IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 August 2023, accepted 16 September 2023, date of publication 20 September 2023,
date of current version 25 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317512

RESEARCH ARTICLE

Deep-Learning-Based Segmentation of Individual Tooth and Bone With Periodontal Ligament Interface Details for Simulation Purposes

PEIDI XU[1], TORKAN GHOLAMALIZADEH[2], FAEZEH MOSHFEGHIFAR[1], SUNE DARKNER[1], AND KENNY ERLEBEN[1]

[1]Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark
[2]3Shape A/S, 1059 Copenhagen, Denmark

Corresponding author: Peidi Xu (peidi@di.ku.dk)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Center for Regional Development, The Scientific Ethics Committee, under Reference No. 21063693.

ABSTRACT The process of constructing precise geometry of human jaws from cone beam computed tomography (CBCT) scans is crucial for building finite element models and treatment planning. Despite the success of deep learning techniques, they struggle to accurately identify delicate features such as thin structures and gaps between the tooth-bone interfaces where periodontal ligament resides, especially when trained on limited data. Therefore, segmented geometries obtained through automated methods still require extensive manual adjustment to achieve a smooth and organic 3D geometry that is suitable for simulations. In this work, we require the model to provide anatomically correct segmentation of teeth and bones which preserves the space for the periodontal ligament layers. To accomplish the task with few accurate labels, we pre-train a modified MultiPlanar UNet as the backbone model using inferior segmentations, i.e., tooth-bone segmentation with no space in the tooth-bone interfaces, and fine-tune the model with a dedicated loss function over accurate delineations that considers the space. We demonstrate that our approach can produce proper tooth-bone segmentations with gap interfaces that are fit for simulations when applied to human jaw CBCT scans. Furthermore, we propose a marker-based watershed segmentation applied on the MultiPlanar UNet probability map to separate individual tooth. This has advantages when the segmentation task is challenged by common artifacts caused by restorative materials or similar intensities in the teeth-teeth interfaces in occurrence of crowded teeth phenomenon. Code and segmentation results are available at https://github.com/diku-dk/AutoJawSegment.

INDEX TERMS Cone-beam computed tomography, deep learning, finite element modeling, human jaws, instance segmentation, learning with limited data, semantic segmentation, transfer learning.

I. INTRODUCTION

Accurate segmentation of medical images of a human jaw, such as cone beam computed tomography (CBCT) scans, is crucial in creating patient-specific preoperative and predictive finite element (FE) models that improve the design of implants and treatments [30]. A key aspect in the development of FE models is having a precise geometric representation of anatomical structures [21].

In the case of developing FE models of the human jaw, in addition to teeth and bone geometries, it is essential to model the connective tissue between them, called periodontal ligament (PDL). In general, PDL layer plays an important role in transferring load from teeth to the bone in orthodontic treatments, and when triggered with enough orthodontic forces, it results in bone remodeling [12]. As a result, accurate segmentations of human jaws must not only depict the shape

The associate editor coordinating the review of this manuscript and approving it for publication was Yi Zhang.

and boundaries of the involved teeth and bone structures but also preserve the space between them (see Fig. 1B and D) for the further modeling of PDLs [12].

Manual segmentation of the CBCT scans with the accurate geometrical representation of a human jaw's anatomies is labor-intensive and extremely time-consuming and depends on the scans' resolution and the annotator's expertise. In addition, it is especially challenging to accurately delineate the teeth and bone boundaries with relatively similar intensities to preserve the PDL from the CBCT scans. Hence, there is a need for automated segmentation tools that can generate accurate geometries for developing FE models.

Automatic segmentation methods commonly utilize Convolutional Neural Networks with an encoder-decoder architecture, of which the most effective is the UNet structure [22] which incorporates skip connections on high-resolution feature maps in the encoding stage to include more fine-grained information. Despite the development of newer models for natural image segmentation such as DeepLabV3+ [3] and transformer-based models [1], [25], [28], UNet remains one of the top performers in 3D medical image segmentation [14]. As a result, using a variation of UNet, such as 3D UNet, for segmenting 3D medical data like computed tomography scans is a straightforward approach that has demonstrated state-of-the-art performance [6], [14]. However, applying 3D convolutions directly to large 3D images may result in memory overflow. To mitigate this, 3D models are typically trained on small patches, which limits their field of view and causes the loss of global information. An alternative with lower memory usage is the MultiPlanar UNet (MPUNet) model proposed by Perslev et al. [20]. This model utilizes a 2D UNet to learn representative 3D semantic information by sampling slices from various orientations.

Most research in the field of auto-segmentation of human jaws aims on the neural networks designs that can accurately separate certain anatomical structures with minimal manual input [4], [5], [7], [9], [10], [15], [26], [27], [29], [31], [33]. Most of these works only focus on teeth segmentation [4], [5], [8], [9], [10], [11], [15], [29], [31], especially on the separation of individual tooth, while others only focus on the bone segmentation [26], [32], [33]. Although individual tooth segmentation is critical for computer-aided analysis towards clinical decision support and treatment planning, PDL layers cannot be retrieved from either tooth or bone segmentation alone, thus cannot be used to model the transferring load from teeth to the bone in orthodontic treatments [12]. Recently, Wang et al. [27] and Cui et al. [7] work on the multiclass segmentation of human jaws to simultaneously segment the bones (i.e., mandible and maxilla) and the teeth. Their models are either trained only over axial slices of CBCT scans [27] or trained on thousands of scans to reach a Dice score above 90% [7]. More crucially, their segmentations ignore the inter-bone gaps and thus are anatomically inaccurate. These anatomically inaccurate segmentations cannot be used to generate 3D models and limits the application in finite element simulations [12].

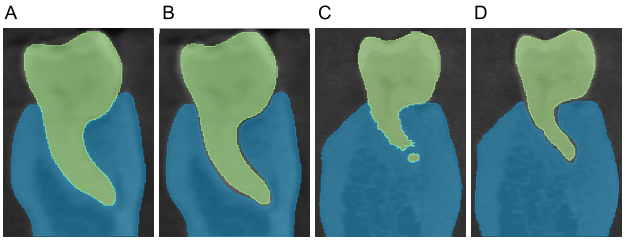

FIGURE 1. Illustration of gap generation. A: Inferior ground truth labels ignoring the space where the periodontal ligament resides. B: The accurate labels of the same patient that considers space for the periodontal ligament. C: Results of the proposed method on a test scan with model trained only on inferior dataset with no gap. D: Fine-tuned model with gap information.

Analog to Xu et al. [30], which accurately delineates the gap in hip joint segmentation for further cartilage simulation studies, we require the deep learning models to provide anatomically correct segmentation of human jaws which preserves the space for the PDL layers between teeth and mandibles as shown in Fig 1. Our approach leverages a standard UNet with batch normalization as the backbone model and incorporates the concept of MultiPlanar [20] to integrate more volumetric features into the model and increase the model's efficiency. In addition, due to the difficulty of manual delineation of the PDL layers, we have very few anatomically accurate teeth-bone label maps of the CBCT scans that detail the gap where PDL resides. Our framework utilizes an interactive learning process to reinforce such gap with limited annotated data by pre-training the MPUNet on a dataset with subpar segmentation to gain a general understanding of the tooth and bone structures. Subsequently, the model is fine-tuned using just a few highly accurate segmentations with a specific loss function that penalizes more on the gap regions.

By combining these techniques, our proposed pipeline is capable of achieving accurate segmentation results that fill in the missing gaps between the tooth-bone interfaces where the PDL is located with few accurate training data, while being both memory and computationally efficient. Our findings are verified using a test set of CBCT scans, where we construct finite element models and numerically evaluate the segmentation performance with the manually corrected segments utilized in biomechanical models.

In addition to the gap generation process, a further task is to separate individual tooth. The UNet output consists of the segmentation of a single class of jaw bones, as well as a single class of teeth that contains all the teeth. Automatic segmentation of individual tooth is critical for computer-aided analysis towards clinical decision support and treatment planning, but this segmentation is further challenged by blurring the boundaries of neighboring teeth and metal artifacts. Therefore, a simple post-processing with Connected Component Decomposition (CCD) over the UNet output will not correctly separate the adjacent teeth, especially if the subject has crowded teeth or is in a biting position.

Deep-learning-based instance segmentation methods, e.g., Mask R-CNN, have shown state-of-the-art performance on 2D natural images [13]. These networks involve region proposals to generate bounding boxes around each instance, with one branch for box regression and object detection and another for semantic segmentation. Cui et al. exploited 3D Mask R-CNN as a base network to realize automatic tooth segmentation and identification from CBCT images [9]. However, region proposals in 3D are extremely time and memory-consuming and require a larger training set than semantic segmentation methods that only deal with voxel labeling. Many of the modifications by Cui et al. that make region proposal work on 3D cases rely on the teeth having similar structure and orientation, thus will fail with, for instance, wisdom teeth and, more fatally, after adding jaw bone classes. Moreover, a threshold of the confidence level on each proposed region needs to be selected manually during inference, which may completely miss an object or generate overlapping instances and hinder biomechanical modeling afterward.

Instead, since there is no occlusion in 3D images, a common way to accomplish instance segmentation in practice is to apply post-processing over semantic segmentation output. For example, Chen et al. proposed to apply watershed on the raw probability map of the output of semantic segmentation models [4]. Besides, they proposed to train a multi-task 3D VNet that learns both the teeth region and the teeth surfaces to gather more information about teeth and better separate neighboring teeth [17]. However, the dense skip connections in VNet and multi-task learning severely increase computational overhead. We follow the same idea of separating individual tooth by applying watershed over UNet probability map that fits into our pipeline. However, we keep a simple single-task problem with MPUNet as the backbone model while enforcing the gap regions for better separation of teeth through a dedicated loss function. To our best knowledge, our work is the first on automatic segmentation of human jaws that separates both individual tooth and bones (maxilla and mandible) while accurately detailing the gaps between them with very few data.

II. MATERIALS AND METHODS

In order to achieve accurate segmentation with a limited number of annotated training images, our strategy involves several key components: (i) We use the MPUNet approach, which segments 3D medical images by breaking them down into 2D views while maintaining as much spatial information as possible. (ii) To prevent overfitting and memory issues, we use a simple yet effective backbone model for segmentation. (iii) The model is pre-trained using data without the gap to learn general semantic features and then fine-tuned using a small set of highly accurate annotated data for gap generation. (iv) A dedicated weighted distance loss is used to emphasize the gap between the teeth and bones and between neighboring teeth. (v) We separate individual tooth by applying marker based watershed segmentation over the UNet output probability map.

Here we divide the pipeline into two parts; the first part focuses on model construction and training, while the second deals with individual tooth segmentation over the model output, corresponding to strategy (v).

A. GENERAL PIPELINE

As a general pipeline, we are inspired by the method proposed by Xu et al. in regard to neural network training [30].

1) MODEL

As a baseline model, we use the MPUNet proposed by Perslev et al. to segment the 3D jaws using 2D UNet while preserving as much 3D spatial information as possible by generating views from different perspectives [20].

2) TRANSFER LEARNING

The accurate segmentation and efficient convergence with limited data rely partially on pre-training the model using a dataset with inferior annotation, followed by fine-tuning with a smaller set of precisely labeled data. Here we also follow [30] by transferring the weight in the last softmax layer and explicitly learning the encoder, which results in much faster convergence and correct encoding of the gap respectively.

3) LOSS FUNCTION WITH WEIGHTED DISTANCE MAP

In the pre-training step, the model is trained using a standard categorical cross-entropy loss, as we observed no improvement using a class-wise weighted cross-entropy loss or the Dice loss. During fine-tuning, to guide the model in learning the space between the teeth and bones where the PDL is located, a voxel-wise weight map $w(x)$ is applied to the loss function based on the distances from the foreground class borders. This approach was first proposed in the original UNet paper, which we have adapted for use with 3D data in a modified form [22], [30]. We define $w(x)$ as follows,

$$w(x) \equiv w_c(x) + w_0\, e^{-\left(\frac{(d_1(x)+d_2(x))^2}{2\sigma^2}\right)}. \tag{1}$$

where d_1 and d_2 represent the distance to the border of the nearest foreground class and the second nearest foreground class respectively. We follow the original UNet paper and set $w_0 = 10$ and $\sigma = 5$. $w_c : \Omega \rightarrow R$ was originally proposed to balance the class frequencies, which we do not enforce; thus, w_c is set to 1 for every class c. During the fine-tuning process, the corresponding slice of the 3D weight map is sampled in conjunction with the images and labels. Then, the weight map is multiplied element-wisely with the cross-entropy loss between predictions and labels on each pixel before reduction and backpropagation. The whole process is illustrated in Fig. 2.

The incorporation of the distance-based weights (Eq. (1)) into the training of the neural network is inspired by the anticipation that in further FE simulations, a similar distance-based metric will be employed to generate space between the

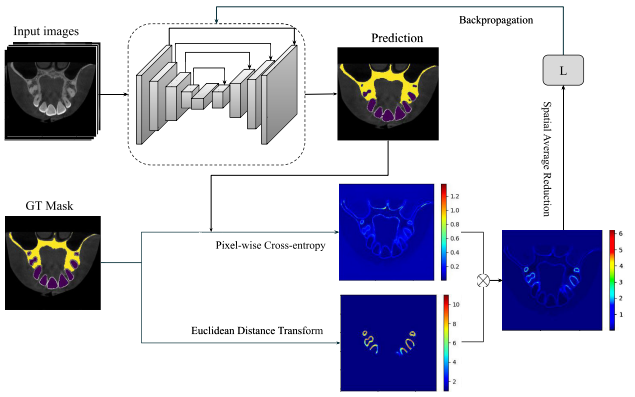

FIGURE 2. Model training pipeline weighted by distance map calculated from Eq. (1). ⊗ denotes element-wise product, which suppresses the general boundary uncertainties while amplifying loss near the gaps. Note that the pixel-wise cross-entropy is visualized after averaging over all the classes. L is the final loss as a scalar after reduction.

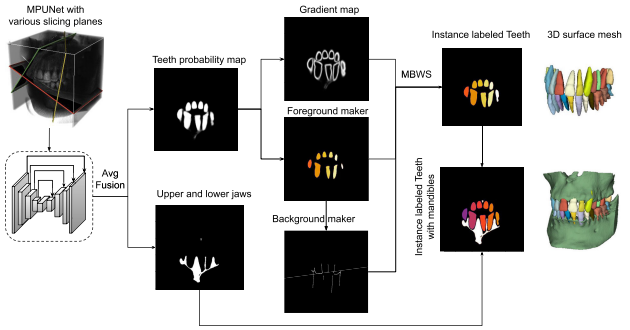

FIGURE 3. Individual tooth segmentation pipeline. MPUNet output from various views are first fused together. The probability map of teeth class is used by MBWS to generate individual tooth segmentation, which is then combined with the segmentation of bone class (maxilla and mandible). Coloring is random.

segmented teeth and bone geometries to locate the PDL. This methodology is outlined in [12].

4) SAMPLING STRATEGIES

Careful sampling and interpolation are crucial for obtaining corresponding 2D slices from a 3D medical image viewed from a random orientation different from the standard RAS axes, which our initial test runs showed evidence of being the actual key to success. Here we follow the same idea of Perslev et al. to sample on isotropic grids within a sphere of diameter m centered at the origin of the scanner coordinate system in the physical scanner space [20]. Pixel dimension $d \in \mathcal{Z}^+$ of the grid and the actual size (diameter) $m \in \mathcal{R}^+$ (controlled by voxel size) in millimeters of the sphere need to be settled before the model training to decide the input size to UNet and its field of view. We differ from [20] by following the modification made by [30] in that these two numbers are chosen differently during training and inference. Briefly, d and m are computed heuristically as the 75 percentile across all axes and images during training but as the maximum value across all axes and all training images together with the current test image during testing. Please refer to [30] for the justification of this modification.

B. INDIVIDUAL TOOTH SEGMENTATION

The segmentation of the CBCT scans is conducted in two steps; first, proper teeth-bone segmentation is performed using the strategy discussed in the previous section with MPUNet; second, the teeth segments are decomposed into individual tooth segments.

Based on our observation, the model is less confident in the contacting interface of the two adjacent teeth in the output of the MPUNet before the argmax step, meaning a lower value in the probability map. Therefore, we apply a Marker Based Watershed Segmentation (MBWS) algorithm over the teeth probability map to separate the wrongly merged neighboring teeth [4]. Watershed is an unsupervised instance segmentation model that refers metaphorically to a geological watershed that separates adjacent drainage basins. Fig. 3 illustrates the whole process, where foreground and background markers are generated to guide the watershed operation based on the output probability map of MPUNet after averaging different views. Details of MBWS with foreground and background generation are explained in Appendix B.

The final result is the union of segmented tooth instances and the bone classes as shown in Fig 3, while the bone class has a higher priority in the intersecting/overlapping regions. Note that the upper and lower bone classes, i.e., mandible and maxilla can be trivially separated by a simple CCD because, unlike teeth, the upper and lower bone classes are always disconnected by a large gap. Here the bone class is labeled in one color for simplicity.

C. DATA AND EXPERIMENTS

We use 13 CBCT scans in this study, where 12 scans belong to 3Shape A/S in-house CBCT dataset, and one scan (P12 in Table 5) is obtained from 3DSlicer's ''Sample Data'' module, titled ''CBCT-MRI Head''. In all scans, the teeth and bone are annotated in both the upper and lower jaw. The scans were acquired from multiple resources from the typical age group of adult male and female ranging between 34 to 64 years old. Further details on sex, age, manufacturer details, and scanner settings are presented in Table 5 in Appendix A. Most of the patients have various dental problems such as dental implants and missing teeth. Besides, the dataset comprised of scans with different voxel sizes and various levels of artifacts such as metal filling artifacts or double contouring artifacts due to the movement of patients in the image acquisition step [16], [18]. Details about the utilized scans and artifacts are listed in Table 4 in Appendix A.

1) TRAIN-TEST SPLIT

Due to the difficulty of concise manual labeling to ensure the gap between teeth and bones, we only have eight cases with an accurate label map detailing the teeth-bone gaps whereas the rest five labeled data are inaccurate due to the missing gaps. These five scans are used for pre-training the

network, while the eight scans with accurate label maps are split equally for the train-test, meaning that only four scans are used to train the network to detail the gaps. Specifications about the data split are listed in Table 4 in Appendix A.

2) PRE-PROCESSING

As presented in Table 4, the original scans are with various voxel sizes and different dimensions. Therefore, all the scans are first upsampled to the smallest voxel size (0.15 mm) in the dataset with a B-Spline interpolation and cropped to an identical dimension of 512^3. We then pre-process the data by applying an intensity standardization based on the equation $X_{\text{scale}} \equiv (x_i - x_{\text{mean}})/(x_{75} - x_{25})$, where x_{25} and x_{75} are the 1st and 3rd quartiles respectively. This transformation scales the intensity based on quartiles and is more robust to outliers, which is especially crucial when working with data involving metal artifacts, in some cases resulting as outliers with extremely high-intensity values. We apply this standardization in two steps: over the 3D volume and then over each sampled slice to MPUNet. No other pre-processing is used to avoid potential errors that can easily propagate in the neural network.

3) EXPERIMENTAL SETUP

The network is trained on NVIDIA GeForce RTX 3090 with a batch size of 10 using the Adam optimizer for 60 epochs with a learning rate of 10^{-5} and reduced by 10% for every two consecutive epochs without performance improvements. We stop training if the performance of five consecutive epochs does not improve. Pre-training takes approximately one day, while fine-tuning takes about 10 hours to converge.

4) AUGMENTATIONS

We apply Random Elastic Deformations to generate images with deformed strength and smoothness [24]. The augmentations are generated on the fly during the training process, and following MPUNet we assign a weight value of 1/3 for the deformed samples [20].

D. ETHICS STATEMENT

The requirement for the ethical committee's approval was waived from ''Center for Regional Development, The Scientific Ethics Committee'' with a reference number 21063693, with the following statement: ''It has been assessed that this is not a health science research project as defined in section II of the committee act, but that it is a non-invasive study containing 3D scan images of jaws and teeth''. Note that, this work only uses an available dataset that already had been collected by 3Shape, and no new scans has been collected just for use of this study. All scans had been acquired as part of a patient's treatment and had already been thoroughly studied by patient's dentist/orthodontist, which is a legal requirement when performing a CBCT scan. Hence, here is no possibility that we can discover additional diseases etc. that the patient had not already been informed about. The patients and the dentists have given written consents for using the scans.

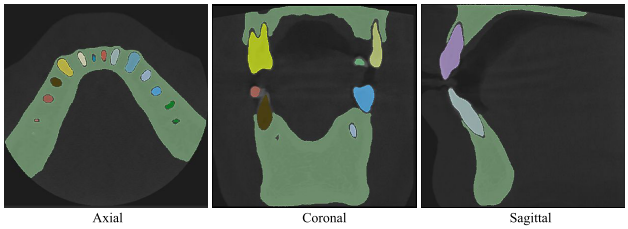

FIGURE 4. Generated gaps for one of the scans in the test set displayed from different views.

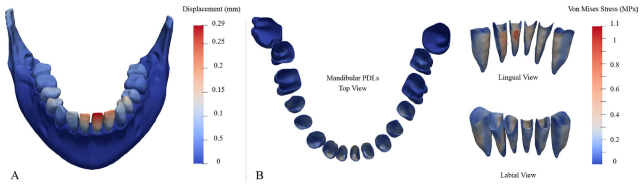

FIGURE 5. Finite element analysis of a tipping scenario. A: Displacement filed of the teeth. B: smooth von Mises stress pattern on the periodontal ligaments.

III. RESULTS AND DISCUSSIONS

In order to produce geometries suitable for finite element (FE) models, the auto-segmentation framework must accurately separate teeth and jaw bones and produce precise results near the boundaries, which are crucial for creating the PDL layers in the jaw. We evaluate the performance of generating the general tooth-bone semantic structures and the gaps between the teeth-bone interfaces in the first subsection. In the second subsection, we evaluate our further task of individual teeth segmentation.

As shown in Figure 4, the enforcement of a distance weight to the loss allows the model to accurately capture the gap. The final result is anatomically accurate and requires minimal manual intervention for subsequent simulations, such as finite element analysis. Segmentation results in 3D are available at https://github.com/diku-dk/AutoJawSegment. As an example, we have generated the PDL geometries on the reconstructed geometries obtained from the segmented CBCT scans with a method proposed in [12] to analyze the stress distributions in a tipping scenario as shown in Fig 5. The results demonstrate a smooth stress pattern, indicating that the output from our method is suitable for finite element (FE) simulations.

A. PERFORMANCE METRICS

Although the commonly adopted measurements of voxel-wise correspondence, e.g., Dice Score, could be misleading regarding the final FE simulations, we still include these measurements as part of the quantitative validation and an ablation study of our several design choices. The Dice Score is defined as

$$\text{Dice}(P, Y) \equiv \frac{2\,|P \cap Y|}{|P| + |Y|}. \tag{2}$$

where P and Y denote the predicted result and ground truth segmentation respectively.

In addition to the standard Dice score, we are particularly interested in evaluating the performance of the model in the surface and gap regions. To assess this, we adopt two additional evaluation metrics. The first metric is the Hausdorff distance (HD), which measures the surface accuracy by calculating the largest distance between the predicted result P and the nearest point on the ground truth Y.

$$\mathrm{HD}(P, Y) \equiv \max(\max_{p\in P} \min_{y\in Y} \|p - y\|_2, \max_{y\in Y} \min_{p\in P} \|p - y\|_2) \tag{3}$$

Secondly, Average Segmentation Surface Distance (ASSD) measures the average distance between the estimated segmentation surface S_P and the ground truth surface S_Y. The surface is computed by subtracting erosion from dilation.

$$\mathrm{ASSD}(P, Y) \equiv \mathrm{mean}(\mathrm{mean}_{d\in S_P} (\mathrm{dist}(d, S_Y)) \times \mathrm{mean}_{g\in S_Y} (\mathrm{dist}(g, S_P))). \tag{4}$$

where $\mathrm{dist}(d, S_Y) \equiv \min_{y\in S_Y} \|d - y\|_2$ denotes the nearest Euclidean distance from point d to surface S_Y.

Although the above two surface measurements better capture the segmentation stability and conciseness than Dice score, they are based on the whole structure with parts that are not that critical for later simulation studies, e.g., the upper surface of the maxilla and lower surface of the mandible. Instead, we are only interested in the parts where two instances meet, i.e., the teeth-bone interfaces. Therefore, we also adopt GapDice proposed in [30] in Eq (6) to measure the average Dice score only around the gap regions.

Given the segmentation results P and the ground truth segmentation Y, the gap region G is defined by thresholding the Euclidean distance transformation map of Y

$$G = \{x | d_1(x) + d_2(x) < \epsilon\} \tag{5}$$

where as defined in Eq (1), d_1 and d_2 represent the distance to the border of the nearest foreground class and the second nearest foreground class in Y, respectively. ϵ is the threshold value, which we set $\epsilon = 5$ as we found it to effectively capture both the gap and boundary regions.

The Dice score between P and Y is then calculated in the standard manner, but only inside G , as defined in Eq (6). Fig 6 shows an indication of such regions.

$$\mathrm{GapDice}(P, Y) \equiv \frac{2\,|P \cap Y \cap G|}{|P \cap G| + |Y \cap G|}. \tag{6}$$

B. QUANTITATIVE RESULTS AND ABLATION STUDY

Table 1 presents the aforementioned performance metrics on the test set, including four images with accurate ground truth segmentations. This experiment is implemented by modifying one of the design choices each time while fixing the others. (i) The strategy described in the Materials and methods Section (ours), (ii) Training the model without pre-training inaccurate data with no gap (NoPretrain), (iii) Training the model without enforcing distance-based weight map (NoWeight), (iv) Using only inaccurate data without fine-tuning the model (NoFineTune), (v) Using a 3D UNet [6] as the backbone model (3DUNet).

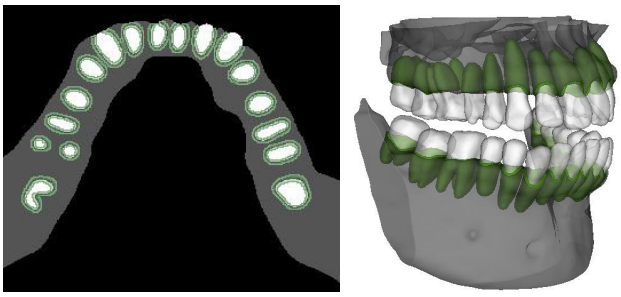

FIGURE 6. The estimated gap region (green) when calculating GapDice for a patient, illustrated in an axial slice (left) and in 3D (right).

TABLE 1. Test results of our model compared with various design choices and other models from the literature.

	Dice (%) ↑	GapDice (%) ↑	ASSD (*mm*) ↓	HD (*mm*) ↓
Ours	**95.14 ± 1.21**	**64.28 ± 4.65**	0.118 ± 0.03	8.538 ± 3.65
NoPretrain	94.05 ± 1.35	59.68 ± 4.69	0.139 ± 0.04	11.91 ± 6.93
NoWeight	94.85 ± 1.19	61.87 ± 4.49	0.127 ± 0.03	8.183 ± 1.62
NoFineTune	95.09 ± 1.34	58.24 ± 3.20	**0.115 ± 0.03**	**7.098 ± 1.48**
3DUNet [6]	64.07 ± 0.52	36.36 ± 4.51	2.010 ± 0.87	45.41 ± 11.4
MSDNet-3 [19], [27]	49.80 ± 2.77	21.85 ± 2.74	1.943 ± 0.599	29.41 ± 6.56
MSDNet-100 [19], [27]	88.91 ± 4.33	59.01 ± 1.89	0.838 ± 0.86	26.95 ± 7.48
MSDNet-200 [19], [27]	89.69 ± 3.74	58.97 ± 2.56	0.800 ± 0.90	28.09 ± 6.32

In addition, we also compare our results with 2D mixed-scale dense CNN (MSDNet) [19] adopted by Wang et al. [27] for the segmentation of human jaws as mentioned in the Introduction. The model is trained only over the extracted axial slices of CBCT scans from the training set, as proposed in [27]. During inference, the 2D prediction results of all the slices of each test scan are concatenated back to 3D for validation. The MSDNet employed by Wang et al. [27] has only a depth of 3, which we found extremely insufficient for such task. We have thus also considered a depth of 100 and 200, as adopted in the original MSDNet paper [19].

The results indicate that the MPUNet (all the first four models) performs significantly better than the standard 3D UNet when dealing with limited data. The 3D UNet fails to learn the general semantic features of tooth-bones with few data compared with MPUNet. Similarly, our model significantly outperforms the MSDNet [19] adopted by Wang et al. [27], even with more depth adjustment. We speculate the poor performance of MSDNet is due to the model's oversimplified structure without downsampling and upsampling phases like in UNet and its insufficiency in learning 3D data from axial slices alone. These drawbacks prevent it from learning appropriate features, especially on data with high noise ratios like the scans in our dataset, compared with the dataset in [27] where the CBCT scans are free of metal artifacts.

Among ablation studies using MPUNet as the backbone model (all the first four models), it is very interesting to

notice that the model without being fine-tuned (NoFineTune) gives a high Dice score and best surface measurements (HD and ASSD). However, since it is anatomically incorrect in that it fails to detail the gap between tooth-bones (cf. Fig. 1c), it has significantly worse performance in the proposed task-specific measurement, i.e., GapDice. This is an indication of why the standard performance metrics that measure voxel-wise correspondence or surface closeness can be misleading regarding the final FE simulation and needs to be resolved for future segmentation works.

Apart from this, our pipeline outperforms in almost all four metrics. Especially, although the difference in the Dice score is not significant (95.14 ± 1.21 vs 94.05 ± 1.35), pre-training on inaccurate data and enforcing the weight map during fine-tuning shows a significantly better GapDice score (64.28 ± 4.65 vs 59.68 ± 4.65) and ASSD (0.118 ± 0.03 vs 0.139 ± 0.04), which is vital for further simulation. Nonetheless, we notice that the GapDice score is significantly lower than the standard Dice score even in our pipeline which has the highest GapDice. Such segmentation errors are mostly due to various artifacts in the scan, as listed in Table 4, which influences the segmentation results. In general, our results in Table 1 have shown good segmentation performance and robustness to the aforementioned artifacts by producing high Dice scores and low surface deviations. However, the concise modeling of details such as PDL layers in noisy scans can be challenging even with our model adaptations to penalize more on the gap regions. A future direction for providing an even more robust network against the mentioned artifacts would be including more data that capture various kinds of artifacts or adding synthetic artifacts to the scans to verify if the model can be trained to learn invariance to the artifacts. Alternatively, deep learning models have been proposed to reduce artifacts as a preprocessing step for the auto-segmentation task [2], [34]. This means we would need to have more data from the scan with artifacts along with the scan of the same patient without artifacts, which is difficult to obtain.

1) RESULTS WITH CROSS-VALIDATION

The aforementioned experiments and results in Table 1 are based on a specific train-test split of the eight scans with accurate label maps. This choice is to preserve a similar level of noise/artifacts in the training data (used for fine-tuning to learn the gap) and test data, as listed in Table 4 in Appendix A. As another common practice in machine learning, here we also conduct a 5-fold cross-validation by randomly dividing the eight scans into training and test sets, ensuring an equal split of four scans in each set as before to analyze performance variations. Table 2 shows the mean and standard deviation of the results with various design choices in correspondence to those in Table 1. Note that the methods with different backbone models (3DUNet and MSDNet) are not included for cross-validation as they have shown to have significantly poorer performance in Table 1. Table 2 generally shows a similar pattern with Table 1 in that our pipeline is able to give significantly better GapDice which is vital for further simulation studies [12].

TABLE 2. Cross-validation with various design choices.

	Dice (%) ↑	GapDice (%) ↑	ASSD (*mm*) ↓	HD (*mm*) ↓
Ours	95.13 ± 1.35	$\mathbf{65.75 \pm 4.64}$	$\mathbf{0.117 \pm 0.06}$	7.76 ± 0.99
NoPretrain	91.63 ± 4.2	58.23 ± 1.35	0.174 ± 0.05	11.93 ± 0.07
NoWeight	$\mathbf{95.27 \pm 2.83}$	63.95 ± 6.11	0.117 ± 0.06	8.183 ± 1.00
NoFineTune	94.48 ± 0.59	58.07 ± 3.02	0.139 ± 0.02	$\mathbf{7.46 \pm 0.4}$

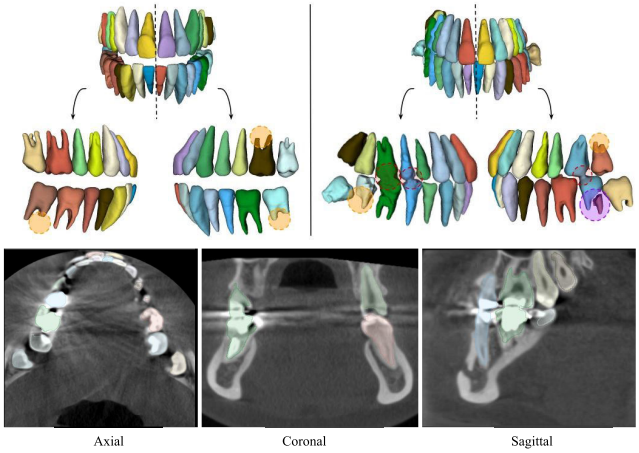

FIGURE 7. Illustration of failure cases. Top: The segmentation results on two different test scans. Different failure cases are illustrated with different colored circles, i.e., orange: inaccurate segmentation of the root apexes; red: connected teeth problem; purple: a single tooth is wrongly segmented with different labels. Colors are randomly assigned to different teeth. Bottom: The scan with overlaid individual tooth labels from the top right case, displayed from different views and showing various artifacts that explain the failures.

C. PERFORMANCE OF INDIVIDUAL TEETH SEGMENTATION

We further evaluate the performance of individual tooth segmentation of our pipeline with the watershed method mentioned above. Fig 3 illustrates that our pipeline can generate visually accurate surface meshes of each tooth and bone even in cases where the CBCT had been acquired in the natural biting position, making the individual tooth segmentation complex as the maxillary and mandibular teeth are touching each other in most of the occlusal surfaces. On the other hand, Fig 7 illustrates several failure cases in one test scan. This test case is filled with various artifacts such as crowded teeth, metal fillings, or dental bridges, as indicated in Fig 7 Bottom. As mentioned in the previous subsection, such artifacts can influence the segmentation in fine detail, e.g., the gap between neighboring teeth and roots, which results in connected teeth and missing root apexes.

Numerical evaluation of the individual tooth and bone segmentation is tricky because, unlike deep-learning-based instance segmentation methods, no soft region proposals are involved in the proposed method, making it impossible to compute a mean Average Precision (mAP). Therefore, our result of individual tooth segmentation is evaluated using the

TABLE 3. Test results of individual teeth segmentation compared with semantic segmentation.

	Dice (%) ↑	GapDice (%) ↑	ASSD (*mm*) ↓	HD (*mm*) ↓
Teeth	95.14 ± 1.21	64.28 ± 4.65	0.118 ± 0.03	8.538 ± 3.65
Individual Tooth	95.05 ± 1.00	64.49 ± 3.93	0.116 ± 0.03	7.840 ± 1.94

same metrics in the previous section as shown in Table 3. In this case, all the predicted teeth instances are mapped back to a binary case and then combined with the bone class. This is an unfair comparison since those metrics cannot reflect its ability to distinguish individual teeth. However, it is interesting to investigate if the further separation of individual teeth does not harm the overall performance, even in this unfair setting. In particular, Table 3 shows that the result after applying individual tooth segmentation gives almost identical results, with a surprising improvement of GapDice and HD. This evaluation scheme could provide insight that the teeth segments deviate negligibly from the prior segmented single tooth class.

We acknowledge that the watershed method involves several hyper-parameters, e.g., the threshold of the UNet probability map and the opening and erosion sizes in Eq (7). The values of these parameters must be tuned beforehand to ensure that neighboring teeth do not share the same foreground marker while avoiding creating multiple markers for the same tooth. Furthermore, one might need to tune these parameters when applying the same pipeline to other problems with different scales or resolutions. Therefore, a future work direction would be to infer those numbers automatically from the studied dataset.

IV. CONCLUSION

Our proposed auto-segmentation framework successfully segments both individual tooth and bones (maxilla and mandible) from CBCT scans of human jaws, with accurate tooth-bone boundaries and the gaps between the teeth roots and sockets. The framework employs a modified version of MPUNet, which is pre-trained on a dataset that does not consider the presence of the PDL layer to learn the general features of tooth-bone geometries. The model is then fine-tuned using a small set of highly accurate segmentations with a dedicated loss function that penalizes the gap regions. This allows the model to better understand the gap where the PDL layer resides and generate anatomically accurate segmentations. We further separate individual teeth by applying watershed segmentation over the MPUNet output. The results of our experiments demonstrate the effectiveness of our framework in detailing critical features, such as the gap between the teeth-bone interfaces and the interproximal regions of the teeth.

A trained segmentation professional has verified our work, and the results show improved numerical results, reaching an overall Dice score above 95% and a significantly higher GapDice than other methods. Our approach can improve anatomically incorrect and poorly annotated datasets with a few accurate labels. One ablation study indicates that the standard performance metrics can be misleading regarding the final FE simulation by producing high-performance metrics but anatomically incorrect results. On the other hand, our results from the finite element (FE) analysis performance test indicate that the models generated produce stress patterns that are smooth and free of artifacts caused by missing gaps in the geometry. As a result, the segmentation outcomes from this study can be applied to generate FE models with minimal adjustments.

APPENDIX A
UTILIZED SCAN DETAILS AND CBCT ARTIFACTS

Image artifacts can be broadly defined as visual effects in reconstructed data that are absent in the real-world object being studied. These artifacts may be the result of various factors, such as subject movement, hardware limitations, the simplified mathematical assumptions used for 3D reconstruction, or their combination. These artifacts, their severity, and voxel sizes can play a significant role in the segmentation task's complexity. Therefore, to provide an overview of observed artifacts in the dataset, we assessed the existence of the common CBCT artifacts, i.e., metal artifacts, noise, blurriness, motion, and aliasing artifacts [16], [18], [23].

The noise artifact can be observed as inconsistent voxel intensities in regions where similar intensities should be present. In addition, double contours can be observed in the CBCT scans that are typically caused due to the patient's movement during the image acquisition process, making it difficult to accurately identify boundaries and delicate structures. Another common effect is the aliasing pattern, which can be seen as lines diverging from the center toward the periphery [16], [18], [23]. Moreover, metal artifacts can be seen as regions with high intensities followed by streaks diverging from the center of the metallic restoration/crown, making it difficult to precisely identify the studied tooth's boundaries. Furthermore, such metal artifacts can cause inaccurate grayscale values in areas not immediately adjacent to the metallic restoration [16], which we refer to as the ghosting effect here.

Table 4 provides details of the utilized scans in this study and represents an overview of the train-test split in this study, as well as the involved artifacts in each scan ranked between zero to two, specifying the artifact's severity level. Further details on sex, age, manufacturer details, and scanner settings are presented in Table 5.

APPENDIX B INDIVIDUAL TOOTH SEGMENTATION DETAILS

Individual tooth is separated by applying the watershed method over the MPUNet probability map of the teeth class. The watershed method considers the intensity value of each voxel as the height, where a high value denotes spikes/hills and a low value denotes valleys. It fills every isolated valley (local minima) with different colored water (labels). As the water rises, depending on the peaks (gradients) nearby, water from other valleys with different colors will merge. The

TABLE 4. Specification of the utilized scans, including details on the original voxel size, number of missing teeth (including the wisdom teeth), different artifact types, and data split for model training. The included artifacts ranked between zero, one, and two to specify the artifact's level in each scan. This variety of artifacts indicates a challenging task for learning an auto-segmentation network. The last column represents the data split for the pre-training (PreT), fine-tuning (FineT), and testing (Test) steps.

Scan	Original scan		Cropped resampled ROI		Missing	Artifacts						Involved
ID	Voxel Size	Dimension	Voxel Size	Dimension	Teeth	Metal	Noise	Motion	Ghosting	Aliasing	Total	set
P1	0.3	400x400x280	0.15	512x512x512	3	2	2	1	2	1	8	Test
P2	0.15	532x532x540	0.15	512x512x512	2	1	1	0	1	1	4	Test
P3	0.15	532x532x540	0.15	512x512x512	4	2	1	0	1	2	6	FineT
P4	0.15	532x532x540	0.15	512x512x512	4	1	1	0	0	1	3	PreT
P5	0.15	400x400x280	0.15	512x512x512	4	1	2	0	1	1	5	FineT
P6	0.3	400x400x280	0.15	512x512x512	3	2	0	1	2	2	7	Test
P7	0.3	400x400x280	0.15	512x512x512	4	1	1	0	1	1	4	FineT
P8	0.3	400x400x280	0.15	512x512x512	0	2	2	0	2	2	8	PreT
P9	0.3	400x400x280	0.15	512x512x512	1	2	0	0	2	2	6	PreT
P10	0.3	400x400x280	0.15	512x512x512	4	2	0	1	2	2	7	PreT
P11	0.2	752x750x400	0.15	512x512x512	1	1	1	0	2	1	5	PreT
P12	0.25	520x406x340	0.15	512x512x512	3	2	2	0	2	2	8	FineT
P13	0.2	501x501x501	0.15	512x512x512	1	2	2	0	2	2	8	Test

TABLE 5. Details of studied cohort and utilized devices for image acquisition including manufacturer information and device settings.

Scan ID	Sex	Age	Tube voltage (kvp)	Tube current (mA)	Manufacturer	Manufacturer model name
P1	F	55	85	4.8	Vatech Company Limited	Implagraphy
P2	N/A	N/A	N/A	N/A	3Shape Medical A/S	X1
P3	N/A	N/A	N/A	N/A	3Shape Medical A/S	X1
P4	N/A	N/A	N/A	N/A	N/A	N/A
P5	F	35	120	18	Xoran Technologies	i-CAT 3D Dental Imaging System
P6	F	64	89	4	Vatech Company Limited	PaX-Flex3D
P7	M	64	85	4.8	Vatech Company Limited	Implagraphy
P8	F	55	85	4.8	Vatech Company Limited	Implagraphy
P9	F	46	89	4	Vatech Company Limited	PaX-Flex3D
P10	M	36	89	4	Vatech Company Limited	PaX-Flex3D
P11	N/A	N/A	90	12	3Shape Medical A/S	X1
P12	M	N/A	N/A	N/A	N/A	N/A
P13	F	34	90	7	Planmeca	Planmeca ProMax

algorithm then tries to prevent the merging by building ''barriers'' locations where water merges until all the peaks are underwater. The barriers then naturally mark the boundary for each instance, which results in instance segmentation of the teeth.

In practice, the primary watershed method usually produces over-segmented results due to its sensitivity to noise or other irregularities in the image, like many local minima. Instead, Marker-Based Watershed Segmentation (MBWS) alleviates this problem by specifying the valley points (*foreground markers*) that are to be merged and barriers (*background markers*) to the model. The whole process is shown in Fig 3, which is explained in the following paragraphs.

1) FOREGROUND MARKERS GENERATION

Instead of working directly on the image here, the *foreground markers* are determined by thresholding the UNet output probability map since the probability map naturally represents how confident the model is in predicting the foreground class, in this case, teeth. Eq (7) below indicates the foreground regions where we first apply a threshold of 0.8 over the probability map on teeth class $P(x)$. We then remove isolated false positives and shrink the thresholded foreground regions by applying an opening, $\circ$, with a structural ball element $E_{5\times5\times5}$ followed by an erosion, $\bullet$, with a structural ball element $E_{3\times3\times3}$ to provide disconnected teeth. Note that the radii should be determined based on the general shape of the instance, in this case, a tooth, to separate neighboring teeth while avoiding introducing undesired disconnectivity inside each tooth.

$$M_f \equiv \{x \mid (P(x) > 0.8) \circ E_{5\times5\times5} \bullet E_{3\times3\times3}\}. \quad (7)$$

2) BACKGROUND MARKERS GENERATION

The *background markers* are generated based on the *foreground markers* M_f generated from the previous step by first applying a distance transform over M_f, which corresponds to the terms of d_1 and d_2 in Eq (1). The final background region is generated by thresholding both the difference and the sum of the two distances, which is indicated in Eq (8). This choice of background markers corresponds to the trimmed perpendicular bisector plane between any two neighboring teeth, thus ensuring neighboring teeth do not get merged by the watershed. The threshold of $d_1 + d_2$ is necessary to ensure that the background marker does not penetrate other foreground regions. The value of 20 is experimental and will need to be tuned for other datasets or voxel sizes.

$$M_b \equiv \{x \mid |d_1(x) - d_2(x)| \leq 1 \wedge |d_1(x) + d_2(x)| \leq 20\}. \quad (8)$$

3) MARKER-BASED WATERSHED SEGMENTATION

As shown in Fig 3, with the selected foreground and background markers, the final MBWS is conducted on the gradient of the UNet probability map due to its good response to weak edge information [4]. The gradient is computed by convolving Gaussian derivative kernel with $\sigma = 2$. Our experience indicates that this preserves root structures better than directly working on the probability map.

4) TRAINING STRATEGY WITH ADDITIONAL WEIGHT-MAP

Although this MBWS to separate individual tooth has shown to be effective, its performance largely depends on the quality of the UNet probability map. More specifically, if the model gives inaccurate results (high probability of being foreground) near the gap between some neighboring teeth, these teeth will share a common foreground marker. Increasing the utilized thresholding value for the foreground or the erosion/opening kernel sizes in Eq (7) can provide different foreground markers for the adjacent teeth. However, this

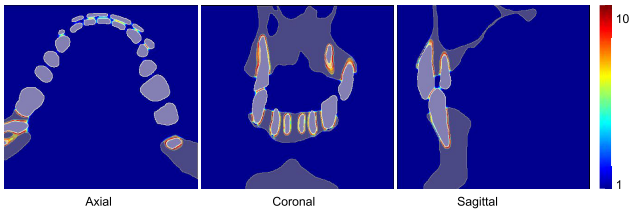

FIGURE 8. The values of the weight map presented as colormap along with the labeled teeth and bone. The proposed weight map enforces gaps not only between the teeth and bone segments where the periodontal ligament lies, but also between nearby teeth.

may also introduce fractions (several foreground markers) inside the same tooth, causing the watershed algorithm to assign several labels to the different parts of the same tooth. Therefore, keeping the morphological and thresholding level is crucial while providing a more accurate result near the interproximal gaps.

As the distance weight map is very effective in learning the gaps in the teeth-bone interfaces, we use this same strategy to learn the interproximal gaps. After applying connected component decomposition over ground truth to get a different label for each tooth, we can follow the same strategy in Eq (1) to enforce a higher weight on both the gap in the teeth-bone interface and in the interproximal regions of the teeth to better separate the adjacent teeth from each other. As shown in Fig 8, the weight map calculated by Eq (1) has a higher value not only between tooth-bone gaps but also between neighboring teeth. Note that such Euclidean transformation in Eq (1) is highly time-consuming because it involves the distance computation to every class and sorting the values afterward. Hence, the time complexity increases with at least $\mathcal{O}(n \log n)$ where n denotes the number of classes. For example, the gap modeling in the teeth-bone interfaces involved only two classes (bone and teeth), but modeling the gaps in the interproximal regions involves approximately 30 classes (number of teeth). Therefore, it is crucial that the weight map is computed before the model training and then sampled together with the corresponding images and labels.

ACKNOWLEDGMENT

The authors would like to thank 3Shape A/S, especially Peter Lempel Søndergaard, for providing this study's CBCT scans, and for their supports in enhancing the segmentation results. The mentioned data was originally acquired for diagnostic purposes unrelated to this study. No other aspect of this work triggered ethical issues.

REFERENCES

[1] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[2] D. F. Bauer, C. Ulrich, T. Russ, A.-K. Golla, L. R. Schad, and F. G. Zöllner, "End-to-end deep learning CT image reconstruction for metal artifact reduction," *Appl. Sci.*, vol. 12, no. 1, p. 404, Dec. 2021.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[4] Y. Chen, H. Du, Z. Yun, S. Yang, Z. Dai, L. Zhong, Q. Feng, and W. Yang, "Automatic segmentation of individual tooth in dental CBCT images from tooth surface map by a multi-task FCN," *IEEE Access*, vol. 8, pp. 97296–97309, 2020.

[5] M. Chung, M. Lee, J. Hong, S. Park, J. Lee, J. Lee, I.-H. Yang, J. Lee, and Y.-G. Shin, "Pose-aware instance segmentation framework from cone beam CT images for tooth segmentation," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103720.

[6] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424–432.

[7] Z. Cui, Y. Fang, L. Mei, B. Zhang, B. Yu, J. Liu, C. Jiang, Y. Sun, L. Ma, J. Huang, Y. Liu, Y. Zhao, C. Lian, Z. Ding, M. Zhu, and D. Shen, "A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images," *Nature Commun.*, vol. 13, no. 1, p. 2096, Apr. 2022.

[8] Z. Cui, C. Li, N. Chen, G. Wei, R. Chen, Y. Zhou, D. Shen, and W. Wang, "TSegNet: An efficient and accurate tooth segmentation network on 3D dental model," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101949.

[9] Z. Cui, C. Li, and W. Wang, "ToothNet: Automatic tooth instance segmentation and identification from cone beam CT images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2019, pp. 6368–6377.

[10] Z. Cui, B. Zhang, C. Lian, C. Li, L. Yang, W. Wang, M. Zhu, and D. Shen, "Hierarchical morphology-guided tooth instance segmentation from CBCT images," in *Information Processing in Medical Imaging*. Cham, Switzerland: Springer, 2021, pp. 150–162.

[11] E. Gardiyanoğlu, G. Ünsal, N. Akkaya, S. Aksoy, and K. Orhan, "Automatic segmentation of teeth, crown–bridge restorations, dental implants, restorative fillings, dental caries, residual roots, and root canal fillings on orthopantomographs: Convenience and pitfalls," *Diagnostics*, vol. 13, no. 8, p. 1487, Apr. 2023.

[12] T. Gholamalizadeh, F. Moshfeghifar, Z. Ferguson, T. Schneider, D. Panozzo, S. Darkner, M. Makaremi, F. Chan, P. L. Søndergaard, and K. Erleben, "Open-full-jaw: An open-access dataset and pipeline for finite element models of human jaw," *Comput. Methods Programs Biomed.*, vol. 224, Sep. 2022, Art. no. 107009.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[14] F. Isensee, P. F. Jaeger, S. A. 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, Feb. 2021.

[15] T. J. Jang, K. C. Kim, H. C. Cho, and J. K. Seo, "A fully automated method for 3D individual tooth identification and segmentation in dental CBCT," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6562–6568, Oct. 2022.

[16] S. R. Makins, "Artifacts interfering with interpretation of cone beam computed tomography images," *Dental Clinics North Amer.*, vol. 58, no. 3, pp. 485–495, Jul. 2014.

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

[18] N. Dwivedi, A. Nagarajappa, and R. Tiwari, "Artifacts: The downturn of CBCT image," *J. Int. Soc. Preventive Community Dentistry*, vol. 5, no. 6, p. 440, 2015.

[19] D. M. Pelt and J. A. Sethian, "A mixed-scale dense convolutional neural network for image analysis," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 2, pp. 254–259, Jan. 2018.

[20] M. Perslev, E. B. Dam, A. Pai, and C. Igel, "One network to segment them all: A general, lightweight system for accurate 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 30–38.

[21] S. Poelert, E. Valstar, H. Weinans, and A. A. Zadpoor, "Patient-specific finite element modeling of bones," *Proc. Inst. Mech. Eng., H, J. Eng. Med.*, vol. 227, no. 4, pp. 464–478, Apr. 2013.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[23] R. Schulze, U. Heil, D. Groß, D. Bruellmann, E. Dranischnikow, U. Schwanecke, and E. Schoemer, "Artefacts in CBCT: A review," *Dentomaxillofacial Radiol.*, vol. 40, no. 5, pp. 265–273, Jul. 2011.

[24] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. ICDAR*, vol. 3, 2003, pp. 958–963.
[25] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 7262–7272.
[26] P.-J. Verhelst, A. Smolders, T. Beznik, J. Meewis, A. Vandemeulebroucke, E. Shaheen, A. Van Gerven, H. Willems, C. Politis, and R. Jacobs, "Layered deep learning for automatic mandibular segmentation in cone-beam computed tomography," *J. Dentistry*, vol. 114, Nov. 2021, Art. no. 103786.
[27] H. Wang, J. Minnema, K. J. Batenburg, T. Forouzanfar, F. Hu, and G. Wu, "Multiclass CBCT image segmentation for orthodontics with deep learning," *J. Dental Res.*, vol. 100, no. 9, pp. 943–949, 2021.
[28] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: BEiT pretraining for all vision and vision-language tasks," 2022, *arXiv:2208.10442*.
[29] 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 *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 939–942.
[30] P. Xu, F. Moshfeghifar, T. Gholamalizadeh, M. B. Nielsen, K. Erleben, and S. Darkner, "Auto-segmentation of hip joints using multiplanar UNet with transfer learning," in *Proc. Workshop Med. Image Learn. Ltd. Noisy Data*. Cham, Switzerland: Springer, 2022, pp. 153–162.
[31] Y. Yang, R. Xie, W. Jia, Z. Chen, Y. Yang, L. Xie, and B. Jiang, "Accurate and automatic tooth image segmentation model with deep convolutional neural networks and level set method," *Neurocomputing*, vol. 419, pp. 108–125, Jan. 2021.
[32] T. Yeshua, S. Ladyzhensky, A. Abu-Nasser, R. Abdalla-Aslan, T. Boharon, A. Itzhak-Pur, A. Alexander, A. Chaurasia, A. Cohen, J. Sosna, I. Leichter, and C. Nadler, "Deep learning for detection and 3D segmentation of maxillofacial bone lesions in cone beam CT," *Eur. Radiol.*, pp. 1–12, May 2023, doi: 10.1007/s00330-023-09726-6.
[33] J. Zhang, M. Liu, L. Wang, S. Chen, P. Yuan, J. Li, S. G.-F. Shen, Z. Tang, K.-C. Chen, J. J. Xia, and D. Shen, "Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization," *Med. Image Anal.*, vol. 60, Feb. 2020, Art. no. 101621.
[34] Y. Zhang and H. Yu, "Convolutional neural network based metal artifact reduction in X-ray computed tomography," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1370–1381, Jun. 2018.

PEIDI XU received the B.Eng. degree in computer science and technology from the Nanjing University of Posts and Telecommunications, in 2018, and the M.Sc. degree in computer science from the Department of Computer Science, University of Copenhagen (DIKU), in 2020, where he is currently pursuing the Ph.D. degree. His research interests include medical image analysis, including image segmentation and the computational modeling of renal blood vessels and flows simulation.

TORKAN GHOLAMALIZADEH received the Ph.D. degree in computer science from the University of Copenhagen, in 2022. During the Ph.D. degree, she focused on developing and analyzing patient-specific computational models obtained from medical images using simulation methods and machine learning. Since 2022, she has been working on AI-based solutions applied to medical images as a Machine Learning Researcher with 3Shape A/S, Copenhagen, Denmark. Her research interests include computational modeling, medical image analysis, and machine learning applied to healthcare.

FAEZEH MOSHFEGHIFAR received the Ph.D. degree, in 2022. She is currently a Postdoctoral Researcher with the Department of Computer Science, University of Copenhagen. Her research interests include the application of state-of-the-art knowledge from the medical and computer science domains to assist medical research.

SUNE DARKNER received the Ph.D. degree, in 2009. He is currently an Associate Professor with the Department of Computer Science, University of Copenhagen. His research interests include density estimation and image similarity, histograms and scale space, geometry, and simulation with applications to digital twins and digital populations.

KENNY ERLEBEN received the Ph.D. degree, in 2005. He is currently a Full Professor with the Department of Computer Science, University of Copenhagen. His research interests include computer simulation and numerical optimization with particular interests in the computational contact mechanics of rigid and deformable objects, inverse kinematics for computer graphics and robotics, computational fluid dynamics, computational biomechanics, foam simulation, and interface tracking meshing.