# Segmenting Maxillofacial Structures in CBCT Volumes

Federico Bolelli[1, ⋆, ✉] Kevin Marchesini[1, ⋆] Niels van Nistelrooij[2, 3] Luca Lumetti[1]
Vittorio Pipoli[1] Elisa Ficarra[1] Shankeeth Vinayahalingam[2] Costantino Grana[1]

[1]University of Modena and Reggio Emilia, Italy
[2]Radboud University Medical Center, the Netherlands
[3]Charité – Universitätsmedizin Berlin, Germany

## Abstract

*Cone-beam computed tomography (CBCT) is a standard imaging modality in orofacial and dental practices, providing essential 3D volumetric imaging of anatomical structures, including jawbones, teeth, sinuses, and neurovascular canals. Accurately segmenting these structures is fundamental to numerous clinical applications, such as surgical planning and implant placement. However, manual segmentation of CBCT scans is time-intensive and requires expert input, creating a demand for automated solutions through deep learning. Effective development of such algorithms relies on access to large, well-annotated datasets, yet current datasets are often privately stored or limited in scope and considered structures, especially concerning 3D annotations. This paper proposes ToothFairy2, a comprehensive, publicly accessible CBCT dataset with voxel-level 3D annotations of 42 distinct classes corresponding to maxillofacial structures. We validate the dataset by benchmarking state-of-the-art neural network models, including convolutional, transformer-based, and hybrid Mamba-based architectures, to evaluate segmentation performance across complex anatomical regions. Our work also explores adaptations to the nnU-Net framework to optimize multi-class segmentation for maxillofacial anatomy. The proposed dataset provides a fundamental resource for advancing maxillofacial segmentation and supports future research in automated 3D image analysis in digital dentistry.*

## 1. Introduction

Medical imaging is currently employed in the clinical practice of dental treatments. Different modalities are available to assist in diagnosis, treatment planning, and surgery:

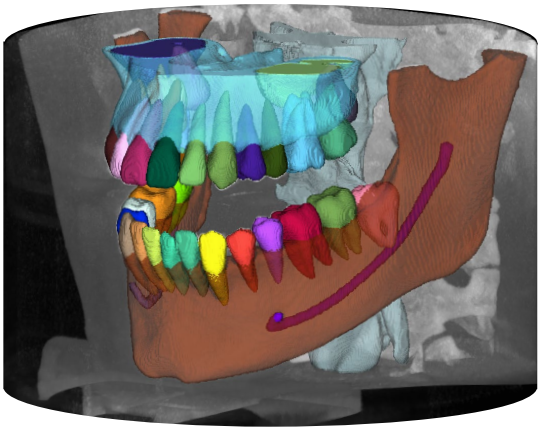


Figure 1. Fully annotated sample from the proposed dataset.

2D panoramic X-rays, 3D Intra-Oral Scans (IOS), and 3D Cone-Beam Computed Tomography (CBCT) [29]. Among the available modalities, CBCT is the only one providing comprehensive 3D volumetric information on all the anatomical structures in the orofacial area, including jawbones, complete teeth, sinuses, and neurovascular canals, becoming a standard modality for maxillofacial image analysis. Segmenting individual structures from CBCT images to reconstruct a precise 3D model is essential in digital dentistry for various clinical applications, including surgical planning and implant placement [16, 53].

Manual segmentation of CBCT scans is a time-consuming and labor-intensive process that requires specialized expertise. For this reason, automated algorithms using deep learning techniques offer a promising solution to alleviate this burden, enhance efficiency, and improve the consistency of segmentation results [41]. However, developing robust deep learning models heavily depends on the availability of large, annotated datasets. Unfortunately, existing datasets often focus on a limited number of anatomical structures —i.e., teeth or alveolar canal— and usually lack 3D annotations. The few available methods trained on

⋆Equal contribution. Authors are allowed to list their names first in their respective CVs.

✉Corresponding author: *federico.bolelli@unimore.it*.

Table 1. Datasets used in literature to segment maxillofacial anatomical structures. Our analysis is limited to 3D imaging modalities, i.e., CBCTs and IOS. ✣ means that only a portion of the 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 | Year | Country | # Train & Validation | # Test | Label Type | | Public |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Train | Test | |
| Inferior Alveolar Canal (IAC) | CBCT | Jaskari et al. [28] | 2020 | Finland | 509 | 128 | 2D | 128 2D, 15 3D | ✗ |
| | | Lahoud et al. [33] | 2022 | Belgium | 205 | 30 | 3D | 3D | ✗ |
| | | Usman et al. [48] | 2022 | South Korea | 510 | 500 | 3D | 3D | ✗ |
| | | Cipriano et al. [11] | 2022 | Italy | 332 | 15 | 332 2D, 76 3D | 15 2D, 15 3D | ✓ |
| | | Chun et al. [10] | 2023 | South Korea | 32 | 18 | 3D | 3D | ✗ |
| | | Bolelli et al. [6] | 2023 | Italy, Netherlands | 443 | 50 | 443 2D, 153 3D | 3D | ✓ |
| Teeth Crown, Teeth Root | CBCT | Cui et al. [16] | 2022 | China | 4 531 | 407 | 3D | 3D | ✣ |
| | | Cui et al. [13] | 2022 | China | 22 | ∼ | 3D | ∼ | ✓ |
| | | Dou et al. [17] | 2022 | China | 35 | 5 | 3D | 3D | ✗ |
| | | Tae Jun Jang et al. [26] | 2022 | China | 66 | 7 | 66 3D, 31 2D | 7 3D, 4 2D | ✗ |
| Teeth Crown | IOS | Ben-Hamadou et al. [3] | 2022 | Tunisia | 1 200 | 600 | 3D | 3D | ✓ |
| | | Shankeeth et al. [49] | 2023 | Netherlands | 1 400 | 350 | 3D | 3D | ✗ |
| IACs, Teeth, Others | CBCT | Ours | 2024 | Italy, Netherlands | 480 | 50 | 3D | 3D | ✓ |

3D datasets leverage very small-sized training and testing sets ($<$ 50 CBCTs) [10, 13, 17], most of the time unavailable to the research community, limiting their generalizability to scans acquired with other imaging protocols and their applicability to diverse patient populations.

In this paper, we introduce a novel and unique dataset with 3D annotations of 42 distinct anatomical classes on maxillofacial CBCTs. The classes include jawbones, left and right Inferior Alveolar Canals (IACs), left and right maxillary sinus, pharynx, upper and lower teeth, including wisdom teeth, bridges, crowns, and dental implants. The proposed comprehensive dataset provides a rich resource for developing and benchmarking segmentation algorithms that require 3D anatomical annotations, providing a strong basis for the future development and evaluation of automated methods for maxillofacial image analysis (Fig. 1).

To validate the dataset, we further explore the performance of various state-of-the-art neural networks used in the literature for medical image segmentation: classical Convolutional Neural Networks (CNNs) like nnU-Net and its variants [24], transformer-based models such as TransU-Net [7] and UNETR++ [47], and hybrid Mamba-based architectures like UMamba [38], Swin-UMamba [34], and VMamba [35].

Specifically, we explore how to adapt and optimize the nnU-Net framework [24], a state-of-the-art, self-configuring segmentation method, to enhance the segmentation of all 42 anatomical structures in our dataset, by adjusting hyperparameters and augmentation strategies.

By leveraging our comprehensive dataset, we aim to advance the development of segmentation algorithms that can handle the complexities of 3D maxillofacial imaging. Our work not only contributes a valuable dataset to the research community but also provides insights into the optimization of neural network architectures for improved segmentation performance in complex anatomical regions.

In summary, our contributions are:

- The largest publicly available CBCT dataset with detailed 3D annotations covering 42 distinct maxillofacial anatomical structures, addressing a critical gap in existing research resources;
- A comprehensive comparative evaluation of state-of-the-art neural network architectures, encompassing convolutional networks, transformers, and hybrid models, to establish performance benchmarks using our dataset;
- Adaptation and optimization of the nnU-Net framework to target multi-class segmentation in complex maxillofacial anatomy, highlighting the impact of customized augmentation and hyperparameters tuning.

## 2. Related Works

### 2.1. Existing datasets

In maxillofacial analysis, existing research predominantly emphasizes developing AI applications that demonstrate high accuracy on carefully selected datasets; however, limited consideration has been given to examining the datasets employed in the training and evaluation phases of AI model development [45]. Current research has primarily used academic databases, which may not fully address AI-specific data needs, including the FAIR (Findability, Accessibility, Interoperability, and Reusability) principles [52]. Only a few dental studies make datasets available to the research community [6, 12, 13, 16], making it essential to seek alternative data sources to mitigate these limitations and to prioritize the creation of high-quality, medically validated, and AI-ready datasets. All of these datasets are confined to specific anatomical structures, particularly the IAC [12] or the teeth complex [16], and others provide binary segmentation classes only, without considering tooth-level labels [16].

Tab. 1 presents a comprehensive comparison of datasets used in state-of-the-art research.

### 2.2. Teeth segmentation

Multiple approaches to segmenting teeth in 3D data have been proposed, each introducing different challenges in clinical and computational applications [8].

**Binary segmentation.** When employing a simple binary semantic segmentation, the primary goal is to separate teeth from surrounding tissues without any distinction between individual teeth. While it is useful for basic assessments, its clinical utility is limited, especially when dealing with adjacent or closely spaced teeth. To improve accuracy, recent methods have utilized CNN-based architectures with custom modules, hybrid loss functions [21], a combination of 2D and 3D networks [20], and post-processing techniques, such as posterior probability maps and dense conditional random fields [42].

**Instance segmentation.** By contrast, when performing instance segmentation, teeth are segmented as separate instances, though they have not yet been assigned specific labels. Instance segmentation methods usually leverage a two-stage approach consisting of detection followed by segmentation. The detection stage locates individual tooth instances through various techniques, including bounding box detection [14, 18], heatmap prediction [54], or offset regression [15–17]. Once identified, localized regions around each tooth instance can be extracted to perform binary segmentation and identify precise boundaries.

**Multi-class segmentation.** With this approach, each tooth must be identified and assigned a specific label [32, 46] according to the FDI (Fédération Dentaire Internationale, also known as World Dental Federation) notation [25]. This is the most challenging of the three approaches, especially in cases with a complex anatomy, such as missing and impacted teeth, and will likely continue to benefit from more robust datasets and enhanced model architectures, motivating once again our proposed dataset.

Current literature can be split into single- and multi-stage methods. While the former leverages a single framework to produce the final classification, multi-stage methods incorporate additional steps. The more commonly employed strategy consists of downsampling the original CBCT or splitting it into multiple regions (e.g., the four dental arch quadrants) to generate a coarse multi-class segmentation, later refining individual tooth segments by isolating volumes of interest and enhancing local details [46, 50].

### 2.3. IAC segmentation

Since the advent of CBCT technology in the early 2000s, substantial effort has been dedicated by the scientific community to developing automated systems for segmenting the inferior alveolar canal from 3D scans [44]. With the rise of deep learning in medical diagnosis, learning-based networks [23, 27, 28] have outclassed traditional computer vision techniques [1, 4, 30, 31, 40, 51].

Early work by Jaskari et al. [28] applied the U-Net [43] architecture on a coarsely annotated dataset, achieving promising results compared to traditional methods, though limited by the lack of dense annotations.

Lahoud [33] advanced this line of research by training a 3D U-Net with both interpolated control points and finer voxel-level annotations, though, again, access to data and code was restricted. Significant progress in the segmentation of the IAC came with the release of the first public CBCT dataset containing 2D and 3D annotations of the mandibular canal [11, 12]. Contextually, authors introduced PosPadUNet3D, a modified 3D U-Net that segments IAC in stages by expanding sparse 2D labels into 3D annotations, thus generating additional synthetic labels and using such generated data in combination with expert-labeled 3D voxel data to train the final U-Net based model.

Usman et al. [48] proposed a two-stage U-Net approach to address class imbalance between the mandibular canal and background, using CNNs to isolate regions of interest (ROIs) before segmentation, while Zhao et al. [55] introduced a Frenet frame-based method to better capture mandibular topology and maintain canal structure during segmentation. More recently, Lv et al. [37] developed a transformer-based model with adaptive image processing and a "deep label fusion" technique to enhance label consistency across sparse public data, building on Cipriano's label expansion methodology.

Finally, Lumetti et al. [36] tackled issues with patch-based learning by using a memory-augmented transformer encoder to improve spatial coherence in the U-Net bottleneck, enriching patch context and improving segmentation.

IAC segmentation remains a challenging area with room for further advancements, largely dependent on access to comprehensive, high-quality 3D datasets.

## 3. The ToothFairy2 Dataset

The proposed dataset, ToothFairy2, represents the most extensive publicly available and fully annotated collection of dental CBCT scans. It comprises 530 3D volumes, of which 480 are entirely accessible for training purposes[1] while the remaining 50 are reserved exclusively for evaluation. Although these 50 scans are not directly accessible for model training, researchers can evaluate their models on this test set via the grand-challenge platform,[2] which automatically runs the submitted models and provides performance results. In this way, we ensure that researchers will not incur the risk of mixing training and testing data and that the reported results will not suffer from the random choice of the testing data subset selection.

The *training data* (480 scans) were acquired by the Af-

[1] https://ditto.ing.unimore.it/toothfairy2/.
[2] https://toothfairy2.grand-challenge.org/.

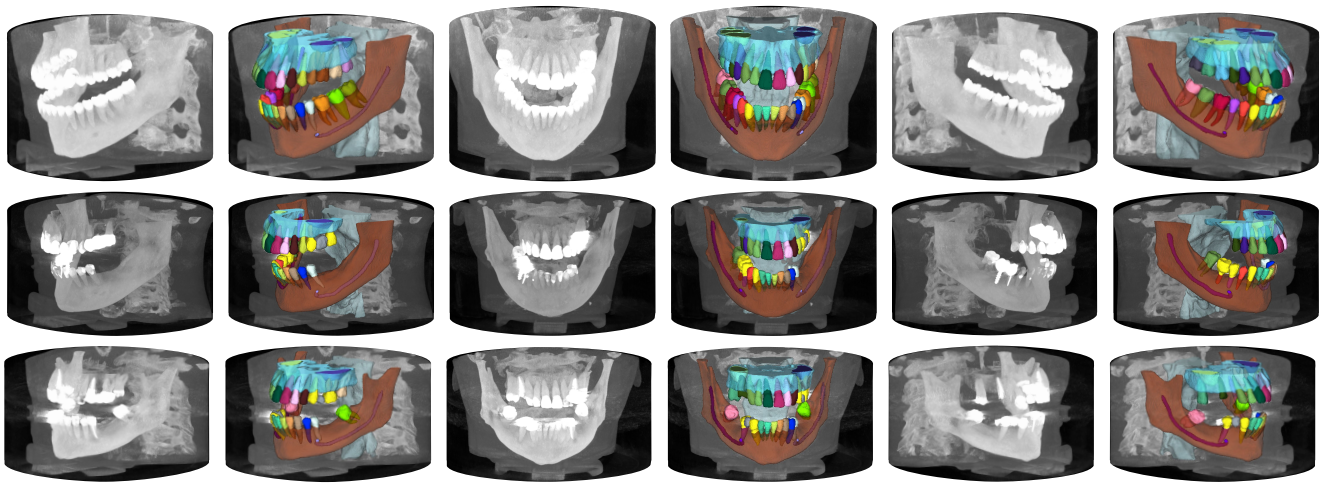


Figure 2. Dataset samples, one patient per line. For each view, i.e., left, frontal, and right, both raw CBCT data and labels are provided.

fidea Center, a pan-European healthcare group that specializes in advanced diagnostics, laboratory analyses, rehabilitation, and cancer diagnosis and treatment. The scans were obtained through Cone Beam Computed Tomography (CBCT) via a *NewTom/NTVGiMK4*, operating at 3 mA, 110 kV, and offering 0.3 mm isotropic voxels (Fig. 2). The *test data* (50 scans) were provided by the Department of Oral and Maxillofacial Surgery at Radboud University Medical Centre, Nijmegen, obtained using a standard CBCT scanning protocol with the *i-CAT 3D Imaging System*. The scans were performed with a field of view measuring 16 cm in diameter and 22 cm in height. The scanning process involved two scans of 20 seconds each, resulting in an isotropic voxel size of 0.4 mm. These were later rescaled to match the voxel size of the training data before annotation.

Every patient was anonymized and we only kept a few personal details —namely gender, age, and year of the scan. Specifically, 58.30% of the patients are female (58.54% in the training set, 56.00% in the test set), all the scans were performed between 2019 and 2024, and volumes belong to patients with ages in range (10-100] with the highest frequencies in ranges (20-30] and (60-70] for the training set and (50-70] for the test set.

### 3.1. Annotation protocol and tools

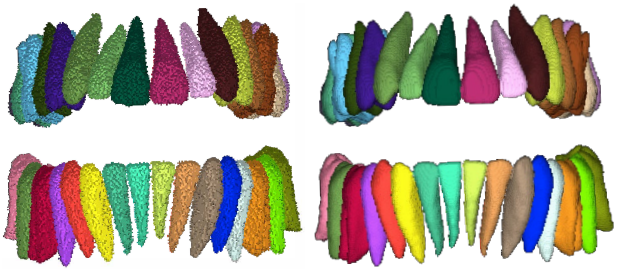
The inclusion of densely labeled 3D data is necessary in order to achieve the full potential of CNNs [22]. 3D annotations refer to the annotation process carried out by medical professionals directly at the voxel level on CBCT scans. This annotation approach involves detailed markings applied to the individual voxels, providing a more precise depiction of the structures involved.

As often happens in medical image annotation, accurate delineation of individual teeth and maxillofacial structures is a complex and time-consuming process, which requires expert eyes to be achieved. To reduce the burden of the annotators while ensuring high-quality labels, we employed a semi-automated approach that was fully supervised by clinicians. The annotation process was performed by 7 different maxillofacial experts. To prevent models trained on our dataset from being rewarded for learning annotator-specific biases and to allow for the understanding of their generalizability, 5 of the annotators focused on the training set only, while the remaining 2 handled the test cases only.

Since labeling voxels directly on 3D views is ambiguous and prone to errors, our annotations have been performed on 2D slices. Moreover, to reduce the jagged effect that usually occurs when annotating 3D objects on 2D, the annotation is performed with multiple iterations, starting from the axial (or transversal) view and later refined by analyzing the scan from the sagittal and coronal (or frontal) planes.

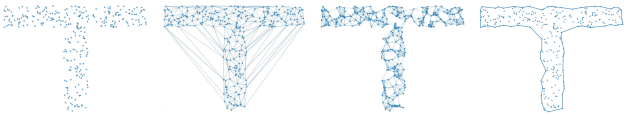
For each patient, a single annotation is provided, including the following classes: upper and lower jawbone, left and right inferior alveolar canal, maxillary sinuses, pharynx, and upper and lower teeth, including wisdom teeth, bridges, crowns, and dental implants. The annotation was produced by a single expert only but validated by a second one to reduce possible mistakes.

In our work, we leverage five different *base models* built on the nnU-Net framework to provide an initial annotation that is later adjusted and refined by expert clinicians. Each of the models has been trained on a specific (group of) classes, including the jawbone (upper and lower), inferior alveolar canal (left and right branches), maxillary sinus (left and right), pharynx, and teeth. Regarding inferior alveolar canal branches and teeth, we leverage two datasets available in the literature, [12] and [16], respectively. Even when annotating the first volumes of our dataset, clinicians were provided with an initial segmentation of teeth and canals and were required to modify, update, or remove wrongly an-



(a) Dense Point Set (b) Voxelized Teeth

Figure 3. Annotations post-processing. The annotations produced as described in Sec. 3 are dense and jagged (a). For this reason, we compute a concave $\alpha$-shape, which, after voxelization, produces the volumes in (b). Such an approach is applied to all of the anatomical structures involved in our dataset but the jawbones, whose annotations contain multiple holes.



(a) Point Set (b) Triangulation (c) $\alpha$-complex (d) $\alpha$-shape

Figure 4. $\alpha$-shape construction process for the point set of (a). In the Delaunay triangulation (b), triangles with circumradius $\leq -\frac{1}{\alpha}$ form a simplicial subcomplex known as $\alpha$-complex (c), whose border is the $\alpha$-shape (d).

notated voxels. Following an iterative approach, as soon as 20 more annotated volumes were available, the base models were retrained from scratch by including the new data. Following such an approach, clinicians are provided with better annotation proposals at each iteration, reducing the annotation effort and time at the increase of annotated volumes. Since jawbone, maxillary sinus, and pharynx classes were not available in any public dataset, these classes have been annotated from scratch in the first 20 volumes without leveraging any automatic proposal.

### 3.2. Annotation refinement

Even though the annotation process is performed on multiple planes, one after the other, artifacts due to 3D annotations performed on 2D views remain a challenge, albeit reduced (Fig. 3a). In order to obtain a smooth polygonal mesh out of expert-produced annotations, we compute a concave $\alpha$-*shape*. The $\alpha$-shape [19] is a generalization of the convex hull aimed at representing the intuitive concept of the shape of a point set. The only parameter of the algorithms is $\alpha \in \mathbb{R}$, which regulates the "crudeness" of the result. Let's defines a *generalized disk of radius* $\frac{1}{\alpha}$ as $D_\alpha$:

$$D_\alpha \begin{cases} \text{The complement of a disc of radius } -\frac{1}{\alpha}, & \text{if } \alpha < 0 \\ \text{A halfplane}, & \text{if } \alpha = 0 \\ \text{A disc of radius } \frac{1}{\alpha}, & \text{if } \alpha > 0 \end{cases} \quad (1)$$

Given a point set $S$ and a value for $\alpha$, the $\alpha$-shape is constructed in this way: an edge is put between two points $p_i$ and $p_j$ whenever there exists a $D_\alpha$ with $p_i$ and $p_j$ lying on its boundary, and which contains the entire $S$. When $\alpha = 0$, this procedure constructs the convex hull; instead, cruder or finer shapes can be respectively obtained using positive or negative $\alpha$ values. Because of the geometrical nature of the alveolar nerve, we are specifically interested in concave $\alpha$-shapes, achievable when $\alpha < 0$.

The most common method for computing a concave $\alpha$-shape consists of taking the border of a simplicial subcomplex extracted from the Delaunay triangulation, containing only triangles with circumradius $\leq -\frac{1}{\alpha}$. An example of the process is depicted in Fig. 4.

The above concepts can be extended to the three-dimensional case by substituting disks and triangles with spheres and tetrahedra, respectively. After creating an $\alpha$-shape polygonal mesh starting from dense annotations, it is converted into a binary raster volume by means of voxelization: the final result is given in Fig. 3b.

### 3.3. Accessibility

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.

**Ethics approval.** The training data received ethical committee approval from Comitato Etico dell'Area Vasta Emilia Nord (Approval Number 1374/2020/OSS/ESTMO SIRER ID 1275 - NAICBCT-D) and can be downloaded under the CC BY-SA license after user registration.

## 4. Experiments

In order to validate the effectiveness of the proposed dataset, experiments have been performed considering two different datasets: the one proposed with this paper and the publicly released subset of the dataset proposed by Cui et al. [15], which is the only dataset publicly available in the literature for multi-class tooth instance segmentation. Many of the dental-related scientific papers promising the release of the dataset upon request [2, 9] have ignored our emails.

### 4.1. Evaluation metrics

In our experiments, two widely accepted metrics for the segmentation task [39] have been employed, namely Dice Similarity Coefficient (DSC in %) and the 95th percentile Hausdorff Distance (HD95 in mm).

The DSC has the same meaning as the IoU (Intersection over Union), but DSC is better suited when the region of interest is significantly smaller than the background. In such

Table 2. Results on the test set of the proposed dataset. Classes are grouped by the main anatomical structure to which they belong. Default nnU-Net plan is employed. The best results are in **bold**, while the second best are underlined.

| | Model | Average | | L/R IAC | | L/R Sinus | | Teeth | | Jawbones | | Pharynx | | Others | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DSC | HD95 | DSC | HD95 | DSC | HD95 | DSC | HD95 | DSC | HD95 | DSC | HD95 | DSC | HD95 |
| CNNs | nnU-Net [24] | 70.92 | 17.86 | 71.34 | 29.11 | 64.81 | 28.39 | 73.17 | 18.32 | 90.31 | 12.53 | **95.66** | 19.23 | 29.50 | 20.95 |
| | nnU-Net ResEnc [24] | 74.16 | 14.48 | 73.01 | 27.81 | 65.71 | 29.99 | 76.48 | 14.26 | 91.77 | 12.53 | 95.26 | 17.75 | 37.10 | 16.32 |
| Transf. | TransU-Net [7] | 70.32 | 20.17 | 81.96 | 11.99 | 59.69 | 59.76 | 72.46 | 15.04 | 90.33 | 49.02 | 87.89 | 42.87 | 27.68 | 22.65 |
| | nnFormer [56] | 76.79 | 5.45 | 72.28 | 10.06 | 75.11 | 8.22 | 79.37 | 2.94 | 91.50 | 20.09 | 90.85 | 24.53 | 18.95 | **13.67** |
| | UNETR++ [47] | 71.43 | 17.23 | 68.87 | 15.10 | 74.51 | 15.48 | 73.70 | 17.91 | 92.38 | 5.44 | 91.60 | 18.96 | 26.21 | 19.70 |
| Mamba | UMamba [38] | **85.05** | 5.28 | **85.26** | 16.23 | 77.02 | 4.35 | **86.58** | **2.23** | 90.05 | 22.17 | 92.18 | 25.25 | **43.89** | 13.97 |
| | VMamba [35] | 73.13 | 5.17 | 60.62 | 9.94 | 73.20 | 9.89 | 75.99 | 3.39 | 90.75 | 7.17 | 88.23 | 23.95 | 14.86 | 14.63 |
| | Swin-UMamba [34] | 79.64 | **2.94** | 72.64 | **2.02** | **87.75** | **2.38** | 80.71 | 2.59 | **94.29** | **4.25** | 93.05 | **2.79** | 25.30 | 14.93 |

a context, more weight is given to the correctly identified region, making the DSC more robust and informative than IoU. The relationship between DSC and IoU is expressed by the following formula:

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

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

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

$$\mathrm{d}_{95}(A, B) = \mathrm{x}^{95}_{a \in A} \left\{ \min_{b \in B} d(a, b) \right\} \tag{3}$$

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

$$\text{HD95(P,GT)} = \max \left\{ d_{95}(P, GT), d_{95}(GT, P) \right\} \tag{4}$$

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.

## 4.2. Compared algorithms

Experimental analysis has been conducted on recently proposed general-purpose state-of-the-art algorithms for segmenting medical 3D volumes, considering all of the top-notch architecture (i.e., CNNs, Transformers, and Mamba-based hybrid solutions). For what concerns CNNs, we include the original nnU-Net [24] configuration making use of the U-Net architecture (nnU-Net), and its variations leveraging residual connections in the encoder (nnU-Net ResEnc). The considered Tranformer-based architectures are TransU-Net [7], UNETR++ [47], and the recently published nnFormer [56]. Finally, we include UMamba [38], VMamba [35], and Swin-UMamba [34] as the representatives of solutions based on state-space models.

## 4.3. Experimental setting

In all of the experiments carried out, we adopted a standardized scheme for hyperparameters configuration. More specifically, we leverage the planning provided by the nnU-Net framework. For intrinsically 3D algorithms, i.e., nnU-Net, TransU-Net, nnFormer, UNETR++, and UMamba, the nnU-Net planning applied to the training set of the proposed dataset suggests a patch size of $80 \times 160 \times 160$, and a batch size of 2. The only exception is the nnU-Net ResEnc, which adopts a patch size of $122 \times 224 \times 256$. On the other hand, models that produce 3D segmentations working slice-by-slice, i.e., VMamba and Swin-UMamba, have a suggested patch size of $384 \times 384$ and a batch size of 19.

All the images in the training set of the proposed dataset have the same spacing ($0.3\,mm$ isotropic), so no resampling is required. The volumes from the secondary dataset employed in our experiments, Cui et al. [16], have a mixed voxel spacing instead, so they have been resampled to match $0.3\,mm \times 0.3\,mm \times 0.3\,mm$.

Models are trained from scratch without any pre-training data. The nnU-Net three-fold cross-validation schema has always been employed for model selection. All the models have been trained for a total of 1000 epochs on a 48GB A40 Nvidia GPU using CUDA 11.8 and PyTorch 2.1.2.

## 4.4. Results

To fully explore the capabilities of state-of-the-art models on the proposed comprehensive dataset, Tab. 2 is provided. For space constraints, results are grouped by anatomical structures by averaging single-class results. The "others" column includes implants, artificial crowns, and bridges. Surprisingly, 2D models, i.e., VMamba and Swin-UMamba, outperform intrinsically 3D ones in many scenarios. Although performance is satisfying, 2D models lack the ability to fully capture and learn the spatial relationships between the 3D structures in the maxillofacial region due to a loss of contextual information and a limited capacity to learn effective representations of complex structures. This is also confirmed by the qualitative results discussed in Sec. 4.7, where it is evident that predictions are more noisy and seem fragmented. Anyway, their local ability to dis-

Table 3. Results on the test set of the proposed dataset when using variations of the nnU-Net ResEnc.

| Model | DSC | HD95 |
|---|---|---|
| Default | 74.16 | 14.48 |
| w/o l/r mirroring | 80.79 | 12.37 |
| w/o l/r mirroring, increased depth | 82.11 | 11.86 |
| w/o l/r mirroring, increased depth, post-processing | 84.99 | 8.57 |

criminate among classes ensures high-level performance.

Overall, Mamba-based architectures are the most effective in this context. Among all, UMamba provides the best performance. Blending convolutions to model precise spatial information and state-space models to learn long-range voxel-level interactions ensures high accuracy. Mamba provides a global context alongside voxel-wise precision, the former missing in traditional convolutional layers due to limited receptive fields and the latter absent in Transformers due to computational complexity. Moreover, it requires less training data to converge, making it more suitable in the medical field where access to data is limited.

Considering the different classes available in the dataset, jawbones and pharynx achieved the highest DSC scores, independently from the considered model. This could be explained by the consistent anatomical shape those structures display throughout the entire dataset. Performance drops when it comes to segmenting classes subject to more variability, like teeth; their position and orientation change considerably from patient to patient, there might be missing teeth, or they can be covered by artificial crowns, or completely replaced by implants or bridges. It is not uncommon for the models to confuse artificial with natural teeth. Independently from the model employed, the worst performances are obtained when segmenting "others" classes, which represent all of the artificial structures that restore or replace natural teeth. One reasonable explanation is that these classes are underrepresented in the proposed dataset.

### 4.5. Improving default nnU-Net configuration

To augment the training dataset and improve performance, the nnU-Net framework applies mirroring along multiple spatial dimensions. Such an augmentation usually improves model performance, providing additional useful training data without compromising the ability to differentiate between left and right organs. In the maxillofacial context, however, there is pronounced symmetry between structures: teeth, nerves, and maxillary sinuses are all symmetrical w.r.t. to the sagittal plane splitting the face. In such a scenario, ensuring precise orientation is fundamental. The application of the left/right mirroring augmentation has exactly the opposite effect: instead of increasing model performance and generalization capabilities, it leads to a drop in the model's ability to establish a reliable left/right distinction, downgrading overall performance (Tab. 3). Additionally, given the complexity of the proposed dataset, we

Table 4. Results on the test set of the proposed dataset when using our plan w/o post-processing.

| | Model | Default Plan | | Our Plan | |
|---|---|---|---|---|---|
| | | DSC | HD95 | DSC | HD95 |
| CNNs | nnU-Net [24] | 70.92 | 17.86 | 78.24 | 13.36 |
| CNNs | nnU-Net ResEnc [24] | 74.16 | 14.48 | 82.11 | 11.86 |
| Transf. | TransUNet [7] | 70.32 | 20.17 | 75.12 | 7.17 |
| Transf. | nnFormer [56] | 76.79 | 5.45 | 79.12 | 5.15 |
| Transf. | UNETR++ [47] | 71.43 | 17.23 | 83.57 | 3.04 |
| Mamba | UMamba [38] | 85.05 | 5.28 | 85.39 | 4.70 |
| Mamba | VMamba [35] | 73.13 | 5.17 | 74.21 | 4.37 |
| Mamba | Swin-UMamba [34] | 79.64 | 2.94 | 81.95 | 2.91 |

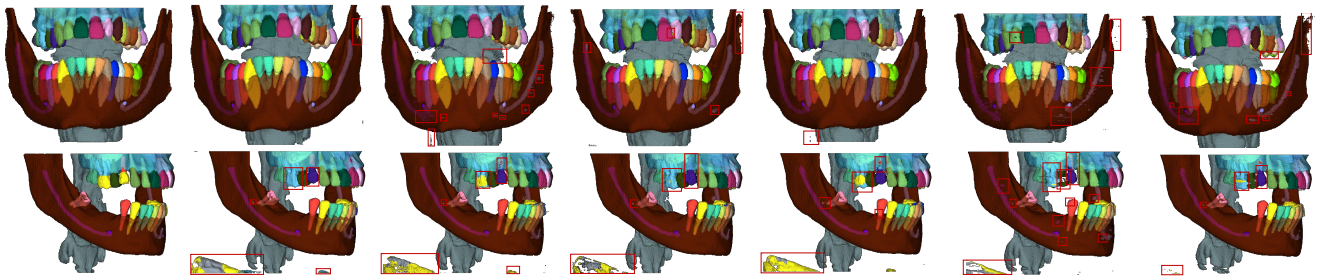
enriched the network topology by introducing an additional layer. Such an enhancement ensures an increased receptive field in the bottleneck of the network, improving contextual information and the model's ability to identify long-range relationships between structures. Finally, considering the nature of the structures involved in the dataset, all of the voxels belonging to the same class should be in a close relationship and connected to each other. For this reason, we introduce a post-processing technique that computes connected components [5] and filters out predictions that are below a given threshold or the smallest prediction if multiple objects for the same class are identified by the model. Thresholds are computed per class using statistics on the training dataset, i.e., minimum connected component volume. Such a procedure ensures us an additional gain of more than 2 DSC points on the proposed dataset.

Finally, in order to confirm the effectiveness of the improved nnU-Net planning on the proposed dataset, i.e., disabling l/r mirroring and the deeper network topology, we apply the same strategy to all of the considered models. Results are reported in Tab. 4.

### 4.6. Comparison with existing datasets

Tab. 5 provides a comparison of state-of-the-art models, one from each category, trained on either the proposed data or a subset of it and Cui et al. [16]. To perform a fair comparison, we selected 82 random scans from both datasets to be used for training. Since [16] includes only tooth annotations, we set all non-teeth classes to background. Selected models are evaluated on our complete test set, the full Cui dataset, and its 15 volumes not used for training.

A column-wise comparison of the results suggests that the volume of data has only a marginal impact on overall tooth segmentation performance but is critical for demographic generalizability. Additionally, models trained on our dataset generalize more effectively to the Cui dataset, while those trained exclusively on Cui experience a dice score reduction of up to 10 points when evaluated on our dataset compared to their own test set. Notably, UMamba demonstrates consistently strong performance across both datasets, highlighting its robust generalization capabilities.



(a) Ground-Truth (b) nnU-Net ResEnc (c) nnFormer (d) UNETR++ (e) UMamba (f) VMamba (g) Swin-UMamba

Figure 5. Qualitative results obtained with the proposed plan w/o post-processing, one patient per line. Red squares identify relevant errors.

Table 5. Dice comparison between our and Cui [16] datasets considering only teeth classes and converting others to background. Our plan w/o post-processing is employed.

| Training Dataset | Model | Testing Dataset | | |
|---|---|---|---|---|
| | | Our (test) | Cui (all) | Cui (15 scans) |
| Our (all the 480 training scans) | nnU-Net [24] | 91.12 | 81.86 | 82.13 |
| | nnFormer [56] | 85.16 | 79.19 | 80.25 |
| | UMamba [38] | 90.97 | 87.60 | 87.90 |
| Our (82 scans from training) | nnU-Net [24] | 90.04 | 79.67 | 81.12 |
| | nnFormer [56] | 82.72 | 74.85 | 77.42 |
| | UMamba [38] | 90.18 | 86.23 | 87.71 |
| Cui (82 scans over 97 available) | nnU-Net [24] | 78.05 | - | 88.52 |
| | nnFormer [56] | 70.63 | | 74.35 |
| | UMamba [38] | 79.95 | | 89.74 |

### 4.7. Qualitative evaluation

Fig. 5 depicts a qualitative comparison of the selected state-of-the-art segmentation models trained on the proposed dataset when using the previously introduced personalized nnU-Net planning. The lack of contextualized 3D receptive fields of the 2D VMamba is evident, especially on elongated objects like the inferior alveolar canal (first line). Transformer-based architectures, UNETR++ and nnFormer, produce more noisy labels. While some can be easily removed by employing post-processing techniques, the majority will remain. Interestingly, all of the 3D models mistakenly label the plastic chin support used to acquire the CBCT with a mix of artificial crown (yellow label, second line) and pharynx (gray label, second line). This has a negative impact on the HD95 of both "other" and pharynx classes. The local discriminative capabilities of 2D models make them stronger, avoiding this misidentification and explaining the lower HD95 values.

All of the models except UNETR++ correctly recognize the first upper premolar implant (red label within the upper jawbone, second line). However, none of them label the artificial crown of the same tooth as such.

## 5. Conclusions and Limitations

In this paper, we presented the largest publicly available annotated CBCT dataset with 3D segmentations for 42 distinct maxillofacial anatomical structures, offering a valuable resource for developing and benchmarking automated segmentation models in digital dentistry and addressing a significant gap in the literature. Our dataset supports detailed multi-class segmentation across complex maxillofacial anatomy, enabling improved model training, evaluation, and generalization for clinical applications such as surgical planning and implant placement. Extensive benchmark evaluations of state-of-the-art neural network architectures on our dataset highlight the potential of existing models in enhancing the accuracy and efficiency of maxillofacial image analysis, also demonstrating the significance of the proposed dataset. Moreover, we demonstrated different techniques that can be adopted w.r.t. the standard nnU-Net planning to improve the overall performance of state-of-the-art models when trained on complex CBCT datasets such as the one proposed in this paper.

**Limitations and future work.** Despite its significant contributions, our dataset has limitations that should be further addressed in the future. Although training and testing data come from different machines and acquisition centers, limiting potential biases in the evaluation, training data was acquired from a single center, potentially affecting the generalizability of models trained on this dataset to other clinical environments or patient demographics. Furthermore, while the dataset includes a wide range of classes, some —implants, crowns, and bridges— are underrepresented, affecting the robustness of model performance. Additionally, some components encountered in maxillofacial imaging, such as dental braces and bone plates, are missing due to their unavailability in the current patient cohort. Future work should focus on expanding this dataset to include multi-center data, ensuring a more representative sample of anatomical and demographic diversity. Including more cases with underrepresented and missing classes would further enhance the dataset's utility, supporting more comprehensive models. We believe that this dataset provides a strong foundation for advancing automated maxillofacial segmentation, and we encourage the research community to build upon it to address the outlined limitations.

## Acknowledgments

This work was supported by the University of Modena and Reggio Emilia and Fondazione di Modena, through the FAR 2024 and FARD-2024 funds (Fondo di Ateneo per la Ricerca 2024) and by the Italian Ministry of Research under the complementary actions to the NRRP "Fit4MedRob - Fit for Medical Robotics" Grant (#PNC0000007).

## References

[1] Fatemeh Abdolali, Reza Aghaeizadeh Zoroofi, Maryam Abdolali, Futoshi Yokota, Yoshito Otake, and Yoshinobu Sato. 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, 2017. 3

[2] Khalid Ayidh Alqahtani, Reinhilde Jacobs, Andreas Smolders, Adriaan Van Gerven, Holger Willems, Sohaib Shujaat, and Eman Shaheen. Deep convolutional neural network-based automated segmentation and classification of teeth with orthodontic brackets on cone-beam computed-tomographic images: a validation study. *European Journal of Orthodontics*, 45(2):169–174, 2023. 5

[3] Achraf Ben-Hamadou, Oussama Smaoui, Ahmed Rekik, Sergi Pujades, Edmond Boyer, Hoyeon Lim, Minchang Kim, Minkyung Lee, Minyoung Chung, Yeong-Gil Shin, et al. 3DTeethSeg'22: 3D Teeth Scan Segmentation and Labeling Challenge. *arXiv preprint arXiv:2305.18277*, 2023. 2

[4] Jonathan Blacher, Scott Van DaHuvel, Vijay Parashar, and John C Mitchell. Variation in Location of the Mandibular Foramen/Inferior Alveolar Nerve Complex Given Anatomic Landmarks Using Cone-beam Computed Tomographic Scans. *Journal of Endodontics*, 42(3):393–396, 2016. 3

[5] Federico Bolelli, Stefano Allegretti, Luca Lumetti, and Costantino Grana. A State-of-the-Art Review with Code about Connected Components Labeling on GPUs. *IEEE Transactions on Parallel and Distributed Systems*, 2024. 7

[6] Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani, Kevin Marchesini, Niels Van Nistelrooij, Pieter Van Lierop, Tong Xi, Yusheng Liu, et al. Segmenting the Inferior Alveolar Canal in CBCTs Volumes: the ToothFairy Challenge. *IEEE Transactions on Medical Imaging*, 2024. 2

[7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2, 6, 7

[8] Xiaokang Chen, Nan Ma, Tongkai Xu, and Cheng Xu. 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 (2):115–131, 2024. 3

[9] Zeyu Chen, Senyang Chen, and Fengjun Hu. CTA-UNet: CNN-transformer architecture UNet for dental CBCT images segmentation. *Physics in Medicine & Biology*, 68(17): 175042, 2023. 5

[10] So-Young Chun, Yun-Hui Kang, Su Yang, Se-Ryong Kang, Sang-Jeong Lee, Jun-Min Kim, Jo-Eun Kim, Kyung-Hoe Huh, Sam-Sun Lee, Min-Suk Heo, et al. Automatic classification of 3D positional relationship between mandibular third molar and inferior alveolar canal using a distance-aware network. *BMC Oral Health*, 23(1):794, 2023. 2

[11] Marco Cipriano, Stefano Allegretti, Federico Bolelli, Mattia Di Bartolomeo, Federico Pollastri, Arrigo Pellacani, Paolo Minafra, Alexandre Anesi, and Costantino Grana. Deep Segmentation of the Mandibular Canal: a New 3D Annotated Dataset of CBCT Volumes. *IEEE Access*, 10:11500–11510, 2022. 2, 3

[12] Marco Cipriano, Stefano Allegretti, Federico Bolelli, Federico Pollastri, and Costantino Grana. 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)*, pages 21137–21146. IEEE, 2022. 2, 3, 4

[13] Weiwei Cui, Yaqi Wang, Qianni Zhang, Huiyu Zhou, Dan Song, Xingyong Zuo, Gangyong Jia, and Liaoyuan Zeng. 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*, pages 191–200. Springer, 2022. 2

[14] Zhiming Cui, Changjian Li, and Wenping Wang. ToothNet: Automatic Tooth Instance Segmentation and Identification From Cone Beam CT Images. In *Computer Vision and Pattern Recognition*, pages 6368–6377, 2019. 3

[15] Zhiming Cui, Bojun Zhang, Chunfeng Lian, Changjian Li, Lei Yang, Wenping Wang, and Min Zhu. Hierarchical Morphology-Guided Tooth Instance Segmentation from CBCT Images. In *Information Processing in Medical Imaging*, pages 150–162, 2021. 3, 5

[16] Zhiming Cui, Yu Fang, Lanzhuju Mei, Bojun Zhang, Bo Yu, Jiameng Liu, Caiwen Jiang, Yuhang Sun, Lei Ma, Jiawei Huang, et al. A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. *Nature Communications*, 13(1):2096, 2022. 1, 2, 4, 6, 7, 8

[17] Wenhan Dou, Shanshan Gao, Deqian Mao, Honghao Dai, Chenhao Zhang, and Yuanfeng Zhou. Tooth instance segmentation based on capturing dependencies and receptive field adjustment in cone beam computed tomography. *Computer Animation and Virtual Worlds*, 33(5):e2100, 2022. 2, 3

[18] Wei Duan, Yufei Chen, Qi Zhang, Xiang Lin, and Xiaoyu Yang. Refined tooth and pulp segmentation using U-Net in CBCT image. *Dentomaxillofacial Radiology*, 50(6): 20200251, 2021. 3

[19] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983. 5

[20] Kang Hsu, Da-Yo Yuh, Sc Lin, Pin-Sian Lyu, Guan-Xin Pan, Yi-Chun Zhuang, Chia-Ching Chang, Hsu-Hsia Peng, Tung-Yang Lee, Cheng-Hsuan Juan, Cheng-En Juan, Yi-Jui Liu, and Chun Jung Juan. 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(1):19809, 2022. 3

[21] Fengjun Hu, Zeyu Chen, and Fan Wu. A novel difficult-to-segment samples focusing network for oral CBCT image segmentation. *Scientific Reports*, 14:5068, 2024. 3

[22] Kuofeng Hung, Andy Wai Kan Yeung, Ray Tanaka, and Michael M Bornstein. Current Applications, Opportunities, and Limitations of AI for 3D Imaging in Dental Research and Practice. *International Journal of Environmental Research and Public Health*, 17(12):4424, 2020. 4

[23] Jae-Joon Hwang, Yun-Hoa Jung, Bong-Hae Cho, and Min-Suk Heo. An overview of deep learning in the field of dentistry. *Imaging Science in Dentistry*, 49(1):1–7, 2019. 3

[24] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 2, 6, 7, 8

[25] ISO. Dentistry - designation system for teeth and areas of the oral cavity. https://www.iso.org/standard/68292.html, 2016. Accessed: 2024-11-11. 3

[26] Tae Jun Jang, Kang Cheol Kim, Hyun Cheol Cho, and Jin Keun Seo. A Fully Automated Method for 3D Individual Tooth Identification and Segmentation in dental CBCT. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6562–6568, 2021. 2

[27] Jorma Järnstedt, Jaakko Sahlsten, Joel Jaskari, Kimmo Kaski, Helena Mehtonen, Ari Hietanen, Osku Sundqvist, Vesa Varjonen, Vesa Mattila, Sangsom Prapayasatok, et al. Reproducibility analysis of automated deep learning based localisation of mandibular canals on a temporal CBCT dataset. *Scientific Reports*, 13(1):14159, 2023. 3

[28] Joel Jaskari, Jaakko Sahlsten, Jorma Järnstedt, Helena Mehtonen, Kalle Karhu, Osku Sundqvist, Ari Hietanen, Vesa Varjonen, Vesa Mattila, and Kimmo Kaski. Deep Learning Method for Mandibular Canal Segmentation in Dental Cone Beam Computed Tomography Volumes. *Scientific Reports*, 10(1):5842, 2020. 2, 3

[29] Touko Kaasalainen, Marja Ekholm, Teemu Siiskonen, and Mika Kortesniemi. Dental cone beam CT: An updated review. *Physica Medica*, 88:193–217, 2021. 1

[30] Dagmar Kainmueller, Hans Lamecker, Heiko Seim, Max Zinser, and Stefan Zachow. 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*, pages 76–83. Springer, 2009. 3

[31] Dirk-Jan Kroon. *Segmentation of the Mandibular Canal in Cone-Beam CT Data*. PhD thesis, University of Twente, Netherlands, 2011. 10.3990/1.9789036532808. 3

[32] Pierre Lahoud, Mostafa EzEldeen, Thomas Beznik, Holger Willems, André Leite, Adriaan Van Gerven, and Reinhilde Jacobs. Artificial Intelligence for Fast and Accurate 3-Dimensional Tooth Segmentation on Cone-beam Computed Tomography. *Journal of Endodontics*, 47(5):827–835, 2021. 3

[33] Pierre Lahoud, Siebe Diels, Liselot Niclaes, Stijn Van Aelst, Holger Willems, Adriaan Van Gerven, Marc Quirynen, and Reinhilde Jacobs. Development and validation of a novel artificial intelligence driven tool for accurate mandibular canal segmentation on CBCT. *Journal of dentistry*, 116:103891, 2022. 2, 3

[34] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 615–625. Springer, 2024. 2, 6, 7

[35] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual State Space Model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 6, 7

[36] Luca Lumetti, Vittorio Pipoli, Federico Bolelli, Elisa Ficarra, and Costantino Grana. Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. *IEEE Access*, pages 1–12, 2024. 3

[37] Jinxuan Lv, Lang Zhang, Jiajie Xu, Wang Li, Gen Li, and Hengyu Zhou. Automatic segmentation of mandibular canal using transformer based neural networks. *Frontiers in Bioengineering and Biotechnology*, 11, 2023. 3

[38] Jun Ma, Feifei Li, and Bo Wang. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 2, 6, 7, 8

[39] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, pages 1–18, 2024. 5

[40] Behnam Moris, Luc J. M. Claesen, Yi Sun, and Constantinus Politis. Automated tracking of the mandibular canal in CBCT images using matching and multiple hypotheses methods. *2012 Fourth International Conference on Communications and Electronics (ICCE)*, pages 327–332, 2012. 3

[41] Archie Morrison, Marco Chiarot, and Stuart Kirby. Mental Nerve Function After Inferior Alveolar Nerve Transposition for Placement of Dental Implants. *Journal-Canadian Dental Association*, 68(1):46–50, 2002. 1

[42] Yunbo Rao, Yilin Wang, Fanman Meng, Jiansu Pu, Jihong Sun, and Qifei Wang. A Symmetric Fully Convolutional Residual Network With DCRF for Accurate Tooth Segmentation. *IEEE Access*, 8:92028–92038, 2020. 3

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 3

[44] A. Schramm, M. Rücker, N. Sakkas, R. Schön, J. Düker, and N.-C. Gellrich. The use of cone beam CT in craniomaxillofacial surgery. *International Congress Series*, 1281: 1200–1204, 2005. 3

[45] Namrata Sengupta, Sachin C Sarode, Gargi S Sarode, and Urmi Ghone. Scarcity of publicly available oral cancer image datasets for machine learning research. *Oral Oncology*, 126:105737, 2022. 2

[46] Eman Shaheen, André Leite, Khalid Ayidh Alqahtani, Andreas Smolders, Adriaan Van Gerven, Holger Willems, and Reinhilde Jacobs. 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, 2021. 3

[47] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. UNETR++: Delving Into Efficient and Accurate 3D Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 43(9):3377–3390, 2024. 2, 6, 7

[48] Muhammad Usman, Azka Rehman, Amal Muhammad Saleem, Rabeea Jawaid, Shi-Sub Byon, Sung-Hyun Kim, Byoung-Dai Lee, Min-Suk Heo, and Yeong-Gil Shin. Dual-Stage Deeply Supervised Attention-Based Convolutional Neural Networks for Mandibular Canal Segmentation in CBCT Scans. *Sensors*, 22(24):9877, 2022. 2, 3

[49] Shankeeth Vinayahalingam, Steven Kempers, Julian Schoep, Tzu-Ming Harry Hsu, David Anssari Moin, Bram van Ginneken, Tabea Flügge, Marcel Hanisch, and Tong Xi. Intra-oral scan segmentation using deep learning. *BMC Oral Health*, 23(1):643, 2023. 2

[50] Yiwei Wang, Wenjun Xia, Zhennan Yan, Liang Zhao, Xiaohe Bian, Chang Liu, Zhengnan Qi, Shaoting Zhang, and Zisheng Tang. 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, 2023. 3

[51] Xueqiong Wei and Yuanjun Wang. Inferior alveolar canal segmentation based on cone-beam computed tomography. *Medical Physics*, 48(11):7074–7088, 2021. 3

[52] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016. 2

[53] Philip Worthington. Injury of the Inferior Alveolar Nerve during Implant Placement: a Literature Review. *International Journal of Oral & Maxillofacial Implants*, 19(5), 2004. 1

[54] Xiyi Wu, Huai Chen, Yijie Huang, Huayan Guo, Tiantian Qiu, and Lisheng Wang. Center-Sensitive and Boundary-Aware Tooth Instance Segmentation and Classification from Cone-Beam CT. In *International Symposium on Biomedical Imaging*, pages 939–942, 2020. 3

[55] Huanmiao Zhao, Junhua Chen, Zhaoqiang Yun, Qianjin Feng, Liming Zhong, and Wei Yang. Whole mandibular canal segmentation using transformed dental CBCT volume in Frenet frame. *Heliyon*, 9(7), 2023. 3

[56] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer. *IEEE Transactions on Image Processing*, 2023. 6, 7, 8