ToothSeg: Robust Tooth Instance Segmentation and Numbering in CBCT using Deep Learning and Self-Correction

Niels van Nistelrooij, Lars Krämer, Steven Kempers, Michel Beyer, Federico Bolelli, Tong Xi, Stefaan Bergé, Max Heiland, Klaus H. Maier-Hein, Shankeeth Vinayahalingam, and Fabian Isensee

***Abstract*— Accurate interpretation of cone-beam computed tomography (CBCT) scans is critical for oral diagnosis and treatment planning. Existing methods for automated tooth segmentation in CBCT face challenges, such as difficulties in generalizing across imaging artifacts and anatomical variations, as well as requiring manual revisions in many cases. To address these limitations, this study introduces ToothSeg, a fully automated approach for tooth instance segmentation and numbering in CBCT using deep learning and self-correction. ToothSeg combines semantic and instance segmentation into a unified method where their respective strengths are complemented. In particular, self-correction is employed when combining the segmentations, resolving merged or split teeth and determining the optimal sequence of tooth numbers for each dental arch. We conducted a comprehensive evaluation using a diverse in-house dataset (n = 1282, 25+ devices) and the publicly available ToothFairy2 challenge dataset (n = 480, 1 device), including an ablation study, a comparison to state-of-the-art methods, and an analysis of challenging cases. Compared to an optimized semantic segmentation model, including instance segmentation and self-correction consistently improved tooth segmentation (True Positive Dice: 93.6% to 94.3%) and tooth detection and numbering (multiclass instance F1: 94.2% to 95.5%). Furthermore, ToothSeg outperformed the other methods on both datasets (True Positive Dice: $\geq$ +0.4%, multiclass instance F1: $\geq$ +1.8%), particularly for challenging cases. This study provides a promising approach for automated tooth segmentation and numbering in CBCT, which is significant for reducing manual workload and supporting scalable, data-driven research in oral and craniofacial health. Code and models are publicly available at https://github.com/MIC-DKFZ/ToothSeg.**

***Index Terms*— 3D Segmentation, Cone-Beam Computed Tomography, Deep Learning, Dental Imaging, Tooth Instance Segmentation and Numbering**

Manuscript submitted January 4, 2026. This work was supported by the Radboud Dental AI Hub and partially funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science.

N. van Nistelrooij and L. Krämer contributed equally.

S. Vinayahalingam and F. Isensee contributed equally.

N. van Nistelrooij, S. Kempers, T. Xi, S. Bergé, and S. Vinayahalingam are with the Department of Oral and Maxillofacial Surgery, Radboud University Medical Center, Nijmegen, Netherlands (e-mail: {Niels.vanNistelrooij, Steven.Kempers, Tong.Xi, Stefaan.Berge, Shankeeth.Vinayahalingam}@radboudumc.nl)

M. Heiland is with the Department of Oral and Maxillofacial Surgery, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Germany (e-mail: {max.heiland}@charite.de)

L. Krämer, K. H. Maier-Hein, and F. Isensee are with Helmholtz Imaging, German Cancer Research Center (DKFZ) and the Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany (e-mail: {lars.kraemer, k.maier-hein, f.isensee}@dkfz-heidelberg.de)

M. Beyer is with the Department of Oral and Cranio-Maxillofacial Surgery, University Hospital Basel and the Medical Additive Manufacturing Research Group (Swiss MAM), Department of Biomedical Engineering, University of Basel, Switzerland (e-mail: michel.beyer@usb.ch)

F. Bolelli is with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Italy (email: federico.bolelli@unimore.it)

I. Introduction

MEDICAL image analysis has advanced significantly through integration of deep learning [1], [2]. It is not surprising, therefore, that data-driven dentistry promises to revolutionize oral diagnosis, treatment decision-making, and patient communication [3]–[5]. Interpreting dental image modalities, such as panoramic radiographs, intraoral scans, and cone-beam computed tomography (CBCT) scans [6], [7], is fundamental. CBCT, in particular, provides a volumetric view of the oral and maxillofacial region at a reduced radiation exposure compared to conventional computed tomography (CT) [8], [9]. CBCT enables 3D evaluation of scanned areas, aiding in the assessment of anatomical structures, implant sites, root canal morphology, and the visualization of impacted teeth, tooth alignment, and tooth localization [10]. Moreover, CBCT examinations often lead to a better-informed treatment plan [11].

Tooth segmentation in CBCT is a fundamental component in daily clinical procedures, but its effectiveness highly depends on the examiner's experience [12], [13]. Therefore, the efficient and precise automated interpretation of dental data provides many opportunities for improving diagnostic accuracy and streamlining treatment planning [14], [15]. A 3D reconstruction of the teeth can be used as a comprehensive overview of a patient's dental anatomy, to support radiographic interpretation and patient communication [16], [17]. Furthermore, numbering of the individual teeth is crucial for precise and consistent documentation. In addition to its direct applications, the efficient and automated interpretation facilitates the analysis of large datasets, providing the opportunity to gain new insights in the setting of clinical research. However, the

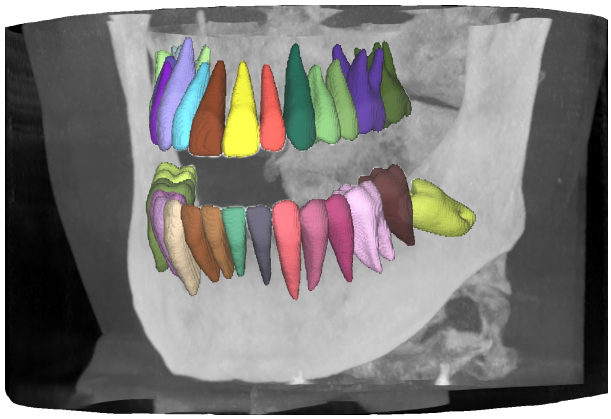

Fig. 1: Tooth instance predictions for a CBCT scan using ToothSeg. By combining semantic and instance segmentation with self-correction, our method ensures precise segmentation and accurate numbering, robustly handling metal-induced artifacts and anatomical variations.

current standard for determining a 3D reconstruction is based on semi-automated segmentation of CBCT, which introduces examiner-dependent variation and is too time-consuming for routine dental practice [18], [19].

Tooth instance segmentation and numbering involves two challenges: accurately segmenting each tooth instance and correctly assigning the appropriate tooth numbers. From a technical perspective, the difficulty in solving this task arises from the complex anatomical structures in the oral and maxillofacial region. Teeth have a large variability in shape, size, and orientation across patients, and imaging artifacts, such as noise and metal-induced distortions, further complicate segmentation. As a result, many existing methods lack robustness and fail to generalize across different datasets, often requiring manual revisions to achieve acceptable accuracy. This dependence on human oversight undermines the goal of fully automated workflows in clinical practice. The reliability of the segmentation method is relevant for clinical application, particularly when dealing with missing teeth, as the lack of contextual information makes it challenging to distinguish between teeth with similar shapes. This is complemented by the inherent shortcomings that arise from CBCT, such as limited soft tissue contrast, higher noise, and scattering artifacts [20], [21]. In addition to the problems on the methodological side, there is a lack of publicly available datasets, and source code of existing methods is not available or not easily usable [22]. Often small ($<$ 100 CBCT scans) and in-house datasets are used that may not reflect the general population, for example by excluding pathological cases [23]. This makes meaningful benchmarking difficult and undermines the reliability of evaluations. Additionally, the adaptability of the methods to other datasets that differ, for example in terms of scanners, fields of view, or patient symptoms and populations, is not evaluated, which leads to overfitting and overengineering for individual use cases [24], [25].

To address these limitations, we present ToothSeg, a fully automated approach using deep learning and self-correction designed to overcome the inherent variability of teeth in CBCT scans. In contrast to existing methods that mostly depend on multistage detect-then-segment pipelines, ToothSeg combines two independent branches for instance and semantic segmentation to mitigate error propagation, achieving a high degree of robustness. Tooth instance segmentation is implemented using a border–core semantic formulation and region-growing, and both branches use baseline nnU-Net models to avoid architectural overfitting. In addition, a probabilistic self-correction approach is employed when combining the branch outputs to automatically resolve tooth numbering inconsistencies and merged or split teeth without dataset-specific adjustments. Our approach eliminates the need for manual revisions in most cases, offering a solution that surpasses the current state-of-the-art on two large datasets. To validate our method's performance, a diverse dataset comprising 1,282 CBCT scans was compiled and annotated [14], [26], [27]. Additionally, we validated our method on the recently released ToothFairy2 challenge dataset (Fig. 1) [28]–[30]. Our method's ability to generalize across varied dental conditions and imaging artifacts demonstrates its potential for widespread clinical application. By eliminating the need for manual revisions in most cases, in combination with accurate tooth numbering, ToothSeg can streamline clinical workflows and reduce diagnostic variability. In addition to the technical contributions, we embrace open science by publishing the source code of our method and all comparison methods. By including validation on a public dataset, we embrace accessibility and reproducibility.

II. Related Work

Tooth segmentation in CBCT scans is typically formulated as one of three tasks. The first is binary semantic segmentation, which separates teeth from surrounding tissues without distinguishing individual teeth. While useful, this approach has limited clinical utility due to difficulties in assessing closely spaced or adjacent teeth. The second task, tooth instance segmentation, identifies individual teeth but does not assign specific labels to them. The third task, tooth instance segmentation and numbering, not only identifies individual teeth but also assigns them unique numbers according to the FDI tooth numbering system [31]. Since a fully automated method for tooth instance segmentation and numbering can provide significant benefits to clinical practice, numerous deep learning methods have been developed [32].

A. Binary semantic segmentation

This task focuses on separating teeth from the background. A publicly available dataset, CTooth, has been introduced, featuring expert annotations on 22 CBCT scans [33]. Various convolutional neural network (CNN) approaches have been proposed for binary segmentation, incorporating advanced network architectures with custom modules and hybrid loss functions [34], as well as combinations of 2D and 3D networks [35]. Additionally, postprocessing techniques have been employed to refine CNN predictions, including the use of posterior probability maps based on grayscale values [36] and dense conditional random fields [37]. Another approach

extended the segmentation task by differentiating between the background, jaws, and teeth, enabling a more detailed 3D reconstruction [38].

B. Tooth instance segmentation

The second task involves identifying individual tooth instances. This is typically performed using a detect-then-segment approach. The detection can be implemented through various techniques, including bounding box detection [39], [40], multiclass segmentation [41], tooth center heatmap prediction [42], or location offset regression [43]–[45]. Once tooth instances are identified, volumes of interest (VOIs) can be cropped from the CBCT scan around each identified tooth, after which another model predicts a binary segmentation of the tooth at the center of each VOI. Additional tasks can be integrated into the method to enhance its effectiveness. For example, a second input channel with predicted tooth boundaries can assist in tooth detection [39], whereas tooth boundary segmentation and tooth apices keypoint detection can provide auxiliary supervision for individual tooth segmentation [42], [43], [45]. Various network architectures incorporating self-attention, dilated convolutions, and dense skip connections, have been explored to improve effectiveness [42], [44]. Additionally, focal loss and boundary-aware Dice loss have been employed to focus more on challenging tooth boundary voxels [34], [42].

C. Tooth instance segmentation and numbering

This task presents two key challenges: precisely segmenting each tooth instance and assigning the appropriate FDI numbers. The FDI system attributes a unique number to each tooth from a set of 32 numbers. Tooth instance segmentation and numbering could thus be implemented using multiclass semantic segmentation with 33 output classes within a single deep neural network, offering greater efficiency compared to multistage approaches [41], [46]. However, the trade-off is often reduced effectiveness, as the model may struggle to specialize in both tooth segmentation and numbering simultaneously. Additionally, individual teeth can be mistakenly split into multiple classes (splitters), or multiple tooth instances may be assigned the same tooth number (mergers). Alternative approaches incorporate tooth numbering into the detect-then-segment approach for tooth instance segmentation. Shaheen et al. [41] downsampled the complete CBCT scan to predict a coarse multiclass semantic segmentation, identifying tooth instances. Then, tooth crops were extracted from the original scan to produce finer segmentations of individual teeth. Lee et al. [47] also downsampled the complete CBCT scan and predicted 32 heatmaps and bounding box sizes to perform tooth detection and numbering. As before, individual teeth were cropped from the original scan to predict a tooth instance segmentation. Wang et al. [15] performed tooth numbering after the detection stage. They used a common encoder with three decoders for binary segmentation, centroid offset regression, and tooth numbering, respectively. The outputs of the first two decoders were combined to determine tooth instances, with volumes of interest (VOIs) cropped around these instances. Features from the third decoder for tooth numbering were then pooled and processed by a multi-layer perceptron (MLP) to assign tooth numbers. One recent method preprocessed a CBCT scan by standardizing its intensity value distribution, after which a first stage predicted a binary tooth segmentation, which was used to inform the second stage for multiclass semantic segmentation [48]. Other studies integrated tooth numbering directly into the model for individual tooth segmentation by processing latent features with global pooling and an MLP [39], [45]. These methods operating on small VOIs around detected tooth instances, often struggle to incorporate broader contextual information necessary for accurate tooth numbering. As a result, they exhibit reduced effectiveness in cases with significant anatomical variations. Such multistage approaches were typically prone to error propagation, where inaccuracies in an early stage could add up and lead to degraded performance in subsequent stages.

III. Method

This study followed the principles of the Declaration of Helsinki. The need for ethical approval was waived by the institutional review board (CMO Arnhem-Nijmegen, file number 2021-13253). Each subject provided their informed consent.

Tooth instance segmentation and numbering involves identifying and segmenting individual teeth, then assigning a specific tooth number to each one. Conceptually, this could be handled as semantic segmentation since each tooth number appears at most once per image. However, semantic segmentation struggles with voxel-wise predictions that lack object-level understanding, often causing teeth to be fragmented with multiple numbers. Despite this, semantic segmentation offers key advantages: its voxel-wise supervision is data-efficient and delivers highly precise predictions.

Therefore, semantic segmentation could be a strong basis for tooth identification but requires reinforcement through dedicated tooth instance segmentation strategies. To address this, we introduce ToothSeg, a hybrid methodology that combines the complementary strengths of semantic and instance segmentation. The proposed approach comprises three main components: a semantic branch responsible for assigning tooth numbers, an instance branch that precisely delineates individual teeth, and an algorithm for combining the semantic and instance predictions that employs self-correction to resolve merged or split teeth and that determines an optimal sequence of tooth numbers. Our model corrects errors inherent to each branch individually, delivering a highly robust solution. Both branches of ToothSeg are based on nnU-Net [49]. In contrast to existing methods, our approach obviates the need for cropping around the volume of interest (VOI). Instead, ToothSeg generates two independent predictions for the entire image, which are subsequently combined with self-correction to produce the final result. Fig. 2 provides an overview of how the semantic and instance branches are combined.

A. Semantic segmentation branch

Within the proposed method, the semantic segmentation branch is responsible for allocating the correct FDI number

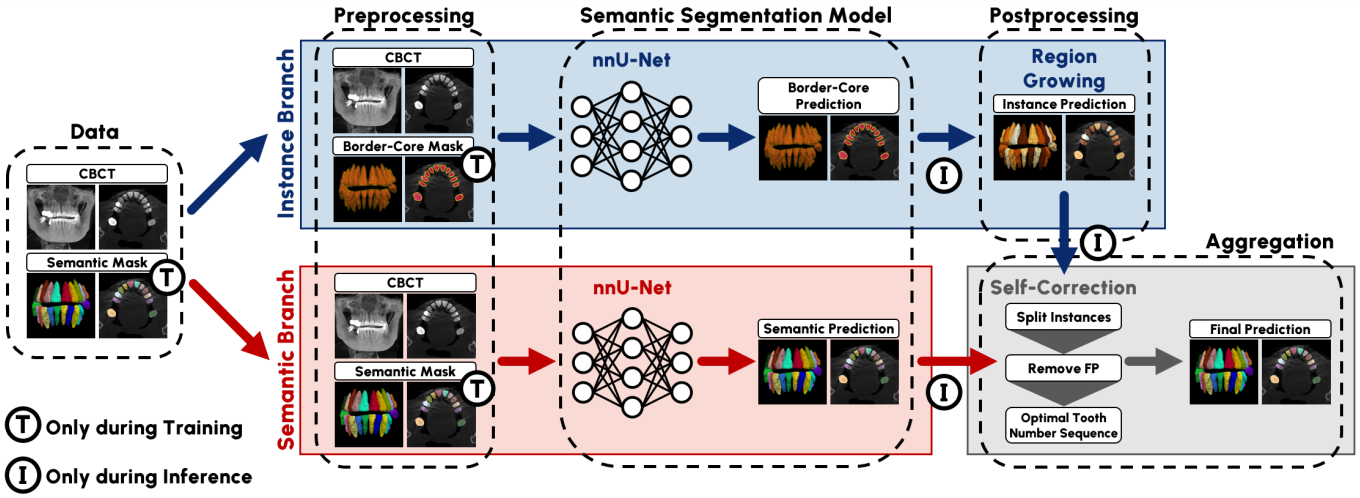

Fig. 2: ToothSeg consists of two independently trained branches for instance and semantic segmentation, combined through self-correction for the final prediction. The instance branch uses a border-core representation, reformulating instance segmentation as a semantic segmentation task. From a border-core prediction, individual instances are obtained by region growing of the core components. The semantic branch is trained using the multiclass semantic ground truth and predicts the tooth numbers of the detected instances. Self-correction splits merged instances and removes false positives. For the final prediction, the optimal sequence of tooth numbers is determined for each dental arch based on the tooth number predictions and the expected position differences between subsequent teeth in the sequence.

to each tooth instance. Our semantic segmentation branch is based on nnU-Net's default 3d_fullres configuration and we extensively modified the default settings to maximize tooth numbering performance.

Tooth numbering performance strongly depends on the context processed by the network. Ideally, the complete dental arch is processed at once, such that contralateral teeth can be compared to ensure consistent tooth numbers. Therefore, we use a large patch size of 256x256x256 voxels and a large voxel size of 0.3mm, to accommodate the larger input, we increase the network's receptive field by adding an additional pooling operation. We use a batch size of 8, compared to nnU-Net's default of 2, to improve gradient quality. During training, extensive data augmentation following nnU-Net's default settings is applied. Due to the need to distinguish left from right teeth in a mostly symmetrical jaw, we find that disabling left-right mirroring drastically improves performance.

B. Instance segmentation branch

We formulate tooth instance segmentation as a three-class semantic segmentation task: background, tooth border, and tooth core. From this representation, tooth cores can be identified as connected components of the core voxels, as the border voxels prevent the interaction of core voxels from adjacent teeth. Based on the identified tooth cores, a precise tooth instance segmentation can be determined by assigning border voxels to the closest tooth core. Ground truth tooth instance segmentation maps are converted to border-core semantic segmentation maps by eroding individual teeth and setting the eroded areas to the tooth border label. The remaining tooth voxels are assigned as tooth cores.

The design of the instance segmentation branch is geared towards precisely outlining the individual teeth while ignoring the tooth numbers, which requires a high resolution with less contextual information. However, a high resolution does considerably increase inference time with diminishing returns for effectiveness. For these reasons, 0.2mm is chosen as voxel size and a patch size of 192x192x192 is used. A border thickness of three voxels provides the optimal value to prevent split or merged tooth instances. Other configuration settings remain consistent with the semantic segmentation branch, except that left-right mirroring is included.

Tooth instances are recovered from the predicted border-core segmentation by first determining the connected components of the core voxels, where small core regions ($\leq$16 mm^3) are removed. Each remaining tooth instance is assigned a unique sequential label, which is then propagated to all corresponding voxels. Subsequently, border voxels are assigned to the core regions by iteratively dilating the cores by one voxel into the predicted border area until all border voxels are accounted for. This method is advantageous over directly computing distances from border voxels to each core, as it prevents skipping across background voxels.

C. Combining semantic and instance predictions

Unlike previous approaches that rely on multistage processes, leading to error accumulation, our self-correcting approach allows for simultaneous correction of semantic and instance errors, resulting in a more reliable system.

Given a CBCT scan of size $W \times H \times D$, the semantic branch determines a probability distribution over 32 tooth numbers and a background class for each voxel ($\mathcal{S} = \mathbb{R}^{W \times H \times D \times 33}$). Conversely, the instance branch determines an index map ($\mathcal{I} = \{0, 1, ..., K\}^{W \times H \times D}$), with the background as 0 and

each instance as an index from 1 to K.

The probabilities from the semantic prediction were aggregated for all M voxels belonging to tooth instance k as $\mathcal{S}_{\mathcal{I}=k} \in \mathbb{R}^{M\times 33}$:

$$\mathcal{S}_{\mathcal{I}=k} := \{\mathcal{S}_{x,y,z} \in \mathbb{R}^{33} | x \in \{1,...,W\}, \\ y \in \{1,...,H\}, z \in \{1,...,D\}, \mathcal{I}_{x,y,z} = k\}. \quad (1)$$

1) Splitting merged instances.: An overall probability distribution of an instance $\left(\overline{\mathcal{S}_{\mathcal{I}=k}} \in \mathbb{R}^{33}\right)$ was computed as the average over the M voxel probability distributions:

$$\overline{\mathcal{S}_{\mathcal{I}=k}} := \frac{1}{M}\sum_{m=1}^{M}\left(\mathcal{S}_{\mathcal{I}=k}\right)_m. \quad (2)$$

Additionally, labels were determined for the instance voxels ($\mathbf{l}^k \in \{0,1,...,32\}^M$) based on their probabilities:

$$\mathbf{l}^k := \left\{\arg\max_c \left(\mathcal{S}_{\mathcal{I}=k}\right)_{m,c} | m \in \{1,...,M\}, c \in \{0,...,32\}\right\}. \quad (3)$$

Subsequently, tooth numbers with an overall probability of at least 0.1 ($\left(\overline{\mathcal{S}_{\mathcal{I}=k}}\right)_c \geq 0.1$) were selected and their *confidences* were computed by determining the average probability over each tooth number's corresponding voxels:

$$p_c^k := \frac{1}{|\{l_m^k \in \mathbf{l}^k | l_m^k = c\}|}\sum_{m=1}^{M}\begin{cases}\left(\mathcal{S}_{\mathcal{I}=k}\right)_{m,c} & \text{if } l_m^k = c \text{ and}\\ 0 & \text{otherwise.}\end{cases} \quad (4)$$

Then, the final tooth numbers with $p_c^k \geq 0.95$ are selected as $\mathbf{c}^k$ and the M instance voxels are split among the final tooth numbers as:

$$\mathbf{l}_*^k := \left\{\arg\max_c \left(\mathcal{S}_{\mathcal{I}=k}\right)_{m,c} | m \in \{1,...,M\}, c \in \mathbf{c}^k\right\}. \quad (5)$$

With the split instance, the original index map ($\mathcal{I}$) is updated by setting the voxels of the first tooth number of $\mathbf{c}^k$ to index k and setting the voxels of the remaining tooth numbers of $\mathbf{c}^k$ to indices following the maximum index of the index map. This process is repeated for each original instance to split all possible merged instances.

2) Removing false-positive instances.: With the updated index map following the splitting of the merged instances, the overall probability distribution of each instance ($\overline{\mathcal{S}_{\mathcal{I}=k}}$) was re-computed. Instances with a background probability of at least 0.95 ($\left(\overline{\mathcal{S}_{\mathcal{I}=k}}\right)_0 \geq 0.95$) were removed.

3) Optimal sequence of tooth numbers.: Following the steps above, the resulting index map represents the prediction of all connected components comprising teeth. To effectively attribute a tooth number to each connected component, an algorithm was developed that merges the semantic and instance predictions and compares the order of tooth numbers to the expected order to mitigate potential mistakes. Previous methods for optimizing the sequence of tooth numbers relied on heuristic functions [50], whereas the current study optimizes the tooth number sequence by employing a probabilistic perspective.

The algorithm operates on one dental arch at a time, so the connected components of the index map were categorized as either upper or lower arch using:

$$\text{arch}^k := \begin{cases}\text{upper} & \text{if } \sum_{c=1}^{16}\left(\overline{\mathcal{S}_{\mathcal{I}=k}}\right)_c \geq \sum_{c=17}^{32}\left(\overline{\mathcal{S}_{\mathcal{I}=k}}\right)_c \text{ and}\\ \text{lower} & \text{otherwise.}\end{cases} \quad (6)$$

The numbers for the tooth instances in each arch were determined independently and the following paragraphs apply to teeth of the upper arch. First, the center position or centroid of each connected component was determined by reorienting and rescaling the predictions into the patient-centered coordinate system and computing the average over the coordinates of the voxels belonging to each instance.

Based on these centroids, a sequence of teeth was determined following the dental arch, starting with the most posterior tooth and iteratively adding the nearest tooth, until all teeth of an arch were included in the sequence.

The overall probability distribution of each instance ($\overline{\mathcal{S}_{\mathcal{I}=k}}$) was normalized by ensuring the probabilities of the arch-specific tooth numbers sum to 1. The cost of assigning a tooth number to the ith instance in the sequence was computed as the negative log of its normalized probability:

$$\mathcal{L}^i(c) := -\log\left(\frac{\left(\overline{\mathcal{S}_{\mathcal{I}=i}}\right)_c}{\sum_{c'=1}^{16}\left(\overline{\mathcal{S}_{\mathcal{I}=i}}\right)_{c'}}\right). \quad (7)$$

The sequence of tooth numbers with the minimum sum of tooth number costs assigns the numbers with the maximum probability, i.e. the *argmax* approach. This methodology is further improved by also considering tooth pair costs.

The ground-truth annotations were used to model the differences between centroids of instances. More specifically, the ground truth annotations of the training set were converted to border-core representation and back to an index map to allow for multiple connected components per tooth (e.g. root remnants). The centroid of each connected component was computed in the patient-centered coordinate system and categorized as belonging to the upper or lower dental arch according to the corresponding FDI number. Then, the differences (left-right, anterior-posterior, and inferior-superior) between the centroids of all pairs of connected components were determined for teeth within the same dental arch and differences were grouped by the pair of corresponding tooth numbers. Subsequently, the differences of each tooth number pair were modeled using a multi-variate Gaussian distribution with mean vector ($\boldsymbol{\mu}_{c_1\to c_2}$) and a full covariance matrix ($\boldsymbol{\Sigma}_{c_1\to c_2}$).

A tooth pair cost was computed as the negative log likelihood of transitioning from a tooth number to another tooth number based on the centroid differences between subsequent teeth in the tooth sequence. First, the centroid differences between each pair of subsequent teeth were determined ($\mathbf{d}_{i\to i+1}$). These differences were used to compute probability densities for each pair of corresponding tooth numbers, which were converted to tooth pair costs as:

$$\mathcal{L}^{i\to i+1}(c_1,c_2) := -\log\mathcal{N}\left(\mathbf{d}_{i\to i+1}|\boldsymbol{\mu}_{c_1\to c_2},\boldsymbol{\Sigma}_{c_1\to c_2}\right). \quad (8)$$

Based on the computed costs, an optimal sequence of tooth numbers was determined ($\mathbf{c}^*$) that minimized the total cost after summing all tooth number and tooth pair costs along the sequence from index 1 to K:

$$\mathbf{c}^* := \underset{\mathbf{c}\in\{1,\ldots,16\}^K}{\arg\min} \sum_{i=1}^{K} 4\,\mathcal{L}^i(c_i) + \sum_{i=1}^{K-1} \mathcal{L}^{i\to i+1}(c_i, c_{i+1}). \quad (9)$$

To find the optimal sequence in polynomial time, a dynamic programming algorithm was developed, see Listing 1. The algorithm requires the tooth number costs for each tooth and the tooth pair costs for each pair of tooth numbers and pairs of subsequent teeth in the sequence. The total cost after selecting the first tooth number is stored in `q` and `p` stores the predecessors to recover the sequence after the algorithm.

The algorithm then iteratively selects a tooth number for the next tooth in the sequence and determines the minimum total cost up to selecting that next tooth number (`min_cost`). After the algorithm exhausts all possible sequences, the final sequence is read from the predecessors from the last to the first tooth number. The `eta` parameter is introduced to balance the contribution of the tooth number and tooth pair costs. This parameter was set to 4 based on an empirical investigation, showing that the semantic predictions were already highly effective for tooth numbering.

The final output is a semantic segmentation map, where all voxels associated with a given tooth are labeled with its corresponding tooth number (FDI notation). This refined merging of the two branches enhances the precision of the segmentation and significantly strengthens the model's robustness by correcting mistakes by either branch.

Listing 1: Pseudocode of dynamic programming algorithm to determine minimum-cost sequence of tooth numbers.

```
def dynamic_programming(
  tooth_costs: NDArray[('K',16), float],
  pair_costs: NDArray[('K'-1,16,16), float],
  eta: float=4.0,
):
  #balance tooth and pair costs
  tooth_costs = eta * tooth_costs

  #memoization
  q = np.zeros_like(tooth_costs)
  q[0] = tooth_costs[0]

  #predecessors
  p = np.zeros_like(q, dtype=int)
  p[0] = np.arange(16)

  #loop over each tooth in the sequence
  for i in range(1, tooth_costs.shape[0]):
    for j in range(16):
      prev_costs = q[i - 1]
      trans_costs = pair_costs[i - 1, :, j]
      costs = prev_costs + trans_costs

      min_cost = costs.min()
      q[i, j] = min_cost + tooth_costs[i, j]
      p[i, j] = costs.argmin()

  #determine final tooth number sequence
  path = [q[-1].argmin()]
  for i in range(1, tooth_costs.shape[0]):
    prev_number = p[-i, path[0]]
    path = [prev_number] + path

  return path
```

IV. Evaluation & Results

To enable meaningful validation, we compiled a comprehensive CBCT dataset of 1,282 scans. The ground truth annotations were generated by experts following the FDI notation system [31]. The dataset includes scans from more than 25 different scanners and various operators with a wide range of fields of view (FOVs), enhancing the dataset's applicability to real-world scenarios (Table I). 49% of patients were male and the median age, inter-quartile range, and age range were 59 years, 18 years, and 16 to 93 years, respectively. The only excluded cases were those containing primary teeth, as their small sample size hindered robust learning and validation. To provide a public benchmark, we also included the ToothFairy2 dataset [28]–[30], which consists of 480 CBCT scans. For both datasets, segmentation focused on the 32 permanent tooth classes.

The models were evaluated by two categories of metrics: instance metrics and multiclass instance metrics. Instance metrics assess segmentation performance by comparing instances without considering tooth numbers. In contrast, multiclass instance metrics also evaluate how accurately FDI numbers are assigned. The three metrics employed for both categories are True Positive Dice (TP Dice), instance F1, and panoptic Dice. TP Dice computes the average Dice score of true-positive predicted instances to validate the voxel-level accuracy of the segmentations. Instance F1 computes the object-level F1 score to represent how reliable objects can be identified. The underlying instance matching between ground truth and prediction is based on greedily pairing with a Dice overlap higher than 0.1. Additionally, in the case of multiclass instance metrics, the FDI numbers must match. Finally, panoptic Dice multiplies TP Dice and instance F1 to evaluate both voxel-level and instance-level accuracy.

We compare our method to recent state-of-the-art approaches, which were included as ReluNet [41], CuiNet [45], WangNet [15], and LiuNet [48] using the following criteria: (1) published in 2021 or later, (2) proposed a fully automated method for tooth instance segmentation and FDI numbering, (3) not surpassed by another method of the same authors, and (4) the most citations compared to other studies published in the same year. All models are trained from scratch on both datasets, using the same dataset splits. Due to the lack of public source code, reference methods were replicated to the best of our knowledge and ability.

TABLE I: Imaging protocols in the 1,282 CBCT scans of the in-house dataset. Each range has an open endpoint.

	Cases	%
Voxel size (mm)		
0.12-0.16	206	16
0.16-0.2	282	22
0.2-0.25	280	22
0.25-0.3	170	13
0.3-0.4	292	23
0.4-	49	3.8
Field of view (cm^3)		
240-400	74	6.1
400-500	249	20
500-1000	326	27
1000-1500	155	13
1500-2500	228	19
2500-	184	15
Tube voltage (kV)		
70-80	5	0.43
80-90	311	27
90-100	527	45
100-110	34	2.9
110-120	23	2.0
120-	266	23
Tube current (mA)		
1-3	26	2.3
3-5	184	16
5-7	505	44
7-9	226	20
9-11	88	7.7
11-	88	7.7

A. Quantitative results

Table II shows the comparison between all five methods on our in-house dataset. For each method, a single model was trained on all train cases ($n = 903$) and validated on a separate test set ($n = 397$). ToothSeg demonstrates superior performance across all metrics, achieving the highest panoptic Dice of 93.35 and 89.32 for instance and multiclass instance metrics, respectively. A large difference of 3.31 and 4.20 to the next best instance and multiclass instance metrics, respectively, can be seen.

TABLE II: **Test set results on our in-house dataset.** Instance metrics ignore tooth numbers and measure the ability to segment teeth accurately. Multiclass instance metrics include tooth numbers to also represent how reliably the correct FDI numbers are assigned.

	Instance metrics			Multiclass instance metrics		
	TP Dice	inst. F1	panop. Dice	TP Dice	inst. F1	panop. Dice
ReluNet	93.26	96.46	89.96	93.30	91.19	85.10
CuiNet	87.42	98.09	85.75	87.98	92.02	80.90
WangNet	91.34	98.58	90.04	91.62	92.94	85.12
LiuNet	87.02	97.46	84.81	87.20	92.34	80.53
ToothSeg	**94.11**	**99.19**	**93.35**	**94.23**	**94.80**	**89.32**

The same findings were found for the ToothFairy2 challenge dataset (Table III). The dataset was randomly split into training and validation with a 70:30 ratio and no methodological modifications were made for all methods. Nevertheless, ToothSeg once again outperforms all other methods, highlighting its robustness and ability to generalize to new datasets without requiring manual adaptation.

TABLE III: **Results on the ToothFairy2 challenge dataset.**

	Instance metrics			Multiclass instance metrics		
	TP Dice	inst. F1	panop. Dice	TP Dice	inst. F1	panop. Dice
ReluNet	93.90	92.75	87.09	93.89	89.04	83.64
CuiNet	91.63	94.97	87.02	92.23	90.42	83.36
WangNet	96.08	94.13	90.45	96.15	90.51	87.04
LiuNet	90.42	93.35	84.41	90.64	90.09	81.69
ToothSeg	**96.49**	**95.71**	**92.35**	**96.60**	**93.50**	**90.32**

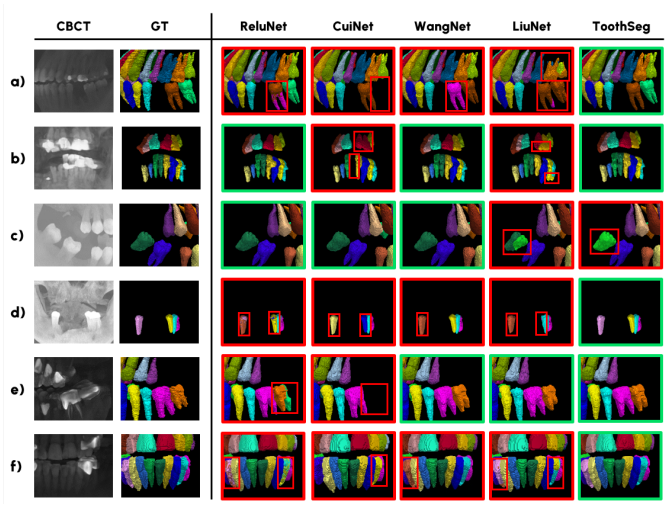

Fig. 3: Comparison of different methods on held-out test cases. The CBCT scan and ground truth (GT) are shown on the left, followed by predictions from ReluNet, CuiNet, WangNet, LiuNet, and ToothSeg (ours). Green frames indicate correct predictions, while red frames and red bounding boxes highlight errors. The six cases include: (a) a third molar, (b) a large FOV, (c) a misaligned tooth, (d) an edentulous maxilla, (e) an implant, and (f) an artifact.

B. Qualitative results

Fig. 3 presents a qualitative comparison between ToothSeg and all reference methods. Two common failure cases in tooth segmentation can be seen: splitters, where a single tooth is incorrectly predicted as two separate instances, and mergers, where two teeth are mistakenly connected into a single predicted instance. Fig. 3 shows that CuiNet struggles to detect all tooth instances, often resulting in completely missing teeth. ReluNet and LiuNet experience difficulties with both splitters and mergers, leading to errors in delineating individual teeth and separating adjacent teeth. WangNet performs better than ReluNet and CuiNet, however, it has problems assigning the correct tooth numbers, which affects the overall accuracy of the segmentation. Conversely, ToothSeg provides the most reliable results, accurately predicting the boundaries of adjacent teeth and assigning FDI numbers with the fewest errors.

C. Ablation study

The ablation study in Table IV underscores the importance of each component in ToothSeg. The impact of individual modifications to our method are compared to a plain nnU-Net baseline without left-right mirroring, which addresses the task as a naive semantic segmentation problem. Left–right

TABLE IV: **Ablation study.** The study begins with the semantic branch which corresponds to the default nnU-Net configuration without left-right mirroring, with components incrementally added from top to bottom. Only in the last step the instance branch is added with self-correction.

	Instance metrics			Multiclass instance metrics		
	TP Dice	inst. F1	panop. Dice	TP Dice	inst. F1	panop. Dice
nnU-Net (no mirror)	92.44	90.17	83.35	91.56	82.96	76.43
↳ patch size 128	92.48	91.49	84.61	91.24	84.53	77.63
↳ voxel size $0.2 \rightarrow 0.3$	93.05	94.79	88.20	92.88	89.41	83.12
↳ patch size 192	93.22	96.65	90.10	93.25	91.67	85.51
↳ patch size 256	93.50	97.47	91.13	93.42	93.10	86.98
↳ batch size $2 \rightarrow 8$	93.62	98.00	91.75	93.63	94.23	88.24
↳ merge instance branch	94.24	98.77	93.1	94.25	95.26	89.78
↳ self-correction	94.28	98.97	93.31	94.27	95.53	90.05

mirroring was disabled because bilateral symmetry in the dentition makes flipped counterparts indistinguishable, resulting in anatomically inconsistent labels and degrading class-specific learning. The experiments are performed on an 80:20 split of the training set from our in-house dataset, ensuring the test set remains held-out. Increasing the voxel size from 0.2mm to 0.3mm and increasing the patch size results in further improvements, suggesting that more contextual information is crucial for assigning the correct tooth numbers. A larger batch size contributes to more stable learning dynamics and potentially better generalization, which was observed in this study. Combining semantic and instance branches leads to a further substantial performance improvement. Finally, incorporating the self-correction mechanism yields the best overall results. Notably, self-correction is essentially cost-free, introducing only negligible runtime overhead. To examine the impact of the self-correction approach in more detail, Fig. 4 shows several cases that illustrate both its advantages and limitations.

For splitters, where a single tooth instance is predicted as multiple instances, the semantic and instance branches can compensate each other's mistakes. If a splitter occurs in the semantic prediction and the instance branch provides a correct prediction, self-correction will assign a single tooth number, effectively correcting the error (Fig. 4a). Conversely, if a splitter occurs in the instance prediction, it can be compensated by a correct prediction of the semantic branch, as all segments of the splitter will be assigned the same tooth number (Fig. 4b). Thus, a false prediction occurs only if the branches fail simultaneously (Fig. 4f).

For mergers, where multiple teeth are predicted as one tooth, combining the semantic and instance predictions can correct mistakes from either branch. An incorrect prediction in the semantic branch can be corrected if the tooth number probabilities of correctly detected instances assign the correct tooth numbers (Fig. 4c) or if the sequence of tooth numbers can be optimized based on the expected centroid differences between the two instances (Fig. 4e). Furthermore, a merged instance prediction can be split into two or more instances if the semantic predictions are *confident* the instance should be split (Fig. 4d).

TABLE V: Comparison of model performance in challenging cases found in the test set ($n = 379$), including those with third molars ($n = 125$), misaligned teeth ($n = 41$), implants or pontics ($n = 174$), metal artifacts ($n = 137$), and a large field of view (FOV) ($n = 11$). Cases not included in these subsamples were categorized as normal cases ($n = 40$).

Method	**Metric***	**Cases**					
		Normal	3rd molar	Misaligned	Implant/pontic	Artifact	Full FOV
ReluNet	*instance*	91.78	90.70	90.33	89.41	89.17	89.30
	multiclass	90.92	85.75	86.38	84.80	83.60	87.10
	numbers	**99.02**	94.40	95.56	94.79	93.71	97.24
CuiNet	*instance*	87.82	85.51	85.33	85.31	85.54	85.91
	multiclass	86.03	80.81	81.59	81.03	80.29	83.24
	numbers	97.81	94.03	95.19	94.42	93.35	95.93
WangNet	*instance*	91.61	90.22	90.13	89.92	89.44	90.33
	multiclass	89.78	85.08	86.08	85.25	84.36	87.63
	numbers	97.88	94.08	95.33	94.51	94.05	96.71
LiuNet	*instance*	87.07	84.60	84.35	84.32	83.87	85.92
	multiclass	86.30	80.58	80.39	80.83	79.00	84.68
	numbers	99.00	**94.92**	95.03	95.63	93.89	**98.13**
ToothSeg	*instance*	**95.08**	**93.50**	**93.43**	**93.04**	**92.85**	**91.78**
	multiclass	**94.02**	**88.84**	**89.55**	**89.28**	**88.78**	**90.03**
	numbers	98.83	94.84	**95.64**	**95.88**	**95.47**	97.99

*The results are evaluated using three metrics: instance panoptic Dice (*instance*), multiclass instance panoptic Dice (*multiclass*), and the multiclass instance F1 divided by the instance F1 (*numbers*).

D. Analysis of challenging cases

To evaluate the performance of the models in various scenarios, we conducted a subsample analysis on the in-house dataset. Challenging subsamples were defined by presence of third molars, implants, pontics, metal artifacts, misaligned teeth, or a field of view (FOV) of at least 16x16 cm. CBCT scans could meet multiple conditions and cases without any of these conditions were categorized as normal to assess model performance in ideal conditions.

Table V shows noticeably higher effectiveness for normal cases, with ToothSeg outperforming reference methods, especially in multiclass instance metrics. This trend is consistently observed across all challenging subsamples, highlighting the robustness and adaptability of ToothSeg. Furthermore, the qualitative results in Fig. 3 present at least one case from each subsample, demonstrating ToothSeg's effectiveness across different dental conditions.

E. Comparison against commercial systems

To compare state-of-the-art research methods with commercial systems for automated CBCT analysis, seven CBCT scans without primary teeth were selected from cases requested from Ileșan et al. [51]. Teeth were manually segmented and labeled using the FDI numbering system. The scans were processed using ToothSeg and the four baseline models trained on the in-house dataset, as well as two commercial systems: Relu® Creator (Relu, Leuven, Belgium) and Diagnocat AI (Diagnocat, Tel Aviv, Israel). The commercial systems were evaluated without prior notice to prevent any special attention to our cases.

As shown in Table VI, ToothSeg outperformed all other methods. Following the self-correction steps, ToothSeg was the only method with a perfect multiclass instance F1, whereas the other methods missed teeth, had false-positive teeth, or predicted incorrect FDI numbers. Among the commercial systems, Relu® Creator was the most effective, improving on the

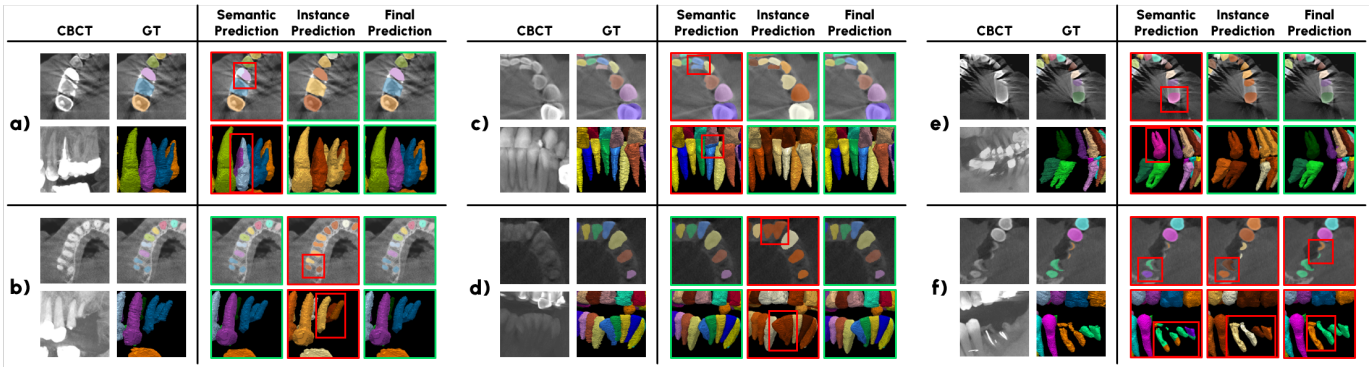

Fig. 4: Illustration of the strengths and limitations of combining predictions from both branches. (a) A splitted semantic prediction is corrected by the instance branch, while (b) shows the reverse. (c) A merged semantic prediction can be resolved by correct instance predictions. (d) For merged instance predictions, a correct semantic prediction alone is insufficient, but instance splitting during self-correction can resolve the error. (e) Self-correction can also fix incorrect tooth numbering by optimizing the sequence of tooth numbers. (f) When both branches fail, neither merged nor splitted errors can be corrected.

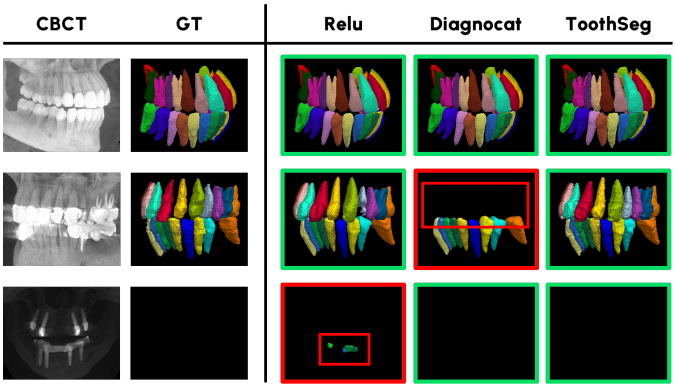

Fig. 5: Qualitative comparison of different methods on test cases against two commercial systems for automated CBCT analysis. The CBCT image and ground truth (GT) segmentation are shown on the left, followed by the predictions of Relu® Creator, Diagnocat AI, and ToothSeg (ours). Green frames indicate correct predictions, while red frames and red bounding boxes highlight errors.

TABLE VI: **Results of seven cases by ToothSeg, four baselines, and two commercial systems.** ToothSeg outperforms all other methods in terms of instance panoptic Dice and multiclass instance panoptic Dice.

	instance panoptic Dice	multiclass panoptic Dice
ReluNet	85.76	86.13
CuiNet	85.26	82.94
WangNet	86.64	86.51
LiuNet	85.61	84.91
Relu	86.76	87.42
Diagnocat	84.50	84.29
ToothSeg	**87.58**	**87.99**

ReluNet baseline, which suggests further development since the publication by Shaheen et al. [41]. Overall performance on these scans was lower than on the in-house dataset, most likely due to severe metal artifacts in four out of seven cases.

Figure 5 provides a qualitative comparison of our method against the commercial systems. While all three methods performed well on normal cases, critical failure cases were identified: Relu® Creator falsely identified teeth in an edentulous jaw, and Diagnocat AI missed all upper teeth in another scan. Although these isolated failure cases provide limited statistical significance, they serve as further evidence of the robustness and reliability of our method.

V. Discussion & Conclusion

This study addresses key challenges in tooth instance segmentation and numbering in cone beam computed tomography (CBCT) by introducing ToothSeg, a fully automated deep learning method that incorporates a dedicated self-correction mechanism. ToothSeg employs a dual-branch design that combines semantic and instance segmentation to generate accurate and anatomically consistent results, even in the presence of imaging artifacts, anatomical variability, or missing teeth. The proposed self-correction strategy resolves common failure modes, such as merged or split teeth, by reconciling discrepancies between the two segmentation branches. This enables robust segmentation and reliable tooth numbering across both routine and challenging clinical cases.

ToothSeg holds significant potential for transforming clinical workflows in dental and oral radiology. Accurate and consistent tooth segmentation and numbering are critical tasks in routine dental care, underpinning diagnosis, treatment planning, and documentation. However, these tasks remain time-consuming and examiner-dependent, contributing to variability across practitioners. By automating this process with high accuracy, ToothSeg reduces clinician workload and promotes standardized outcomes. The metrics used in this study evaluate tooth instance segmentation (TP Dice) and tooth detection and numbering (instance F1). The results reported improvements in these metrics, which directly reflects fewer misclassifications, such as merged or split teeth or incorrect FDI numbers. In practical terms, this leads to measurable time savings during clinical work, as clinicians spend less time verifying and editing segmentation outputs. The resulting increase in automation efficiency enhances the usability of CBCT systems, particularly in general dental practices with limited

access to oral radiology specialists. Moreover, combining the clinician's experience with the model's ability to consistently evaluate all tooth surfaces may further enhance diagnostic quality compared to manual clinical workflows. The method demonstrates strong performance across a wide range of imaging conditions, including noise, metal-induced artifacts, missing teeth, and anatomical variations. This robustness ensures reliable results, particularly in complex cases, and supports consistent application across clinical sites, users, and patient populations. By enabling accurate tooth segmentation and numbering at minimal cost, requiring no specialized hardware, and publishing its source code, ToothSeg further supports standardized documentation, longitudinal monitoring, and clinical communication. Overall, ToothSeg contributes to more efficient, reliable, and accessible dental imaging, laying the groundwork for scalable, data-driven approaches in oral and craniofacial healthcare.

Compared with prior hybrid segmentation-numbering approaches for CBCT, ToothSeg differs both from a technical and clinical perspective. Previous methods often adopted multi-stage pipelines involving separate networks for detection and segmentation, which introduced error propagation and required dataset-specific tuning. In contrast, ToothSeg integrates these tasks into a dual-branch approach with self-correction steps, enabling greater robustness across datasets and imaging conditions. Clinically, earlier systems often required manual preprocessing steps, such as cropping or isolating tooth regions before inference, whereas ToothSeg can process complete CBCT scans fully automatically. This design not only streamlines the clinical workflow but also facilitates interactive revisions when needed. While a similar concept has been introduced independently for intraoral scans [50], ToothSeg represents the first application to volumetric data, validated across multiple datasets and real-world imaging conditions.

Future research could extend the introduced method to the segmentation of other anatomical structures in 3D medical scans, where the combination of semantic and instance segmentation could provide similar benefits. For example, individual vertebra segmentation and numbering could be a promising application for our proposed approach [52]. Moreover, an investigation into the time saved on manual revisions could be conducted to compare the efficient use of commercially available systems to ToothSeg.

Although ToothSeg demonstrates strong overall performance, several limitations remain. The method relies on both the semantic and instance segmentation branches, which can compensate for each other's errors to a certain extent. However, when their predictions differ substantially, especially in cases with low image quality or unusual anatomy, segmentation errors such as missing or merged teeth may still occur. Tooth numbering can also be uncertain in situations involving extracted teeth, inclined molars, or root fragments, where anatomical context is limited. The training dataset is large and diverse, but certain clinically important groups, such as children, patients with dental prostheses, or those with advanced dental conditions, are not sufficiently represented. Although there are indications that the method may generalize to these cases, a reliable conclusion cannot be made without further evaluation. Expanding the dataset to include more of these cases, along with targeted data augmentation and comprehensive validation, could improve generalizability and extend the scope of application. Currently, scans with primary teeth were not included, as too few were collected to be able to develop an effective model. Thus, ToothSeg is not applicable for pediatric patients. As this limitation is due to insufficient data, future work can focus on incorporating adequate samples from underrepresented groups to enable reliable extension of the method or the training of specialized models.

Many methods for tooth instance segmentation and numbering in CBCT have been published in the literature. However, due to a lack of source code, no validation on public datasets, and the diverse technical implementations, only four studies from 2021 to 2024 were replicated and included in the comparison to ToothSeg. The comparison of ToothSeg to two commercial systems was based on only seven CBCT scans, hindering the generalizability of the findings. Furthermore, current commercial systems provide segmentations of additional structures, such as the lower jaw and inferior alveolar canals, whereas ToothSeg only predicts teeth. Nevertheless, the approaches introduced by ToothSeg can be easily integrated into these systems for more robust tooth segmentation. Subsequently, a more comprehensive investigation to compare the ToothSeg approaches to current commercial systems must be undertaken to discover the benefits and limitations before clinical integration. In addition, a prospective clinical validation is needed to ensure responsible use of the system, as well as efficient integration into existing clinical workflows. Lastly, several potential barriers for real-world deployment remain, such as the cost and time for model inference, the integration with existing systems, training of dental practitioners, and the potentially lower accuracy for scans from underrepresented patient populations.

In summary, ToothSeg represents a significant step forward in fully automated tooth instance segmentation and numbering for CBCT scans. Its self-correcting approach, applicability across datasets, and its ability to handle varying imaging conditions and challenging cases make it a robust and reliable tool for clinical use. While there are areas for refinement, ToothSeg offers a promising foundation for future developments in both dental and broader medical imaging applications. By releasing our source code and providing validation on a public dataset, we hope to contribute to the advancement of open science in dental imaging.

Data availability statement

The in-house dataset of 1,282 CBCT scans cannot be made public due to privacy concerns, but can be reasonably requested from the corresponding author. The ToothFairy2 challenge dataset can be found at `https://ditto.ing.unimore.it/toothfairy2/`. For this study, only the tooth numbers were included as 32 foreground class labels.

The source code for this article is published at `https://github.com/MIC-DKFZ/ToothSeg`, including ToothSeg, the reference methods, and all checkpoints trained on the ToothFairy2 challenge dataset.

REFERENCES

[1] R. Najjar, "Redefining radiology: A review of artificial intelligence integration in medical imaging," *Diagnostics*, vol. 13, no. 17, 2023.

[2] L. Pinto-Coelho, "How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications," *Bioengineering*, vol. 10, no. 12, 2023.

[3] V. Allareddy, S. Rengasamy Venugopalan, R. P. Nalliah, J. L. Caplin, M. K. Lee, and V. Allareddy, "Orthodontics in the era of big data analytics," *Orthodontics & Craniofacial Research*, vol. 22, no. S1, pp. 8–13, 05 2019.

[4] M. M. Meghil, P. Rajpurohit, M. E. Awad, J. McKee, L. A. Shahoumi, and M. Ghaly, "Artificial intelligence in dentistry," *Dentistry Review*, vol. 2, no. 1, p. 100009, 03 2022.

[5] F. Schwendicke and J. Krois, "Data dentistry: How data are changing clinical care and research," *Journal of Dental Research*, vol. 101, no. 1, pp. 21–29, 01 2022.

[6] Y.-W. Chen, K. Stanley, and W. Att, "Artificial intelligence in dentistry: Current applications and future perspectives," *Quintessence international*, vol. 51, no. 3, pp. 248–257, 02 2020.

[7] J. Tomášik, M. Zsoldos, Ľ. Oravcová, M. Lifková, G. Pavleová, M. Strunga, and A. Thurzo, "Ai and face-driven orthodontics: A scoping review of digital advances in diagnosis and treatment planning," *AI*, vol. 5, no. 1, pp. 158–176, 01 2024.

[8] M. Loubele, R. Bogaerts, E. Van Dijck, R. Pauwels, S. Vanheusden, P. Suetens, G. Marchal, G. Sanderink, and R. Jacobs, "Comparison between effective radiation dose of cbct and msct scanners for dentomaxillofacial applications," *European Journal of Radiology*, vol. 71, no. 3, pp. 461–468, 09 2009.

[9] R. Schulze and N. Drage, "Cone-beam computed tomography and its applications in dental and maxillofacial radiology," *Clinical Radiology*, vol. 75, no. 9, pp. 647–657, 2020.

[10] S. Friedlander-Barenboim, W. Hamed, A. Zini, N. Yarom, I. Abramovitz, H. Chweidan, T. Finkelstein, and G. Almoznino, "Patterns of cone-beam computed tomography (cbct) utilization by various dental specialties: A 4-year retrospective analysis from a dental and maxillofacial specialty center," *Healthcare*, vol. 9, no. 8, p. 1042, 08 2021.

[11] H. Gaêta-Araujo, A. F. Leite, K. d. F. Vasconcelos, and R. Jacobs, "Two decades of research on cbct imaging in dmfr - an appraisal of scientific evidence," *Dentomaxillofacial Radiology*, vol. 50, no. 4, p. 20200367, 02 2021.

[12] O. S. Alothmani, L. T. Friedlander, B. D. Monteith, and N. P. Chandler, "Influence of clinical experience on the radiographic determination of endodontic working length," *International Endodontic Journal*, vol. 46, no. 3, pp. 211–216, 2013.

[13] J. Brown, R. Jacobs, E. Levring Jäghagen, C. Lindh, G. Baksi, D. Schulze, and R. Schulze, "Basic training requirements for the use of dental cbct by dentists: a position paper prepared by the european academy of dentomaxillofacial radiology," *Dentomaxillofacial Radiology*, vol. 43, no. 1, p. 20130291, 11 2013.

[14] A. Swaity, B. Elgarba, N. Morgan, S. Ali, S. Shujaat, E. Borsci, I. Chilvarquer, and R. Jacobs, "Deep learning driven segmentation of maxillary impacted canine on cone beam computed tomography images," *Scientific Reports*, vol. 14, no. 1, p. 369, 01 2024.

[15] Y. Wang, W. Xia, Z. Yan, L. Zhao, X. Bian, C. Liu, Z. Qi, S. Zhang, and Z. Tang, "Root canal treatment planning by automatic tooth and root canal segmentation in dental cbct with deep multi-task feature learning," *Medical Image Analysis*, vol. 85, p. 102750, 04 2023.

[16] D. Forst, S. Nijjar, C. Flores-Mir, J. Carey, M. Secanell, and M. Lagravère, "Comparison of in vivo 3d cone-beam computed tomography tooth volume measurement protocols," *Progress in orthodontics*, vol. 15, no. 1, p. 69, 12 2014.

[17] S. Sabancı, E. Şener, R. I. Turhal, B. O. Gürses, F. Gövsa, U. Tekin, A. Baltacı, H. Boyacıoğlu, and P. G uneri, "Is manual segmentation the real gold standard for tooth segmentation? a preliminary in vivo study using conebeam computed tomography images," *Meandros Medical and Dental Journal*, vol. 22, no. 1, pp. 263–273, 05 2021.

[18] J. Wallner, I. Mischak, and J. Egger, "Computed tomography data collection of the complete human mandible and valid clinical ground truth models," *Scientific Data*, vol. 6, no. 1, p. 190003, 01 2019.

[19] J. Wang, Z. Huang, X. Yang, W. Jia, and T. Zhou, "Three-dimensional reconstruction of jaw and dentition cbct images based on improved marching cubes algorithm," *Procedia CIRP*, vol. 89, no. 1, pp. 239–244, 01 2020.

[20] L. Yu, T. J. Vrieze, M. R. Bruesewitz, J. M. Kofler, D. R. DeLone, J. F. Pallanch, E. P. Lindell, and C. H. McCollough, "Dose and image quality evaluation of a dedicated cone-beam ct system for high-contrast neurologic applications," *American Journal of Roentgenology*, vol. 194, no. 2, pp. W193–W201, 02 2010.

[21] E. Venkatesh and S. V. Elluru, "Cone beam computed tomography: basics and applications in dentistry," *Journal of Istanbul University Faculty of Dentistry*, vol. 51, no. 3, pp. S102–S121, 11 2017.

[22] C.-M. Mörch, S. S. Atsu, W. Cai, X. Li, S. A. Madathil, X. Liu, V. Mai, F. Tamimi, M.-A. Dilhac, and M. Ducret, "Artificial intelligence and ethics in dentistry: A scoping review," *Journal of Dental Research*, vol. 100, no. 13, pp. 1452–1460, 06 2021.

[23] S. Uribe, A. Sofi-Mahmudi, E. Raittio, I. Maldupa, and B. Vilne, "Dental research data availability and quality according to the fair principles," *Journal of Dental Research*, vol. 101, no. 11, pp. 1307–1313, 06 2022.

[24] A. Holtkamp, K. Elhennawy, J. E. Cejudo, J. Krois, S. Paris, and F. Schwendicke, "Clinical medicine generalizability of deep learning models for caries detection in near-infrared light transillumination images," *Journal of Clinical Medicine*, vol. 10, no. 5, p. 961, 03 2021.

[25] J. Krois, A. Cantu, A. Chaurasia, R. Patil, P. Chaudhari, R. Gaudin, S. Gehrung, and F. Schwendicke, "Generalizability of deep learning models for dental image analysis," *Scientific Reports*, vol. 11, no. 1, p. 6102, 03 2021.

[26] R. C. Fontenele, M. d. N. Gerhardt, J. C. Pinto, A. Van Gerven, H. Willems, R. Jacobs, and D. Q. Freitas, "Influence of dental fillings and tooth type on the performance of a novel artificial intelligence-driven tool for automatic tooth segmentation on cbct images - a validation study," *Journal of Dentistry*, vol. 119, no. 1, p. 104069, 04 2022.

[27] K. Orhan, E. Bilgir, I. S. Bayrakdar, M. Ezhov, M. Gusarev, and E. Shumilov, "Evaluation of artificial intelligence for detecting impacted third molars on cone-beam computed tomography scans," *Journal of Stomatology, Oral and Maxillofacial Surgery*, vol. 122, no. 4, pp. 333–337, 09 2021.

[28] L. Lumetti, V. Pipoli, F. Bolelli, E. Ficarra, and C. Grana, "Enhancing patch-based learning for the segmentation of the mandibular canal," *IEEE Access*, 2024.

[29] M. Cipriano, S. Allegretti, F. Bolelli, M. Di Bartolomeo, F. Pollastri, A. Pellacani, P. Minafra, A. Anesi, and C. Grana, "Deep segmentation of the mandibular canal: a new 3d annotated dataset of cbct volumes," *IEEE Access*, vol. 10, pp. 11 500–11 510, 2022.

[30] M. Cipriano, S. Allegretti, F. Bolelli, F. Pollastri, and C. 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*, 2022, pp. 21 137–21 146.

[31] ISO, "Dentistry - designation system for teeth and areas of the oral cavity," https://www.iso.org/standard/68292.html, 2016, accessed: 2024-10-29.

[32] X. Chen, N. Ma, T. Xu, and C. 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*, vol. 238, no. 2, pp. 115–131, 02 2024.

[33] W. Cui, Y. Wang, Y. Li, D. Song, X. Zuo, J. Wang, Y. Zhang, H. Zhou, B. Chong, L. Zeng, and Q. Zhang, "Ctooth+: A large-scale dental cone beam computed tomography dataset and benchmark for tooth volume segmentation," in *Data Augmentation, Labelling, and Imperfections*, 09 2022, pp. 64–73.

[34] F. Hu, Z. Chen, and F. Wu, "A novel difficult-to-segment samples focusing network for oral cbct image segmentation," *Scientific Reports*, vol. 14, p. 5068, 03 2024.

[35] K. Hsu, D.-Y. Yuh, S. Lin, P.-S. Lyu, G.-X. Pan, Y.-C. Zhuang, C.-C. Chang, H.-H. Peng, T.-Y. Lee, C.-H. Juan, C.-E. Juan, Y.-J. Liu, and C. J. 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*, vol. 12, no. 1, p. 19809, 11 2022.

[36] S. Lee, S. Woo, J. Yu, J. Seo, J. Lee, and C. Lee, "Automated cnn-based tooth segmentation in cone-beam ct for dental implant planning," *IEEE Access*, vol. 8, pp. 50 507–50 518, 02 2020.

[37] Y. Rao, Y. Wang, F. Meng, J. Pu, J. Sun, and Q. Wang, "A symmetric fully convolutional residual network with dcrf for accurate tooth segmentation," *IEEE Access*, vol. 8, pp. 92 028–92 038, 05 2020.

[38] H. Wang, J. Minnema, K. Batenburg, T. Forouzanfar, F. Hu, and G. Wu, "Multiclass cbct image segmentation for orthodontics with deep learning," *Journal of Dental Research*, vol. 100, no. 9, pp. 943–949, 03 2021.

[39] Z. Cui, C. Li, and W. Wang, "Toothnet: Automatic tooth instance segmentation and identification from cone beam ct images," in *Computer Vision and Pattern Recognition*, 06 2019, pp. 6368–6377.

[40] W. Duan, Y. Chen, Q. Zhang, X. Lin, and X. Yang, "Refined tooth and pulp segmentation using u-net in cbct image," *Dentomaxillofacial Radiology*, vol. 50, no. 6, p. 20200251, 09 2021.

[41] E. Shaheen, A. Leite, K. A. Alqahtani, A. Smolders, A. Van Gerven, H. Willems, and R. Jacobs, "A novel deep learning system for multi-class tooth segmentation and classification on cone beam computed tomography. a validation study," *Journal of Dentistry*, vol. 115, p. 103865, 12 2021.

[42] X. Wu, H. Chen, Y. Huang, H. Guo, T. Qiu, and L. Wang, "Center-sensitive and boundary-aware tooth instance segmentation and classification from cone-beam ct," in *International Symposium on Biomedical Imaging*, 04 2020, pp. 939–942.

[43] Z. Cui, B. Zhang, C. Lian, C. Li, L. Yang, W. Wang, and M. Zhu, "Hierarchical morphology-guided tooth instance segmentation from cbct images," in *Information Processing in Medical Imaging*, 06 2021, pp. 150–162.

[44] W. Dou, S. Gao, D. Mao, H. Dai, C. Zhang, and Y. Zhou, "Tooth instance segmentation based on capturing dependencies and receptive field adjustment in cone beam computed tomography," *Computer Animation and Virtual Worlds*, vol. 33, no. 5, p. e2100, 08 2022.

[45] Z. Cui, Y. Fang, L. Mei, B. Zhang, B. Yu, J. Liu, C. Jiang, Y. Sun, L. Ma, H. Jiawei, Y. Liu, Y. Zhao, C. Lian, Z. Ding, and M. Zhu, "A fully automatic ai system for tooth and alveolar bone segmentation from cone-beam ct images," *Nature Communications*, vol. 13, no. 1, p. 2096, 04 2022.

[46] P. Lahoud, M. EzEldeen, T. Beznik, H. Willems, A. Leite, A. Van Gerven, and R. Jacobs, "Artificial intelligence for fast and accurate 3-dimensional tooth segmentation on cone-beam computed tomography," *Journal of Endodontics*, vol. 47, no. 5, pp. 827–835, 2021.

[47] J. Lee, M. Chung, M. Lee, and Y.-G. Shin, "Tooth instance segmentation from cone-beam ct images through point-based detection and gaussian disentanglement," *Multimedia Tools and Applications*, vol. 81, pp. 1–16, 05 2022.

[48] Y. Liu, R. Xie, L. Wang, H. Liu, C. Liu, Y. Zhao, S. Bai, and W. Liu, "Fully automatic ai segmentation of oral surgery-related tissues based on cone beam computed tomography images," *International Journal of Oral Science*, vol. 16, no. 34, 05 2024.

[49] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 02 2021.

[50] S. Zhuang, G. Wei, Z. Cui, and Y. Zhou, "Robust hybrid learning for automatic teeth segmentation and labeling on 3d dental models," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.

[51] R. R. Ileşan, M. Beyer, C. Kunz, and F. M. Thieringer, "Comparison of artificial intelligence-based applications for mandible segmentation: From established platforms to in-house-developed software," *Bioengineering*, vol. 10, no. 5, 2023.

[52] J. van der Graaf, M. van Hooff, C. Buckens, M. Rutten, J. van Susante, R. Kroeze, M. De Kleuver, B. Ginneken, and N. Lessmann, "Lumbar spine segmentation in mr images: a dataset and a public benchmark," *Scientific Data*, vol. 11, 03 2024.