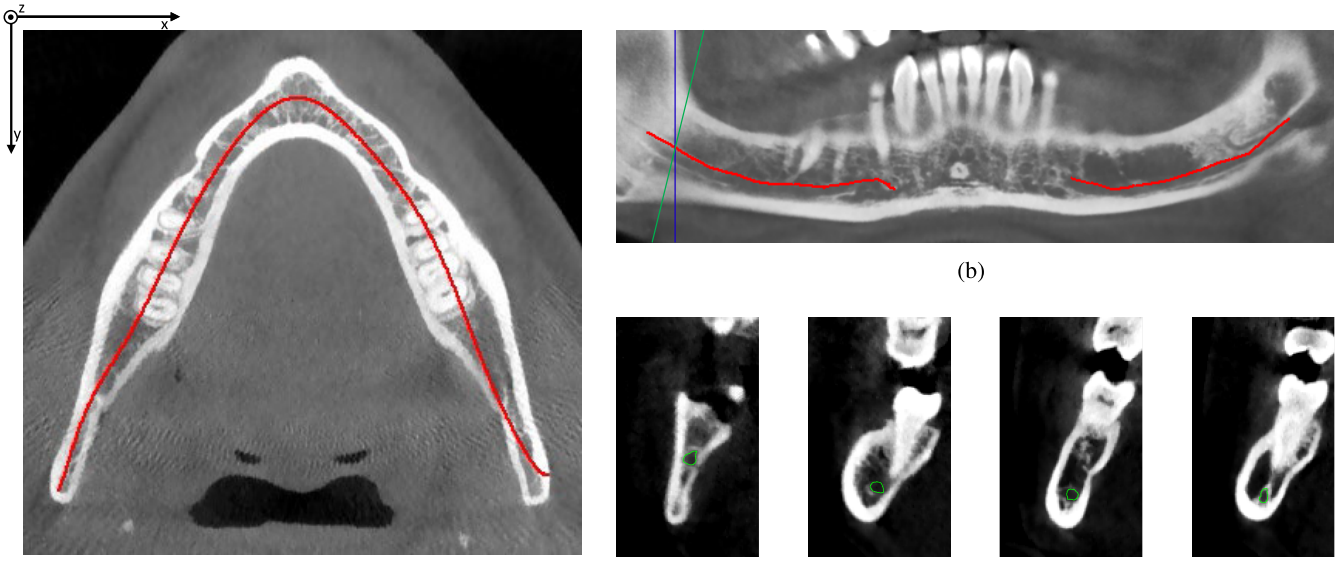
Usman et al. [14] proposed a two-stage approach also based on the U-Net architecture. Their methodology was formulated on the hypothesis that the predominant challenge in segmenting the inferior alveolar canal relates to the class imbalance between the mandibular canal and the background. To address this challenge, they initially applied a CNN to identify and isolate volume regions where the canal is likely to be located (Regions of Interest, ROIs), thereby substantially reducing background interference. Subsequently, in the second phase, they leverage U-Net architecture to perform the segmentation of the mandibular canal exclusively within the ROIs.

Similarly to [14], Zhao et al. [21] devised an algorithm based on a two-stage approach. The authors proposed a whole mandibular canal segmentation using transformed dental CBCT volumes in the Frenet frame. They first extracted the mandibular centerline via automatic segmentation of the mandible and localization of both the mandibular foramen and mental foramen. The sub-volumes containing the mandibular canal information were then obtained using a double reflection

[2]https://github.com/AImageLab-zip/ToothFairy

TABLE I
COMPARISON BETWEEN THE DATASETS USED IN LITERATURE FOR THE SEGMENTATION OF THE ALVEOLAR CANAL

Dataset	Year	Country	# Train (# Validation)	# Test	Labels		Public Availability
					Train	Test	
Jaskari et al. [10]	2020	Finland	457 (52)	128	2D	128 2D, 15 3D	✗
Lahoud et al. [12]	2022	Belgium	166 (39)	30	3D	3D	✗
Usman et al. [14]	2022	South Korea	510	500	3D	3D	✗
Chun et al. [15]	2023	South Korea	32	18	3D	3D	✗
Cipriano et al. [13]	2022	Italy	324 (8)	15	332 2D, 76 3D	15 2D, 15 3D	✓
ToothFairy	2023	**Italy, Netherlands**	**435 (8)**	**50**	**443 2D, 153 3D**	**3D**	✓



(a) (b) (c) (d) (e) (f)

Fig. 3. Exemplification of the annotation procedure. (a) is an axial slice extracted from the original volume. The red line crossing the mandible is called *panoramic base curve* and is used to generate the *panoramic view* (b), also identified as *panorex*. The panorex allows to "easily" determine the position of the IAC, by marking its upper boundary, the red line of (b). The blue line in (b) depicts the plane orthogonal to the panoramic base curve of (a) at a given point. Instead, the green line of (b) represents the plane orthogonal to the IAC curve (depicted in the same figure) at a given point. (c), (d), (e), and (f) are examples of *cross-sectional views*, i.e., images obtained interpolating voxels of the original volume laying on the plane identified by the green line in (b). The mandibular canal is here highlighted with green circles.

method based on the Frenet frame. The transformed sub-volumes were fed into the 3D segmentation network (again, U-Net based), and the clDice loss was used to constrain the topology of the mandibular canal. Last, the prediction masks were inversely transformed back into the original CBCT images to obtain final segmentations.

Another verifiable approach (i.e., tested on public data and/or providing source code) was contributed by Lv et al. [22], and leveraged a Transformer-based architecture paired with the clDice loss. The authors also employed adaptive histogram equalization to enhance input image contrast, image cropping to isolate the mandibular foramen, deep residual convolutions to amplify the model's sensitivity to fine details, and proposed a "deep label fusion" technique to gather additional information from the sparse labels available in the public datasets. The proposed "deep label fusion" resembles the "deep label expansion" proposed by Cipriano et al. [13], although with different details.

Recently, to overcome limitations related to patch-based learning, Lumetti et al. [23] proposed a memory-augmented Transformer encoder that effectively harnesses absolute spatial coordinates in the learning process. Such a U-Net-based model capitalizes on the inherent capacity of Transformers to model interactions between all pairs of elements within a given sequence, with the aim of enhancing the flow of information among the elements of the U-Net bottleneck, thus increasing contextualization. The authors leveraged such a flow of information to effectively inject contextual information related to the processed patches, i.e., their position within the entire volume, mitigating the patch-based learning-related issues.

The segmentation of the IAC is still an open problem with many development opportunities, but the improvements are closely related to the release of new publicly available datasets, possibly with highly detailed 3D labeling. Tab. I presents an overview of the datasets used in the aforementioned studies.

III. DATASET

This section presents the *ToothFairy* dataset that accompanies the challenge, providing a comprehensive overview of the data acquisition process and the annotation steps.

A. Data Sources and Demographics

The *training data* was acquired by the Affidea Center, a pan-European healthcare group located in Modena that specializes in advanced diagnostics, laboratory analyses, rehabilitation,

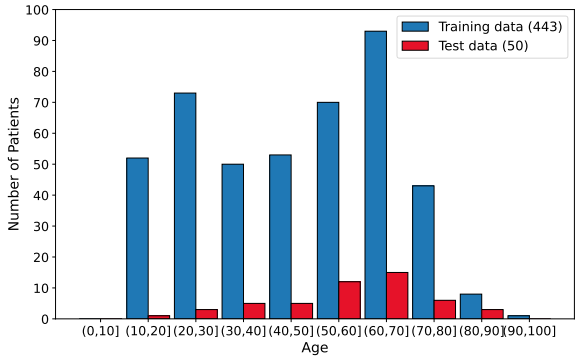


Fig. 4. Demographic information of patients included in the ToothFairy dataset.

and cancer diagnosis and treatment. The scans were obtained by means of Cone Beam Computed Tomography (CBCT) via a *NewTom/NTVGiMK4*, operating at 3 mA, 110 kV, and offering 0.3 mm cubic voxels. The *test data* originates from the Department of Oral and Maxillofacial Surgery at Radboud University Medical Centre, Nijmegen, obtained using a standard CBCT scanning protocol with the *i-CAT 3D Imaging System*. The scans were performed in "Extended Field" mode, with a field of view (FOV) measuring 16 cm in diameter and 22 cm in height. The scanning process involved two scans of 20 seconds each, resulting in a voxel size of 0.4 mm.

Every patient was properly anonymized, but we have access to a few personal details —namely gender, age, and year of the scan. Specifically, 59.02% of the patients are female (59.82% in the training set, 52.00% in the test set), all the scans were performed between 2019 and 2020, and volumes belong to patients with ages in range (10-100] with the highest frequencies in ranges (20-30] and (60-70] for the training set and (50-70] for what concerns the test set (Fig. 4).

B. Annotation Protocol and Tools

In recent years, 2D annotation has become the standard in dento-maxillofacial radiological images. This technique involves marking the canal's position on a panoramic view of the dental arch (Fig. 3b), which is derived from an axial slice of the CBCT volume (Fig. 3a). Then, the resulting annotation is mapped back to the original 3D volume, obtaining what, from this point forward, will be referred to as the *sparse annotation*. This approach was widely adopted by medical experts due to its speed and simplicity. However, in many cases, such as IAC segmentation, 2D annotations lack significant internal information about the bone structure and the canal position. The inclusion of densely labeled 3D data is necessary in order to achieve the full potential of CNNs, as stated in [24]. 3D *dense annotation* refers to the annotation process carried out by medical professionals directly at the voxel level on CBCT scans. This annotation approach involves detailed markings applied to the individual voxels, providing a more precise depiction of the IAC. However, the acquisition of large amounts of these highly detailed annotations is extremely difficult and time-consuming. For this reason, the Inferior Alveolar Canal Annotation Tool, IACAT in short, was designed to support and assist medical experts in the annotation task [25], [26].

The 3D annotation process involves approximating the alveolar canal course with a one-pixel thick curve used to generate a Catmull-Rom spline (red line of Fig. 3b). This tool leverages spline points to produce views that are orthogonal to the canal, referred to as Cross-Sectional Views, or CSVs in short (Fig. 3c - Fig. 3f). Within each CSV, annotators must draw a closed Catmull-Rom spline at the location of the IAC. Eventually, the coordinates of the control points on these splines are employed to produce the ultimate ground-truth volume through the application of the α-shape algorithm [27]. The tool supports experts throughout each annotation step, in particular by automatically generating the initial approximation curve (manually adjustable when required). The updated version of the tool [26] also incorporates the PosPadUNet3D segmentation model proposed in [20] to assist in the annotation of CSVs. Moreover, it employs localized contrast stretching to enhance the visibility of darker regions, unveiling portions of the IAC that might remain concealed even to medical experts.

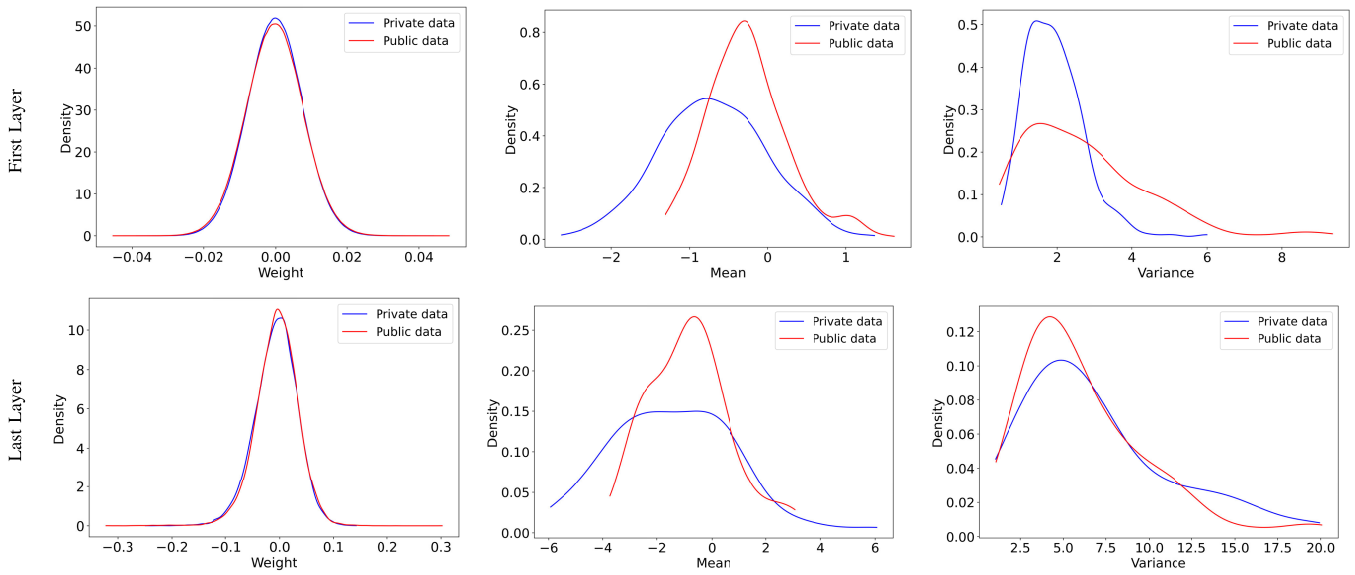
TABLE II
SUMMARY DESCRIPTION OF THE MAXILLO AND TOOTHFAIRY DATASETS. THE PRIMARY DATASET INCLUDES VOLUMES WITH BOTH DENSE AND SPARSE ANNOTATIONS, WHILE THE SECONDARY DATASET CONTAINS ONLY SPARSELY ANNOTATED VOLUMES. REGARDING THE CHALLENGE DATASET (TOOTHFAIRY), 91 DENSELY ANNOTATED VOLUMES ARE SHARED WITH THE MAXILLO DATASET (40 UNDERWENT RE-SEGMENTATION THROUGH IACAT 2.0, WHILE THE ANNOTATIONS FOR THE OTHER 51 REMAINED UNALTERED), AND 62 VOLUMES WERE DENSELY ANNOTATED FROM SCRATCH USING IACAT 2.0

Field	Dataset	
	Maxillo	ToothFairy
Primary dataset	91	153
Secondary dataset	256	290
File format	Numpy, DICOM	Numpy, DICOM
Values scale	Hounsfield	Hounsfield
Max volume shape	[171, 423, 462]	[178, 423, 463]
Min volume shape	[148, 272, 334]	[148, 265, 312]
Avg volume shape	[169, 337, 381]	[169, 342, 370]

The annotation process for the ToothFairy dataset engaged five medical experts with more than five years of experience. This process resulted in a total of 493 annotated volumes, 443 designated for the training set and 50 volumes reserved for the test set. Within the training set, all 443 volumes are equipped with sparse annotations, and among these, only 153 additionally feature 3D dense annotations. Participants in the challenge could choose whether and how to split the training data to create a validation set. Furthermore, they were allowed to use external data for the training, but none of them did so. Regarding the test set, all 50 volumes are densely annotated, an essential requirement for conducting a meaningful evaluation of the algorithms. Such a dense annotation ensures that the evaluation accurately reflects the performance of the models in capturing complex 3D structures.

C. Ground-Truth Validation and Inter-Agreements

As each volume was labeled by a single medical expert only, annotations were reviewed and validated by at least another expert. If any inaccuracy was identified, an agreement between the two annotators was reached to provide a more accurate mask than a single ground truth.



(a) Weights (b) Running Mean (c) Running Variance

Fig. 5. Parameter distributions of convolutional and instance normalization layers on public/train (blue) and private/test (red) ToothFairy datasets. The 3D nnU-Net has been trained on each dataset separately. The distributions were sampled at two different depths of the network: the first encoder layer (top row), and the last encoder layer (bottom row).

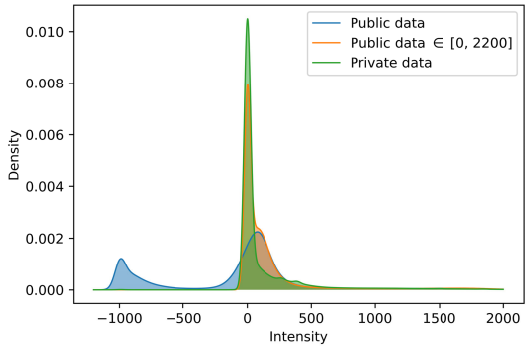


Fig. 6. Intensity distributions of scans from the public and private data. The distributions were estimated using Gaussian kernel density estimation (KDE) and evaluated for intensity values between -1,200 and 2,000. The intensity distribution of the private data was shifted compared to the public data, partly because most private scans did not have negative intensity values. Clipping the intensity values of the public scans between 0 and 2,200 decreases the shift in distribution.

Moreover, to assess the inter-agreement of the annotators involved, two different medical experts independently labeled the 15 cases in the public test set,[3] and an average Dice score of 0.81 was recorded. Such a number is useful when interpreting the downstream performance of a method trained on the ToothFairy dataset, challenge methods included.

D. Error Sources Related to the Annotation

During the analysis of the annotated data, it was observed that the annotations of a few patients exhibit disconnection. This can be attributed to a lack of density difference in certain areas of the jawbone. Such disconnections may arise due to factors like CBCT acquisition noise or patient-specific conditions. In a few cases, this posed extreme challenges or rendered it impossible to provide a comprehensive 3D annotation, resulting in the presence of holes in the ground-truth masks.

However, all the sparse labels that were annotated using a different approach are complete and do not contain any missing parts. In the default split that we provide with the dataset, it is ensured that all ground truth samples in the *public test set*[3] exhibit exactly two connected components, representing one canal each.

E. Maxillo and ToothFairy Datasets

The *ToothFairy* dataset is actually an extension of the previously released Maxillo dataset by Cipriano et al. [13]. The details of both datasets are reported in Tab. II. The Maxillo dataset contains 343 CBCTs, with 3D annotations provided for only 91 of them, created using IACAT. As a result, to create the challenge dataset 62 volumes were densely annotated from scratch using IACAT 2.0. Of the 91 volumes shared with the Maxillo dataset, it is noteworthy that 40 underwent re-segmentation through IACAT 2.0, while the annotations for the other 51 volumes remained unaltered. When releasing the ToothFairy challenge, the corresponding dataset was (and still is when writing this article) the most extensive and publicly available in the literature for what concerns the Inferior Alveolar Canal (IAC) segmentation in CBCT volumes. This dataset contains both 2D and 3D annotations.

F. Ethics Approval and Data Availability

The data of the training set received ethical committee approval from Comitato Etico dell'Area Vasta Emilia Nord (Approval Number 1374/2020/OSS/ESTMO SIRER ID 1275 - NAICBCT-D) and can be shared for research purposes. The training volumes can be downloaded under CC BY-SA license at https://ditto.ing.unimore.it/ after user registration.

[3]This public test set is different from the challenge test set, and it is not used for submission ranking. Any participant can choose whether to use this data (e.g., as internal validation by using both public train and validation as training data).

TABLE III

FINAL TEST PHASE LEADERBOARD. THE ORDER OF THE OVERALL RANKING IS DETERMINED AS THE AVERAGE RANK FOR EACH METHOD'S MEAN DICE AND HD95 METRICS. PARTICIPANTS FROM EIGHT UNIQUE COUNTRIES AND THREE CONTINENTS WERE REPRESENTED ON THE FINAL LEADERBOARD AND SOME TEAMS PARTICIPATED WITH MORE THAN ONE METHOD. THE TOP THREE METHODS SHOW A HIGH EFFECTIVENESS WITH A DICE SCORE ABOVE 0.75 AND A HD95 METRIC BELOW 10. THE METHODS WITH "INF" MEAN HD95 METRIC PREDICTED AN EMPTY SEGMENTATION FOR ONE OR MORE SCANS IN THE PRIVATE DATA. ⋆ REPRESENTS THE ORGANIZERS' PROPOSED BASELINE APPROACH AT THE TIME OF CHALLENGE CREATION. MEMBERS OF THE ORGANIZERS' INSTITUTES COULD PARTICIPATE, BUT WERE NOT ELIGIBLE FOR AWARDS

Final Rank	ID	Participant	Country	Dice		HD95	
				Score	Rank	Score	Rank
1	A	Liu et al.	China	0.796 ± 0.093	1	4.49 ± 6.08	1
2	B	Wang et al.	China	0.769 ± 0.106	2	7.45 ± 10.90	2
3	C	Wodzinski et al.	Switzerland	0.760 ± 0.088	3	9.22 ± 14.40	3
4	D	Kirchoff et al.	Germany	0.739 ± 0.146	4	13.90 ± 28.70	5
5	E	Huang	China	0.715 ± 0.110	8	13.10 ± 13.50	4
6	F	Su	China	0.734 ± 0.099	5	18.00 ± 25.30	8
7	G	Ye	China	0.728 ± 0.094	7	18.90 ± 25.00	9
8	H	Yang	China	0.731 ± 0.151	6	28.60 ± 47.40	12
8	I	Han et al.	South Korea	0.700 ± 0.148	11	15.70 ± 24.80	7
10	J	Szczepański et al.	Poland	0.675 ± 0.208	13	14.90 ± 29.30	6
11	K	Wu	China	0.712 ± 0.142	10	19.40 ± 35.60	10
11	L	Zheng	China	0.713 ± 0.129	9	19.50 ± 28.00	11
⋆13	M	Lumetti et al.	Italy	0.699 ± 0.151	12	38.60 ± 48.00	13
14	N	Han et al.	South Korea	0.643 ± 0.179	14	40.20 ± 49.70	14
15	O	Huang	China	0.642 ± 0.187	15	inf	17
16	P	Szczepański et al.	Poland	0.613 ± 0.201	16	inf	17
17	Q	Pang	China	0.612 ± 0.178	17	inf	17
18	R	Gamal	Egypt	0.326 ± 0.172	21	42.70 ± 43.90	15
18	S	Li	China	0.327 ± 0.218	20	65.90 ± 48.10	16
18	T	Caselles Ballester et al.	Spain	0.507 ± 0.209	19	inf	17
21	U	Kirchoff et al.	Germany	0.565 ± 0.259	18	inf	21

On the other hand, the ToothFairy test set has no ethical committee approval for public release. However, it is accessible through the Grand Challenge platform via the *post challenge phase*[4] and represents a common benchmark to allow for a long-term fair comparison of future proposals.

G. Distribution Shift

1) Gaussian Kernel Density Estimator: As mentioned, the scans of the private dataset were collected from an external medical center, which used a different CBCT machine than that employed for the scans of the public dataset. The change in the CBCT scanner resulted in a distribution shift when comparing the intensities of scans from the public and private datasets (Fig. 6). This shift in distribution could have affected the effectiveness of the participating methods if they had not accounted for this change in their implementation. For example, clipping the intensity values between 0 and 2,200 makes the distribution shift from the public to the private dataset considerably less severe.

2) Parameter Distribution of Network Layers: To evaluate the relationship between the test and training sets of the ToothFairy dataset, another practical approach is employed. Previous works adopted a similar strategy to highlight dataset shifts due to variations in scene and illumination conditions [28] or different perspectives of the same target [29].

Basically, we analyze the parameter distribution of both convolutional layers and instance normalization layers (INL), sampled at the beginning and at the end of nnU-Net. Results obtained considering the two datasets are reported in Fig. 5.

Two main observations can be drawn. The weights of the high-level and low-level convolutional layers exhibit nearly identical distributions since they are mainly influenced by local features. Given that different scans target the same subject, the local information extracted from the images is more consistent than the global appearance. Consequently, the convolutional weights are less sensitive to dataset-specific variations.

On the other hand, the instance normalization statistics show different distributions, particularly in the high-level features, such as those represented by the first layer. These variations arise due to the specific characteristics of each dataset. The instance normalization layers capture and summarize the dataset-specific mean and variance of features, so they are more sensitive to dataset variations.

IV. METHODS

The ToothFairy dataset has received almost 600 data download requests from unique users. 19 teams uploaded their submission onto the leaderboard. 190 submissions were seen in the preliminary phase, and 30 submissions for the final test phase.[5] Tab. III provides a brief synopsis of all the participating teams. Below, we present the algorithms proposed by the four best-performing teams of the challenge Sec. IV-B is provided to highlight the commonalities and distinctive elements of the algorithms submitted to the

[4]https://toothfairy.grand-challenge.org/evaluation/post-challenge-phase-test-your-algorithm/leaderboard/

[5]Numbers were collected on the 2nd of February, 2024.

TABLE IV
COMPARISON OF THE BEST-PERFORMING APPROACHES IN TERMS OF ARCHITECTURE, AUGMENTATIONS, AND PRE-/POST-PROCESSING STRATEGIES EMPLOYED, LOSSES AND TRAINING APPROACH ADOPTED, AND WHETHER THEY USED THE SPARSE LABELS AVAILABLE IN THE SECONDARY DATASET OR NOT

ID	Model	Augmentations	Pre-processing	Post-processing	Training Approach	Loss(es)	Sparse Labels
A	nnU-Net with KD	Strong augm. Test-time augm.	CCL filtering Uniform resampling Z-score norm.	✗	Semi-super.	Cross-entropy Dice	✓
B	nnU-Net with Geometric fine-tune	✗	✗	✗	Supervised	Focal Dice	✓
C	3D ResUNet	Strong augm. Elastic transf.	Resampling CCL analysis Intensity clipping	CCL analysis	Supervised	Soft Dice Focal	✗
D	nnU-Net	nnU-Net augm. Black rect. transf.	nnU-Net norm. Z-score norm.	✗	Supervised	Focal clDice Modified HD	✗

challenge, with particular attention to the four best-performing solutions.

A. Participating Methods

A - Liu et al. The proposed segmentation framework utilizes nnU-Net [30], employing a self-training approach for semi-supervised semantic segmentation. A connectivity-based selective re-training strategy is introduced to enhance the reliability of pseudo-labels. Pre-processing involves labeled data filtering through connected component labeling [31] and boundary refinement, re-sampling for uniform spacing, and intensity normalization using z-score normalization. The framework is inspired by ST++ [32] and employs nnU-Net for both teacher and student models. A selective re-training scheme prioritizes the reliability of unlabeled samples by evaluating the stability of evolving pseudo-masks across training iterations. Model checkpoints are saved, and the difference in predictions for unlabeled images is used to measure reliability. The top k data with the highest stability score and two connected components are selected to generate pseudo-labels. Strong data augmentations, including test-time augmentations, are applied to mitigate overfitting and improve the student model generalization.

B - Wang et al. The method employs nnU-Net and a Focal Dice Loss [33], [34]. To enhance voxel classification precision, fine-tuning is conducted in geometric space, targeting points near the surface of the IAN structure using an occupancy network [35], [36]. The fine-tuning process involves sampling points around the IAN, extracting features, and utilizing an occupancy network to classify a point as inside or outside. Network pre-training addresses data scarcity by using the dataset with sparse annotations, employing label propagation with Gaussian heat maps generated from the sparse annotations.

C - Wodzinski et al. The method begins with pre-processing, re-sampling input volumes, and clipping and normalizing intensities. Augmentation during training involves various transformations with random order and probabilities, implemented using the TorchIO library [37]. Elastic transformation is performed offline before training due to its computational complexity. A 3D ResUNet-based [38] neural network is employed, taking the full re-sampled volumes as input and producing segmentations with the same shape. The objective function combines Soft Dice Loss and Focal Loss [33] with equal weights. Fully supervised training uses dense annotations, with AdamW optimizer and an exponentially decaying learning rate scheduler. Inference involves re-sampling, intensity clipping, and normalization, followed by prediction and connected component analysis [39] to remove artifacts.

D - Kirchhoff et al. The method leverages the nnU-Net framework. Pre-processing involves intensity normalization, experimenting with both default normalization and z-score normalization. Different loss functions, including focal loss [33], clDice loss [40], and a modified Hausdorff loss [41] are experimented with to address segmentation challenges. nnU-Net data augmentation strategies are enhanced with increased probabilities of individual augmentations and adding specialized transforms such as "blank rectangle transform." During inference, ensembling is employed by combining models with a 5-fold cross-validation. Input images are pre-processed on-the-fly and predicted in a sliding window fashion.

A summary comparison between the top-performing solutions is reported in Tab. IV.

B. Observations

First of all, it must be noticed that, although with different approaches, most of the top-performing solutions leverage the nnU-Net *framework* to perform segmentation. Once again, this confirms the effectiveness of such a model for medical image segmentation tasks [42], [43], [44]. It is worth noting that nnU-Net (No New U-Net) is not a new model architecture with respect to U-Net, but it "simply" defines a task-agnostic configuration policy to be used in the adaptation of the U-Net network to a specific use case, without requiring an extensive empirical research in the design choices.

For what concerns the *pre-processing*, all the participants made use of re-sampling, and intensity normalization and

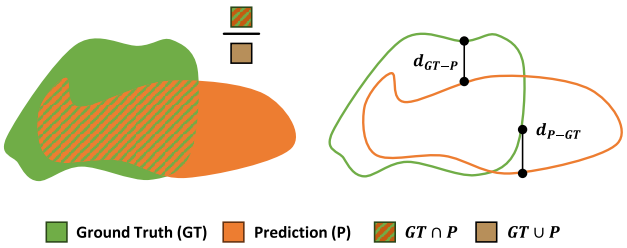


Fig. 7. Visual representation of evaluation metrics. On the left is depicted the intersection over union (IoU), on the right, the d_{95} distances used to calculate the Hausdorff Distance (HD) between two regions of points.

exploited different kinds of augmentations to push algorithm performance. This is a common approach in modern machine learning models that requires large amounts of quality annotated data. Currently, data augmentation is the most effective way of reducing the amount of high-quality manually annotated data required, still achieving satisfactory performance. The main goal of data augmentation is to increase the volume, quality, and diversity of training data while keeping the annotation cost and time relatively small [45].

Specific *training* challenges, such as disconnected components in the segmentation labels and class imbalance, were tackled differently by participants. However, most of them leverage model ensembling to improve prediction during the inference phase. Basically, multiple (weak) learners are fitted on the training set and provide one prediction each. The final outcome of the ensemble is computed by combining the results from all of the learners. Regardless of the combination strategy (max-voting, averaging, stacking, blending, etc), ensembling is proven to be an effective strategy to push model performance by reducing the biases of the trained estimators involved. It has been extensively used in the literature, especially for challenges [46], [47].

Among the main differences between ToothFairy participants are the different uses of *loss functions* and the *training strategies* involved. Liu et al. and Wang et al. both experimented with a combination of loss functions (cross-entropy loss and dice loss), while Wodzinski et al. focused on addressing disconnected components in the segmentation. On the other hand, Kirchhoff et al. explored various loss functions and optimized the data augmentation pipeline, picking up the combination that best performed on the public test set.

Concerning the training strategies, Liu et al. and Wang et al. employed self-training and fine-tuning in geometric space, respectively. Wodzinski et al. and Kirchhoff et al. focused on improving the baseline with different loss functions and data augmentation strategies.

V. Evaluation

A. Evaluation Metrics

The metrics used to rank the submitted proposals are the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95), Fig. 7, two metrics commonly used in image segmentation [48]. The DSC has practically the same meaning as the IoU (Intersection over Union), but the

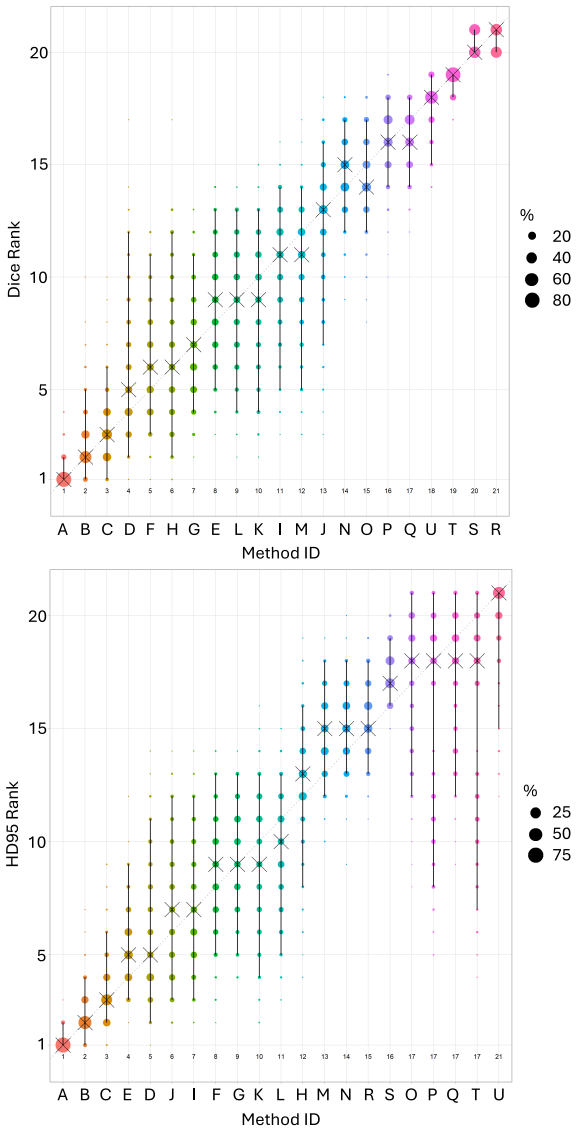


Fig. 8. Algorithms are color-coded, and the area of each blob at position (A_i, *rank j*) is proportional to the relative frequency A_i achieved rank j across $b = 1000$ bootstrap samples. The median rank for each algorithm is indicated by a black cross. 95**%** bootstrap intervals across bootstrap samples are indicated by black lines.

first one is better suited when the region of interest is much smaller than the background. In such a scenario, DSC can be more robust and informative than IoU since more weight is given to the correctly identified region. The DSC metric, and its relationship with the IoU, are expressed by the following formula:

$$\mathrm{DSC(P,GT)} = \frac{2 \times |P \cap GT|}{|P| + |GT|} = \frac{2 \times IoU}{1 + IoU} \tag{1}$$

where P is the model prediction and GT is the ground truth.

On the other hand, the HD95 computes the maximum distance between two sets of points, considering the 95th percentile of these distances. In general, the 95th percentile of the distances between boundary points in A and B is defined as follows:

$$\mathrm{d_{95}}(A, B) = \mathrm{x}^{95}_{a \in A} \left\{ \min_{b \in B} d(a, b) \right\} \tag{2}$$

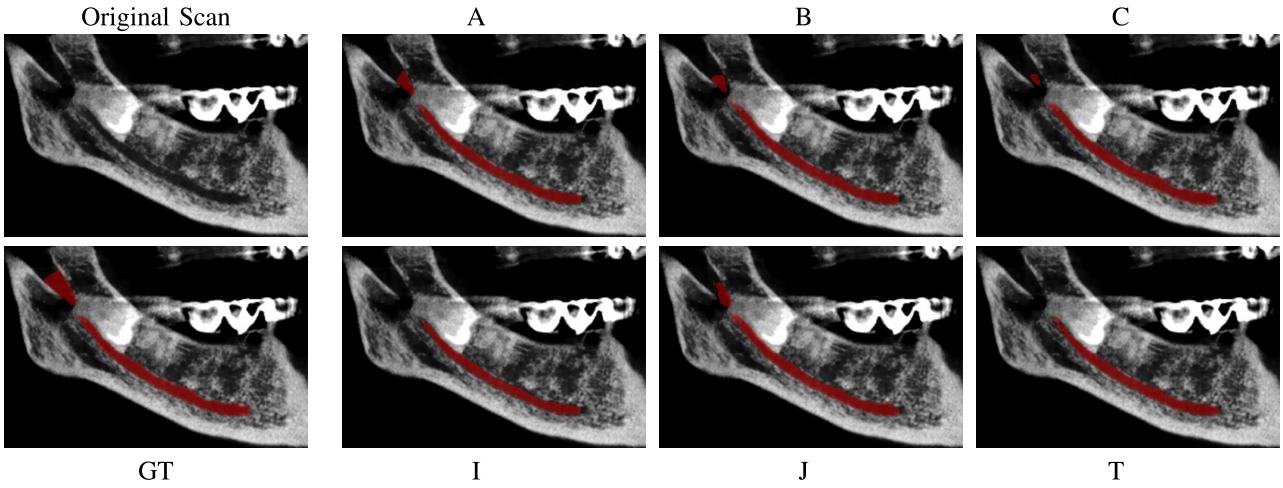


Fig. 9. Visualization of IAN predictions for a case with a risk of damaging the IAN when extracting a molar. The top-left image of each case shows a projection of the CBCT scan to reveal the IAN in one view and the bottom-left image is the ground-truth segmentation. The six following images show the predictions of the methods A, B, C, I, J, and T, from left to right, top to bottom.

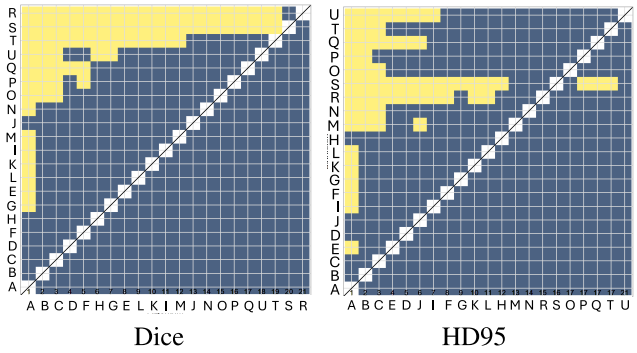


Fig. 10. Significance maps for Dice (left) and HD95 (right). x- and y-axes report method IDs. In this charts, the incidence matrices of pairwise significant test results for the one-sided Wilcoxon signed rank test at a 5% significance level are depicted. Yellow shading indicates that metric values from the algorithm on the x-axis were significantly superior to those from the algorithm on the y-axis, blue color indicates no significant difference.

where $x^{95}_{a\in A}\{\}$ denotes the 95th percentile of the elements in the set enclosed within the brackets. Given the set formed by the pixels in the predicted mask (P) and the set of pixels belonging to the ground truth (GT), the Hausdorff distance is determined as the maximum value of the two distances between P and GT and GT and P at the 95th percentile:

$$\text{HD95(P,GT)} = \max\left\{d_{95}(P, GT), d_{95}(GT, P)\right\} \quad (3)$$

By using the 95th percentile, this metric provides a robust evaluation that is less sensitive to outliers or extreme differences between the sets of points.

B. Ranking Protocol

All of the CBCTs in the ToothFairy dataset belonged to different patients and were annotated by one expert only. Both the aforementioned metrics provide homogeneous numbers on different patients, meaning that they can be averaged later to provide the final rank (one for each metric).

To ensure robustness in the final ranking, the recommendations provided by Maier et al. [49] have been followed. In general, there are two contrasting approaches to aggregate metrics across the test cases. The first approach, known as metric-based aggregation, involves initially aggregating metric values across all test cases (e.g., using mean or median) and then ranking the algorithms on the aggregated value. The second approach is case-based aggregation, which starts by calculating a rank for each test case, and then the final rank is determined by aggregating the ranks of the test cases. According to [49], the single-metric rankings (specifically for DSC and HD95 in our case) demonstrate higher statistical robustness when employing metric-based aggregation and using the mean instead of the median for aggregation.

For the aforementioned reasons, the ranking schema of our challenge involves the following steps:

1) Calculate the Dice score (averaged across all volumes), the HD95 (averaged across 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 is considered a definitive tie.

In order to test and visualize the ranking stability when using the selected metrics, Fig. 8 and Fig. 10 are provided. Both figures are generated by *ChallengeR* [50], a standard tool for analyzing and visualizing challenge outcomes. Reported results are obtained by performing random sampling with replacement (bootstrapping) 1,000 times. Fig. 8 employs a blob plot to visualize the ranking stability based on bootstrap sampling. Best algorithms consistently confirm their rank with lower variability w.r.t. other methodologies that demonstrate a much higher variability.

On the other hand, Fig. 10 depicts the incidence matrices of pairwise significant test results for the one-sided Wilcoxon signed rank test at a 5% significance level. Yellow shading

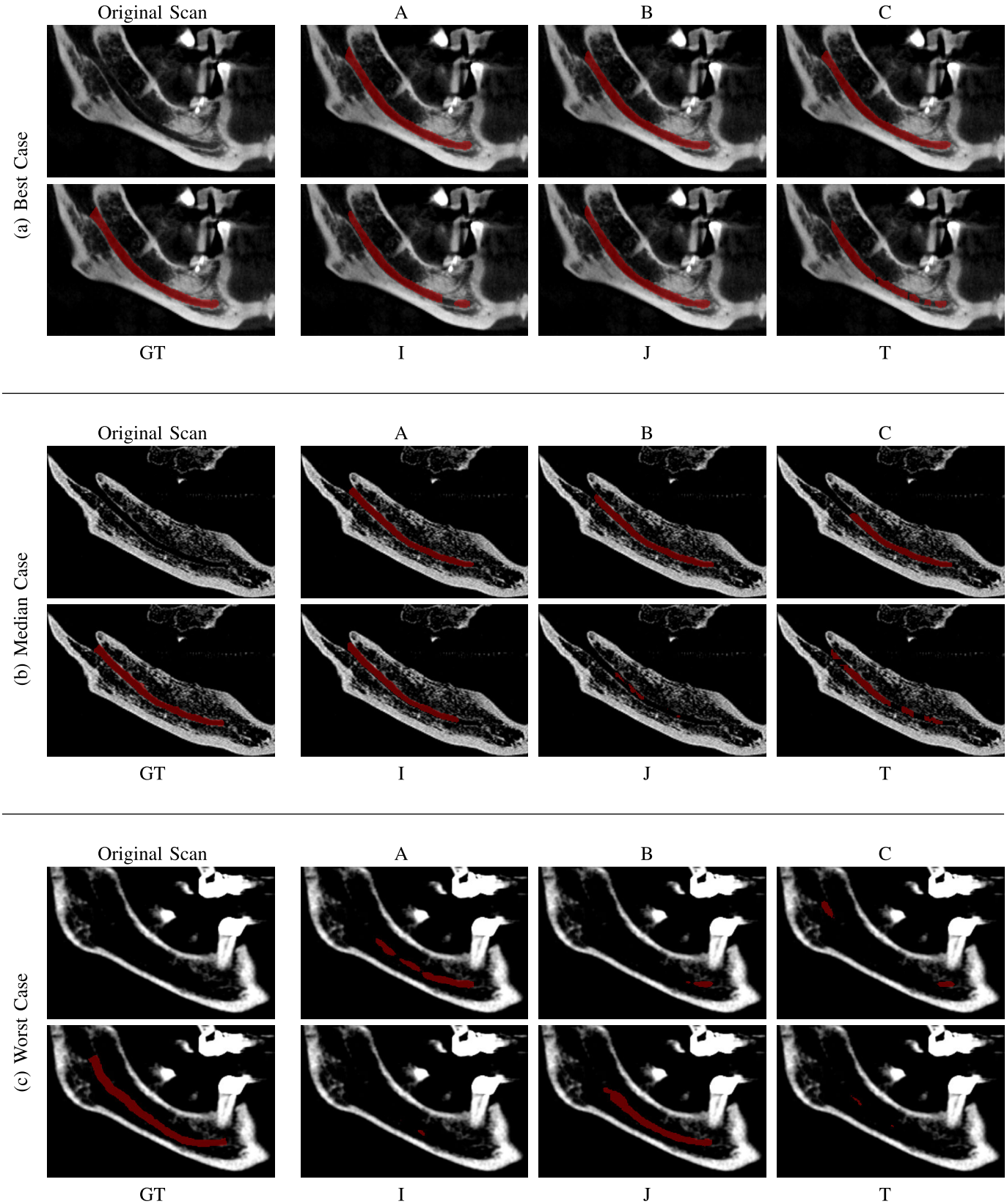


Fig. 11. Visualization of IAN predictions for the cases with the best (a), median (b), and worst (c) effectiveness. The top-left image of each case shows a projection of the CBCT scan to reveal the IAN in one view and the bottom-left image is the ground-truth segmentation. The six following images show the predictions of the methods A, B, C, I, J, and T, from left to right, top to bottom.

indicates that metric values from the algorithm on the x-axis were significantly superior to those from the algorithm on the y-axis, and blue color indicates no significant difference. As also confirmed by the slight difference in performance scores reported in Tab. III, the dominance of the best-performing method is not statistically significant w.r.t. to other

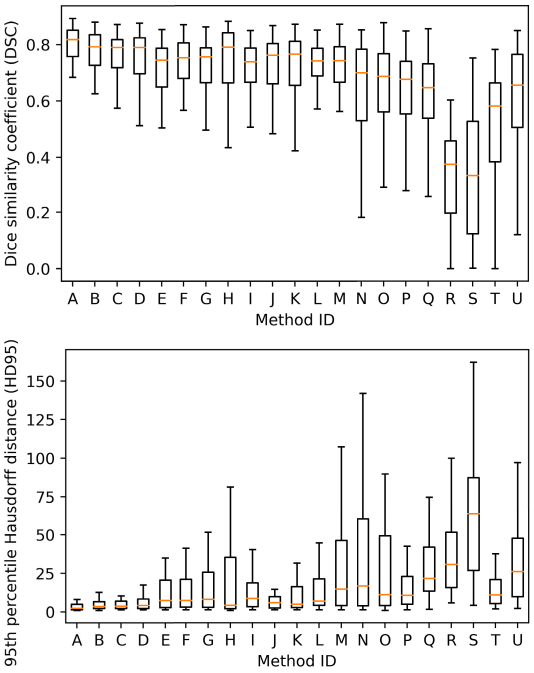


Fig. 12. Dice similarity coefficient (DSC) and the 95th percentile Hausdorff distance (HD95) metrics for each participant. The effectiveness of the participating algorithms varies greatly and the method at rank 1 is the most effective in terms of DSC and HD95. Outliers were removed to improve readability.

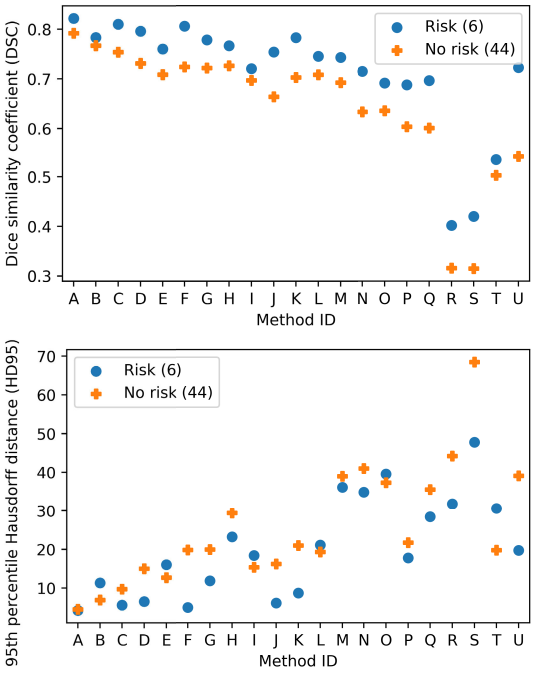


Fig. 13. Mean Dice similarity coefficient (DSC) and mean 95th percentile Hausdorff distance (HD95) for cases with, respectively without, a risk of damaging the IAN when extracting a molar.

top-5-performing algorithms, but it is compared with all the others.

C. AWS Infrastructure

All the algorithms submitted by participants were run on Grand Challenge AWS infrastructure, which is able to scale up storage and compute capabilities on demand. A `g4dn.xlarge` instance (Nvidia T4, 16GB GPU memory, 4 CPU, 16GB CPU memory) or `g4dn.2xlarge` instance (Nvidia T4, 16GB GPU memory, 8 CPU, 32GB CPU memory) is used based on the configuration selected. During execution, the docker container does not have access to the internet, thus preventing any exfiltration of test set-related information.

VI. Results

Results of four representative cases from the final test set show a large variety in effectiveness among the participating methods (Fig. 11). The least effective results, Fig. 11c, were most likely the result of a limited contrast in the mandible's cancellous tissue, making it difficult to delineate the IAC.

The effectiveness for a case with a risk of IAN damage when extracting a molar (Fig. 9) was high, and the segmentations capture the displacement of the IAN well, at least those of the best-performing proposals.

The effectiveness varied greatly between participating methods, particularly in terms of HD95 (Fig. 12). The ranking of the methods does not follow a monotonic relation with the DSC and HD95 metrics. The lowest-ranked methods predicted empty segmentations of the IAC for some cases, which resulted in an infinite HD95 metric (Tab. III). After

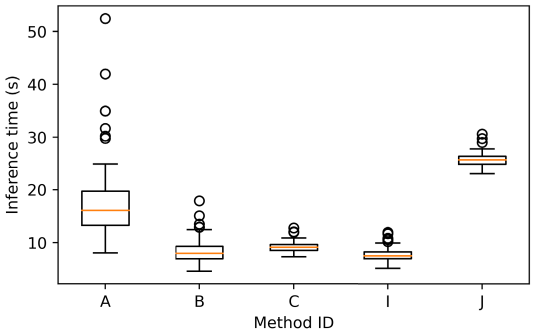


Fig. 14. Inference times of methods from participants that shared their algorithm at rank 1, 2, 3, 8, and 10, respectively. The shared algorithms can predict a segmentation of the IAN in under one minute.

removing such outliers in Fig. 12, the final ranking does not match the mean metrics.

Four participants in the top 10 that shared the source code could predict a segmentation of the IAC in under one minute (Fig. 14). The top-ranked method required more processing time than the second and third-ranked methods.

The effectiveness on cases in the final test set with a risk of damaging the IAN when extracting a molar was compared to cases without a risk (Fig. 13). All methods achieved a better DSC for the cases with a risk compared to the cases without a risk. This may be explained by the improved contrast between the IAC and its surrounding tissues; separating the IAN and molar is easier than separating the IAC and cancellous bone tissue. Such behavior is partially confirmed also when considering HD95 metric (Fig. 13, bottom chart). In this case, for 15 out of 21 participating methods, the HD95 is better (lower) in cases with a risk. Apart from method T, which ranked at the bottom of the leaderboard, all the other five methods have a small difference in performance when

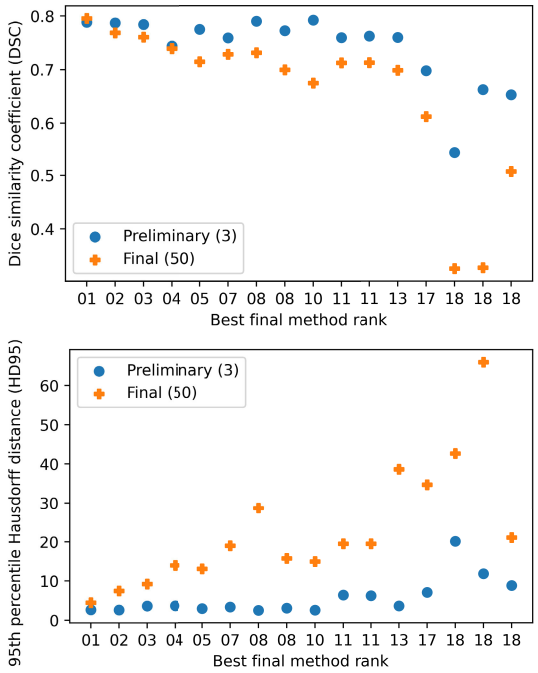


Fig. 15. Most effective mean Dice similarity coefficient (DSC) and mean 95th percentile Hausdorff distance (HD95) for participants in the preliminary and final test phases. The participant ranking changed considerably from the preliminary phase to the final phase and only the mean DSC of the top-ranked participant improved.

comparing the HD95 on patients with and without risk. This could be explained by considering the different optimization strategies employed by participants, always including the Dice loss or its variations.

Lastly, the effectiveness of the participants was compared between the preliminary and final test phases (Fig. 15). Only the DSC metric for the top-ranked participant improved from the preliminary test phase to the final test phase. The three cases in the preliminary test phase were also present in the final test phase. Participants may thus have overfitted on these preliminary cases, resulting in a decrease in effectiveness for the remaining 47 cases.

VII. Discussion

In this section, we will discuss the most relevant techniques employed in the algorithms submitted to the challenge. Several strategies implemented in the participating methods can be combined to create a highly effective segmentation approach. The selected techniques aim to maximize the accuracy and robustness of the segmentation model.

A. Data Cleaning and Pre-processing

Since the segmentation algorithms are data-driven, it is very important to ensure the quality and reliability of the input data. Firstly, the top-performing participants paid attention to the (possible) shift in distribution from the public to the private data (Fig. 6) when developing their methods. For example, the top-ranked method employed strong data augmentation, including strong intensity transformations, for predicting an effective segmentation of the IAC independent of the underlying intensity distribution. Another effective approach used by the third-ranked method was to clip the intensity values to $[-100, 900]$, which disallowed the model to learn features based on large negative intensities that were not present in the private data. It is worth noting that the effectiveness of a deep learning system developed using data from a single center does not readily generalize to data from external centers. Ideally, a challenge should provide public access to data from multiple centers to ensure reliable model development, while keeping data from another center private to obtain a dependable model evaluation. Unfortunately, this was impractical in the ToothFairy context because no other center publicly released CBCT scans of the lower jawbone.

Regarding the cleaning of the data, the most relevant strategy was employed by the first classified team. They remove dense annotations with discontinuities or unreliable annotations of terminal segments. This ensured that only high-quality annotations were used for training and evaluation, improving the segmentation accuracy on the private set.

B. Network Architecture and Training Strategies

The complex three-dimensional shape of the alveolar canal, along with the need for high accuracy, required the adoption of a 3D segmentation network for canal identification. In fact, all the top-performing methods either predicted a segmentation of the fully re-sampled scan or predicted segmentations of the IAC in patches and then aggregated these patch predictions. None of them relied on 2D segmentations performed at the slice level. The adoption of the *3d fullres* nnU-Net, a specific configuration of the popular nnUNet that involves feeding the entire 3D volume into the network, allowed for extremely good performance (third-ranked method), but required high computational resources.

Given the limited size of the densely annotated dataset, the most appropriate learning technique employed in this scenario is semi-supervised learning. This approach involves iteratively augmenting the training data by incorporating unlabeled images and pseudo-labels, which contribute to enhancing the training process. To further improve this method, the top-performing team used a strategy that selectively considers the most reliable pseudo-labels, based on the stability of predictions across epochs. This prevented pseudo-labeled data from negatively affecting the performance of the model.

One possible direction to further improve the current methods is a two-stage approach. Specifically, a segmentation of the mandible or IAC can be predicted in the first stage to project the CBCT scan to a number of parallel plain radiographs, as seen in Fig. 11. Following this projection, a proposed method for IAC segmentation on a plain radiograph can be used as the second stage. Lastly, the 2D segmentations can be back-projected to the CBCT scan for the actual result.

Another two-stage approach effectively explored in the literature [14], [21] consists of predicting a coarse segmentation of the IAC (first stage) to reduce the input volume to the smallest bounding box containing the IAC. Dealing with smaller volumes will allow to fit memory requirements and perform fine grain segmentation (second stage) by feeding the model with the entire data, avoiding patch-based learning strategies that usually degrade performance.

Different participants used a combination of loss functions to train the neural networks, typically employing the Dice loss in combination with Focal or modified Hausdorff losses.

C. Data Augmentation

A common method was to use augmentations for data paired with dense labels. The top-performing team also employed strong data augmentations on volumes with a pseudo-label only, to mitigate overfitting noisy labels and decouple similarity between teacher and student predictions. Some of the augmentation techniques used were color, noise, painting jitter, rotation, flipping, and cropping. Test-time augmentation has also been applied during inference to enhance the performance by ensembling predictions.

D. Fine-Tuning and Post-Processing

After training the segmentation model, several teams performed a fine-tuning step. This involves using interpolated features from the final feature maps and the output layer to predict the occupancy of points close to the boundary of the alveolar canal. This finer-resolution sampling improved the localization accuracy of the canal. Finally, a technique that can be employed to enhance the consistency of the segmentation output is the removal of small connected components from the segmentation mask (similar to that suggested for pre-processing). Removing these small components ensures that only the larger regions remain in the final segmentation mask, leading to a more accurate and consistent segmentation output.

In summary, the key strategies adopted by the challenge participants and demonstrated to be effective in the context of the ToothFairy challenge are:

- Removal of dense annotations with discontinuities from the training set;
- Adoption of intensity normalization strategies based on data augmentation or performed by clipping intensity values and applying z-score normalization based on foreground voxel values;
- Selection of state-of-the-art 3D-based models (e.g., nnU-Net) instead of 2D slice-based solutions;
- Feed of the model with the entire (eventually re-sampled) input volume instead of adopting patch-based solutions whenever allowed by the memory constraint of the training infrastructure;
- Leverage the availability of sparse annotations employing semi-supervised learning strategies and pseudo-labels;
- Adoption of substantial data augmentation techniques during both the training and test phases;
- Combine different loss functions (e.g., Dice loss, Focal loss, and modified Hausdorff loss) as a regularization technique to improve the generalization of the models;
- Post-process the output segmentation masks to improve the consistency of the segmentation output, e.g., by removing small connected components.

VIII. Limitation and Future Works

Today research regarding CBCT images has mainly focused on teeth segmentation [51], [52], [53]. Many challenges have been thrown into this task, which is particularly complex due to the peculiar 3D anatomical shape of dental structures. Furthermore, some challenges have also focused on identifying dental lesions, such as periapical cysts and pulp calcification, or identifying mandibular fractures on CBCTs or OPGs [51], [54], [55]. However, the ultimate practical implementation would involve combining the aforementioned methods with accurate segmentation of the IAC.

Identifying relationships between structures is the crucial step for surgical planning, and automating this key step could be very useful, especially when combined with 3D assessment of bone volumes [15], [56]. The result could help in the automation of subsequent reconstructive and rehabilitation steps, such as the positioning of endosseous implants and bone regenerative surgery. These processes could be made more complicated if the normal anatomy is altered due to the presence of a pathological lesion or fracture. An anatomical variant, such as a bifid IAC, is another potential confounding factor. Overcoming these obstacles represents an intriguing challenge, as it would allow further expand the use of automation in clinical practice, increasing their validation.

Shortcomings in existing works about IAC segmentation (Sec. II) are mainly related to network architectures and data.

A. Scarcity of Data

Many studies use datasets with limited size [12], [15], or lack of comprehensive 3D annotations [11], relying instead on 2D slices, which do not capture the full anatomical complexity of the IAC. Additionally, there is a reliance on proprietary datasets that are not publicly available, limiting the reproducibility and verification of results [10], [12], [14], [15]. Increasing the size and diversity of publicly available 3D annotated datasets would significantly enhance the robustness and generalizability of segmentation models. Our dataset partly addresses this need, as it is the largest publicly released dataset with 3D annotations for the IAN, providing a significant resource for the research community. Another way to address the scarcity of data should be to explore semi-supervised learning architectures, which have shown promising results by leveraging limited annotated data [57]. These methods can effectively use both labeled and unlabeled (or sparsely labeled) volumes to improve model performance, as already demonstrated by the top-ranked solution.

B. Architectural Shortcomings

From an architectural perspective, many models used for IAC segmentation lack the ability to effectively handle the inherent noise and artifacts present in CBCT images, which can lead to suboptimal performance. For instance, to satisfy the requirement for extensive data cleaning and pre-processing, top-performing participants in the ToothFairy challenge employed strong data augmentation and specific intensity transformations to handle distribution shifts between public and private data. Moreover, there is a tendency to adapt general-purpose segmentation architectures without adapting them to the specific challenges posed by IAC segmentation, such as the fine and tubular nature of the canal. Tailoring segmentation architectures to address the specific anatomical

and imaging challenges of the IAC can lead to more accurate and reliable models [58], [59].

IX. Conclusion

In this paper, we present the design and outcomes of the 1st ToothFairy challenge, jointly organized by the University of Modena and Reggio Emilia and the Radboud University Medical Center located in Nijmegen. The challenge was organized within the MICCAI 2023 conference and aimed at addressing the scarcity of publicly available datasets for IAC segmentation. Through the collaboration with expert doctors for the annotation process, we released the biggest publicly available dataset for IAC segmentation. Eighteen research teams from around the world submitted their algorithms for evaluation, and the first classified method is the one of Liu et al. with a Dice score of 0.796 and an HD95 of 4.49, winning a price of €1, 500. In general, all the top-performing solutions used the nnU-Net architecture, proving its effectiveness for medical image segmentation. The adoption of various pre-processing techniques, such as re-sampling, intensity normalization, and data augmentation, has significantly contributed to improving the algorithms' performance.

Overall, the ToothFairy challenge provided a useful reference point in the domain of IAC segmentation, offering researchers an opportunity to benchmark their algorithms against state-of-the-art solutions. The second edition of the challenge, already accepted to be held at MICCAI 2024, will extend the segmentation task to additional anatomical structures that are in close relation with the IAC (i.e., mandible, teeth, maxillary bone, and the pharynx). However, the 1st ToothFairy challenge will remain open for new submission on the Grand Challenge website and can be used as a benchmark for future IAC segmentation algorithms.

Author Contributions

Conceptualization: Federico Bolelli, Shankeeth Vinayahalingam, Alexandre Anesi, and Costantino Grana; Methodology: Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Bram van Ginneken, Alexandre Anesi, and Costantino Grana; Software–Webpages and Evaluation Scripts: Federico Bolelli and Luca Lumetti; Software–challenge Submissions: Yusheng Liu, Rui Xin, Tao Yang, Lisheng Wang, Haoshen Wang, Chenfan Xu, Zhiming Cui, Marek Wodzinski, Henning M̈uller, Yannick Kirchhoff, Maximilian R. Rokuss, Klaus Maier-Hein, Jaehwan Hann, Wan Kim, Hong-Gi Ahnn, Tomasz Szczepański, Michal K.Grzeszczyk, Przemyslaw Korzeniowski, Vicent Caselles-Ballester, Xavier Paolo Burgos-Artizzu, and Ferran Prados Carrasco; Validation: Luca Lumetti Niels, van Nistelrooij, and Kevin Marchesini; Formal Analysis: Federico Bolelli, Luca Lumetti, Niels van Nistelrooij, and Kevin Marchesini; Investigation: Luca Lumetti, Niels van Nistelrooij, and Kevin Marchesini; Resources: Federico Bolelli, Bram van Ginneken, Alexandre Anesi, and Costantino Grana; Writing–original Draft: Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Kevin Marchesini, and Niels van Nistelrooij; Writing–Review and Editing: Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Kevin Marchesini, Niels van Nistelrooij, Tong Xi, Yusheng Liu, Zhiming Cui, Marek Wodzinski, Yannick Kirchhoff, Jaehwan Han, Tomasz Szczepański, Vicent Caselles-Ballester, Stefaan Berge', Alexandre Anesi, and Costantino Grana; Visualization: Federico Bolelli, Luca Lumetti, Niels van Nistelrooij, and Kevin Marchesini; Data Curation: Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalinga, Mattia Di Bartolomeo, Arrigo Pellacani, and Pieter van Lierop; Supervision: Bram van Ginneken, Alexandre Anesi, and Costantino Grana; Project Administration: Federico Bolelli and Costantino Grana; and Funding Acquisition: Federico Bolelli, Bram van Ginneken, Alexandre Anesi, and Costantino Grana.

Appendix

This section provides all the challenge organization's technical elements that did not fit with the main flow of the article but need to be reported for an exhaustive description.

Challenge Timetable: The training data was released on the 30th of March, 2023, enabling participants to access them for analysis and model development. During the "Preliminary Phase," participants had the opportunity to submit their models from the 1st of July, 2023, until the 16th of August, 2023. In this case, the test was performed on a subset of 3 test volumes out of the total 50. These preliminary results provided participants with feedback about the quality of their submissions and the docker produced, without the possibility of crafting an ad hoc solution to fit the entire test set.

The "Final Phase" took place from the 8th to the 16th of August, 2023, and only two algorithm submissions per team were allowed. Participants were invited to publish their algorithms on GitHub and share their research papers with the organizing team by the 31st of August, 2023. This step was mandatory for the top three classified teams. The results of the challenge were released on the 1st of September, 2023, and in the following days, winners were officially announced.

The discussion about the ToothFairy challenge results and the presentation of the top-performing works was held on the 8th of October, during the associated workshop days within the MICCAI 2023 event. The first three teams were awarded the following prizes: €1, 500 for the first classified, €1, 000 for the runner-up, and €500 the third-place best.

Publication Policies: For the entire duration of the challenge, participants had access to the challenge repository on GitHub,[6] where they could (and still can) find the source code of the challenge evaluation script and a docker template with a baseline algorithm to be replaced for participating.

As mentioned before, all the participants were required to public their code (mandatory for the top ranking methods), and it is now shared with the research community within the challenge GitHub repository.[6]

All the members of the team participating in the challenge qualify as authors of the submission. This paper resumes challenge results and includes a description of the main proposals: at most three members for each team have been

[6]https://github.com/AImageLab-zip/ToothFairy

included as co-authors; an exception has been made only for the first classified team.

All the participants are allowed to submit their own results in any venue (conferences, workshops, etc) with embargo restriction: 6 months after the MICCAI 2023 event.

Acknowledgment

Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani, Bram van Ginneken, Alexandre Anesi, and Costantino Grana are member of the Challenge Organizing Team.

Federico Bolelli, Luca Lumetti, Kevin Marchesini, and Costantino Grana are with the Department of Engineering "Enzo Ferrari," University of Modena and Reggio Emilia, 41121 Modena, Italy (e-mail: federico.bolelli@unimore.it; luca.lumetti@unimore.it; kevin.marchesini@unimore.it; costantino.grana@unimore.it).

Shankeeth Vinayahalingam, Pieter van Lierop, Tong Xi, and Stefaan Berge' are with the Department of Oral and Maxillofacial Surgery, Radboud University Medical Center, 6500 HB Nijmegen, The Netherlands (e-mail: shankeeth.vinayahalingam@radboudumc.nl; Pietervanlierop@hotmail.com; tong.xi@radboudumc.nl; stefaan.berge@radboudumc.nl).

Mattia Di Bartolomeo is with the Department of Oral and Maxillofacial Sciences, Sapienza University of Rome, 00185 Rome, Italy (e-mail: mattia.dibartolomeo@uniroma1.it).

Arrigo Pellacani is with the Unit of Dentistry and Maxillo-Facial Surgery, Surgery, Dentistry, Maternity and Infant Department, University of Verona, 37129 Verona, Italy (e-mail: pellacani.arrigo@aou.mo.it).

Niels van Nistelrooij is with the Department of Oral and Maxillofacial Surgery, Radboud University Medical Center, 6500 HB Nijmegen, The Netherlands, and also with the Department of Oral and Maxillofacial Surgery, Charité-Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität zu Berlin, 12203 Berlin, Germany (e-mail: niels.vannistelrooij@radboudumc.nl).

Yusheng Liu, Rui Xin, Tao Yang, and Lisheng Wang are with the Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: lys_sjtu@126.com; xr1999@sjtu.edu.cn; yangtao22@sjtu.edu.cn; lswang@sjtu.edu.cn).

Haoshen Wang, Chenfan Xu, and Zhiming Cui are with the School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China (e-mail: wanghsh2022@shanghaitech.edu.cn; xuchf2023@shanghaitech.edu.cn; cuizhm@shanghaitech.edu.cn).

Marek Wodzinski is with the Department of Measurement and Electronics, AGH University of Kraków, 30-059 Kraków, Poland, and also with the Institute of Informatics, University of Applied Sciences Western Switzerland (HES-SO), 2800 Delémont, Switzerland (e-mail: wodzinski@agh.edu.pl).

Henning Müller is with the Institute of Informatics, University of Applied Sciences Western Switzerland (HES-SO), 2800 Delémont, Switzerland, and also with the Medical Faculty, University of Geneva, 1205 Geneva, Switzerland (e-mail: Henning.Mueller@hevs.ch).

Yannick Kirchhoff is with the Division of Medical Image Computing, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany, also with the Faculty of Mathematics and Computer Science, Heidelberg University, 69120 Heidelberg, Germany, and also with the HIDSS4Health–Helmholtz Information and Data Science School for Health, 76131 Karlsruhe/Heidelberg, Germany (e-mail: yannick.kirchhoff@dkfz-heidelberg.de).

Maximilian R. Rokuss is with the Division of Medical Image Computing, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany, and also with the Faculty of Mathematics and Computer Science, Heidelberg University, 69120 Heidelberg, Germany (e-mail: maximilian.rokuss@dkfz-heidelberg.de).

Klaus Maier-Hein is with the Division of Medical Image Computing, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany, and also with the Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, 69120 Heidelberg, Germany (e-mail: K.Maier-Hein@dkfz-heidelberg.de).

Jaehwan Han, Wan Kim, and Hong-Gi Ahn are with Osstem Implant Company Ltd., Seoul 07789, Republic of Korea (e-mail: jh.han@osstem.com; daiz0128@osstem.com; kkwa999@osstem.com).

Tomasz Szczepański, Michal K. Grzeszczyk, and Przemyslaw Korzeniowski are with the Sano Centre for Computational Medicine, 30-054 Kraków, Poland (e-mail: tmk.szczepanski@gmail.com; m.grzeszczyk@sanoscience.org; p.korzeniowski@sanoscience.org).

Vicent Caselles-Ballester is with the eHealth Center, Universitat Oberta de Catalunya, 08018 Barcelona, Spain (e-mail: vcasellesb@uoc.edu).

Xavier Paolo Burgos-Artizzu is with Movumtech S.L., 15003 Bajos, Spain (e-mail: xpburgos@movumtech.com).

Ferran Prados Carrasco is with the eHealth Center, Universitat Oberta de Catalunya, 08018 Barcelona, Spain, and also with the Center for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, and the Queen Square MS Centre, Department of Neuroinflammation, University College London, WC1E 7JE London, U.K. (e-mail: fpradosc@uoc.edu).

Bram van Ginneken is with the Department of Radiology, Radboud University Medical Center, 6500 HB Nijmegen, The Netherlands (e-mail: Bram.vanGinneken@radboudumc.nl).

Alexandre Anesi is with the Department of Medical and Surgical Sciences for Children and Adults, Cranio and Maxillo-Facial Unit, University of Modena and Reggio Emilia, 41121 Modena, Italy (e-mail: alexandre.anesi@unimore.it).

References

[1] K. Garlapati, D. B. G. Babu, N. C. S. K. Chaitanya, H. Guduru, A. Rembers, and P. Soni, "Evaluation of preference and purpose of utilisation of cone beam computed tomography (CBCT) compared to orthopantomogram (OPG) by dental practitioners—A cross-sectional study," *Polish J. Radiol.*, vol. 82, pp. 248–251, Feb. 2018.

[2] P. Tapia, G. Matus-Miranda, F. Díaz, and P. Arrué, "Preservation of the inferior alveolar vasculonervous bundle in mandibular resective therapies: Systematic review and report of two cases," *Medicina Oral, Patologia Oral Y Cirugia Bucal*, vol. 29, no. 2, p. 26239, Oct. 2023.

[3] S. Tereshchuk, S. Y. Ivanov, and V. Sukharev, "Inferior alveolar nerve preservation during resection and reconstruction of the mandible for benign tumors as a factor improving patient's quality of life," *J. Cranio-Maxillofacial Surg.*, vol. 50, no. 5, pp. 393–399, May 2022.

[4] J. Iwanaga, Y. Matsushita, T. Decater, S. Ibaragi, and R. S. Tubbs, "Mandibular canal vs. Inferior alveolar canal: Evidence-based terminology analysis," *Clin. Anatomy*, vol. 34, no. 2, pp. 209–217, Mar. 2021.

[5] J. D. Nguyen and H. Duong, *Anatomy, Head and Neck: Alveolar Nerve*. Saint Petersburg, FL, USA: StatPearls Publishing, 2023.

[6] P. Pitros, I. Jackson, and N. O'Connor, "Coronectomy: A retrospective outcome study," *Oral Maxillofacial Surg.*, vol. 23, no. 4, pp. 453–458, Dec. 2019.

[7] J. S. Schenkel, C. Jacobsen, C. Rostetter, K. W. Grätz, M. Rücker, and T. Gander, "Inferior alveolar nerve function after open reduction and internal fixation of mandibular fractures," *J. Cranio-Maxillofacial Surg.*, vol. 44, no. 6, pp. 743–748, Jun. 2016.

[8] M. Schlund, P. Grall, J. Ferri, and R. Nicot, "Effect of modified bilateral sagittal split osteotomy on inferior alveolar nerve neurosensory disturbance," *Brit. J. Oral Maxillofacial Surgery*, vol. 60, no. 8, pp. 1086–1091, Oct. 2022.

[9] A. Morrison, M. Chiarot, and S. Kirby, "Mental nerve function after inferior alveolar nerve transposition for placement of dental implants," *J.-Can. Dental Assoc.*, vol. 68, no. 1, pp. 46–50, 2002.

[10] J. Jaskari et al., "Deep learning method for mandibular canal segmentation in dental cone beam computed tomography volumes," *Sci. Rep.*, vol. 10, no. 1, p. 5842, Apr. 2020.

[11] G. H. Kwak et al., "Automatic mandibular canal detection using a deep convolutional neural network," *Sci. Rep.*, vol. 10, no. 1, p. 5711, Mar. 2020.

[12] P. Lahoud et al., "Development and validation of a novel artificial intelligence driven tool for accurate mandibular canal segmentation on CBCT," *J. Dentistry*, vol. 116, Jan. 2022, Art. no. 103891.

[13] M. Cipriano et al., "Deep segmentation of the mandibular canal: A new 3D annotated dataset of CBCT volumes," *IEEE Access*, vol. 10, pp. 11500–11510, 2022.

[14] M. Usman et al., "Dual-stage deeply supervised attention-based convolutional neural networks for mandibular canal segmentation in CBCT scans," *Sensors*, vol. 22, no. 24, p. 9877, Dec. 2022.

[15] S.-Y. Chun et al., "Automatic classification of 3D positional relationship between mandibular third molar and inferior alveolar canal using a distance-aware network," *BMC Oral Health*, vol. 23, no. 1, p. 794, Oct. 2023.

[16] D. Kainmueller, H. Lamecker, H. Seim, M. Zinser, and S. Zachow, "Automatic extraction of mandibular nerve and bone from cone-beam CT data," in *Proc. 12th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, London, U.K. Cham, Switzerland: Springer, Sep. 2009, pp. 76–83.

[17] F. Abdolali, R. A. Zoroofi, M. Abdolali, F. Yokota, Y. Otake, and Y. Sato, "Automatic segmentation of mandibular canal in cone beam CT images using conditional statistical shape model and fast marching," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 12, no. 4, pp. 581–593, Apr. 2017.

[18] X. Wei and Y. Wang, "Inferior alveolar canal segmentation based on cone-beam computed tomography," *Med. Phys.*, vol. 48, no. 11, pp. 7074–7088, Nov. 2021.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, Munich, Germany. Cham, Switzerland: Springer, 2015, pp. 234–241.

[20] M. Cipriano, S. Allegretti, F. Bolelli, F. Pollastri, and C. Grana, "Improving segmentation of the inferior alveolar nerve through deep label propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21105–21114.

[21] H. Zhao, J. Chen, Z. Yun, Q. Feng, L. Zhong, and W. Yang, "Whole mandibular canal segmentation using transformed dental CBCT volume in frenet frame," *Heliyon*, vol. 9, no. 7, Jul. 2023, Art. no. e17651.

[22] J. Lv, L. Zhang, J. Xu, W. Li, G. Li, and H. Zhou, "Automatic segmentation of mandibular canal using transformer based neural networks," *Frontiers Bioeng. Biotechnol.*, vol. 11, Nov. 2023, Art. no. 1302524.

[23] L. Lumetti, V. Pipoli, F. Bolelli, E. Ficarra, and C. Grana, "Enhancing patch-based learning for the segmentation of the mandibular canal," *IEEE Access*, vol. 12, pp. 79014–79024, 2024.

[24] K. Hung, A. W. K. Yeung, R. Tanaka, and M. M. Bornstein, "Current applications, opportunities, and limitations of AI for 3D imaging in dental research and practice," *Int. J. Environ. Res. Public Health*, vol. 17, no. 12, p. 4424, Jun. 2020.

[25] C. Mercadante et al., "A cone beam computed tomography annotation tool for automatic detection of the inferior alveolar nerve canal," in *Proc. 16th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, vol. 4. Setúbal Municipality, Portugal: SciTePress, 2021, pp. 724–731.

[26] L. Lumetti, V. Pipoli, F. Bolelli, and C. Grana, "Annotating the inferior alveolar canal: The ultimate tool," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, Jan. 2023, pp. 525–536.

[27] H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 4, pp. 551–559, Jul. 1983.

[28] L. Wang, D. Li, H. Liu, J. Peng, L. Tian, and Y. Shan, "Cross-dataset collaborative learning for semantic segmentation in autonomous driving," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 3, pp. 2487–2494.

[29] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "MS-Net: Multi-site network for improving prostate segmentation with heterogeneous MRI data," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2713–2724, Sep. 2020.

[30] F. Isensee et al., "NnU-net: Self-adapting framework for U-Net-based medical image segmentation," 2018, *arXiv:1809.10486*.

[31] S. Allegretti, F. Bolelli, M. Cancilla, F. Pollastri, L. Canalini, and C. Grana, "How does connected components labeling with decision trees perform on GPUs?" in *Proc. Comput. Anal. Images Patterns*, Jan. 2019, pp. 39–51.

[32] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "ST++: Make self-trainingWork better for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4258–4267.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[34] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[35] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 165–174.

[36] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4455–4465.

[37] F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Comput. Methods Programs Biomed.*, vol. 208, Sep. 2021, Art. no. 106236.

[38] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[39] L. Cabaret, L. Lacassagne, and D. Etiemble, "Parallel light speed labeling: An efficient connected component algorithm for labeling and analysis on multi-core processors," *J. Real-Time Image Process.*, vol. 15, no. 1, pp. 173–196, Jun. 2018.

[40] S. Shit et al., "ClDice—A novel topology-preserving loss function for tubular structure segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16555–16564.

[41] D. Karimi and S. E. Salcudean, "Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 499–513, Feb. 2020.

[42] Y. He, V. Nath, D. Yang, Y. Tang, A. Myronenko, and D. Xu, "SwinUNETR-V2: Stronger Swin transformers with stagewise convolutions for 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2023, pp. 416–426.

[43] J. Kalkhof and A. Mukhopadhyay, "M3D-NCA: Robust 3D segmentation with built-in quality control," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2023, pp. 169–178.

[44] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Commun.*, vol. 15, no. 1, p. 654, Jan. 2024.

[45] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, Dec. 2022, Art. no. 100258.

[46] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 241–258, Apr. 2020.

[47] F. Pollastri et al., "A deep analysis on high-resolution dermoscopic image classification," *IET Comput. Vis.*, vol. 15, no. 7, pp. 514–526, Oct. 2021.

[48] L. Maier-Hein et al., "Metrics reloaded: Recommendations for image analysis validation," *Nature Methods*, vol. 21, no. 2, pp. 195–212, Feb. 2024.

[49] L. Maier-Hein et al., "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nat. Commun.*, vol. 9, no. 1, p. 5217, 2018.

[50] M. Wiesenfarth et al., "Methods and open-source toolkit for analyzing and visualizing challenge results," *Sci. Rep.*, vol. 11, no. 1, p. 2369, Jan. 2021.

[51] S. Li et al., "Transformer-based tooth segmentation, identification and pulp calcification recognition in CBCT," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2023, pp. 706–714.

[52] S. Kim, I. Song, and S. J. Baek, "Automatic segmentation of internal tooth structure from CBCT images using hierarchical deep learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2023, pp. 703–713.

[53] S. Park, S. Kim, I. Song, and S. J. Baek, "3D teeth reconstruction from panoramic radiographs using neural implicit functions," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2023, pp. 376–386.

[54] B. Kirnbauer, A. Hadzic, N. Jakse, H. Bischof, and D. Stern, "Automatic detection of periapical osteolytic lesions on cone-beam computed tomography using deep convolutional neuronal networks," *J. Endodontics*, vol. 48, no. 11, pp. 1434–1440, Nov. 2022.

[55] D.-M. Son, Y.-A. Yoon, H.-J. Kwon, C.-H. An, and S.-H. Lee, "Automatic detection of mandibular fractures in panoramic radiographs using deep learning," *Diagnostics*, vol. 11, no. 6, p. 933, May 2021.

[56] P. Liu, Y. Sun, X. Zhao, and Y. Yan, "Deep learning algorithm performance in contouring head and neck organs at risk: A systematic review and single-arm meta-analysis," *Biomed. Eng. OnLine*, vol. 22, no. 1, p. 104, Nov. 2023.

[57] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, Feb. 2020.

[58] Y. Qi, Y. He, X. Qi, Y. Zhang, and G. Yang, "Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6070–6079.

[59] Y. Kirchhoff et al., "Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures," in *Proc. 18th Eur. Conf. Comput. Vis.*, Milan, Italy, Jan. 2024, pp. 218–234.

Open Access funding provided by 'Università degli Studi di Modena e Reggio Emilia' within the CRUI CARE Agreement