Medical Image Analysis (2026)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Multi-Structure Segmentation in CBCT Volumes: the ToothFairy2 Challenge

Federico Bolelli[1,2,*], Luca Lumetti[1,2], Niels van Nistelrooij[1,2], Shankeeth Vinayahalingam[2], Mattia Di Bartolomeo[2], Kevin Marchesini[2], Arrigo Pellacani, Ettore Candeloro, Gabriele Rosati, Tong Xi, Fabian Isensee, Yannick Kirchhoff, Lars Krämer, Maximilian Rokuss, Constantin Ulrich, Klaus Maier-Hein, Yuxian Jiang, Yusheng Liu, Lisheng Wang, Haoshen Wang, Siyu Chen, Zhiming Cui, Pengcheng Shi, Zhaohong Pan, Xiaokun Liang, Qi Ma, Ender Konukoglu, Marek Wodzinski, Henning Müller, Haipeng Mai, Xiaobing Dang, Shrajan Bhandary, Radu Grosu, Stefaan Bergé, Alexandre Anesi[2], Costantino Grana[2]

[a] *For space and visual constraints, affiliations are reported at the end of the document.*

ARTICLE INFO

Article history:
Received XX Month XXX
Received in final form XX Month XXX
Accepted XX Month XXX
Available online XX Month XXX

Communicated by XXXX

Keywords: Multi-class Segmentation, Tooth, CBCT, ToothFairy

ABSTRACT

Cone-beam computed tomography (CBCT) is widely used for dento-maxillofacial diagnostics and treatment planning, and comprehensive multi-structure segmentation remains time-consuming, limiting large-scale, reproducible research. In this article, we present ToothFairy2, a MICCAI 2024 challenge on multi-structure segmentation in maxillofacial CBCT. The accompanying dataset comprises 530 CBCT volumes (480 public training, 50 hidden test) with expert 3D annotations of 42 classes, including maxilla, mandible, crowns, bridges, implants, inferior alveolar canals, maxillary sinuses, pharynx, and teeth using the International Tooth Numbering System (FDI). 26 international teams participated in ToothFairy2, and their methods were run and evaluated for voxel-wise multi-class segmentation using a standardized protocol. This report extends the evaluation of teeth to also investigate the current capabilities of tooth detection and FDI numbering. Furthermore, ranking stability was analyzed to assess the robustness of the final challenge outcome.

Overall, challenge participants achieved consistently high performance for large, high-contrast structures such as jawbones, pharynx, and most teeth, while maxillary sinuses, dental restorations, and fine structures remain challenging due to class imbalance and metal artifacts. Analysis of tooth-related metrics further revealed that assigning correct FDI numbers was more challenging than delineating individual teeth. By releasing CBCT data, 3D annotations, baseline models, and evaluation code, ToothFairy2 establishes a long-term benchmark to drive the development of automated methods for robust, clinically meaningful multi-structure segmentation in maxillofacial CBCT.

1. Introduction

Clinical background and role of CBCT. Various medical imaging techniques are routinely employed for the diagnosis and treatment of maxillofacial conditions. Two-dimensional (2D) panoramic X-rays, 3D intra-oral scans (IOS), and 3D cone-beam computed tomography (CBCT) are widely accessible and fundamental to contemporary clinical workflows (Kaasalainen et al., 2021). Among these, CBCT has become a pivotal modality in dental and maxillofacial diagnostics and treatment planning, and its use is rapidly expanding across head and neck surgical specialties. CBCT usually offers isotropic volumetric information on all orofacial structures

*Corresponding author. *e-mail*: federico.bolelli@unimore.it
[1]Equal contribution.
[2]Member of the challenge organizing team.

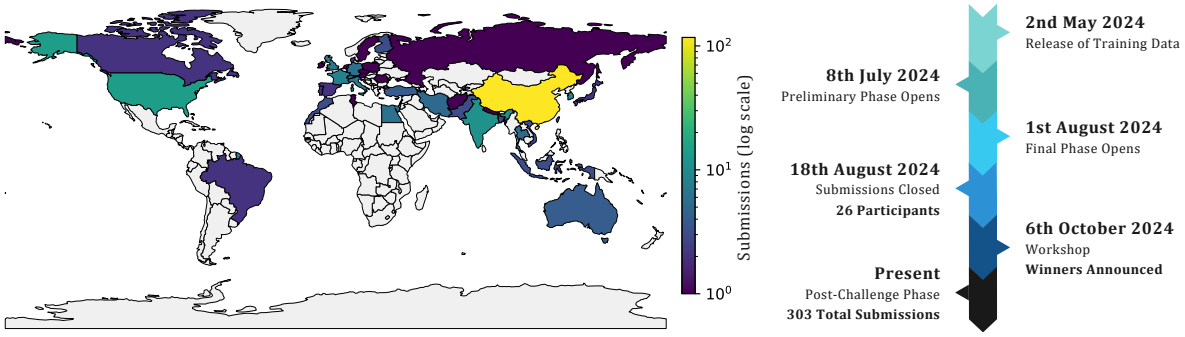

Fig. 1: On the left, the challenge submissions heatmap based on participant locations. On the right, the ToothFairy2 Challenge timeline.

with shorter acquisition times and lower radiation doses than conventional CT, while preserving excellent resolution of high-density structures such as bone and teeth. These characteristics make CBCT an ideal tool for dentistry and surgical contexts where sub-millimeter precision is crucial and impacts postoperative quality of life, including orthognathic surgery, trauma management, implantology, and airway analysis (Worthington, 2004; Cui et al., 2022b). In these scenarios, precise 3D models of anatomical structures are essential for preoperative planning, intraoperative guidance, and postoperative evaluation.

Critical maxillofacial structures in CBCT. One of the key structures frequently assessed in CBCT scans is the inferior alveolar canal (IAC), a bony channel extending from the mandibular foramen to the mental foramen and housing the inferior alveolar neurovascular bundle, including the inferior alveolar nerve (IAN). The IAC is often in proximity to impacted wisdom teeth, other impacted teeth, or cystic lesions, and must be preserved during surgeries. IAN injury can lead to persistent sensory impairment and a significant reduction in quality of life, making precise preoperative localization of the IAC mandatory in many dental and maxillofacial procedures. The IAC, however, is only one of several structures whose accurate identification and preservation are critical. Tooth identification is a key step in numerous surgical and non-surgical interventions, and CBCT is frequently used to detect artificial structures such as dental implants, crowns, and bridges. In implantology, accurate assessment of jawbone geometry must be performed in relation to surrounding critical structures—adjacent teeth, existing implants, the maxillary sinus, and the IAC for mandibular procedures. Moreover, CBCT allows visualization of relevant soft-tissue structures such as the pharynx, further enhancing its value for diagnosis and preoperative planning (Worthington, 2004; Cui et al., 2022b).

Need for automated 3D segmentation and limitations of existing datasets. Despite its clinical utility, manual segmentation of CBCT scans to delineate these structures is labor-intensive, time-consuming, and requires specialized expertise. Deep learning-based segmentation methods have therefore emerged as a promising approach to reduce expert workload, increase efficiency, and improve consistency of segmentations (Morrison et al., 2002). However, training robust models with good generalization requires large, high-quality, and fully annotated datasets (Litjens et al., 2017; Pollastri et al., 2021b; Isensee et al., 2021). Existing maxillofacial datasets are typically limited in scope, focusing on a small number of structures (e.g., only teeth or only the IAC) and often lack full 3D volumetric annotations. Moreover, certain structures require specialized annotation workflows; for instance, the inferior alveolar canal, due to its elongated and tortuous course, is difficult to delineate reliably from axial slices alone and can remain challenging even with multiplanar (sagittal/coronal) views. Consequently, dedicated tools for IAC annotation have been proposed to reduce annotation time and improve accuracy (Mercadante et al., 2021; Lumetti et al., 2023). Approaches that do exploit 3D CBCT data are usually trained on relatively small cohorts, often < 50 scans (Cui et al., 2022a; Chun et al., 2023; Dou et al., 2022), and many datasets are not publicly available (Hao et al., 2025). This combination of limited scale, restricted anatomical coverage, and closed access hinders reproducibility, complicates fair comparison of automated methods, and limits generalization across different acquisition protocols and populations.

From ToothFairy to ToothFairy2. To address some of these limitations, the ToothFairy Challenge and dataset (Bolelli et al., 2025a) were introduced in 2023 as the first public benchmark for IAC segmentation in CBCT. The dataset includes 443 scans with expert annotations of the IAC, provided as 2D masks for all scans and 3D annotations for a subset of 153 scans, and has already enabled the development and evaluation of deep learning models for IAC segmentation. However, its labels are restricted to the IAC and ignore other clinically relevant orofacial structures that are routinely evaluated in clinical practice.

In this work, we introduce the ToothFairy2 Challenge, launched in 2024 and hosted by the MICCAI 2024 conference, with a substantially broader clinical scope and technical ambition. An overview of the challenge timeline and the geographic distribution of participating teams is shown in Fig. 1. Rather than focusing on a single anatomical target with mixed 2D/3D annotations, ToothFairy2 defines a multi-class, fully 3D, automated segmentation task covering 42 distinct anatomical structures commonly visible in maxillofacial CBCT volumes. These structures include both hard and soft tissues, such as the mandible, maxilla, upper and lower teeth (including restorations and implants), the pharynx, and bilateral instances of the IAC and maxillary sinuses. The associated dataset comprises

530 3D volumes, of which 480 are publicly available for training, and 50 are held out as a private, non-public test set for long-term benchmarking. By providing dense voxel-level annotations and a standardized evaluation framework, ToothFairy2 pushes the development of clinically meaningful, generalizable, and reproducible 3D multi-class segmentation algorithms for maxillofacial CBCT.

Scope and contributions. Beyond describing the dataset and challenge design, this article analyzes the methods proposed by ToothFairy2 participants and provides an in-depth comparison of their performance and algorithmic traits. Specifically, we:

- **describe the ToothFairy2 challenge design**, including the rationale for the training-test split, the task definition, and the adopted evaluation metrics;
- **introduce the ToothFairy2 dataset** with full 3D multi-structure annotations and a private non-public test set accessible only via *Grand Challenge*, establishing a common benchmark for fair comparison of future methods;
- **compare state-of-the-art methods for 3D multi-class CBCT segmentation** proposed by challenge participants, highlighting design choices, strengths, and limitations, and outlining directions for further research;
- **release an open-source repository** containing the challenge evaluation software, implementations of the submitted models, and their pre-trained weights, to facilitate transparency and reproducibility.[3]

The remainder of this paper is organized as follows. Sec. 2 reviews existing maxillofacial datasets and related work on CBCT segmentation. Sec. 3 details the ToothFairy2 dataset and annotation protocol. Sec. 4 describes the participating methods, and Sec. 5 presents the evaluation protocol. Results and a comprehensive discussion are given in Sec. 6, while Sec. 7 outlines limitations and future work, and Sec. 8 concludes the paper.

2. Related Work

This section reviews prior work on automatic analysis of dental and maxillofacial CBCT imaging, focusing on tasks that closely match the scope of the ToothFairy2 challenge. We cover teeth and inferior alveolar canal segmentation, summarizing binary, instance-level, and multi-class approaches, as well as recent CNN-, Transformer-, and Mamba-based 3D medical image segmentation models, including emerging foundational and interactive frameworks. We also compare existing maxillofacial datasets and their limitations, motivating the comprehensive ToothFairy2 benchmark proposed in this work.

2.1. Existing Datasets

In maxillofacial imaging, AI research has largely focused on pushing model performance on carefully curated cohorts, while the underlying data resources themselves have received comparatively little critical analysis (Sengupta et al., 2022). Historically, most studies have relied on institutional or academic collections that are difficult to access externally and are not explicitly designed with the FAIR (Findable, Accessible, Interoperable, Reusable) principles in mind (Wilkinson et al., 2016). As a result, only a limited number of studies release the accompanying datasets publicly, and only a few studies provide results for a standardized benchmark (Cui et al., 2022b; Cipriano et al., 2022b; Cui et al., 2022a; Bolelli et al., 2025a,b), underscoring the need for high-quality, clinically validated, and AI-ready maxillofacial datasets.

In recent years, the number of openly available dental and maxillofacial CBCT collections has begun to grow (Liu et al., 2025; Wang et al., 2025a; Li, 2024; Huang et al., 2024). However, most of these resources remain tailored to narrow clinical indications or specific anatomical subregions, such as the IAC (Cipriano et al., 2022b), the pterygopalatine canal (Li et al., 2026), or the dentition (Cui et al., 2022b; Hao et al., 2026). Furthermore, many existing datasets either lack expert annotations or provide only weak labels. Typical limitations include a restricted field of view, relatively small cohort sizes, and coarse labeling schemes, e.g., binary masks instead of instance-level labels for individual teeth (Cui et al., 2022b). A comparative overview of the datasets most commonly used in state-of-the-art studies is reported in Tab. 1.

2.2. Teeth Segmentation

Teeth segmentation in 3D dental and maxillofacial imaging has been tackled with a variety of methodological paradigms, each tailored to different clinical objectives and levels of anatomical detail (Chen et al., 2024). Existing approaches can be grouped into binary, instance, and multi-class formulations.

Binary segmentation. In a binary semantic segmentation setting, the goal is to separate all dental structures from surrounding tissues without discriminating between individual teeth. Although adequate for basic assessments, its clinical utility is limited, particularly when neighboring teeth are in close proximity. To improve performance, recent work has adopted CNN backbones with custom modules and hybrid loss terms (Hu et al., 2024), combined 2D/3D networks (Hsu et al., 2022), and post-processing steps such as posterior-probability maps and dense conditional random fields (Rao et al., 2020).

Instance segmentation. In contrast, instance segmentation methods treat each tooth as an independent instance, without yet assigning anatomical labels. Most approaches in the literature follow a two-stage pipeline: a detection stage that localizes each tooth using bounding-box regression (Cui et al., 2019; Duan et al., 2021), heat-map peaks (Wu et al., 2020), or offset-vector regression (Cui et al., 2021; Dou et al., 2022; Cui et al., 2022b), followed by binary segmentation on cropped regions to delineate precise boundaries. More recently, Hao et al. (2026) proposed a frequency-driven approach that leverages bipositional encodings within the VMamba (Liu et al., 2024b) architecture to enhance spatial localization and robustness to the intrinsic high-noise, low contrast nature of CBCT scans.

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

Table 1: Datasets used in literature to segment maxillofacial anatomical structures. Our analysis is mainly focused on 3D imaging modalities, i.e., CBCTs and IOS. ✣ means that only a portion of the training data is available to the research community, i.e., 148 over 4 531. Among these, only 97 are effectively usable for the training of automatic algorithms.

Anatomical Structure(s)	Image Modality	Authors	Country	# Labels	# Train & Val	# Test	Label Type		Public
							Train	Test	
Inferior Alveolar Canal (IAC)	CBCT	Jaskari et al. (2020)	Finland	2	509	128	2D	128 2D, 15 3D	✗
		Lahoud et al. (2022)	Belgium	1	205	30	3D	3D	✗
		Usman et al. (2022)	South Korea	2	510	500	3D	3D	✗
		Cipriano et al. (2022a)	Italy	1	332	15	332 2D, 76 3D	15 2D, 15 3D	✓
		Chun et al. (2023)	South Korea	2	32	18	3D	3D	✗
		Bolelli et al. (2025a)	Italy, Netherlands	1	443	50	290 2D, 153 2D + 3D	3D	✓
Teeth	X-Ray	Hao et al. (2026)	China	1	5 000	1 225	2D	2D	✓
	X-Ray & CBCT	Wang et al. (2026)	China	32	2 710	40	330 3D, 2 380 2D	20 3D, 20 2D	✓
Teeth Crown, Teeth Root	CBCT	Cui et al. (2022b)	China	34	4 531	407	3D	3D	✣
		Cui et al. (2022a)	China	1	22	~	3D	~	✓
		Dou et al. (2022)	China	32	35	5	3D	3D	✗
		Jang et al. (2021)	China	32	66	7	66 3D, 31 2D	7 3D, 4 2D	✗
Teeth Crown	IOS	Ben-Hamadou et al. (2023)	Tunisia	32	1 200	600	3D	3D	✓
		Vinayahalingam et al. (2023)	Netherlands	32	1 400	350	3D	3D	✗
IACs Palatine Canal	CBCT	Li et al. (2026)	China	2	153	38	3D	3D	✓
IACs, Teeth, Others	CBCT	Ours	Italy, Netherlands	42	480	50	3D	3D	✓

Multi-class segmentation. In multi-class segmentation, each tooth must be distinctly segmented and assigned a specific label (Lahoud et al., 2021; Shaheen et al., 2021; van Nistelrooij et al., 2025), typically following the FDI (Fédération Dentaire Internationale, or World Dental Federation) notation (ISO, 2016). Among the three segmentation strategies, this is the most challenging, particularly in anatomically complex scenarios involving missing, impacted, or supernumerary teeth. Progress in this area critically depends on the availability of richer datasets and more sophisticated model designs, further underscoring the need for our proposed challenge and dataset.

The existing body of work can be broadly categorized into single-stage and multi-stage approaches. Single-stage methods employ an end-to-end architecture to directly predict the final labels, whereas multi-stage pipelines introduce intermediate steps. A widely adopted strategy involves down-sampling the CBCT volume or dividing it into anatomical subregions (e.g., the four dental arch quadrants) to perform an initial coarse segmentation, followed by per-tooth refinement using region-of-interest extraction and local detail enhancement (Shaheen et al., 2021; Wang et al., 2023; Rekik et al., 2025; Farhat et al., 2025).

2.3. IAC Segmentation

Since the introduction of CBCT technology in the early 2000s, considerable research has focused on developing automated methods for segmenting the inferior alveolar canal (IAC) from 3D scans (Schramm et al., 2005). With the advent of deep learning in medical image analysis, data-driven approaches (Hwang et al., 2019; Jaskari et al., 2020; Järnstedt et al., 2023; Di Bartolomeo et al., 2023; Lumetti et al., 2024b) have significantly outperformed traditional computer vision techniques (Kainmueller et al., 2009; Kroon, 2011; Moris et al., 2012; Blacher et al., 2016; Abdolali et al., 2017).

One of the earliest deep learning approaches, introduced by Jaskari et al. (2020), employed the U-Net (Ronneberger et al., 2015) model on a coarsely annotated dataset, showing promising improvements over conventional methods despite the lack of dense voxel-level annotations.

To address the extreme class imbalance between the mandibular canal and surrounding structures, Usman et al. (2022) introduced a two-stage U-Net pipeline that first identifies regions of interest (ROIs) before applying segmentation. Complementarily, Zhao et al. (2023) proposed a method based on Frenet frames to better preserve the anatomical continuity of the canal during segmentation. Building on Cipriano's label expansion framework, Lv et al. (2023) developed a transformer-based model incorporating adaptive image processing and a "deep label fusion" strategy to enforce consistency across sparsely labeled data.

More recently, Lumetti et al. (2024a) addressed limitations of patch-wise training by integrating a memory-augmented transformer encoder into the U-Net bottleneck, thereby enriching spatial context and improving segmentation quality.

Despite recent progress, IAC segmentation remains a complex and open research challenge. Continued advancement in this domain will strongly depend on access to comprehensive, high-quality, and publicly available 3D datasets.

2.4. General-purpose 3D Medical Segmentation Architectures

Methods submitted to the ToothFairy2 challenge build on general-purpose 3D medical image segmentation frameworks, rather than architectures designed specifically for dental CBCT. We therefore briefly review the main classes of models that underpin the participant solutions.

Convolutional architectures. Convolutional encoder-decoder networks remain a strong baseline for volumetric segmentation. In particular, the nnU-Net framework (Isensee et al., 2021) automatically configures a U-Net-like architecture, preprocessing, and training scheme for a given task and has become a widely adopted standard across diverse 3D benchmarks.

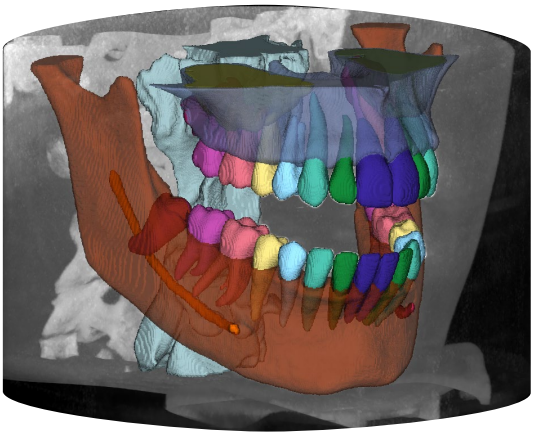

Fig. 2: CBCT sample with annotations from the ToothFairy2 Dataset.

Building on this, nnU-Net ResEnc (Isensee et al., 2024) introduces residual encoder blocks and deeper architectures, and has repeatedly shown improved performance in multi-structure segmentation tasks.

Transformer- and Mamba-based architectures. Inspired by advances in natural language processing, several methods integrate self-attention or state-space models into U-Net-like designs to better capture long-range dependencies in 3D volumes. Examples include TransU-Net (Chen et al., 2021), UNETR++ (Shaker et al., 2024), and the more recent nnFormer (Zhou et al., 2023) for Transformer-centric designs, and UMamba (Ma et al., 2024c), VMamba (Liu et al., 2024b), Swin-UMamba (Liu et al., 2024a), and Taming-Mamba (Lumetti et al., 2025) models for Mamba-based state-space models, which offer a more computationally efficient alternative to full self-attention in large volumes.

Foundation and interactive models. More recently, foundation models and promptable interactive frameworks have been proposed for medical image segmentation, such as nnInteractive (Isensee et al., 2025b), MedSAM (Ma et al., 2024a,b), and VISTA3D (He et al., 2024), which leverage training across many heterogeneous datasets and tasks. While ToothFairy2 focuses on fully automatic, task-specific segmentation, such models are increasingly used to accelerate annotation and to support future interactive benchmarks.

As mentioned, these general-purpose architectures form the backbone of many submissions to ToothFairy2. In Sec. 4, we analyze how different instantiations of nnU-Net and related variants, combined with task-specific pre- and post-processing, impact performance on our multi-structure CBCT benchmark.

3. The ToothFairy2 Dataset

This section introduces the *ToothFairy2 dataset* provided for the challenge, detailing the data collection methodology and the procedures followed for annotation.

Data sources. The *training data* (480 scans) was collected at the Affidea Center, a pan-European healthcare organization based in Modena, specializing in advanced diagnostics, laboratory services, rehabilitation, and oncology care. The scans were acquired using cone-beam computed tomography with a *NewTom/NTVGiMK4* scanner, operating at 3 mA and 110 kV, and producing isotropic voxels of 0.3 mm. The *test data* (50 scans) was provided by the Department of Oral and Maxillofacial Surgery at Radboud University Medical Center in Nijmegen, the Netherlands. These scans were obtained using the *i-CAT 3D Imaging System* following a standard CBCT protocol in "Extended Field" mode, with a field of view (FOV) of 16 cm in diameter and 22 cm in height. Each patient underwent two consecutive 20-second scans, resulting in a voxel resolution of 0.4 mm, which was subsequently rescaled to align with the training data prior to annotation.

Patient cohorts. All patient data was anonymized, retaining only limited demographic information—specifically gender, age, and the year of the scan. In total, 58.30% of the patients are female, with a distribution of 58.54% in the training set and 56.00% in the test set. The scans were acquired between 2019 and 2024, and patient age ranges from 11 to 100 years old. In the training set, the most represented ages are in the (20-30] and (60-70] ranges, while the test set is dominated by the (50-70] range.

Each CBCT scan in the ToothFairy2 dataset corresponds to a distinct patient and has been annotated by a single expert.

Annotation protocol and tools. The dataset is fully annotated in 3D, with slice-wise annotations provided by 7 maxillofacial specialists (>5 years of experience), making it the largest maxillofacial dataset with volumetric annotations available as of the challenge release date. The slice-wise approach was chosen because voxel labeling directly in 3D views usually leads to ambiguity and higher error rates. To reduce the jagged contours often produced when segmenting 3D structures in 2D slices, annotation was carried out in multiple stages, starting from the axial (transversal) view and then refined through sagittal and coronal (frontal) planes. The annotated classes (and their corresponding IDs) are: lower and upper jawbone (1 and 2), left and right inferior alveolar canals (3 and 4), left and right maxillary sinuses (5 and 6), pharynx (7), bridges (8), crowns (9), implants (10), and both upper and lower teeth, including wisdom teeth (following FDI notation, 11-48). Most structures appear in every scan, with the exception of crowns, bridges, implants, some missing teeth, and sinuses, present in only ∼15% of the training scans, due to the vertical field of view being reduced during anonymization to remove patient-identifying regions. The distribution of the classes across the scans is detailed in Fig. 3.

To limit the risk of overfitting to annotator-specific biases and to properly evaluate generalization, annotations were split so that five experts worked on the training set, while the other two were dedicated exclusively to the test set. To ease the annotation workload while maintaining high label quality, we employed a semi-automated strategy under full clinical supervision, providing clinicians with preliminary predictions that they could adjust, correct, or replace. Specifically, five distinct models based on the nnU-Net framework were used to generate initial segmentations, which were then refined by experts. Each base model was trained to segment particular anatomical structures or groups, including the jawbone, inferior alveolar canals (left/right), maxillary sinuses, pharynx, and teeth. The process followed an iterative cycle: after every 20 newly an-

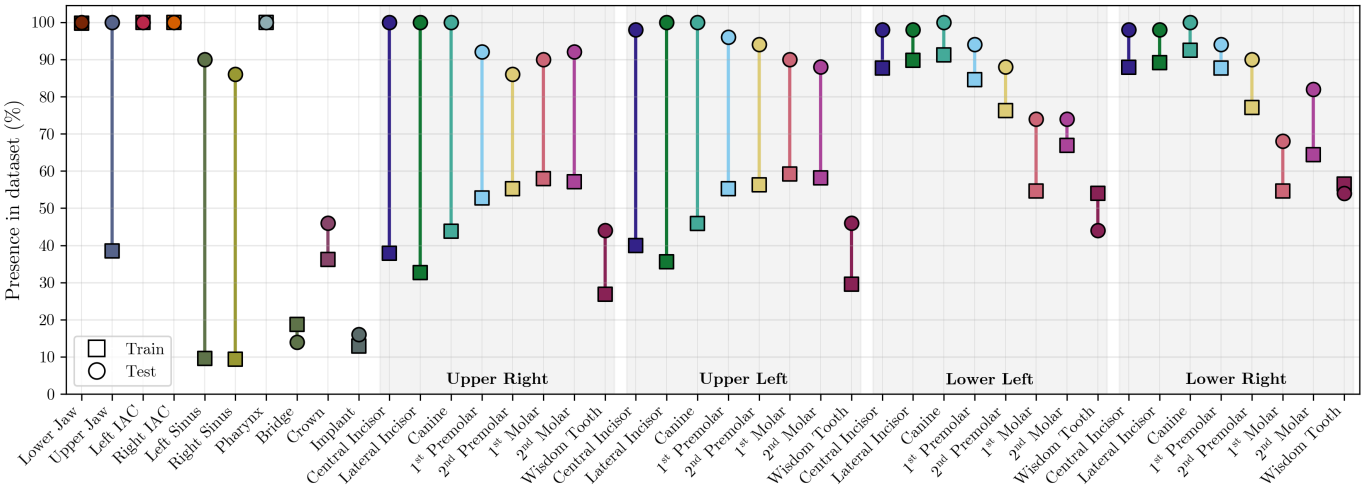

Fig. 3: Difference in class prevalence for the ToothFairy2 dataset between train □ and test ○ sets. Colors are the same employed in the visualization of Fig. 2.

notated volumes, the base models were retrained from scratch with the updated data, progressively improving segmentation quality and reducing manual effort in subsequent iterations. For canals and dental structures, we leveraged publicly available datasets (Cipriano et al., 2022b; Cui et al., 2022b) as training sets. Since no public datasets provide labels for the jawbone, maxillary sinuses, or pharynx, these structures were manually annotated from scratch in the first 20 volumes.

Error sources related to the annotation. Despite expert supervision during the refinement of the semi-automated labels, some error sources can affect the quality of volumetric annotations (Nagarajappa et al., 2015). For teeth, crowns, bridges, and implants, one of the main challenges is the presence of metal artifacts in CBCT scans, which introduce streaking and distort voxel intensities, making accurate identification difficult. Similarly, tooth roots are often poorly distinguishable from surrounding alveolar bone due to limited contrast, leading to potential boundary uncertainties. For the inferior alveolar canals, local disconnections may appear where the canal is not clearly separable from the jawbone because of acquisition noise or patient-specific bone density variations. Annotating soft-tissue structures, such as the pharynx, introduces further variability due to motion artifacts from swallowing or breathing during acquisition, which can cause blurred or irregular boundaries. Jawbones, generally easier to delineate, can still lead to jagged or inconsistent surfaces when small cavities or trabecular structures are present.

Compared to annotating from scratch, the semi-automated strategy reduced variability by providing consistent model-based priors, which the annotators could correct. To further mitigate errors, annotations were performed slice-wise with multi-view refinement and were repeatedly cross-checked across the different views. Then, we also implemented a post-validation step: after completing the manual refinement process, we retrained the same base models, originally used to generate preliminary annotations, on the enlarged set with refined annotations. These models were then used to perform inference on the entire dataset, and the resulting predictions were compared with the corresponding ground-truth masks. In cases where substantial discrepancies were detected (e.g., missing anatomical parts, disconnected structures, or unexpected boundaries), the scans were flagged and returned to the clinical experts for re-evaluation. This iterative validation loop allowed us to identify and correct subtle inconsistencies that might otherwise have remained undetected, thereby further improving the reliability of the final annotations.

Still, some residual imperfections remain unavoidable; in particular, annotations of crowns, bridges, and implants may sometimes overestimate or underestimate boundaries, especially when metallic artifacts overlap with adjacent teeth. Thus, while ToothFairy2 provides a comprehensive and clinically validated ground truth, users of the dataset should be aware of these error sources and take them into account when designing segmentation algorithms.

Dataset format and comparison with ToothFairy. The ToothFairy2 dataset is an extension of the previously released ToothFairy dataset by Bolelli et al. (2025a), increasing the number of scans from 443 to 530. The original ToothFairy dataset contains a mix of 2D and 3D annotations of the IAC, marked as a single class label. Along with new data, the same CBCT scans were used in the ToothFairy2 dataset, where the annotations are fully 3D and semantic. For accessibility and ease of use, the resulting annotated volumes have been packed following the nnU-Net dataset format, which comprises three different components: raw images, corresponding segmentation maps, and a `dataset.json` file specifying the metadata. The class IDs are an extension of the FDI notation and include a total of 42 classes. Final 3D volumes and labels are provided in the `.mha` format for compliance with the Grand Challenge platform.

Ethics approval and data availability. The training data received the approval of the ethics committee of the Comitato Etico dell'Area Vasta Emilia Nord (Approval Number 1374/2020/OSS/ESTMO SIRER ID 1275 - NAICBCT-D) and can be downloaded under the CC BY-SA license after user registration at `https://ditto.ing.unimore.it/`. The Tooth-

Table 2: Final test phase leaderboard (valid submissions only). The ranking is determined following the Challenge ranking schema reported in Sec. 5.2.

Final Rank	ID	First Author	Country	Mean Position
1	A	F. Isensee, Y. Kirchhoff	Germany	4.6
2	B	Y. Jiang	China	4.8
3	C	H. Wang	China	5.2
4	D	A. Gao	China	6.1
5	E	P. Shi	China	7.3
6	F	K. Dmitriev	USA	7.9
6	G	C. Ma	China	7.9
8	H	Y. Yang	China	8.2
9	I	Z. Pan	China	8.5
10	J	M. Wang	China	8.6
11	K	H. Agrawal	Finland	9.9
12	L	Q. Ma	Switzerland	10.2
13	M	H. Wu	China	11.1
14	N	L. Lee	China	12.9
15	O	M. Wodzinski	Switzerland	13.3
16	P	F. Xiao	China	16.2
17	Q	L. Daza	Germany	17.7
18	R	M. Haipeng	China	19.1
19	S	W. Xulong	China	19.2
20	T	S. Bhandary	Austria	19.8
21	U	J. Ma	China	19.9
22	V	J. Yang	China	20
23	W	J. Han	South Korea	20.3
24	X	A. Qayyum	United Kingdom	22.5
25	Y	A. B. George	India	24.3
26	Z	C. Chen	China	24.5

Fairy2 test set is accessible through Grand Challenge via the *post challenge phase* and represents a common benchmark to allow for a long-term fair comparison of future methods.[4] The ToothFairy2 dataset has received 1 423 data download requests from unique users at the time of writing this article. The live download statistics are available on the dataset webpage.[5]

Although the CC BY-SA license permits redistribution through third-party repositories, we currently maintain our institutional platform as the single authoritative source in order to avoid version fragmentation and to ensure consistent updates, changelogs, and maintenance. Should other research groups build upon or extend the dataset, we would warmly encourage them to share the resulting data through the same platform as well, so as to preserve a single authoritative distribution point and facilitate consistent long-term stewardship.

4. Methods

A total of 29 unique teams uploaded their algorithm onto Grand Challenge starting from July 8th, 2024, to the present day, with 164 submissions seen in the preliminary phase and 132 submissions for the final test phase.[6] The number of submissions by country is depicted on Fig. 1. Moreover, Tab. 2 provides a brief synopsis of teams participating in the final phase and their final ranking.

[4] `https://toothfairy2.GrandChallenge.org/evaluation/post-challenge-phase-test-your-algorithm/leaderboard/`

[5] `https://ditto.ing.unimore.it/dataset_view/?dataset=toothfairy2`

[6] Numbers have been collected on July 31, 2025.

4.1. Participating Methods

Here, we describe the algorithms proposed by the best-performing teams. Sec. 4.2 highlights the commonalities and distinctive elements, and Tab. 3 summarizes them.

4.1.1. F. Isensee, Y. Kirchhoff et al.

The proposed segmentation framework (Isensee et al., 2025a) utilizes an ensemble of two nnU-Net ResEnc (Isensee et al., 2024), an improved version of the original nnU-Net (Isensee et al., 2021) architecture, enhanced with residual connections. Pre-processing followed the standard nnU-Net CT intensity clipping transformation. The automatic planning of training hyperparameters was then manually adjusted, increasing the patch size to $160 \times 320 \times 320$ voxels to accommodate nearly the entire CBCT volume (median dimensions: $169 \times 347 \times 371$ voxels). This larger context allowed the network to learn better spatial relationships essential for correct tooth classification. The architecture depth was increased from six to seven resolution stages to leverage the expanded input size, with six residual blocks added. Training was done following a five-fold cross-validation and extended from 1 000 to 1 500 epochs. Data augmentation was also refined: left-right mirroring was disabled during training. This decision was based on observed performance drops when mirroring was enabled, likely due to the teeth and jaw's strong left-right symmetry in the sagittal plane. Finally, a post-processing strategy was introduced to reduce false positives, as they strongly negatively impacted the challenge metrics. For each class and evaluation metric, optimized cutoff thresholds were selected using cross-validation on the training set, removing small-volume predictions that affected performance. For the challenge submission, an ensemble of two models trained on the complete training set with different random seeds was utilized, along with the cutoff thresholds for post-processing.

4.1.2. Y. Jiang et al.

Proposed an ensemble of nnU-Net (Isensee et al., 2021) models, with ad hoc pre- and post-processing (Jiang et al., 2025). Initially, a data filtering heuristic is used to remove training data cases having label disconnections, possible mislabeling, or fewer than five annotated structures. This resulted in a final training dataset of 415 samples, filtering out 65 cases. The nnU-Net training procedure was then accomplished with five-fold cross-validation, along with z-score intensity normalization, using batches with a patch size of $80 \times 160 \times 160$ voxels for a total of 500 epochs. The authors proposed a post-processing approach called *Adaptive Structure Optimization* (ASO) to find the optimal filtering size coefficient (FSC) for each label, to remove small erroneous predictions due to noise (false positives), without causing an excessive amount of false negatives. The method learns the best FSC for each structure (dataset label) by comparing the differences of connected components (Allegretti et al., 2019; Bolelli et al., 2018; Cancilla et al., 2021) between the predictions on the evaluation set and the ground truth. The final FSC values for all structures are determined based on the results from five cross-validation splits, resulting

Table 3: Comparison of the best-performing approaches in terms of architecture, augmentations, pre-/post-processing, losses, and number of labels predicted. RAI = Right-Anterior-Inferior, CCA = Connected Components Analysis, ASO = Adaptive Structure Optimization.

ID	Model(s)	Augmentations	Pre-processing	Post-processing	Loss(es)	#Labels
A	2 × nnU-Net ResEnc	Default nnU-Net L/R mirroring disabled	Default nnU-Net Larger patch size	CCA size filtering	Dice Cross-Entropy	42
B	5 × nnU-Net	Default nnU-Net	Default nnU-Net	ASO cut-offs	Dice Cross-Entropy	42
C	2-stage nnU-Net cascade	Default nnU-Net L/R mirroring disabled (Stage 2)	Default nnU-Net Reoriented to RAI Tooth cropped via CCA	CCA to merge/split teeth Fine-tuned on metal artefacts	Dice (Stage 1 & 2) Cross-Entropy (Stage 1 & 2) Tversky (Stage 2, finetuning)	5 (Stage 1) 35 (Stage 2)

in 42 optimized coefficients used in the official test set submission. Inference utilizes the five-model ensemble to aggregate class probabilities across five model outputs, whereafter each voxel is assigned the label of the highest class probability, followed by the ASO post-processing.

4.1.3. H. Wang et al.

This method employs a multi-stage approach with two consecutive nnU-Net (Isensee et al., 2021) models. The first stage predicts five anatomical structures with no right/left differentiation, namely teeth, jawbones, pharynx, inferior alveolar canals, and sinuses. After introducing a mid-sagittal vertical plane to distinguish right from left, the second network is provided with a crop of the CBCT volume along with a crop of the predictions from the first stage as prior for tooth locations, to predict each single tooth label out of the 35 possible classes. The two models are trained in a similar fashion with five-fold cross-validation; the difference concerns the target labels and patch sizes: respectively 5 labels and $80 \times 160 \times 160$ voxels for the first stage, and 35 labels and $80 \times 160 \times 192$ voxels for the second stage. The authors noticed that the standard training procedure for the second network increased recall at the cost of more false positives, especially for crowns, bridges, and implants, which negatively influenced the metrics. Therefore, to lower the rate of false positives, they fine-tuned the second-stage nnU-Net on all training scans that included these three tooth classes, while using a Tversky Loss (Salehi et al., 2017) with $\alpha = 0.7$ and $\beta = 0.3$. Inference is executed with one model checkpoint per stage, where the results of the first-stage model are fed to the second-stage fine-tuned nnU-Net to predict exact teeth labels.

4.2. Observations

To begin with, it is noteworthy that all top-three solutions adopted the nnU-Net framework, either in its original form or the improved residual variant, nnU-Net ResEnc. This underlines the continued dominance of nnU-Net in medical image segmentation tasks (He et al., 2023; Kalkhof and Mukhopadhyay, 2023; Ma et al., 2024c). As highlighted by Isensee et al. (2021), nnU-Net (No New U-Net) does not propose an architectural innovation over the original U-Net; rather, it provides a task-agnostic configuration framework that streamlines the adaptation of U-Net to diverse medical image analysis challenges without the need for extensive trial-and-error tuning.

Pre-processing. All teams employed the standard nnU-Net preprocessing pipeline, including re-sampling, CT intensity clipping, z-score normalization, and spatial data augmentations. However, some teams introduced refinements aligned with the dataset's specific anatomical structure. For instance, they disabled left-right mirroring due to the bilateral symmetry of the maxillofacial region, which, when augmented, degraded performance by confusing laterality-specific labels (e.g., teeth, IACs, and sinuses).

Beyond standard routines, Y. Jiang et al. performed dataset curation by filtering out noisy or incomplete annotations, thereby sacrificing quantity for higher training data quality. H. Wang et al. carried out a complementary approach: they constructed a focused training subset to improve performance on difficult tooth-related structures like crowns, bridges, and implants.

Training strategies. Ensembling was a key strategy across all top-performing methods, with each team training multiple models (ranging from two to five) under cross-validation protocols. Teams also customized training hyperparameters: *(i)* F. Isensee, Y. Kirchhoff et al. used the recent nnU-Net ResEnc, increasing input patch sizes and architecture depth to capture broader spatial context; *(ii)* H. Wang et al. adopted a two-stage pipeline, first predicting coarse anatomical regions, then refining tooth-specific labels with guided inputs from the first stage.

Post-processing. Post-processing proved crucial for improving final predictions. Both F. Isensee, Y. Kirchhoff et al. and Y. Jiang et al. used filtering techniques to remove small-volume false positives that degraded evaluation scores. Notably, Y. Jiang et al. introduced the *Adaptive Structure Optimization* (ASO) algorithm to learn label-specific size thresholds. H. Wang et al., in contrast, embedded error mitigation into the architecture itself—fine-tuning the second-stage network with a Tversky loss focused on hard-to-segment dental classes.

Summary. All winning methods built upon a strong nnU-Net baseline and differentiated themselves through careful data handling, architectural scaling, task-specific post-processing, and ensemble designs. Their common emphasis on reducing false positives, either via post hoc filtering or architectural guidance, was key to achieving high scores under the challenge metrics.

5. Evaluation

5.1. Evaluation Metrics

The metrics used (Fig. 4) to rank the submitted methods are the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95), two metrics commonly used in image segmentation (Maier-Hein et al., 2024). The DSC has practically the same meaning as the IoU (Intersection over Union), but the first one is better suited when the region of interest is

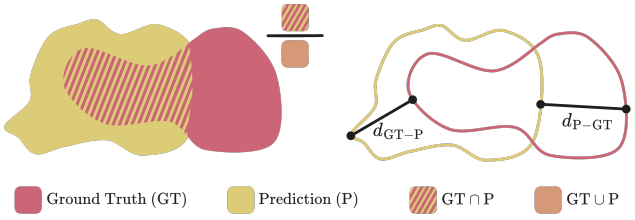

Fig. 4: Visual representation of evaluation metrics. On the left is depicted the intersection over union (IoU), on the right, the d_{95} distances used to calculate the Hausdorff Distance (HD) between two regions of points.

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:

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

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:

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

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 voxels in the predicted mask (P) and the set of voxels 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:

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

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.

It is worth mentioning that Metrics Reloaded[7] (Maier-Hein et al., 2024) recommendations have been employed to select the most appropriate metrics for the challenge. Among the suggested metrics there were the Dice score and the Normalized Surface Distance (NSD). Compared to the Hausdorff Distance (HD), the NSD is less sensitive to the outliers. We opted for the HD95 because it is more clinically indicated. To provide the reader with an example, measuring the distance between the inferior alveolar nerve and the tooth roots requires an upper bound rather than an average. This allows practitioners to carefully plan tooth removal or other surgical procedures while minimizing the risk of nerve damage.

5.2. Ranking Protocol

The two metrics described above yield consistent values across different patients, allowing them to be averaged later in order to derive the final ranking (one ranking per metric/class).

[7] https://metrics-reloaded.dkfz.de/

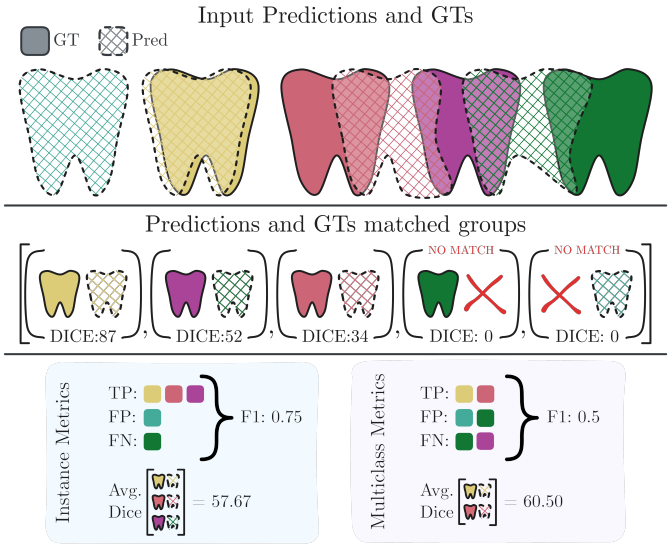

Fig. 5: Additional teeth-wise metrics (see Sec. 5.3). Each predicted tooth is uniquely matched to the ground-truth with the highest Dice (only if Dice > 0.1); unmatched predictions/GT count as FP/FN. Instance metrics ignore FDI labels, whereas multiclass metrics require matching FDI numbers for a true positive.

To guarantee robustness in the ranking procedure, the guidelines outlined by Maier-Hein et al. (2018) were adopted. Broadly, there are two opposing strategies for aggregating metrics across test cases. The first, referred to as metric-based aggregation, consists of first combining the metric values across all test cases (e.g., by computing the mean or median) and then ranking algorithms based on the aggregated score. The second, known as case-based aggregation, entails computing a rank for each individual test case, after which the final ranking is obtained by aggregating the per-case ranks. As reported by Maier-Hein et al. (2018), single-metric rankings (in our case DSC and HD95 for each task/class) exhibit greater statistical robustness when metric-based aggregation is applied, with the mean preferred over the median.

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

1. For each class and for each volume, calculate the Dice score and the HD95. Also compute the maximum used memory (*Mem*) and the total execution time (*Time*) for all cases;
2. Average the Dice and HD95 obtained for each class across volumes, obtaining Dice_c and HD95_c for each class c (42 classes in total).
3. Rank the Dice_c, HD95_c, maximum used memory, and running time, independently (86 rankings in total);
4. Average the rankings obtained at point 3 for all Dice_c and HD95_c to produce the final rank;
5. If two or more final ranks obtained at point 4 are equal, compute the average of the rankings obtained at point 3 for *Mem* and *Time* to break ties;
6. If two or more ranks are still equal, it is a tie.

If an algorithm produces no output for a CBCT volume, the evaluation treats the predicted segmentation as an all-zero volume. Consequently, the Dice score for that case is 0 for each

Table 4: Results on the test set of the ToothFairy2 dataset for the best five Challenge participants. Classes are grouped by the main anatomical structure to which they belong. For brevity, only the team's ID and (shared) first author's names were used. Best results per group and metric (table lines) are reported in bold.

Group	Metric	(A) F. Isensee, Y. Kirchhoff et al.	(B) Y. Jiang et al.	(C) H. Wang et al.	(D) A. Gao et al.	(E) P. Shi et al.
Overall	DSC	**0.925 ± 0.171**	0.917 ± 0.175	0.911 ± 0.188	0.911 ± 0.171	0.902 ± 0.207
	HD95	18.869 ± 106.469	17.647 ± 102.822	**17.564 ± 96.074**	18.576 ± 106.897	26.794 ± 130.444
L/R IAC	DSC	0.897 ± 0.047	0.891 ± 0.040	**0.898 ± 0.037**	0.872 ± 0.054	0.892 ± 0.040
	HD95	1.606 ± 2.034	1.548 ± 1.118	**1.474 ± 0.911**	2.022 ± 2.146	1.557 ± 1.028
L/R Sinus	DSC	**0.891 ± 0.214**	0.813 ± 0.265	0.837 ± 0.265	0.836 ± 0.238	0.886 ± 0.226
	HD95	33.724 ± 142.266	**30.996 ± 109.426**	38.994 ± 138.839	37.028 ± 141.887	49.945 ± 156.102
Teeth	DSC	**0.930 ± 0.171**	0.925 ± 0.170	0.915 ± 0.191	0.918 ± 0.166	0.904 ± 0.211
	HD95	17.878 ± 106.926	16.930 ± 106.305	18.603 ± 101.543	**16.892 ± 106.509**	25.851 ± 132.694
Jawbones	DSC	0.931 ± 0.086	0.946 ± 0.065	0.956 ± 0.050	0.938 ± 0.050	**0.957 ± 0.044**
	HD95	14.492 ± 27.625	6.901 ± 17.914	**4.202 ± 12.159**	12.258 ± 21.708	4.693 ± 14.132
Pharynx	DSC	**0.955 ± 0.027**	0.944 ± 0.039	0.951 ± 0.032	0.947 ± 0.033	0.936 ± 0.051
	HD95	**3.911 ± 3.182**	5.308 ± 5.571	4.061 ± 4.466	4.168 ± 4.075	9.982 ± 23.277
Others	DSC	**0.896 ± 0.241**	0.887 ± 0.240	0.894 ± 0.229	0.883 ± 0.259	0.851 ± 0.288
	HD95	38.479 ± 148.126	38.172 ± 136.603	**16.670 ± 77.282**	43.731 ± 155.026	58.273 ± 178.434

class, and the HD95 is set to the volume diagonal. In addition, the inference time for the case is recorded as 10 minutes.

5.3. Additional Post-challenge Metrics

To provide a more detailed evaluation of the submitted algorithms, we computed additional *post-challenge metrics* to assess both instance-level and semantic (multi-class) segmentation performance, focusing exclusively on tooth structures (FDI classes 11-48).

In fact, instance segmentation can be geometrically accurate yet semantically incorrect: a method may delineate individual surfaces (instances) correctly while misassigning their IDs (classes), or conversely predict the correct ID but with imperfect boundaries. This is particularly challenging for teeth due to strong left-right symmetry and fine-grained class distinctions, compounded by shape changes and appearance variation introduced by dental restorations and other artificial objects.

To account for this, the following metrics were employed in an instance-versus-multiclass fashion, using only tooth labels. A visualization explanation is also reported in Fig. 5.

Instance F1 (*Objective F1*). Represents how reliable individual teeth can be identified, using the F1-score as a balance of true-positive, false-negative, and false-positive predictions. Tooth instances in the predictions and annotations were matched by iteratively finding the tooth pair with the highest DSC overlap of at least 0.1, where each tooth is matched at most once. Additionally, in the case of multi-class instance metrics, the tooth FDI labels must match.

True Positives DSC (*TP-DSC*). Validates the voxel-level accuracy of the tooth segmentations by computing the average DSC score between pairs of matched teeth from the predictions and annotations. These true-positive instances were determined using the iterative matching described for *Instance F1*, with additionally matching FDI labels for the currently described multi-class *TP-DSC*.

Panoptic-DSC. Defined as:

$$\text{Panoptic-DSC} = \text{Instance F1} \times \text{TP-DSC}, \tag{4}$$

multiplies *Instance F1* and *TP-DSC* to evaluate both instance-level and voxel-level accuracy.

Foreground-DSC. This metric considers all the tooth classes (11-48) as a single instance, where the classic DSC formula is then used to compute the corresponding metric score.

To better understand in which setting each of the above metrics is most useful, Fig. 6 illustrates possible segmentation errors. These examples emphasize that no single metric captures all error types, and complementary metrics are required for a complete assessment of tooth segmentation. In this figure, cases (b) and (g) show perfect or slightly misaligned predictions, for which all metrics remain high. From case (c) onward, divergences appear:

c) *Missing tooth:* TP-DSC does not change since it averages only over matched tooth pairs, but Instance F1, Panoptic-DSC, and Foreground-DSC all decrease;
d) *Extra tooth (false positive):* TP-DSC remains perfect on matched teeth, but Instance F1 and Panoptic-DSC decrease due to the additional tooth; Foreground-DSC also drops;
e) *One prediction covering two teeth:* Foreground-DSC may not change since voxel coverage is good, but TP-DSC, Instance F1, and Panoptic-DSC decrease because of incorrect instance boundaries;
f) *Two predictions for one tooth:* Foreground-DSC is high, but only one prediction can be matched, lowering TP-DSC, Instance F1, and Panoptic-DSC;
g) *Oversegmentation with small false positives:* main teeth are correct, so TP-DSC is high, with the extra prediction reducing the other metrics;
h) *Label swap:* geometry is perfect, so Foreground-DSC remains high, but incorrect FDI numbers reduce multi-class TP-DSC and multi-class Instance F1, highlighting the need for label-sensitive evaluation.

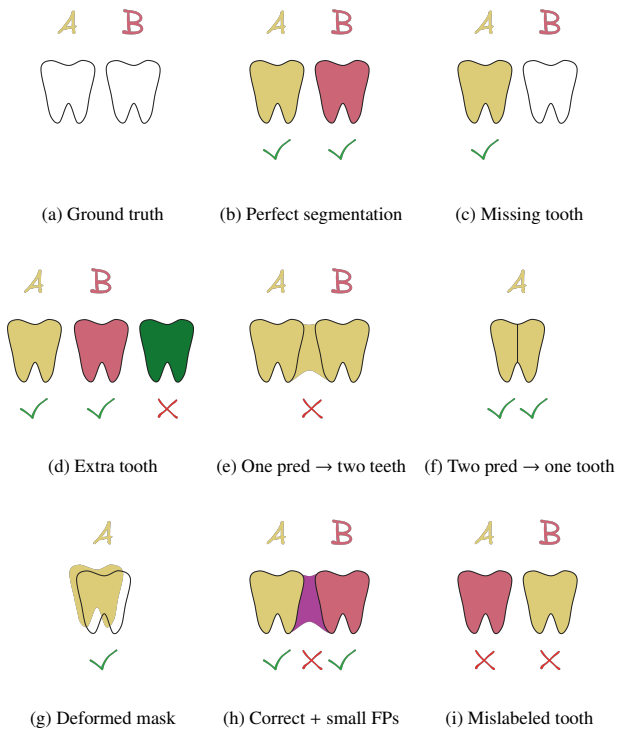

Fig. 6: Segmentation example scenarios in 2D under different error or correctness conditions. Each colored tooth represents an instance, with its own colored label. If the instance segmentation is correct, a green check mark is displayed; otherwise, a red cross is shown. Explanations for the above nine cases in the figure are provided in Sec. 5.3.

5.4. AWS Infrastructure

All participant-submitted algorithms were executed on the Grand Challenge AWS (Amazon Web Services) infrastructure, which provides elastic scaling of storage and computing resources. Depending on the chosen configuration, either a `g4dn.xlarge` instance (Nvidia T4 GPU, 16 GB GPU memory, 4 CPUs, 16 GB RAM) or a `g4dn.2xlarge` instance (Nvidia T4 GPU, 16 GB GPU memory, 8 CPUs, 32 GB RAM) was employed. To prevent leakage of test set data, the containers were run without internet access.

6. Results and Discussion

In this section, we present the key results from the challenge submissions. Several of the proposed strategies are complementary and can be combined to build stronger segmentation models, to improve both accuracy and robustness.

6.1. Overall Challenge Results

Tab. 4 summarizes the performance of the top five participants on the ToothFairy2 test set, reporting results for overall segmentation as well as grouped anatomical structures: IACs (left/right), maxillary sinuses (left/right), teeth, jawbones (upper/lower), pharynx, and the class "others," which includes crowns, bridges and implants. For each group, the reported DSC and HD95 values were obtained by averaging the metrics across the individual classes within that group. The best average DSC was achieved by F. Isensee, Y. Kirchhoff et al. with a score of 0.925, closely followed by Jiang et al. (0.917) and Wang et al. and Gao et al. (0.911). Similarly, the HD95 values were comparable among the top three methods, but in inverse order of performance, since this metric must be minimized, while Shi et al., even though they achieved a high DSC, showed a less effective HD95 metric, indicating that good volumetric overlap does not imply precise boundary identification.

Performance varied across anatomical groups. For the inferior alveolar canals (IACs), all methods achieved high DSCs of approximately 0.89 across the five best methods. The maxillary sinuses proved to be more challenging, with lower DSC and greater variability across participants, mainly because they are present only in about 10% of the training scans. On the sinus, large HD95 values across all methods highlight the difficulties in segmenting their boundaries under scarce training data. Notably, the ResEnc version of nnU-Net, employed by method A, demonstrated its strength in such low-data scenarios, outperforming the standard variant employed by the others.

The teeth group showed relatively strong overall performance, with a DSC between 0.90 and 0.93. Here, method A ranked highest, as its use of class-specific filtering and the disabling of left/right mirroring produced more consistent teeth segmentation; the 2-point lower DSC of method C suggests that, in this context, employing a specialized network only for tooth segmentation and classification is not the ideal solution. The high HD95 values for teeth across all methods are strongly affected by the fact that, in the most common case, when a tooth is misclassified in a distant location (e.g., the opposite sagittal side), the distance metrics increase drastically.

For jawbones, DSC scores were the highest across all structures, with C and E obtaining the best values, suggesting that bony structures with clearer boundaries are better captured by the lighter standard nnU-Net variants as compared to the ResEnc version. The "others" class, which includes crowns, bridges, and implants, is strongly affected by CBCT metal artifacts, often leading to blurred or missing boundaries: for this reason the HD95 metric is usually the highest across the classes. Method C achieved a markedly lower HD95 of 16.670 compared to all other teams (ranging from 38 up to 58), thanks to their fine-tuning strategy on these three classes, with a Tversky loss specifically designed to reduce false positives in restorative dental structures.

These results are consistent with the methodological choices described in Sec. 4. F. Isensee, Y. Kirchhoff et al. achieved the first final rank by leveraging nnU-Net ResEnc with larger receptive fields and an optimized post-processing strategy, which proved to be highly effective for this segmentation task. Y. Jiang et al. placed very close to the first rank despite using a lighter architecture, thanks to their data selection and the adoption of Adaptive Structure Optimization post-processing, which reduced false positives and contributed to their strong performance on jawbone segmentation. H. Wang et al. reached the third place with a two-stage approach, achieving the best overall performance on the "others" class by specifically fine-tuning their network on crowns, bridges, and implants.

In contrast, lower-ranked teams explored more diverse methodologies, including a video foundation model trained on

Table 5: Performance comparison of participants on two groups of post-challenge aggregated metrics described in Sec. 5.3. Higher is better; the best results (per column) are in bold, the second-best are underlined. Teeth Class DSC is the same as the mean Teeth DSC in Tab. 4. Since these experiments required re-running all the algorithms, only submissions that published a Docker image have been included in the analysis.

Rank	ID	Teeth Class DSC	Teeth Foregr. DSC	Instance Metrics			Multiclass Instance Metrics		
				Instance F1	TP-DSC	Panoptic-DSC	Instance F1	TP-DSC	Panoptic-DSC
1	A	**0.931**	**0.958**	**0.9910**	**0.9471**	**0.9386**	**0.9795**	**0.9482**	**0.9288**
2	B	0.925	0.953	0.9903	0.9399	0.9308	0.9780	0.9411	0.9205
3	C	0.915	0.953	0.9780	0.9379	0.9173	0.9594	0.9395	0.9016
5	E	0.905	0.951	0.9776	0.9358	0.9149	0.9556	0.9395	0.8976
9	I	0.894	0.954	0.9729	0.9401	0.9146	0.9525	0.9433	0.8982
12	L	0.883	0.920	**0.9910**	0.8975	0.8894	0.9689	0.9008	0.8732
15	O	0.804	0.889	0.9614	0.8422	0.8097	0.9291	0.8419	0.7843
18	R	0.500	0.801	0.8688	0.5826	0.6590	0.7976	0.7872	0.6271

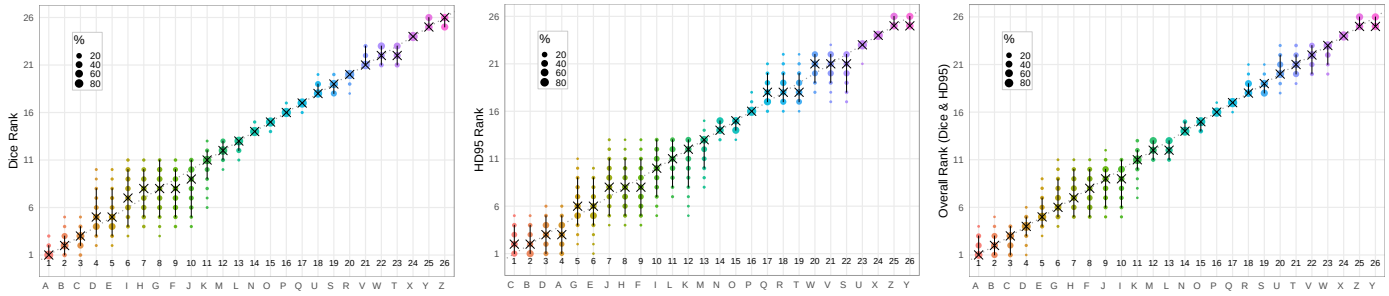

Fig. 7: Visualization of ranking stability when considering only (a) Dice and (b) HD95 scores or (c) both of them. Algorithms are color-coded, and the area of each blob at position (A_i, *rank j*) is proportional to the relative frequency A_i achieved rank j across $b = 1000$ bootstrap samples. The median rank for each algorithm is indicated by a black cross. 95% bootstrap intervals across bootstrap samples are indicated by vertical black lines.

RGB data (Ma et al., 2025), alternative architectures outside of nnU-Net (Wodzinski and Müller, 2025), a transformer-based backbone with language integration (Daza and Schnabel, 2025), and a segmentation pipeline from multi-axial to 3D (Mai et al., 2025). Overall, the results suggest that strong performance is primarily associated with the nnU-Net framework, particularly when combined with ensembling, large input patches and deep architectures to capture extensive spatial context, and task-specific adaptations such as disabling left-right mirroring and applying class-specific optimization or post-processing strategies for challenging structures.

6.2. Ranking Stability

In order to test and visualize the ranking stability when using the Challenge-selected metrics, Fig. 7 is provided. Charts are generated by an ad-hoc variation of *ChallengeR* (Wiesenfarth et al., 2021), a standard tool for analyzing and visualizing challenge outcomes. Reported results are obtained by performing random sampling with replacement (bootstrapping) 1 000 times. The charts use a blob plot to visualize ranking stability from bootstrap sampling.

The ranking stability criterion was selected to account for the fact that small metric differences can translate into rank changes, especially when multiple methods perform similarly. The distribution of ranks across 1 000 bootstrap samples, summarized by median rank and 95% intervals, offers a transparent way to separate clearly leading methods from those whose ordering is sensitive to sampling variability. Such analysis confirms that top-scoring algorithms consistently confirm their rank with lower variability w.r.t. other methodologies.

6.3. Tooth-related Post-challenge Metrics

As introduced in Sec. 5.3, it is crucial to evaluate the proposed solutions not only on voxel-wise segmentation accuracy, but also on their ability to distinguish between highly similar maxillofacial structures. This is particularly evident for teeth, where a robust differentiation between adjacent teeth is required despite left-right symmetry.

Tab. 5 reports the comparison between instance and multi-class post-challenge metrics. As expected, instance-level metrics yield higher values and less variance across participants, while the multi-class group shows larger discrepancies. This confirms that most methods achieve good tooth instance separation, but struggle when assigning correct FDI labels: predictions are largely accurate in voxel coverage but weaker in semantic attribution. As shown, teeth Foreground-DSC is consistently high (> 0.9), even for lower-ranked submissions, whereas teeth Class DSC drops significantly. For example, the ninth-ranked method achieves the second-best foreground score, yet its class DSC falls below 0.90. This demonstrates that detecting all teeth voxels as a single structure is relatively easy, but identifying each individual tooth correctly is more challenging.

Further insight comes from the per-metric breakdown in Tab. 5. Instance and multi-class TP-DSC remain similar, indicating that once a tooth is correctly detected, the voxel-wise DSC is usually high. In contrast, Instance F1 diverges between the two groups, revealing that false positives and false negatives become more common once matching FDI labels are enforced. This aligns with the high HD95 scores observed for aggregated teeth classes in Tab. 4, which point to left-right swaps of symmetric teeth (like case (i) in Fig. 6). These swaps explain the

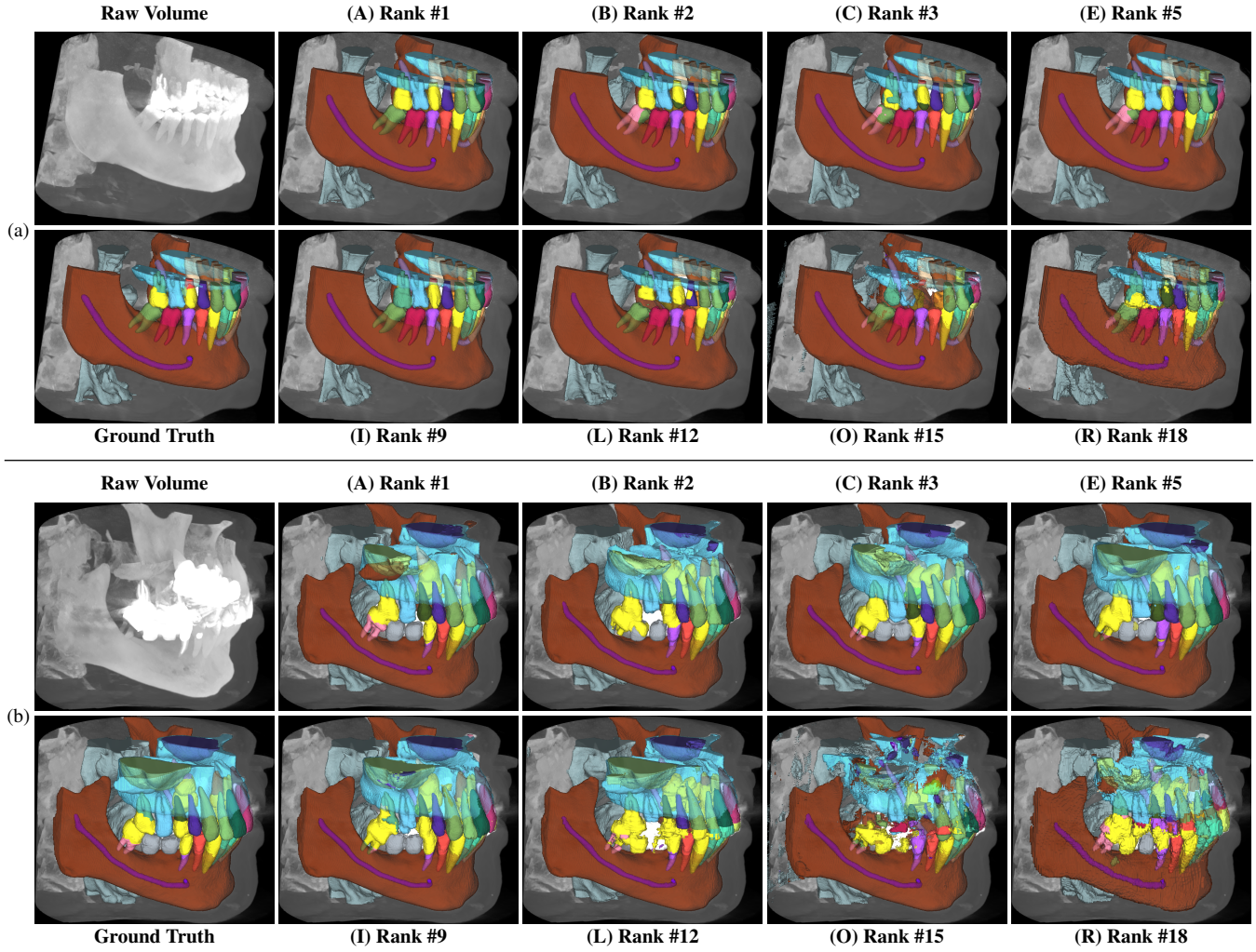

Fig. 8: 3D visualizations of two CBCT volumes from the test set, as well as the ground truth and predicted maxillofacial structures. Case (a) shows a smaller field of view with an implant for tooth 15, and case (b) shows a larger field of view, including the maxillary sinuses, a bridge, and several crowns. The predictions are ordered by the method's rank in the final test phase, including the methods ranked first (F. Isensee, Y. Kirchhoff et al.), second (Y. Jiang et al.), third (H. Wang et al.), fifth (P. Shi et al.), ninth (Z. Pan et al.), twelfth (Q. Ma et al.), fifteenth (M. Wodzinski et al.), and eighteenth (M. Haipeng et al.). This analysis includes only participants who provided the submission Docker image.

mismatch between high voxel-wise accuracy and lower class-level performance.

Finally, it should be noted that the top-performing methods that achieved high scores in almost all metrics, such as the method proposed by F. Isensee, Y. Kirchhoff et al., adopted strategies to reduce confusion between the left and right. Two of the top three explicitly disabled sagittal mirroring, while another employed a two-stage approach with sagittal splitting and orientation normalization to infer side information. Thus, even the best solutions relied on spatial priors to overcome label ambiguity, highlighting the fundamental difficulty of tooth identification more than segmentation.

6.4. Visualizations

Fig. 8 shows two cases comparing ground-truth and predicted segmentations from several participating methods. The first case (Fig. 8a) is a CBCT scan with a limited vertical field of view (FOV), as the maxillary sinuses and root apices of the upper teeth were not included. Tooth 17 (upper right second molar) has a crown, while tooth 15 (upper right second premolar) contains both an implant and a crown. Method A produced the best results, missing only the implant as a result of post-processing, where small components were removed to reduce false positives. Method B split tooth 47 into labels 47 and 48 and misinterpreted the implant as a crown extension, while Method C correctly identified the implant but also split tooth 47 and failed to segment the crown of tooth 17. Method E similarly split tooth 47 but accurately predicted the restorations. Method I detected only a few voxels of the crown of tooth 17 yet was otherwise precise. Method L correctly predicted the mandibular teeth but missed the inferior parts of the maxillary crowns. Method O yielded noisy predictions, particularly in the maxilla, while Method R produced downsampled outputs that substantially limited accuracy.

The second case (Fig. 8b) is a scan with a larger vertical FOV that includes the maxillary sinuses and a dental bridge. All methods struggled with the structures above the maxillary teeth, since this region was poorly represented in the training

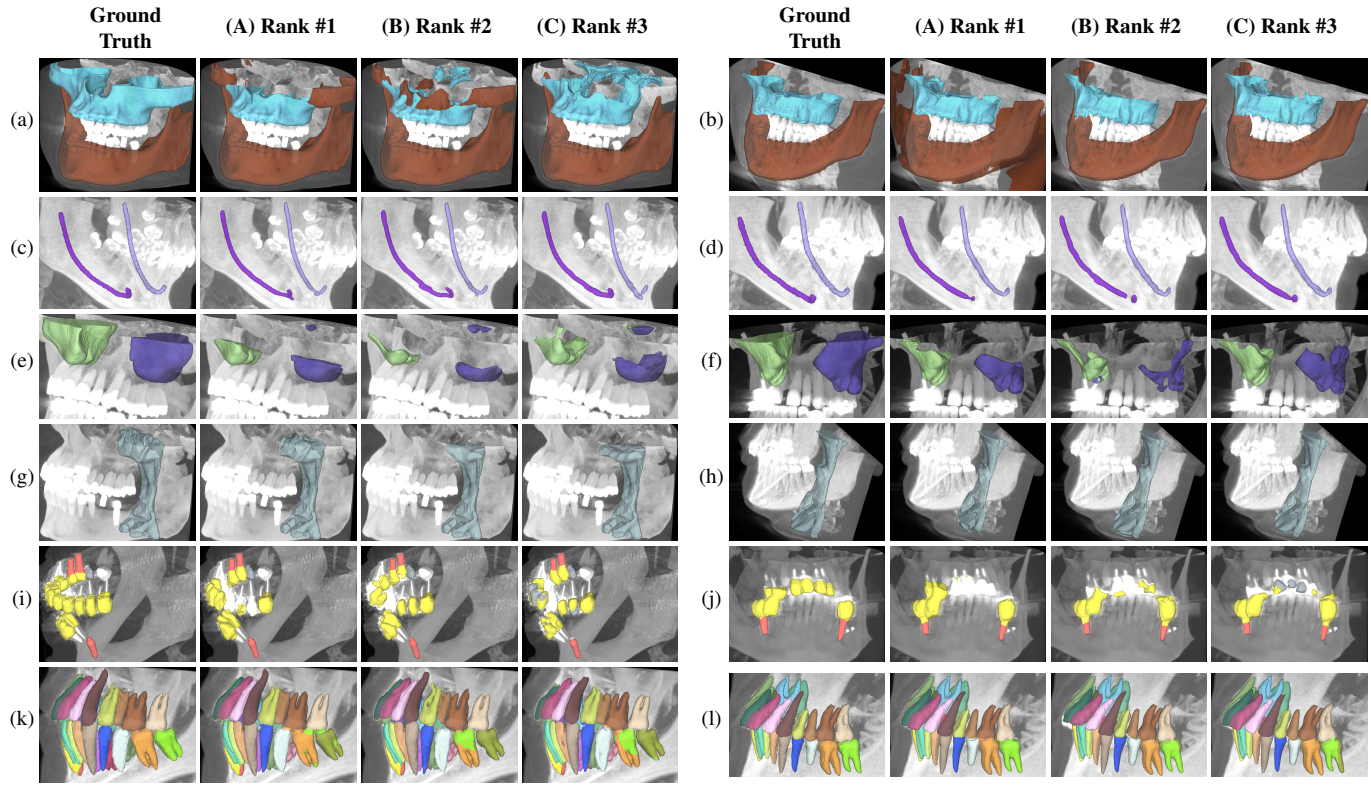

Fig. 9: 3D visualizations of ground-truth and predicted maxillofacial structures in test scans. Each row shows two representative cases for a structure, including the jawbones (a-b), mandibular canals (c-d), maxillary sinuses (e-f), pharyngeal airway (g-h), dental restorations (i-j), and natural teeth (k-l). For each case, the ground-truth segmentation is shown, as well as the predicted segmentation of the participating methods ranked first (F. Isensee, Y. Kirchhoff et al.), second (Y. Jiang et al.), and third (H. Wang et al.) in the final test phase. Only one category of maxillofacial structures is shown at a time to improve readability.

scans. Clear differences also appear in the predictions of teeth and tooth restorations. Methods A and C missed the crown of tooth 15, while Method B omitted the roots of tooth 48 and the crown of tooth 16, and incorrectly predicted part of the bridge as a crown. Method E missed the crowns of teeth 15 and 45, and Method I showed imprecise crown boundaries. Methods L and O showed noisier predictions, and the segmentations of method R are additionally predicted in a downsampled resolution.

Fig. 9 illustrates the predicted segmentations of the top three methods across the evaluated dental structures. For the jawbones (a-b), method A occasionally oversegmented by including parts of the skin as mandible. In the maxilla superior to the teeth, all methods struggled, likely because this region was underrepresented in the training data. The IAC segmentations (c-d) often displayed discontinuities and diverging paths; however, compared to the ground truth, the predicted canals tended to exhibit a more uniform radius. The maxillary sinuses (e-f) were also inaccurately segmented, particularly in large-FOV scans, showing similar limitations as the maxilla. Predictions of the pharyngeal airway (g-h) frequently underestimated the nasopharynx, leading to incomplete segmentations. Tooth restorations such as implants, crowns, and bridges (i-j) showed mixed performance across methods, heavily affected by metal-induced scattering artifacts typical of CBCT imaging. Finally, tooth segmentations (k-l) revealed common errors, including split teeth (>1 predicted class for a single ground-truth tooth), merged teeth (1 class spanning >1 ground-truth tooth), and incorrect tooth labels. These issues arise from the semantic segmentation paradigm used by all participants, which lacks explicit instance-awareness, preventing the models from distinguishing individual teeth reliably.

7. Limitations and Future Work

Dataset. Although the ToothFairy2 dataset has shown strong scientific interest and enabled impressive algorithm performance in dental and maxillofacial semantic segmentation, several limitations remain. First, although training and testing data were acquired from different centers, the training set originates from a single institution, potentially limiting the generalizability of trained models across diverse clinical environments and patient demographics. Second, important clinical classes such as implants, crowns, and bridges are underrepresented, affecting model robustness in real-world applications, and adjacent implants and crowns are not separated per tooth. Moreover, common components such as dental braces and bone plates are absent from the dataset due to limited availability, despite their frequent occurrence in practice.

The ToothFairy2 dataset includes annotations for 42 distinct anatomical structures in the oral and maxillofacial region, with 35 of these structures related to teeth. As the final challenge ranking was determined by an average over the rankings for

each anatomical structure, methods with effective tooth predictions could be ranked unfavorably high, despite less effective results for other structures. This is corroborated by the current results, with the top three methods achieving the first, second, and third-highest Panoptic-DSC scores, whereas the overall top method ranked sixth on jawbones. Therefore, the impact of the effectiveness of tooth-related structures on the final ranking may have been too large, and the final ranking may thus not be an accurate representation of which methods achieved the overall most effective results. Nevertheless, the ranking per group of anatomical structures (Tab. 4) should be accurate for determining model effectiveness on a particular use case.

Another challenge lies in the growing need to segment complex and fine-grained anatomical structures. ToothFairy2 primarily focused on teeth and major maxillofacial structures, but did not include clinically relevant details such as pulp chambers, root canals, or small neurovascular canals. These restrict the dataset's applicability in orthodontics and surgical planning, where precision at the sub-structural level is critical.

To address these gaps, we introduced ToothFairy3, which expands upon ToothFairy2 through *(i)* improved annotations, *(ii)* the inclusion of 52 new scans from an additional acquisition device, and *(iii)* 35 additional labels, covering pulp cavities, incisive canals, and the lingual canal. This expansion not only mitigates some of the representational limitations but also introduces more challenging thin and elongated structures. Furthermore, ToothFairy3 emphasizes computational efficiency by incorporating inference time as a primary evaluation metric and proposes a novel interactive segmentation task for the inferior alveolar canal. Together, these additions aim to push research toward more generalizable, clinically viable, and time-efficient segmentation procedures and will be discussed in the corresponding challenge report.

Ultimately, future efforts should focus on building larger multi-center datasets with broader demographic and anatomical diversity, balancing underrepresented classes, and incorporating missing clinical components. Such directions will be critical for developing robust and deployable AI systems capable of supporting the wide spectrum of maxillofacial procedures, from reconstructive surgery to daily dental practice. In addition, future work should assess and, when needed, correct the calibration of predictive confidence/uncertainty, as modern deep networks can be overconfident and uncertainty estimates may be unreliable at the subject level (Guo et al., 2017; Jungo and Reyes, 2019; Pollastri et al., 2021a).

Ranking. Although following standard and consolidated approaches (Maier-Hein et al., 2018), a key limitation of our evaluation protocol is that the final ordering is obtained by aggregating ranks across multiple tasks/targets, which makes the leaderboard sensitive to the number and type of submissions, particularly when the same group submits multiple times (Piérard et al., 2025). From a theoretical perspective, performance-based rankings should ideally satisfy consistency properties (e.g., the relative order between two methods should not change simply because other methods enter or leave the pool). To reduce instability, we considered only the latest submission per team. However, despite enforcing this rule on the Grand Challenge platform, we later detected likely circumvention via multiple accounts linked to the same group. These cases were evident because they yielded identical metrics across many classes and test cases: an implausible outcome for independently trained models. We contacted the accounts for clarification, but received no response; therefore, we removed all duplicate submissions from the final-test leaderboard before computing the final ranking.

Future editions of the challenge will strengthen *(i)* team/account verification and submission policies, *(ii)* automatic detection of near-duplicate submissions, and *(iii)* rank-stability audits (e.g., leave-one-out sensitivity). Methodologically, we will also investigate ranking procedures more closely aligned with the axiomatic view of performance-based ranking (e.g., preorder-based formulations and explicit "importance" weighting) to reduce dependence on participant pool composition (Piérard et al., 2025).

8. Conclusions

In this article, we presented, described, and commented on the ToothFairy2 challenge, jointly organized by the University of Modena and Reggio Emilia and the Radboud University Medical Center located in Nijmegen, the Netherlands. After the success of the ToothFairy Challenge, presented at MICCAI 2023, the ToothFairy2 challenge, organized the following year at MICCAI 2024, aimed to address the data scarcity in publicly available maxillofacial CBCT datasets, while at the same time improving the diversity of 3D annotated structures.

ToothFairy2, along with its successor ToothFairy3, remains the biggest publicly available dataset for maxillofacial semantic segmentation to date. 29 research teams across the world submitted their algorithm for evaluation, with the winner in first place being the solution proposed by F. Isensee, Y. Kirchhoff et al. with an overall DSC of 0.925 and HD95 of 18.869 across all labels, winning a €1 500 prize. Similarly for ToothFairy, the participant solutions relied mostly on the robust and efficient nnU-Net architecture and its improved version, nnU-Net ResEnc, along with the adoption of simple yet effective pre- and post-processing methods, such as disabling sagittal mirroring augmentation and filtering for small unconnected structures that heavily affected metrics.

The ToothFairy challenges remain a reference point in the domain of semantic volumetric segmentation related to maxillofacial data, offering research teams around the globe an open and transparent benchmark to evaluate their state-of-the-art algorithms. The ToothFairy and ToothFairy2 Challenges, although already concluded, will remain open for new submissions on the Grand Challenge platform[8][9], to support our mission to facilitate the integration of artificial intelligence in maxillofacial medical imaging applications.

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

[9] https://toothfairy2.GrandChallenge.org/evaluation/post-challenge-phase-test-your-algorithm/leaderboard/

Author Contributions

Conceptualization (Federico Bolelli, Luca Lumetti, Niels van Nistelrooij, Shankeeth Vinayahalingam, Alexandre Anesi, Costantino Grana); **Methodology** (Federico Bolelli, Luca Lumetti, Niels van Nistelrooij, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Kevin Marchesini, Alexandre Anesi, Costantino Grana); **Software - webpages and evaluation scripts** (Federico Bolelli, Luca Lumetti); **Software - challenge submissions** (Fabian Isensee, Yannick Kirchhoff, Klaus H. Maier-Hein, Lars Krämer, Maximilian Rokuss, Constantin Ulrich, Yuxian Jiang, Yusheng Liu, Lisheng Wang, Haoshen Wang, Siyu Chen, Zhiming Cui, Pengcheng Shi, Zhaohong Pan, Xiaokun Liang, Qi Ma, Ender Konukoglu, Marek Wodzinski, Henning Müller, Haipeng Mai, Xiaobing Dang, Shrajan Bhandary, Radu Grosu); **Validation** (Luca Lumetti, Niels van Nistelrooij, Kevin Marchesini); **Formal analysis** (Federico Bolelli, Luca Lumetti, Niels van Nistelrooij, Kevin Marchesini, Ettore Candeloro, Gabriele Rosati); **Investigation** (Luca Lumetti, Niels van Nistelrooij, Kevin Marchesini); **Resources** (Federico Bolelli, Alexandre Anesi, Costantino Grana); **Writing - original draft** (Federico Bolelli, Luca Lumetti, Niels van Nistelrooij, Mattia Di Bartolomeo, Kevin Marchesini, Ettore Candeloro, Gabriele Rosati); **Writing - review and editing** (Federico Bolelli, Luca Lumetti, Niels van Nistelrooij, Shankeeth Vinayahalingam, Kevin Marchesini, Tong Xi, Stefaan Bergé, Alexandre Anesi, Costantino Grana); **Visualization** (Federico Bolelli, Luca Lumetti, Niels van Nistelrooij, Kevin Marchesini, Ettore Candeloro, Gabriele Rosati); **Data curation** (Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani); **Supervision** (Tong Xi, Stefaan Bergé, Alexandre Anesi, Costantino Grana); **Project administration** (Federico Bolelli, Costantino Grana); **Funding acquisition** (Federico Bolelli, Shankeeth Vinayahalingam, Costantino Grana).

Appendix

This section provides all the technical elements of the challenge organization 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 May 2nd, 2024, enabling participants to access it for analysis and model development. During the *Preliminary Phase*, participants had the opportunity to submit their models from July 8th to August 18th, 2024. In this phase, tests were performed on a subset of five volumes extrapolated from the training set, allowing participants full access to the output logs without violating the privacy of testing data. These results provided valuable feedback on the quality of their submissions and the Docker images produced.

The challenge *Final Phase* took place from August 1st to August 18th, 2024. During this period, each team was allowed to submit a maximum of two algorithm entries. Participants were invited to publish their algorithms on GitHub and share their research papers with the organizing team by August 31st, 2024. This step was mandatory for the top three teams to remain eligible for the prize and for all other participants who wished to be included as co-authors of this publication.

The results of the challenge were released on September 15th, 2024, and the winners were officially announced during the on-site workshop event on October 6th.

A discussion of the ToothFairy2 results and a presentation of the top-performing methods were held on October 6th, during the associated workshop sessions at the MICCAI 2024 conference, Marrakech, Morocco. The first three teams were awarded the following prizes: €1 500 for the winning team, €1 000 for the runner-up, and €500 for the third-place team. SeeThrough S.r.l., an Italy-based manufacturer of medical devices (including CBCT scanners), was the official sponsor of the challenge and funded all monetary prizes.

Participation policies. Participating teams were permitted to use additional datasets and/or pre-trained networks, provided that such use was clearly disclosed in the submission and that the resources were publicly available at the time of submission.

Members of the organizers' institutions were allowed to participate in the challenge but were not eligible for awards. Submissions from organizing teams are clearly identified as such on the online leaderboards.

Publication policies. For the entire duration of the challenge, participants had access to the challenge repository on GitHub,[10] where they could (and still can) find the source code of the evaluation script and a Docker template containing a baseline algorithm to be replaced for participation.

As mentioned earlier, all participants were asked to publish their code (a mandatory condition for the top-ranking methods), and these implementations are now shared with the research community through the challenge GitHub repository.[10]

All members of teams participating in the challenge qualify as authors of their respective submissions. This paper summarizes the challenge results and includes descriptions of the main contributions; up to three members per team have been included as co-authors, with the sole exception of the first-ranked team, for which all members were listed.

All participants were also invited to submit a description of their algorithms to the workshop associated with the challenge. Accepted papers were published in the corresponding proceedings (Wang et al., 2025b).

Participants were allowed to submit their own results in any venue (conferences, workshops, etc) with embargo restriction: 6 months after the MICCAI 2024 event. An embargo exception was given only for the ToothFairy2-associated workshop.

Funding and Acknowledgments

This work was supported by the University of Modena and Reggio Emilia and Fondazione di Modena through the "Fondo di Ateneo per la Ricerca - FAR 2024" (CUP E93C24002080007), and by the Italian Ministry of Research,

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

under the complementary actions to the NRRP "Fit4MedRob - Fit for Medical Robotics" (PNC0000007).

This work was also supported by the Radboud Dental AI Hub (no funding). Niels van Nistelrooij and Shankeeth Vinayahalingam are an employee and co-founder, respectively, of Ardim B.V. (Nijmegen, the Netherlands), an AI start-up in the field of ultrasound technology for hip dysplasia.

Fabian Isensee, Yannick Kirchhoff, Klaus H. Maier-Hein, Lars Krämer, Maximilian Rokuss, and Constantin Ulrich are funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science. The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health - Helmholtz Information and Data Science School for Health".

Marek Wodzinski and Henning Müller gratefully acknowledge the Polish HPC infrastructure PLGrid support within the computational grant no. PLG/2024/017079.

The challenge organizers gratefully acknowledge SeeThrough S.r.l. for providing the funding for the challenge's monetary prizes.

Disclosures

During the preparation of this work, the authors used ChatGPT in order to polish specific sections of the manuscript. Authors provided all scientific content and references, reviewed and verified all GPT-edited text. The authors take full responsibility for the content of the published article.

Affiliations

Federico Bolelli, *Luca Lumetti*, *Kevin Marchesini*, *Ettore Candeloro*, *Gabriele Rosati* and *Costantino Grana* are with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Italy.

Niels van Nistelrooij, *Shankeeth Vinayahalingam*, *Tong Xi*, and *Stefaan Bergé* are with the Department of Oral and Maxillofacial Surgery, Radboud University Medical Center, the Netherlands.

Mattia Di Bartolomeo is with the Department of Oral and Maxillofacial Sciences, Sapienza University of Rome, Italy.

Arrigo Pellacani is with the Cranio-Maxillo-Facial Unit, University Hospital of Modena, Italy.

Fabian Isensee, *Yannick Kirchhoff*, *Klaus Maier-Hein*, *Lars Krämer*, *Maximilian Rokuss*, and *Constantin Ulrich* are with the German Cancer Research Center (DKFZ), Heidelberg, Division of Medical Image Computing, Germany.

Fabian Isensee and Lars Krämer are with the Helmholtz Imaging, German Cancer Research Center (DKFZ), Heidelberg, Germany.

Yannick Kirchhoff is with HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany, and from the Faculty of Mathematics and Computer Science, Heidelberg University, Germany.

Klaus Maier-Hein is with the Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Germany.

Maximilian Rokuss is with the Faculty of Mathematics and Computer Science, Heidelberg University, Germany.

Yuxian Jiang, *Yusheng Liu*, and *Lisheng Wang* are with the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, China.

Haoshen Wang, *Siyu Chen*, and *Zhiming Cui* are with the School of Biomedical Engineering, ShanghaiTech University, China.

Pengcheng Shi is with the Harbin Institute of Technology (Shenzhen), Electronic & Information Engineering School, China, Shenzhen.

Zhaohong Pan and *Xiaokun Liang* are with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China.

Zhaohong Pan is with the University of Chinese Academy of Sciences, China.

Qi Ma and *Ender Konukoglu* are with the Biomedical Image Computing group, ETH Zurich, Switzerland.

Henning Müller and *Marek Wodzinski* are with the HES-SO Valais, Information Systems Institute, Switzerland.

Marek Wodzinski is with the AGH University of Krakow, Department of Measurement and Electronics, and the Sano Centre for Computational Medicine, Poland.

Henning Müller is with the University of Geneva, Medical Faculty, Switzerland.

Haipeng Mai and *Xiaobing Dang* are with the Guangdong Janus Biotechnology Co., Ltd.

Shrajan Bhandary and *Radu Grosu* are with the Cyber-Physical Systems, Technische Universität Wien, Vienna.

Alexandre Anesi is with the Department of Medical and Surgical Sciences for Children & Adults, Cranio and Maxillo-Facial Unit, University of Modena and Reggio Emilia, Italy.

References

Abdolali, F., Zoroofi, R.A., Abdolali, M., Yokota, F., Otake, Y., Sato, Y., 2017. Automatic segmentation of mandibular canal in cone beam CT images using conditional statistical shape model and fast marching. International Journal of Computer Assisted Radiology and Surgery 12, 581–593.

Allegretti, S., Bolelli, F., Cancilla, M., Pollastri, F., Canalini, L., Grana, C., 2019. How does Connected Components Labeling with Decision Trees perform on GPUs?, in: Computer Analysis of Images and Patterns, Springer. pp. 39–51. doi:10.1007/978-3-030-29888-3_4.

Ben-Hamadou, A., Smaoui, O., Rekik, A., Pujades, S., Boyer, E., Lim, H., Kim, M., Lee, M., Chung, M., Shin, Y.G., Leclercq, M., Cevidanes, L., Carlos P., J., Zhuang, S., Wei, G., Cui, Z., Yuanfeng, Z., Dascalu, T., Ibragimov, B., Yong, T.H., Ahn, H.G., Kim, W., Han, J.H., Choi, B., van Nistelrooij, N., Kempers, S., Vinayahalingam, S., Strippoli, J., Thollot, A., Setbon, H., Trosset, C., Ladroit, E., 2023. 3DTeethSeg'22: 3D Teeth Scan Segmentation and Labeling Challenge. arXiv preprint arXiv:2305.18277 .

Blacher, J., Van DaHuvel, S., Parashar, V., Mitchell, J.C., 2016. Variation in Location of the Mandibular Foramen/Inferior Alveolar Nerve Complex Given Anatomic Landmarks Using Cone-beam Computed Tomographic Scans. Journal of Endodontics 42, 393–396.

Bolelli, F., Baraldi, L., Cancilla, M., Grana, C., 2018. Connected Components Labeling on DRAGs, in: 24th International Conference on Pattern Recognition (ICPR), pp. 89–93. doi:10.1109/ICPR.2018.8545505.

Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., van Nistelrooij, N., van Lierop, P., Xi, T., Liu, Y., Xin, R., Yang, T., Wang, L., Wang, H., Xu, C., Cui, Z., Wodzinski, M., Müller, H., Kirchhoff, Y., Rokuss, M.R., Maier-Hein, K., Han, J., Kim, W., Ahn, H.G., Szczepański, T., Grzeszczyk, M.K., Korzeniowski, P., Caselles-Ballester, V., Paolo Burgos-Artizzu, X., Prados Carrasco, F., Berge', S., van Ginneken, B., Anesi, A., Grana, C., 2025a. Segmenting the Inferior Alveolar Canal in CBCTs Volumes: The ToothFairy Challenge. IEEE Transactions on Medical Imaging 44, 1890–1906. doi:10.1109/TMI.2024.3523096.

Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C., 2025b. Segmenting Maxillofacial Structures in CBCT Volumes, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Cancilla, M., Canalini, L., Bolelli, F., Allegretti, S., Carrión, S., Paredes, R., Gómez, J.A., Leo, S., Piras, M.E., Pireddu, L., Badouh, A., Marco-Sola, S., Alvarez, L., Moreto, M., Grana, C., 2021. The DeepHealth Toolkit: A Unified Framework to Boost Biomedical Applications, in: 2020

25th International Conference on Pattern Recognition (ICPR). doi:https://doi.org/10.1109/ICPR48806.2021.9411954.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306 .

Chen, X., Ma, N., Xu, T., Xu, C., 2024. Deep learning-based tooth segmentation methods in medical imaging: A review. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine 238, 115–131. doi:10.1177/09544119231217603.

Chun, S.Y., Kang, Y.H., Yang, S., Kang, S.R., Lee, S.J., Kim, J.M., Kim, J.E., Huh, K.H., Lee, S.S., Heo, M.S., Yi, W.J., 2023. Automatic classification of 3D positional relationship between mandibular third molar and inferior alveolar canal using a distance-aware network. BMC Oral Health 23, 794.

Cipriano, M., Allegretti, S., Bolelli, F., Di Bartolomeo, M., Pollastri, F., Pellacani, A., Minafra, P., Anesi, A., Grana, C., 2022a. Deep Segmentation of the Mandibular Canal: a New 3D Annotated Dataset of CBCT Volumes. IEEE Access 10, 11500–11510.

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

Cui, W., Wang, Y., Zhang, Q., Zhou, H., Song, D., Zuo, X., Jia, G., Zeng, L., 2022a. CTooth: A Fully Annotated 3D Dataset and Benchmark for Tooth Volume Segmentation on Cone Beam Computed Tomography Images, in: International Conference on Intelligent Robotics and Applications, Springer. pp. 191–200.

Cui, Z., Fang, Y., Mei, L., Zhang, B., Yu, B., Liu, J., Jiang, C., Sun, Y., Ma, L., Huang, J., Liu, Y., Zhao, Y., Lian, C., Ding, Z., Zhu, M., Shen, D., 2022b. A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. Nature Communications 13, 2096.

Cui, Z., Li, C., Wang, W., 2019. ToothNet: Automatic Tooth Instance Segmentation and Identification From Cone Beam CT Images, in: Computer Vision and Pattern Recognition, pp. 6368–6377. doi:10.1109/cvpr.2019.00653.

Cui, Z., Zhang, B., Lian, C., Li, C., Yang, L., Wang, W., Zhu, M., 2021. Hierarchical Morphology-Guided Tooth Instance Segmentation from CBCT Images, in: Information Processing in Medical Imaging, pp. 150–162. doi:10.1007/978-3-030-78191-0_12.

Daza, L., Schnabel, J., 2025. DiENTeS: Dynamic ENTity Segmentation with Local-Global Transformers, in: Wang, Y., Qian, D., Wang, S., Ben-Hamadou, A., Pujades, S., Lumetti, L., Grana, C., Bolelli, F. (Eds.), Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data, Springer Nature Switzerland, Cham. pp. 21–29. doi:10.1007/978-3-031-88977-6_3.

Di Bartolomeo, M., Pellacani, A., Bolelli, F., Cipriano, M., Lumetti, L., Negrello, S., Allegretti, S., Minafra, P., Pollastri, F., Nocini, R., Colletti, G., Chiarini, L., Grana, C., Anesi, A., 2023. Inferior Alveolar Canal Automatic Detection with Deep Learning CNNs on CBCTs: Development of a Novel Model and Release of Open-Source Dataset and Algorithm. Applied Sciences 13, 3271.

Dou, W., Gao, S., Mao, D., Dai, H., Zhang, C., Zhou, Y., 2022. Tooth instance segmentation based on capturing dependencies and receptive field adjustment in cone beam computed tomography. Computer Animation and Virtual Worlds 33, e2100. doi:10.1002/cav.2100.

Duan, W., Chen, Y., Zhang, Q., Lin, X., Yang, X., 2021. Refined tooth and pulp segmentation using U-Net in CBCT image. Dentomaxillofacial Radiology 50, 20200251. doi:10.1259/dmfr.20200251.

Farhat, M., Ben-hamadou, A., Rekik, A., Abida, O., Smaoui, O., 2025. IoSR: End-to-End Intraoral Scans Repairing, in: 36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025, BMVA. pp. 1–12.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On Calibration of Modern Neural Networks, in: International Conference on Machine Learning, PMLR. pp. 1321–1330.

Hao, J., Nalley, A., Yeung, A.W.K., Tanaka, R., Ai, Q.Y.H., Lam, W.Y.H., Shan, Z., Leung, Y.Y., AlHadidi, A., Bornstein, M.M., Tsoi, J.K.H., McGrath, C., Hung, K.F., 2025. Characteristics, licensing, and ethical considerations of openly accessible oral-maxillofacial imaging datasets: a systematic review. npj Digital Medicine 8, 412.

Hao, J., Zhu, Y., He, L., Liu, M., Tsoi, J.K.H., Hung, K.F., 2026. T-Mamba: A unified framework with Long-Range Dependency in dual-domain for 2D & 3D Tooth Segmentation. IEEE Transactions on Multimedia .

He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D., Li, W., 2024. VISTA3D: A Unified Segmentation Foundation Model For 3D Medical Imaging. arXiv:2406.05285.

He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D., 2023. SwinUNETR-V2: Stronger Swin Transformers with Stagewise Convolutions for 3D Medical Image Segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer. pp. 416–426.

Hsu, K., Yuh, D.Y., Lin, S., Lyu, P.S., Pan, G.X., Zhuang, Y.C., Chang, C.C., Peng, H.H., Lee, T.Y., Juan, C.H., Juan, C.E., Liu, Y.J., Juan, C.J., 2022. Improving performance of deep learning models using 3.5D U-Net via majority voting for tooth segmentation on cone beam computed tomography. Scientific Reports 12, 19809. doi:10.1038/s41598-022-23901-7.

Hu, F., Chen, Z., Wu, F., 2024. A novel difficult-to-segment samples focusing network for oral CBCT image segmentation. Scientific Reports 14, 5068. doi:10.21203/rs.3.rs-3748343/v1.

Huang, Y., Liu, W., Yao, C., Miao, X., Guan, X., Lu, X., Liang, X., Ma, L., Tang, S., Zhang, Z., Zhan, J., 2024. A multimodal dental dataset facilitating machine learning research and clinic services. Scientific Data 11, 1291.

Hwang, J.J., Jung, Y.H., Cho, B.H., Heo, M.S., 2019. An overview of deep learning in the field of dentistry. Imaging Science in Dentistry 49, 1–7.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18, 203–211.

Isensee, F., Kirchhoff, Y., Kraemer, L., Rokuss, M., Ulrich, C., Maier-Hein, K.H., 2025a. Scaling nnU-Net for CBCT Segmentation, in: Wang, Y., Qian, D., Wang, S., Ben-Hamadou, A., Pujades, S., Lumetti, L., Grana, C., Bolelli, F. (Eds.), "Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data", Springer Nature Switzerland, Cham. pp. 13–20.

Isensee, F., Rokuss, M., Krämer, L., Dinkelacker, S., Ravindran, A., Stritzke, F., Hamm, B., Wald, T., Langenberg, M., Ulrich, C., Deissler, J., Floca, R., Klaus, M.H., 2025b. nnInteractive: Redefining 3D Promptable Segmentation. arXiv preprint arXiv:2503.08373 .

Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F., 2024. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. arXiv:2404.09556.

ISO, 2016. Dentistry - designation system for teeth and areas of the oral cavity. https://www.iso.org/standard/68292.html. Accessed: 2024-11-11.

Jang, T.J., Kim, K.C., Cho, H.C., Seo, J.K., 2021. A Fully Automated Method for 3D Individual Tooth Identification and Segmentation in dental CBCT. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 6562–6568.

Järnstedt, J., Sahlsten, J., Jaskari, J., Kaski, K., Mehtonen, H., Hietanen, A., Sundqvist, O., Varjonen, V., Mattila, V., Prapayasatok, S., Nalampang, S., 2023. Reproducibility analysis of automated deep learning based localisation of mandibular canals on a temporal CBCT dataset. Scientific Reports 13, 14159.

Jaskari, J., Sahlsten, J., Järnstedt, J., Mehtonen, H., Karhu, K., Sundqvist, O., Hietanen, A., Varjonen, V., Mattila, V., Kaski, K., 2020. Deep Learning Method for Mandibular Canal Segmentation in Dental Cone Beam Computed Tomography Volumes. Scientific Reports 10, 5842. doi:10.1038/s41598-020-62321-3.

Jiang, Y., Liu, Y., Ji, C., Wang, L., 2025. Enhanced Multi-structure Segmentation in CBCT Images with Adaptive Structure Optimization, in: Wang, Y., Qian, D., Wang, S., Ben-Hamadou, A., Pujades, S., Lumetti, L., Grana, C., Bolelli, F. (Eds.), Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data, Springer Nature Switzerland, Cham. pp. 30–40.

Jungo, A., Reyes, M., 2019. Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer. pp. 48–56.

Kaasalainen, T., Ekholm, M., Siiskonen, T., Kortesniemi, M., 2021. Dental cone beam CT: An updated review. Physica Medica 88, 193–217.

Kainmueller, D., Lamecker, H., Seim, H., Zinser, M., Zachow, S., 2009. Automatic Extraction of Mandibular Nerve and Bone from Cone-Beam CT Data, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2009: 12th International Conference, London, UK, September 20-24, 2009, Proceedings, Part II 12, Springer. pp. 76–83.

Kalkhof, J., Mukhopadhyay, A., 2023. M3D-NCA: Robust 3D Segmentation

with Built-in Quality Control, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer. pp. 169–178.

Kroon, D.J., 2011. Segmentation of the Mandibular Canal in Cone-Beam CT Data. Ph.D. thesis. University of Twente. Netherlands. doi:10.3990/1.9789036532808. 10.3990/1.9789036532808.

Lahoud, P., Diels, S., Niclaes, L., Van Aelst, S., Willems, H., Van Gerven, A., Quirynen, M., Jacobs, R., 2022. Development and validation of a novel artificial intelligence driven tool for accurate mandibular canal segmentation on CBCT. Journal of Dentistry 116, 103891.

Lahoud, P., EzEldeen, M., Beznik, T., Willems, H., Leite, A., Van Gerven, A., Jacobs, R., 2021. Artificial Intelligence for Fast and Accurate 3-Dimensional Tooth Segmentation on Cone-beam Computed Tomography. Journal of Endodontics 47, 827–835. doi:10.1016/j.joen.2020.12.020.

Li, G., Lu, Y., Wu, G., Wang, L., Ma, R., 2026. PMCanalSeg: A dataset for automatic segmentation of the pterygopalatine and mandibular canals from 3D CBCT images. Scientific Data .

Li, X., 2024. 3D multimodal dental dataset based on CBCT and oral scan. doi:10.6084/m9.figshare.26965903.v3.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88.

Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., Zheng, H., Wang, S., 2024a. Swin-UMamba: Mamba-Based UNet with ImageNet-Based Pretraining, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer. pp. 615–625.

Liu, K., Elbatel, M., Chu, G., Shan, Z., Sum, F.H.K.M.H., Hung, K.F., Zhang, C., Li, X., Yang, Y., 2025. FDTooth: Intraoral Photographs and CBCT Images for Fenestration and Dehiscence Detection. Scientific Data 12, 1007.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y., 2024b. VMamba: Visual State Space Model. Advances in neural information processing systems 37, 103031–103063.

Lumetti, L., Marchesini, K., Pipoli, V., Ficarra, E., Grana, C., Bolelli, F., 2025. Taming Mambas for 3D Medical Image Segmentation. IEEE Access doi:https://doi.org/10.1109/ACCESS.2025.3570461.

Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C., 2024a. Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. IEEE Access , 1–12doi:10.1109/ACCESS.2024.3408629.

Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C., 2024b. Location Matters: Harnessing Spatial Information to Enhance the Segmentation of the Inferior Alveolar Canal in CBCTs, in: 2024 27th International Conference on Pattern Recognition (ICPR). doi:https://doi.org/10.1007/978-3-031-78104-9_8.

Lumetti, L., Pipoli, V., Bolelli, F., Grana, C., 2023. Annotating the Inferior Alveolar Canal: the Ultimate Tool, in: Image Analysis and Processing - ICIAP 2023, pp. 525–536. doi:10.1007/978-3-031-43148-7_44.

Lv, J., Zhang, L., Xu, J., Li, W., Li, G., Zhou, H., 2023. Automatic segmentation of mandibular canal using transformer based neural networks. Frontiers in Bioengineering and Biotechnology 11.

Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B., 2024a. Segment Anything in Medical Images. Nature Communications 15, 654. doi:10.1038/s41467-024-44824-z.

Ma, J., Li, F., Kim, S., Asakereh, R., Le, B.H., Nguyen-Vu, D.K., Pfefferle, A., Wei, M., Gao, R., Lyu, D., et al., 2024b. Efficient MedSAMs: Segment Anything in Medical Images on Laptop. arXiv preprint arXiv:2412.16085 .

Ma, J., Li, F., Wang, B., 2024c. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. arXiv preprint arXiv:2401.04722 .

Ma, Q., Sun, G., Tombak, G.I., Jain, S., Huber, N.B., Gool, L.V., Konukoglu, E., 2025. Video Foundation Model for Medical 3D Segmentation, in: Wang, Y., Qian, D., Wang, S., Ben-Hamadou, A., Pujades, S., Lumetti, L., Grana, C., Bolelli, F. (Eds.), Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data, Springer Nature Switzerland, Cham. pp. 72–88. doi:10.1007/978-3-031-88977-6_8.

Mai, H., Dang, X., Chen, J., Guo, J., Chen, X., 2025. A Multi-Axial Network for Oral Structure Segmentation, in: Wang, Y., Qian, D., Wang, S., Ben-Hamadou, A., Pujades, S., Lumetti, L., Grana, C., Bolelli, F. (Eds.), Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data, Springer Nature Switzerland, Cham. pp. 49–62. doi:10.1007/978-3-031-88977-6_6.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., Feldmann, C., Frangi, A.F., Full, P.M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B.A., März, K., Maier, O., Maier-Hein, K., Menze, B.H., Müller, H., Neher, P.F., Niessen, W., Rajpoot, N., Sharp, G.C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A.A., van der Sommen, F., Wang, C.W., Weber, M.A., Zheng, G., Jannin, P., Kopp-Schneider, A., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nature Communications 9, 5217.

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., Wiesenfarth, M., Kavur, A.E., Sudre, C.H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M.B., Cardoso, M.J., Cheplygina, V., Cimini, B.A., Collins, G.S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D.A., Hoffman, M.M., Huisman, M., Jannin, P., Kahn, C.E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B.A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A.L., Mattson, P., Meijering, E., Menze, B., Moons, K.G.M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C.I., Shetty, S., van Smeden, M., Summers, R.M., Taha, A.A., Tiulpin, A., Tsaftaris, S.A., Van Calster, B., Varoquaux, G., Jäger, P.F., 2024. Metrics reloaded: Recommendations for image analysis validation. Nature Methods , 1–18.

Mercadante, C., Cipriano, M., Bolelli, F., Pollastri, F., Di Bartolomeo, M., Anesi, A., Grana, C., 2021. A Cone Beam Computed Tomography Annotation Tool for Automatic Detection of the Inferior Alveolar Nerve Canal, in: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, pp. 724–731. doi:10.5220/0010392307240731.

Moris, B., Claesen, L.J.M., Sun, Y., Politis, C., 2012. Automated tracking of the mandibular canal in CBCT images using matching and multiple hypotheses methods. 2012 Fourth International Conference on Communications and Electronics (ICCE) , 327–332.

Morrison, A., Chiarot, M., Kirby, S., 2002. Mental Nerve Function After Inferior Alveolar Nerve Transposition for Placement of Dental Implants. Journal-Canadian Dental Association 68, 46–50.

Nagarajappa, A.K., Dwivedi, N., Tiwari, R., 2015. Artifacts: The downturn of CBCT image. Journal of International Society of Preventive and Community Dentistry 5, 440–445.

van Nistelrooij, N., Krämer, L., Kempers, S., Beyer, M., Bolelli, F., Xi, T., Bergé, S., Heiland, M., Maier-Hein, K.H., Vinayahalingam, S., Isensee, F., 2025. ToothSeg: Robust Tooth Instance Segmentation and Numbering in CBCT using Deep Learning and Self-Correction. IEEE Journal of Biomedical and Health Informatics doi:https://doi.org/10.1109/JBHI.2025.3650444.

Piérard, S., Halin, A., Cioppa, A., Deliège, A., Van Droogenbroeck, M., 2025. Foundations of the Theory of Performance-Based Ranking, in: Computer Vision and Pattern Recognition, pp. 14293–14302.

Pollastri, F., Maroñas, J., Bolelli, F., Ligabue, G., Paredes, R., Magistroni, R., Grana, C., 2021a. Confidence Calibration for Deep Renal Biopsy Immunofluorescence Image Classification, in: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 1298–1305. doi:https://dx.doi.org/10.1109/ICPR48806.2021.9412685.

Pollastri, F., Parreño, M., Maroñas, J., Bolelli, F., Paredes, R., Ramos, D., Grana, C., 2021b. A Deep Analysis on High Resolution Dermoscopic Image Classification. IET Computer Vision 15, 514–526. doi:10.1049/cvi2.12048.

Rao, Y., Wang, Y., Meng, F., Pu, J., Sun, J., Wang, Q., 2020. A Symmetric Fully Convolutional Residual Network With DCRF for Accurate Tooth Segmentation. IEEE Access 8, 92028–92038. doi:10.1109/ACCESS.2020.2994592.

Rekik, A., Ben-Hamadou, A., Smaoui, O., Bouzguenda, F., Pujades, S., Boyer, E., 2025. TSegLab: Multi-stage 3D dental scan segmentation and labeling. Computers in Biology and Medicine 185, 109535. doi:10.1016/j.compbiomed.2024.109535.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer. pp. 234–241.

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3D fully convolutional deep networks, in: Inter-

national workshop on machine learning in medical imaging, Springer. pp. 379–387.

Schramm, A., Rücker, M., Sakkas, N., Schön, R., Düker, J., Gellrich, N.C., 2005. The use of cone beam CT in cranio-maxillofacial surgery. International Congress Series 1281, 1200–1204. doi:10.1016/j.ics.2005.03.224.

Sengupta, N., Sarode, S.C., Sarode, G.S., Ghone, U., 2022. Scarcity of publicly available oral cancer image datasets for machine learning research. Oral Oncology 126, 105737.

Shaheen, E., Leite, A., Alqahtani, K.A., Smolders, A., Van Gerven, A., Willems, H., Jacobs, R., 2021. A novel deep learning system for multi-class tooth segmentation and classification on cone beam computed tomography. a validation study. Journal of Dentistry 115, 103865. doi:10.1016/j.jdent.2021.103865.

Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Shahbaz Khan, F., 2024. UNETR++: Delving Into Efficient and Accurate 3D Medical Image Segmentation. IEEE Transactions on Medical Imaging 43, 3377–3390. doi:10.1109/TMI.2024.3398728.

Usman, M., Rehman, A., Saleem, A.M., Jawaid, R., Byon, S.S., Kim, S.H., Lee, B.D., Heo, M.S., Shin, Y.G., 2022. Dual-Stage Deeply Supervised Attention-Based Convolutional Neural Networks for Mandibular Canal Segmentation in CBCT Scans. Sensors 22, 9877. doi:10.3390/s22249877.

Vinayahalingam, S., Kempers, S., Schoep, J., Hsu, T.M.H., Moin, D.A., van Ginneken, B., Flügge, T., Hanisch, M., Xi, T., 2023. Intra-oral scan segmentation using deep learning. BMC Oral Health 23, 643.

Wang, C., Zhang, Y., Wu, C., Liu, J., Huang, X., Wu, L., Wang, Y., Feng, X., Lu, Y., Wang, Y., 2025a. MMDental-A multimodal dataset of tooth CBCT images with expert medical records. Scientific Data 12, 1172.

Wang, Y., Li, Z., Wu, C., Liu, J., Zhang, Y., Ni, J., Luo, Q., Chen, J., Zhang, H., Liu, J., et al., 2026. Miccai sts 2024 challenge: Semi-supervised instance-level tooth segmentation in panoramic x-ray and cbct images. Medical Image Analysis , 103986.

Wang, Y., Qian, D., Wang, S., Ben-Hamadou, A., Pujades, S., Lumetti, L., Grana, C., Bolelli, F. (Eds.), 2025b. Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data. volume 15571 of *Lecture Notes in Computer Science*. Springer, Cham. doi:10.1007/978-3-031-88977-6.

Wang, Y., Xia, W., Yan, Z., Zhao, L., Bian, X., Liu, C., Qi, Z., Zhang, S., Tang, Z., 2023. Root canal treatment planning by automatic tooth and root canal segmentation in dental CBCT with deep multi-task feature learning. Medical Image Analysis 85, 102750. doi:10.1016/j.media.2023.102750.

Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Scientific Reports 11, 2369.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 1–9.

Wodzinski, M., Müller, H., 2025. Automatic Multi-structure Segmentation in Cone Beam Computed Tomography Volumes Using Deep Encoder-Decoder Architectures, in: Wang, Y., Qian, D., Wang, S., Ben-Hamadou, A., Pujades, S., Lumetti, L., Grana, C., Bolelli, F. (Eds.), Supervised and Semi-supervised Multi-structure Segmentation and Landmark Detection in Dental Data, Springer Nature Switzerland, Cham. pp. 63–71. doi:10.1007/978-3-031-88977-6_7.

Worthington, P., 2004. Injury of the Inferior Alveolar Nerve during Implant Placement: a Literature Review. International Journal of Oral & Maxillofacial Implants 19.

Wu, X., Chen, H., Huang, Y., Guo, H., Qiu, T., Wang, L., 2020. Center-Sensitive and Boundary-Aware Tooth Instance Segmentation and Classification from Cone-Beam CT, in: International Symposium on Biomedical Imaging, pp. 939–942. doi:10.1109/ISBI45749.2020.9098542.

Zhao, H., Chen, J., Yun, Z., Feng, Q., Zhong, L., Yang, W., 2023. Whole mandibular canal segmentation using transformed dental CBCT volume in Frenet frame. Heliyon 9.

Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y., 2023. nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer. IEEE Transactions on Image Processing .