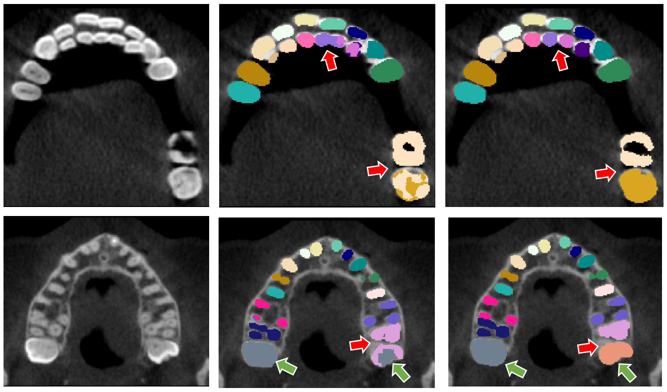
Semantic Graph Attention with Explicit Anatomical Association Modeling for Tooth Segmentation from CBCT Images

Pengcheng Li†, Yang Liu†, Zhiming Cui, Feng Yang, Yue Zhao✉, Chunfeng Lian✉,*Member, IEEE*, Chenqiang Gao

***Abstract*—Accurate tooth identification and delineation in dental CBCT images are essential in clinical oral diagnosis and treatment. Teeth are positioned in the alveolar bone in a particular order, featuring similar appearances across adjacent and bilaterally symmetric teeth. However, existing tooth segmentation methods ignored such specific anatomical topology, which hampers the segmentation accuracy. Here we propose a semantic graph-based method to explicitly model the spatial associations between different anatomical targets (i.e., teeth) for their precise delineation in a coarse-to-fine fashion. First, to efficiently control the bilaterally symmetric confusion in segmentation, we employ a lightweight network to roughly separate teeth as four quadrants. Then, designing a semantic graph attention mechanism to explicitly model the anatomical topology of the teeth in each quadrant, based on which voxel-wise discriminative feature embeddings are learned for the accurate delineation of teeth boundaries. Extensive experiments on a clinical dental CBCT dataset demonstrate the superior performance of the proposed method compared with other state-of-the-art approaches.**

***Index Terms*—Tooth Segmentation, Semantic Graph Attention, Anatomical Association Modeling, CBCT images.**



(a) Original image (b) Conventional method [39] (c) The proposed method

Fig. 1. Typical and challenging cases of automatic tooth segmentation from CBCT images. The red arrows and green arrows show the CBCT image with similar shapes across adjacent teeth and the symmetric teeth, respectively. In these cases, the conventional method led to the confusing delineation of tooth boundaries and misclassification of symmetric teeth; in contrast, our method obtained accurate identification and segmentation results due to the explicit modeling of teeth associations.

I. Introduction

Digital dental technology is widely used in clinical orthodontics and implantation. Accurate tooth segmentation from cone-beam computed tomography (CBCT) images is the foundation for various downstream tasks in the clinical process, e.g., analyzing the relationship of adjacent teeth for dental implants, and measuring the tooth root and crown ratio to accelerate the orthodontic treatment cycle. Since different teeth have similar shape appearance, designing a fully automatic tooth identification and segmentation system is urgently demanded.

Accurate tooth identification and segmentation in CBCT images acquired from dental patients with dramatically changing conditions is a challenging task. Such as the typical examples shown in Fig. 1, adjacent and bilaterally symmetric teeth may have similar appearances and blurry boundaries, or some teeth even conjunct with each other in the cross-sectional CBCT slices, and low-quality images (e.g., metal artifacts) further hinder the segmentation performance. To address these challenges, various methods have been proposed, which can be roughly categorized into two groups, i.e., the *traditional knowledge-based* and the *deep learning-based* methods.

The knowledge-based methods usually predefined hand-crafted features (e.g., according to the prior knowledge of

†These authors contributed equally to this work.

✉Corresponding authors: Yue Zhao (e-mail: zhaoyue@cqupt.edu.cn) and Chunfeng Lian (e-mail: chunfeng.lian@xjtu.edu.cn).

This work is supported by the Chongqing University of Posts and Telecommunications Ph.D. Innovative Talents Project (BYJS202105), the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant (No. KJZD- KJQN202000647, KJQN202100646), the National Natural Science Foundation of China (No. 82101058, 62101431, 61906025, 62176035), the Intelligent Medicine Research Project of Chongqing Medical University (No. ZHYX202101).

Pengcheng Li, Feng Yang, Yue Zhao and Chenqiang Gao are with School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; Chongqing Key Laboratory of Signal and Information Processing, Chongqing 400065, China.

Yang Liu is with Department of Orthodontics, Stomatological Hospital of Chongqing Medical University, Chongqing 401147, China; Chongqing Key Laboratory for Oral Diseases and Biomedical Sciences, Chongqing 401147, China.

Zhiming Cui is with School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China.

Chunfeng Lian is with School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China.

tooth anatomy), based on which the conventional machine learning (e.g., random forest [1] and clustering [2]) or image processing (e.g., edge detection [3]–[5], region growth [6], [7], and active contour [8]–[12]) techniques are further applied to conducting segmentation. Such traditional methods are typically semi-automatic (i.e., require human intervention). More importantly, the hand-crafted features of teeth derived from the healthy population cannot robustly handle the CBCT images for patients with varying conditions.

The state-of-the-art methods typically adopt fully convolutional networks (FCNs) [13] as a fundamental component to hierarchically learn and integrate local-to-global features for tooth delineation, either in an instance-segmentation fashion [14]–[16] or semantic-segmentation fashion [17]. The instance segmentation methods [14]–[16] first detect or localize the positions of different teeth and then delineate each tooth independently. Such detection-based methods may misclassify adjacent and symmetric teeth into the same category, as they could have very similar appearances in CBCT images. The semantic segmentation methods perform the dense voxel-to-voxel prediction for the concurrent delineation of all teeth [17]. However, due to the limited receptive field for a relatively large input image, these semantic segmentation networks may inaccurately label one tooth as multiple categories, i.e., the learned feature representations may fail to differentiate between adjacent teeth in a fine-scale. That is, the performance of existing deep learning methods is still limited in the specific task of tooth segmentation, primarily due to their simple use of general network architectures (from the computer vision community) that ignore the natural associations between anatomical locations of different teeth (especially the adjacent and symmetric ones).

In this paper, we propose a novel semantic graph attention method that *explicitly models the spatial associations between different anatomical targets*, with application to automatic tooth segmentation from CBCT images. Considering that the human teeth have a particular order on the alveolar bone, featuring similar appearances both across bilaterally symmetric quadrants and between adjacent teeth, our method thoroughly removes such inter-class confusions in a coarse-to-fine fashion for precise tooth segmentation. Specifically, a light-weight FCN is first implemented to roughly segment all teeth as four quadrants to reduce the effects of the bilaterally symmetric similarity. Then, in the framework of graph convolutional networks (GCNs) [18]–[20], we propose a specific *semantic graph attention* (SGA) mechanism to learn a discriminative feature embedding, where the confusion in delineating adjacent teeth from each quadrant is minimized. Given a set of feature maps preserving relatively high-resolution spatial information, our SGA explicitly models the anatomical topology of different classes (i.e., teeth) to learn jointly the prototype representations (of each class) and the corresponding voxel-wise semantic attentions, based on which the voxel-wise discriminative capacity of the input features is comprehensively enhanced for the fine-grained segmentation purpose.

The main contributions of this paper are *three-fold*:

- A novel deep learning method is proposed for accurate tooth segmentation from dental CBCT images acquired from patients needing oral treatments. Our method adopts a coarse-to-fine segmentation framework to minimize the adjacent and bilaterally symmetric confusions in learning-based automatic segmentation, which can robustly handle large-scale appearance changes across the images for different patients.
- A semantic graph attention mechanism based on GCNs is designed to explicitly model the anatomical topology of teeth under segmentation. It learns an anatomical-topology-aware feature embedding with fine-grained discriminative capacity, leading to the more accurate delineation of the boundaries between adjacent targets with similar appearances.
- Comprehensive experiments on a clinical dental CBCT dataset show that the proposed method outperformed other state-of-the-art methods, especially for those challenging cases with missing or crowded teeth, justifying the promising performance of our approach in automatic tooth segmentation.

II. Related Work

In recent years, a proliferation of methods has been proposed for tooth segmentation from dental CBCT images. These methods can be broadly divided into two categories in the literature: traditional knowledge-based methods and modern deep learning-based methods.

A. Knowledge-based Method

Handcraft features were utilized to realize CBCT images analysis in the knowledge-based methods. A semi-automatic segmentation method based on a fast threshold was proposed to distinguish single tooth effectively, and alveolar bone from dental CBCT images [21]. However, it is difficult for this model to accurately identify tooth categories and roots by using a thresholding strategy. To segment tooth roots from CBCT images, prior knowledge such as image intensity projection was also proposed for dental CBCT images based on region growing [22]. An improved region growing method utilized an automatic canny filter without Gaussian blur was proposed to segment each tooth [23]. However, all these methods are time-consuming and highly sensitive to noise. To reduce the noise influence and to use prior shape information to guide tooth segmentation, an interactive segmentation method based on Graph Cut theory was subsequently proposed to obtain 3D tooth models from dental CBCT images [24]. This method completes the entire CBCT data segmentation process by propagating the threshold and tooth shape to adjacent slices. Along this research line, active contour methods [12], [25] are then proposed to tackle the unclear boundary issues caused by densely crowded teeth. This method uses a level set algorithm to segment teeth on each slice, and then the threshold method and fast watershed algorithm were both employed to segment the upper and lower teeth. However, these traditional knowledge-based methods heavily rely on the intensity or anatomical heuristics (e.g., assumptions). These methods break down in cases where patient conditions do not satisfy the pre-existing assumptions, e.g., intensity variations caused by metal artifacts and significant anatomical shape variations (e.g., missing or misaligned teeth), which leads to inaccurate tooth category identification.

On the contrary, the proposed semantic graph attention mechanism takes advantage of the inherent tooth arrangement order in each quadrant, which can capture the correlation of different tooth categories according to the tooth anatomy arrangement, and generate stable segmentation results.

B. Deep Learning-based Method

1) CNN-based Approaches: Contrary to previous methods that use prior knowledge to build handcrafted models, deep convolutional networks rely on large amounts of training data to automatically learn a model that outputs feasible predictions. In recent years, although deep learning (or deep neural network) [13], [26], [27] has started to dominate the solving of various problems in medical image analysis [28]–[30], it did not go far enough for dental CBCT image segmentation tasks, and there are limited studies reported to develop automatic tooth segmentation algorithms [14]–[17]. A deep convolutional method based on 3D Mask R-CNN [31] was proposed for tooth instance segmentation [14]. Approaching tooth segmentation as an instance segmentation problem entails treating all teeth as instances of the same class of objects. However, an anatomical identification of the segmented teeth is often also needed, e.g., for further analysis orthodontic correction step or tooth implanting. A coarse-to-fine segmentation network was proposed to address the localizing and segmenting problem in dental CBCT images [17]. This method treats the tooth segmentation task as a multi-class (33) semantic segmentation problem and first trains the coarse step model on a large weakly labeled dataset and subsequently fine-tuning it on a smaller scale albeit precisely labeled dataset. A two-stage localization and segmentation method was suggested to identify the tooth label on 20 CBCT cases [15], then a label optimization strategy is designed to solve the problem of incorrect classification of adjacent teeth or overlapping teeth.

However, the above deep learning-based methods do not employ the inherent arrangement rules of the teeth so that the same tooth may be recognized as multiple categories. The proposed method utilizes an adjacency matrix and graph convolution to model the spatial associations between tooth categories and obtain better segmentation and identification results.

2) Graph-based Approaches: Graph convolutional networks [32] pioneers a new research line for semantic image representation, and this has been applied in many medical image analysis tasks. Liu et al. proposed a bipartite graph convolutional network to bestow existing methods with cross-view reasoning ability of radiologists in mammogram mass detection [33]. Tian et al. proposed an interactive segmentation method based on a graph convolutional network to refine the prostate segmentation results in MR images [34]. GCNs has also attracted attention in the tooth image analysis research community. Sun et al. postulated an end-to-end network based on GCNs to annotate individual teeth and gingiva of dental models [35]. Ma et al. proposed SRF-Net based on prior ordered position information to calculate the adjacency similarity feature vectors for dental model classification [36]. Lian et al. explored a multi-scale graph-constrained module to extract fine-grained local geometric features from dental mesh data, and they combined 3D coordinates and normal vectors to improve the dental mesh segmentation performance [37]. Zhang et al. proposed a two-stream graph convolutional network to learn multi-view information from different geometric attributes in dental model segmentation [38].

However, all these existing methods are applied on dental mesh models, and to our best of knowledge, there is no existing research proposed on dental CBCT imaging with graph convolutional networks.

In this paper, we propose a coarse-to-fine tooth segmentation method based on graph convolution networks. A semantic graph attention module is designed to avoid misclassification by shape similarity, and a global-context attention module is explored to capture the global contextual information. Compared with the proposed method in the tooth segmentation literature, our method can explicitly model the tooth spatial relationship, which is imperative to identify the tooth categories.

III. METHODS

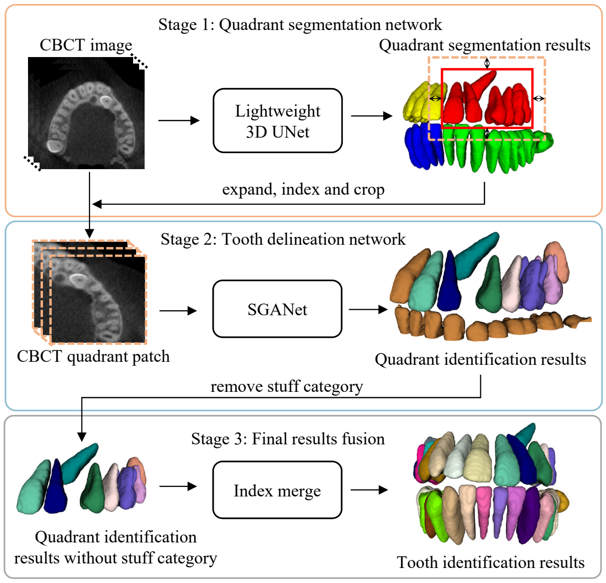

Stage 1: Quadrant segmentation network
CBCT image
Quadrant segmentation results
Lightweight 3D UNet
expand, index and crop
Stage 2: Tooth delineation network
SGANet
CBCT quadrant patch
Quadrant identification results
remove stuff category
Stage 3: Final results fusion
Index merge
Quadrant identification results without stuff category
Tooth identification results

Fig. 2. The schematic diagram of the proposed method. The quadrant-segmentation network in the first stage is employed to divide all teeth into four quadrants. The SGANet in the second stage is utilized to identify and delineate each tooth.The final results fusion in the third stage merged the results of all the four quadrants based on the recorded indices (from the first stage) to reconstruct the segmentation in the original image space.

A. Overview

Fig. 2 shows a schematic diagram of the proposed method for fully automatic tooth segmentation from dental CBCT images in a coarse-to-fine fashion. Briefly, our method contains three main steps. 1) a lightweight network based on 3D

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2022.3179128, IEEE Transactions on Medical Imaging

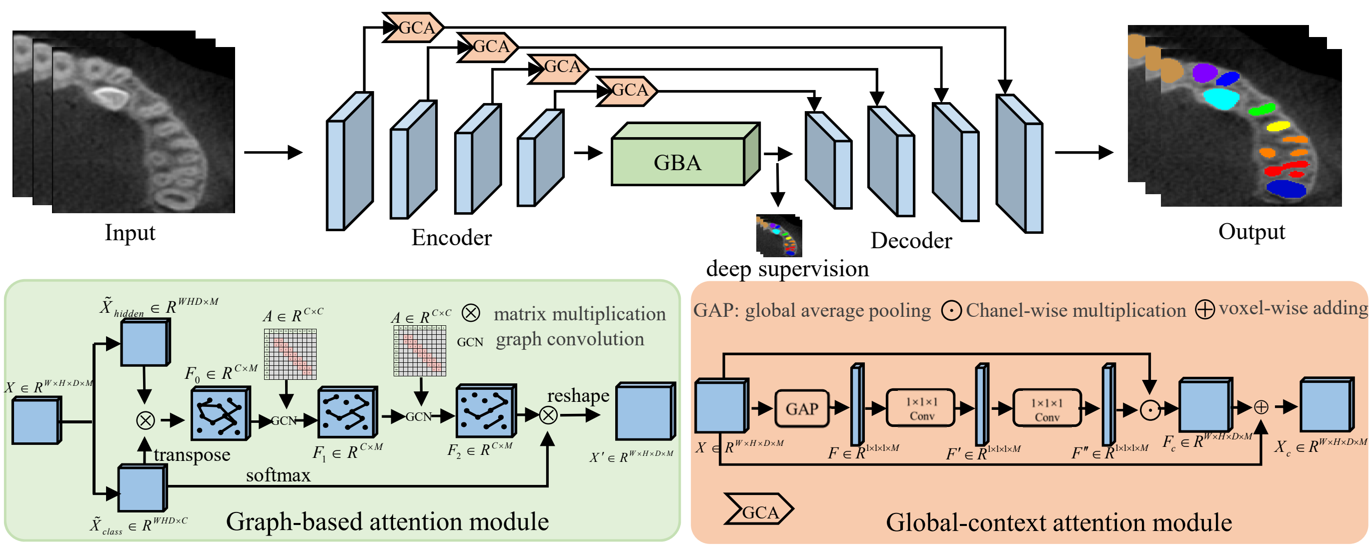

GCA
GCA
GCA
GCA
Input
Encoder
GBA
deep supervision
Decoder
Output
$\tilde{X}_{hidden} \in R^{WHD\times M}$
$A \in R^{C\times C}$
$A \in R^{C\times C}$
⊗ matrix multiplication
GCN graph convolution
$X \in R^{W\times H\times D\times M}$
$F_0 \in R^{C\times M}$
GCN
GCN
reshape
transpose
$F_1 \in R^{C\times M}$
$F_2 \in R^{C\times M}$
$X' \in R^{W\times H\times D\times M}$
softmax
$\tilde{X}_{class} \in R^{WHD\times C}$
Graph-based attention module
GAP: global average pooling ⊙ Chanel-wise multiplication ⊕ voxel-wise adding
GAP
1×1×1 Conv
1×1×1 Conv
$X \in R^{W\times H\times D\times M}$
$F \in R^{1\times 1\times 1\times M}$
$F' \in R^{1\times 1\times 1\times M}$
$F'' \in R^{1\times 1\times 1\times M}$
$F_c \in R^{W\times H\times D\times M}$
$X_c \in R^{W\times H\times D\times M}$
GCA
Global-context attention module

Fig. 3. The architecture of the proposed SGANet. The whole network consists of three main parts. The encoder-decoder module is employed to extract tooth features and recover spatial resolution gradually. The graph-based attention module is designed to explicitly model the spatial associations between different teeth, which avoid adjacent tooth misclassification caused by similar shape. The global-context attention module is explored to enhance multi-scale image details for the voxel-wise classification.

UNet [39] segments four quadrants of teeth from an input CBCT image. It limits the effect of bilaterally symmetric similarity on accurate teeth delineation in subsequent steps. 2) Based on the outputs of the 3D UNet, a more sophisticated segmentation network integrating semantic graph attention mechanism (called SGANet) applies to annotate the teeth in each quadrant. 3) The four quadrant identification results are fused to form a complete tooth segmentation result. SGANet explicitly considers the anatomical topology of different teeth in learning fine-grained discriminative features, which minimizes the confusion for delineating the boundary between adjacent teeth. More details of our method are elaborated in the following subsections.

B. Preliminaries: Graph convolution network

Given the input features $X \in \mathbb{R}^{N\times C}$, where N is the number of nodes in the features defined on the regular grid space $\Omega = \{1,\dots,W\} \times \{1,\dots,H\} \times \{1,\dots,D\}$, and C is the feature dimension. The graph representation G can be obtained from the input features, and formulated as $G = \{V, E, A\}$ with V as its nodes, E as its edges, and A as its adjacency matrix. Formally, the graph convolution can be defined as Eq. 1:

$$\hat{X} = \sigma(AXW), \tag{1}$$

where $\sigma(\cdot)$ is the non-linear activation function, $A \in \mathbb{R}^{N\times N}$ is the adjacency matrix measuring the relations of nodes in the graph, $W \in \mathbb{R}^{C\times C'}$ is the weight matrix. Graph convolution can efficiently spread information between different nodes and edges to obtain features with larger receptive fields based on the defined adjacency matrix.

C. Quadrant Identification

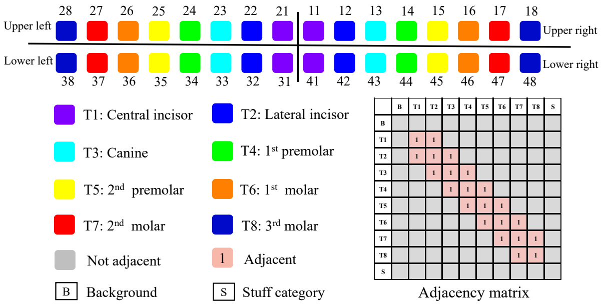
For simplicity, a modified 3D UNet is adopted to participate the teeth into four quadrants, although other more advanced architectures could also be used in this task.

The encoder of the modified 3D UNet contains five convolutional blocks, with each having one $3 \times 3 \times 3$ convolution followed by batch normalization (BN) and ReLU activation. The first convolutional block has 16 channels, which are orderly doubled in the subsequent blocks. Along with the duplication of channel numbers, the feature map size is halved by $2 \times 2 \times 2$ convolutions with stride 2. The decoder has a roughly symmetric architecture compared with the encoder. In the forward path of the decoder, trilinear up-sampling operations are combined with skip connections, convolutions, BN, and ReLu to gradually combine local-to-global features to learn high-resolution feature maps that are equal size to the input image. Finally, a 5-channel $1 \times 1 \times 1$ convolution with softmax activation is applied to assigning each voxel in the input image into a specific class (i.e., four quadrants and the background).

D. Semantic Graph Attention for Teeth Delineation

After the rough segmentation of the four quadrants of teeth, we then design a more advanced deep network for the fine-grained identification and delineation of each tooth in each quadrant, such as the workflow shown in Fig. 3.

Our fine-grained segmentation network SGANet can take any general FCNs as the backbone; being consistent with the quadrant-segmentation network, we still adopt 3D UNet in our implementation for simplicity. In contrast to general FCNs that do not explicitly consider the spatial associations between different targets under segmentation, our method attempts to model such anatomical topology of teeth via the proposed semantic graph attention (SGA) mechanism. Specifically, to



	B	T1	T2	T3	T4	T5	T6	T7	T8	S
B										
T1		1	1							
T2		1	1	1						
T3			1	1	1					
T4				1	1	1				
T5					1	1	1			
T6						1	1	1		
T7							1	1	1	
T8								1	1	
S										

Fig. 4. The tooth categories in ISO tooth numbering system. The teeth are evenly distributed in four quadrants, each quadrant includes eight teeth, and different colors indicate different categories (best viewed in color). The adjacency matrix is constructed by the tooth numbering system, and T1 to T8 denotes tooth category while B and S denotes background and stuff category, respectively.

learn discriminative voxel-wise feature embeddings for fine-grained teeth delineation, the SGA has two critical components: 1) A graph-based attention (GBA) module to minimize the confusion in differentiating between adjacent teeth with similar appearances; 2) a series of global-context attention (GCA) modules to more effectively combine local details with semantic feature embeddings in a segmentation-oriented fashion.

1) Graph-based Semantic Attention: Like the examples shown in Fig. 1, general FCNs may fail to precisely delineate the boundaries between adjacent teeth. One important reason could be that the semantic feature embeddings learned by the FCN encoder are typically unconscious of both the precise positions of each tooth and the spatial associations between different teeth. To improve the voxel-wise discriminative capacity of the segmentation encoder, our graph-based attention module first learns a high-level prototype representation for each class (i.e., different teeth and the background). Then, it further updates the relatively high-resolution encoder feature embeddings by aggregating the class-level prototype representations according to the corresponding semantic attentions.

Specifically, based on the International Standards Organization (ISO) dental standard (as shown in Fig. 4), each specific quadrant has a maximum of eight teeth under segmentation, including the central incisor, lateral incisor, canine, first premolar, second premolar, first molar, second molar, and third molar. Besides, since the localization of the quadrants could be rough in Section III-C, some teeth from other quadrants may also exist in the cropped images of the quadrant under segmentation, and they are referred to as the stuff category in this paper inspired by [40]. Therefore, there are ten possible classes for a quadrant, with eight having specific spatial associations (i.e., the teeth), while the other two are relatively isolated (i.e., the background and stuff category). Based on this observation, we define a $C \times C$ adjacency matrix $\mathbf{A}$ to describe the anatomical topology of the ten classes (i.e., $C = 10$). As shown in Fig. 4, $\mathbf{A}_{i,j} = 1$ if the ith and jth targets under segmentation are spatially adjacent; otherwise, $\mathbf{A}_{i,j} = 0$.

Let $\mathbf{X} \in \mathbb{R}^{W \times H \times D \times M}$ be the feature maps produced by the 3D UNet encoder, where $(W \times H \times D)$ and M denote the size of a feature map and the number of channels, respectively. As shown in Fig. 3, we enhance the discriminative capacity of $\mathbf{X}$ by leveraging the graph-based attention module in terms of $\mathbf{A}$. To this end, $\mathbf{X}$ is first mapped by two parallel $1 \times 1 \times 1$ convolutional layers with C and M channels, respectively. The corresponding outputs can be defined as

$$\mathbf{X}_{\text{class}} = \sigma(\varphi(\mathbf{X}; \mathbf{W}_C)); \ \mathbf{X}_{\text{hidden}} = \sigma(\varphi(\mathbf{X}; \mathbf{W}_M)), \tag{2}$$

where $\varphi(\cdot;\cdot)$ denotes a convolutional operation parameterized by $\mathbf{W}_C$ or $\mathbf{W}_M$, and $\sigma(\cdot)$ is an activation function. In Eq. (2), each channel of $\mathbf{X}_{\text{class}} \in \mathbb{R}^{W \times H \times D \times C}$ implies the possibility of $\mathbf{X}$'s elements belonging to a specific class. The tensor $\mathbf{X}_{\text{hidden}} \in \mathbb{R}^{W \times H \times D \times M}$ encodes $\mathbf{X}$'s hidden representations.

Based on Eq. (2), we can straightforwardly calculate the initial prototype representations of different anatomical targets (i.e., different teeth), denoted as $\mathbf{F}^0 \in \mathbb{R}^{C \times M}$, by aggregating the voxel-wise hidden representations $\mathbf{X}_{\text{hidden}}$ in terms of $\mathbf{X}_{\text{class}}$, such as

$$\mathbf{F}^0 = \widetilde{\mathbf{X}}_{\text{class}}^T \times \widetilde{\mathbf{X}}_{\text{hidden}}, \tag{3}$$

where $\widetilde{\mathbf{X}}_{\text{class}} \in \mathbb{R}^{WHD \times C}$ and $\widetilde{\mathbf{X}}_{\text{hidden}} \in \mathbb{R}^{WHD \times M}$ are reshaped from $\mathbf{X}_{\text{class}}$ and $\mathbf{X}_{\text{hidden}}$, respectively. That is, the initial prototype representation of each class can be regarded as the weighted average of all voxels' hidden representations.

Given the initial $\mathbf{F}^0$, we further adopt GCNs in terms of the adjacency matrix $\mathbf{A}$ to enhance the semantic information encoded in each prototype representation, by explicitly modeling the spatial associations between different anatomical targets. Specifically, each layer in the forward path of our GCN can be defined as

$$\mathbf{F}^i = \sigma(\mathbf{A}\mathbf{F}^{i-1}\mathbf{W}_i), \tag{4}$$

where $\mathbf{W}_i$ is the learnable weight matrix of the i-th GCN layer, and $\sigma(\cdot)$ is an activation function.

Let the output of the GCN be $\widetilde{\mathbf{F}} \in \mathbb{R}^{C \times M}$. We then update the feature embedding of the segmentation encoder as

$$\widetilde{\mathbf{X}}^* = \sigma_{\text{softmax}}(\widetilde{\mathbf{X}}_{\text{class}}) \times \widetilde{\mathbf{F}}, \tag{5}$$

where $\sigma_{\text{softmax}}(\cdot)$ denotes the softmax normalization across the channel. According to Eq. (5), the refined encoder feature representations are defined as the aggregation of different prototype representations in terms of the learned spatially varying attention coefficients. Finally, as the input of the segmentation decoder, $\widetilde{\mathbf{X}}^* \in \mathbb{R}^{WHD \times M}$ is reshaped back to the size of $W \times H \times D \times M$.

2) Global-Context Attentions: Based on the outputs of the encoder, our fine-grained segmentation network further designs an encoder (symmetric and skip-connected to the encoder) to hierarchically combine semantic information with enhanced multi-scale image details for the voxel-wise classification at the input image resolution.

To increase receptive fields of the network and more completely capture the global context of the input image, we insert a global-context attention module into each skip-connection part, as inspired by SE-Net [41]. To be more specific, we apply a 3D version of the modified SE module to the feature maps of each encoder stage to obtain the global contextual information, based on which the channel-wise attention is learned for the input feature maps. The global average pooling compresses the global spatial information into a channel descriptor and counts the channel information by the global average pooling, which can be defined as

$$\mathbf{F} = f_{\text{squeeze}}(\mathbf{X}) = \frac{1}{W \times H \times D} \sum_{i=1}^{W} \sum_{j=1}^{H} \sum_{k=1}^{D} \mathbf{X}(i, j, k). \quad (6)$$

Finally, the output features of each global-context attention module $\mathbf{X}_{gca}$can be defined as

$$\mathbf{X}_{gca} = \mathbf{X} + \sigma_{sigmoid}(\mathbf{F}'' \cdot \mathbf{X}), \quad (7)$$

where $\mathbf{F}''$ is the weighted vector and obtained by two consecutive convolutional layers, and '$\cdot$' denotes channel-wise multiplication. Such global-context attention modules jointly formulate a contextual-information-aware selection of multi-scale image details, which are important complementary of the SGA block for fine-grained teeth delineation.

IV. Experiments and Results

A. Datasets

We validate the proposed method on a dataset collected from dental clinics, which contains teeth missing, crowding, oblique, or metal artifacts. The dataset contains 350 in-house clinical dental CBCT scans collected and established by expert doctors from the Stomatological Hospital of Chongqing Medical University with a manually-annotated label. To be specific, three interns used ITK-SNAP [42] to perform pixel-level annotation on the tooth CBCT images, which were then reviewed by expert physicians, especially in the boundaries of the tooth roots. All dental CBCT images used in this paper are approved by the Institutional Review Board and analyzed anonymously. The voxel resolution of all dental CBCT image data varies from 0.25 mm × 0.25 mm × 0.25 mm to 0.4 mm × 0.4 mm × 0.4 mm, and the size of each slice is 400×400. The slice number varies from 280 to 328, the intensity value is between -1000 and 8000, and the distribution is uneven. To eliminate intensity outliers (e.g., points with extremely high or low gray values), we clip the gray level outside the range of 0.5% to 99.5% intensity histogram. Then all dental CBCT intensity values are normalized to $[0, 1]$ by the z-score method. To reduce the computation cost, we crop all dental CBCT images to a fixed size of $128 \times 128 \times 128$ using an overlapping sliding window strategy. In the training process, on-the-fly data augmentation by random translation and rotation is employed to increase the variability of different data representations.

B. Experimental Settings

1) Implementations: The training and testing phases of the proposed quadrant segmentation network and the tooth delineation network are implemented on 2 NVIDIA TESLA V100 GPUs based on PyTorch with an end-to-end fashion. We set the batch size and iteration step to 8 and 10000 to get the gradient update, respectively. The initial learning rate is set to $3e - 4$, and we decay the learning rate by a factor of 0.1 per 2500 steps. Finally, we choose Adam optimizer to minimize the loss function.

For the quadrant segmentation network, we convert the 32 types of tooth semantic labels into four quadrants in the training phase, and predict the four tooth quadrants and employ morphological operations to remove discrete segmentation results in the testing phase. We expand the maximum connected domain of each quadrant by 15 slices to ensure that all teeth in the current quadrant are included.

For the tooth delineation network, we train the SGANet to identify the 8 types of tooth labels in each quadrant, respectively. In the testing phase, the CBCT quadrant patches are obtained by the output guidance of the quadrant segmentation network. The cropped index is employed to recover the final segmentation results. Since the stuff category is used to identify teeth that are mis-included in (or do not belong to) a specific quadrant, for each quadrant, the segmentation of stuff category will be directly discarded. Finally, the four quadrant segmentation results are merged together by the cropped index.

We use the cross-entropy loss as a classification loss, and introduce multi-class soft Dice loss [46] to prompt the network to focus on the tooth area and denoted by

$$\mathcal{L}_{msdice} = 1 - \frac{2 \sum\limits_{i=1}^{N} \sum_{c=1}^{C} p_i^c g_i^c}{\sum_{i=1}^{N} \sum_{c=1}^{C} p_i^c + g_i^c}, \quad (8)$$

where N is decided by the number of voxels in the whole volume, C denotes the semantic classes, p_i^c and g_i^c depicts the i-th voxel prediction and ground truth class, respectively.

In the training process, a deep-supervised learning strategy is further employed to obtain a reliable graph representation produced by the graph-based attention module. To be specific, we upsample the reshaped prototype representations X' to the resolution of the input patch and use the loss functions as same as the final outputs like Eq. 8. Therefore, it is able to obtain a reliable graph representation produced by the graph-based attention module, the semantic image representation can improve the segmentation accuracy. Finally, the overall loss function $\mathcal{L}$ is defined as

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{ds}, \quad (9)$$

and $\mathcal{L}_{seg}$ is defined as

$$\mathcal{L}_{seg} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{msdice}, \quad (10)$$

among them, λ and α are weighting factors, and According to the experimental experience [26], we set both λ and α to 0.5.

2) Evaluation Metrics: Five metrics are utilized to evaluate the proposed tooth segmentation method.

Firstly, we use the Dice similarity coefficient (DSC), Jaccard similarity (JS), and Hausdorff distance (HD) as evaluation indicators to evaluate the segmentation performance. DSC and JS are functions for evaluating the similarity or overlap of two samples, defined as

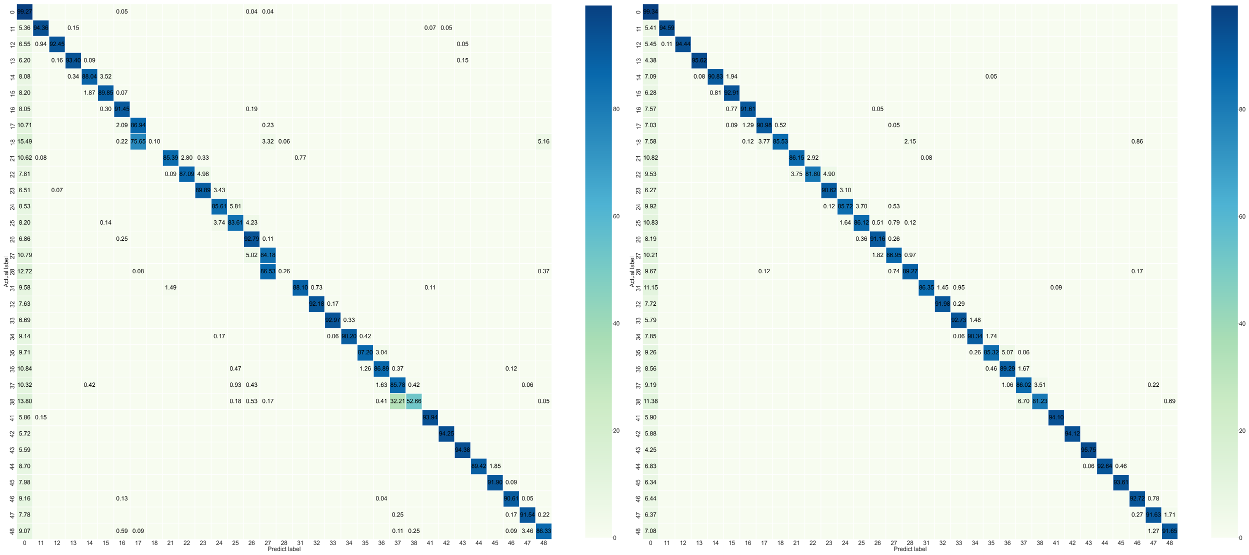
$$DSC = \frac{2|P \cap G|}{|P| + |G|}, JS = \frac{|P \cap G|}{|P \cup G|}, \quad (11)$$

where P indicates the predicted segmentation results and G denotes the ground truth labels. Hausdorff distance is depicted in Eq. 12, it evaluates the symmetrical distance between the

TABLE I

THE QUANTITATIVE FIVE-FOLD CROSS-VALIDATION SEGMENTATION RESULTS (MEAN ± STANDARD DEVIATION) COMPARISON WITH BOTH CNN AND GCN-BASED METHODS.

Methods		DSC	Jaccard	Precesion	Recall	HD (mm)
CNN-based	3D UNet [39]	88.14 ± 0.77	80.60 ± 1.01	90.34 ± 0.29	87.93 ± 1.32	4.90 ± 1.61
	attention UNet [43]	90.28 ± 0.67	83.45 ± 0.69	91.08 ± 1.08	90.59 ± 1.48	2.95 ± 1.33
	C2Fseg [17]	89.65 ± 1.12	82.5 ± 1.53	91.31 ± 0.54	89.13 ± 1.65	3.66 ± 1.85
	MWTNet [44]	90.05 ± 0.69	82.96 ± 0.84	91.23 ± 0.52	89.89 ± 1.23	2.41 ± 1.22
GCN-based	MGU-Net [45]	89.71 ± 0.73	82.49 ± 0.84	91.91 ± 0.17	88.62 ± 1.07	3.59 ± 0.94
	3D GCSN [20]	90.00 ± 0.39	82.84 ± 0.41	90.76 ± 0.76	90.31 ± 0.97	4.14 ± 1.32
	Ours	$\mathbf{91.13\pm0.45}$	$\mathbf{84.80\pm0.57}$	$\mathbf{92.13\pm0.29}$	$\mathbf{91.23\pm0.89}$	$\mathbf{1.00\pm0.27}$



(a) 3D UNet

(b) Ours

Fig. 5. The confusion matrix comparison of 3D UNet and the proposed method. 3D UNet is prone to misidentify teeth with similar shapes, especially identify the second and third molars, while the proposed method has better performance in identifying adjacent tooth classes with similar shapes.

tooth predictions P and the ground truth G, and $d(\cdot)$ represents the Euclidean distance.

$$d_{\mathrm{H}}(P,G)=\max\left\{\sup_{p\in P}\inf_{g\in G}d(p,g),\sup_{g\in G}\inf_{p\in P}d(p,g)\right\},\quad(12)$$

where the sup and inf denote supremum and infimum, respectively. The higher the value of H, the lower the matching degree of the two samples. We employ 95% HD to eliminate the impact of a very small subset of the outliers, which is based on the calculation of the 95-th percentile of the distances between boundary points in P and G.

Then we apply both Precision and Recall to evaluate the tooth voxel detection and classification performance respectively as same as [47].

C. Comparison with State-of-the-art Methods

1) Competing Methods: We compare the proposed method with five state-of-the-art methods, which contain CNN-based methods [17], [39], [43] and GCN-based segmentation methods [20], [45], and report the five-fold cross-validation results (mean±std) on the collected dental CBCT datasets. Note that we utilized the same network backbone in all methods for a fair comparison.

- 3D UNet [39] is replaced by our segmentation backbone network, and the input channel was set to 16.
- Attention UNet [43] is an attention gating (AG) model for medical imaging, which can automatically focus on target structures of different shapes and sizes.
- C2FSeg [17] is a coarse-to-fine segmentation network with two cascaded V-Net [46] for multi-class tooth segmentation.
- MWTNet [44] is a multi-task 3D fully convolutional network (FCN) with a post-processing marker-controlled watershed transform (MWT) for tooth segmentation.
- MGU-Net [45] is a multi-scale GCN-assisted segmentation network for Retinal OCT images.
- 3D GCSN [20] employs 3D graph convolutional segmentation network for multi-class segmentation of vertebrae.

2) Quantitative Results: Table I shows the quantitative comparison results. It can be seen that the proposed method outperforms the compared CNN-based and GCN-based methods, which demonstrates the effectiveness of the proposed method. Compared with the coarse-to-fine method C2Fseg [17] on the collected dataset, the proposed method outperforms approximately 1.48% of DSC, 2.3% of Jaccard index, 2.1% of Recall, and 2.66 mm of 95% Hausdorff distance, which is a significant improvement in the tooth segmentation task.

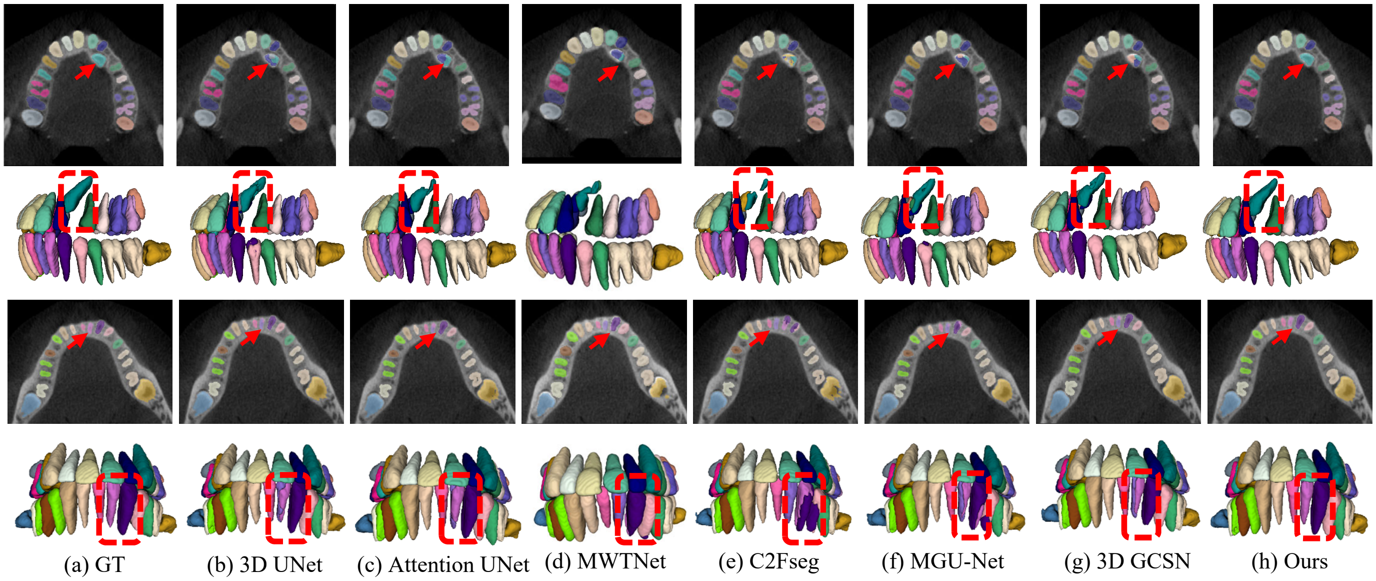


Fig. 6. Qualitative comparison with state-of-the-art methods. The first and second row shows the challenging canine (dislocation and oblique growth) segmentation results and corresponding 3D visualization results. The third and fourth row shows the segmentation results when missing the central incisor. In the competing methods, the missing tooth category is added to its adjacent teeth (i.e., one tooth is marked in 2 colors), and our method maintains accurate identification and segmentation results.

We attribute this to that our quadrant segmentation network can effectively avoid the tooth intra-class similarity between different quadrants. Compared with the GCN-based methods MGU-Net and 3D GCSN [20], [45], the Hausdorff distance of the proposed method is reduced from 4.14 mm to 1 mm. MGU-Net uses a learnable adjacency matrix for global contextual information reasoning, and we attribute that the learnable adjacency matrix cannot accurately describe the relationship between the teeth arrangement. 3D GCSN constructs the adjacency matrix for all categories at the same time in the vertebra segmentation task since the vertebra arrangement has consistency. However, due to the tooth crowding and oblique, the tooth arrangement between the four quadrants may be different. Consequently, we not only construct the adjacency matrix in each quadrant but also employ a global-context attention module to extract distinguishing features between tooth categories. So the proposed method achieves better segmentation performance and maintains a reasonable tooth boundary. To further evaluate the effectiveness of tooth identification, we compare the confusion matrix of 3D UNet and the proposed method in Fig. 5. From the confusion matrix, we can see that the proposed method has a better performance in tooth category identification, while 3D UNet has large mistakes in recognizing adjacent teeth with similar shapes (i.e., the second and third molars).

3) Qualitative Results: Fig. 6 shows some qualitative challenging case segmentation results. When the teeth are dislocated and grown oblique, we observe that the CNN-based method cannot segment the tooth root completely. Although the GCN-based method can segment the entire tooth, it is hard to give the tooth semantic label accurately, and the proposed method can maintain the tooth shape and label precisely. We attribute this to the proposed method reducing the similarity between tooth categories in a two-stage manner while 3D GCSN encodes the correlation between categories simultaneously. In addition, we show the segmentation results when missing the lower central incisor. We can see that the proposed method would not introduce tooth misclassification while the other methods are labeling the same tooth as multiple tooth categories. We attribute these results to the fact that the semantic graph attention mechanism constructs an exemplary graph representation and explicitly models anatomical association between different teeth, therefore avoiding misclassification caused by the similarity in tooth shapes.

D. Ablation study

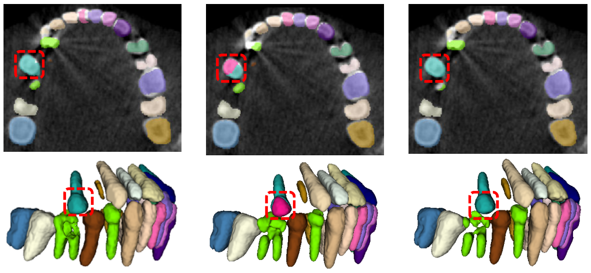
We conduct the ablation study on the proposed method to evaluate the effects of each novel component. In Table II, we present segmentation results of different configurations: (1) BNet is the segmentation backbone network; (2) BNet-GCA denotes adding the global-context attention module into BNet; (3) BNet-GBA/g represents adding the graph-based attention module with removing the graph convolution network; (4) BNet-GBA indicates adding the complete graph-based attention module into BNet; (5) BNet-GBA-s is adding the graph-based attention with the stuff category embedding into BNet; (6) FullNet indicates the combination of GCA module and GBA module with the stuff category embedding.

1) Effects of GCA: To validate the effectiveness of our introduced global-context attention module, we insert this block into each stage of the baseline encoder and show the quantitative results. It can be found that after adding the GCA module, the dice metric is 90.44% while the Jaccard Similarity is 83.28%, which shows more than 2% gains compared with BNet, and the 95% Hausdorff distance is 2.83 mm, which reduces the error by more than 2 mm. We attribute that the

TABLE II

ABLATION STUDY OF THE EFFECTS OF DIFFERENT MODULES. BNET DENOTES THE BASELINE NETWORK.

Methods	DSC	Jaccard	Precision	Recall	HD (mm)
BNet	88.14	80.60	90.34	87.93	4.90
BNet-GCA	90.44	83.28	90.56	91.18	2.83
BNet-GBA/g	90.56	83.47	90.69	91.24	3.73
BNet-GBA	90.79	83.77	90.91	91.82	2.79
BNet-GBA-s	90.91	84.29	91.35	91.97	1.84
FullNet	**91.13**	**84.80**	**92.13**	**91.23**	**1.00**



(a) GT (b) Without stuff (c) With stuff

Fig. 7. The effect of the stuff category on tooth segmentation. The stuff category can help reduce tooth labeling errors especially at the occluded areas.

GCA module can extract the global contextual information in both shallow and deep layers, and it pays more attention to the tooth semantic categories.

2) Effects of GBA: In Table II, it can be found that BNet-GBA/g obtain 90.56% of DSC, 83.47% of Jaccard Similarity, and the 95% Hausdorff distance is 3.73 mm, which shows incremental performance compared with BNet. Further, BNet-GBA verifies the effectiveness of graph convolution. We can see that the overall performance has been improved after adding graph convolution, which indicates that graph convolution is an essential part of the GBA module. In addition, the 95% Hausdorff distance is reduced from 3.73 mm to 2.79 mm, and this indicates the graph convolution in the GBA module is able to refine boundary segmentation results. From Table II, we obviously see that the GBA module is demonstrated to be superior in extracting the local spatial information in the deep semantic features and yet concurrently maintaining the interdependence between the teeth anatomical association in the graph space. BNet-GBA-s indicates that the introduced stuff category has further improvement. Fig. 7 shows that the stuff category helps improve the accuracy of tooth identification, especially at the occluded areas. Even if there is an error in the prediction of the tooth region in the first stage, the error can be corrected by the introduced stuff category in the fine segmentation stage. We attribute this to that the introduced stuff category is able to improve the discrimination ability of the current quadrant teeth. Furthermore, we can see that the FullNet obtain the best performance.

3) Effects of the two-stage training strategy: We compare the training and testing time with both one-stage and two-stage methods in the Table III. The training time is counted based on 10,000 iterations of different models, and the testing time is calculated by the average prediction time of whole testing cases. As listed in Table III, it highlights that the one-stage method uses less training time. Among the two-stage segmentation methods, the proposed method has a shorter training time than C2FSeg, which we attribute to the residual convolution block in C2FSeg increasing the computational overhead. In the test phase, the average testing time of each CBCT image in our method is 4.04 seconds, which is faster than C2FSeg. Although the inference time is slower than the 3D UNet, the segmentation accuracy is significantly outperformed by this method.

TABLE III

TRAINING AND TESTING TIME COMPARISON.

Method	Category	DSC	Training (h)	Testing (s)
3D UNet [39]	One-stage	88.14	18.7	**3.76**
C2FSeg [17]	Two-stage	89.65	28.7	6.27
Ours		**91.13**	20.2	4.04

We remove the first stage quadrant segmentation network and directly construct the adjacency matrix by all tooth categories the effects of our two-stage training strategy. Table IV indicates that modeling all tooth categories together is not as effective as modeling each quadrant's teeth individually. We attribute this to the inconsistent adjacency relationship when modeling all teeth.

TABLE IV

QUANTITATIVE COMPARISON OF THE EFFECTS OF TWO-STAGE TRAINING STRATEGY.

Methods	DSC	Jaccard	Precesion	Recall	HD (mm)
One-stage	90.59	83.09	91.76	90.64	3.24
Two-stage	**91.13**	**84.80**	**92.13**	**91.23**	**1.00**

TABLE V

QUANTITATIVE COMPARISON OF THE DIFFERENT INPUT RESOLUTION SETTINGS.

Input resolution	DSC	Jaccard	Precision	Recall	HD (mm)
64×64×64	89.16	81.79	88.91	90.48	2.86
96×96×96	90.86	84.31	91.88	90.82	2.29
128×128×128	**91.13**	**84.80**	92.13	**91.23**	**1.00**
144×144×144	91.04	84.71	**93.45**	89.71	1.10

4) Effects of the input resolution: To further check the effects of input image resolution on the segmentation performance, we evaluate the performance changes in terms of different patch sizes. Specifically, the input size was isotropically set to 64, 96, 128, and 144, and the corresponding results are summarized in Table V. From the table, we can see that when the input resolution increases from 64 to 128, the segmentation performance improves monotonously, while a larger input size (i.e., 144) does not bring significant change. The reason behind this could be that the 128×128×128 image patches are big enough to capture the global tooth information in each quadrant, while larger inputs can only lead to additional computational overhead.

V. DISCUSSION

1) Analysis of different quadrants: We further analyzed the performance of tooth segmentation in different quadrants.

Fig. 8 shows the quantitative results of our method on different tooth categories in the four quadrants. It can be seen from the figure that the segmentation performance of teeth in the first quadrant and the second quadrant (upper teeth) is generally better than that in the third quadrant and the fourth quadrant (lower teeth). To be specific, the upper central incisors (11, 21) and the lateral incisor (12, 22) show better segmentation results than the lower central incisors (31, 41) and the lateral incisor (32, 42), respectively. This is due to the upper incisors having a larger size than the lower incisors, and the lower incisors are more prone to be crowded and disordered. In addition, we found that the third molar in each quadrant showed lower segmentation performance. This is because the shape of the third molar varies greatly between different patients, and a high-class imbalance problem exists in this kind of tooth category.

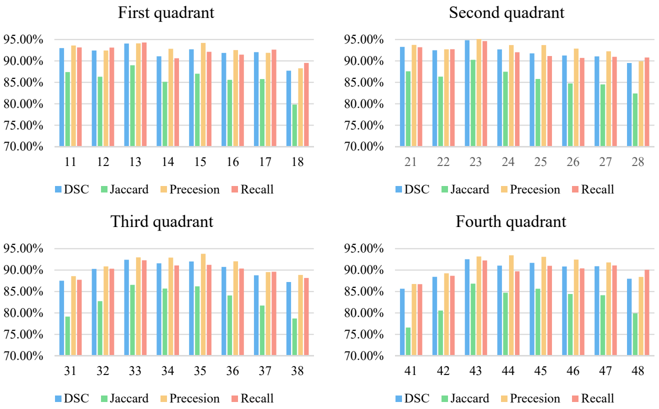


Fig. 8. The segmentation results under different quadrants.

2) Relation to self-attention mechanisms: The proposed method can be regarded as a further extension based on the spatial and category attention mechanism, which has the ability to explore the anatomical association information to explicitly model medical organs. To further demonstrate the effects of our semantic graph attention mechanism, we have compared it with the currently widely used methods based on self-attention mechanisms [48], [49], as Table VI listed. It can be found that the segmentation results of the proposed method are better than those based on the non-local mechanism [48] and the dual attention channel-spatial mechanism [49]. We consider the adjacency relationship helps to improve the ability to discriminate the tooth semantic information.

3) Effects of metal artifacts and missing teeth: We further verify the effects of metal artifacts and missing teeth cases on the proposed method. Table VII indicates that the proposed method can still maintain good segmentation performance when segmenting challenging metal artifacts and missing teeth cases.

Furthermore, Fig. 9 shows the quantitative results, and we can see that when the CBCT images are scanned with metal artifacts, 3D UNet [39] can not effectively identify the teeth affected by the artifacts. The proposed method has the potential to achieve better segmentation results and has the capability to identify accurate teeth boundaries even if there are numerous metal artifacts, and this can be attributed to the global-context module to update the voxel relationship by capturing the global contextual information. In addition, when the surrounding teeth are missing, 3D UNet easily introduces the missing tooth category into the isolated tooth, however, the proposed method introduces the anatomical association information of the tooth category, and the adjacency matrix is defined according to the global tooth categories, which removes local position restriction. Thus it remains robust even when teeth are missing.

TABLE VI
COMPARISON WITH THE SELF-ATTENTION MECHANISMS.

Methods	DSC	Jaccard	Precesion	Recall	HD (mm)
BNet	88.14	80.60	90.34	87.93	4.90
Non-local [48]	90.40	83.55	90.07	**91.83**	2.59
CBAM [49]	90.36	83.57	90.55	91.19	2.28
Ours	**91.13**	**84.80**	**92.13**	91.23	**1.00**

TABLE VII
QUANTITATIVE COMPARISON IN THE CASE OF METAL ARTIFACTS AND MISSING TEETH.

Methods	DSC	Jaccard	Precesion	Recall	HD (mm)
3D UNet [39]	87.59	79.09	89.20	87.64	5.24
Ours	**90.54**	**83.33**	**91.28**	**90.80**	**2.10**

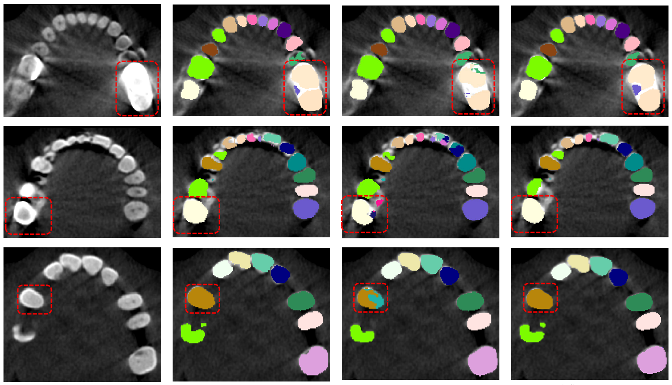
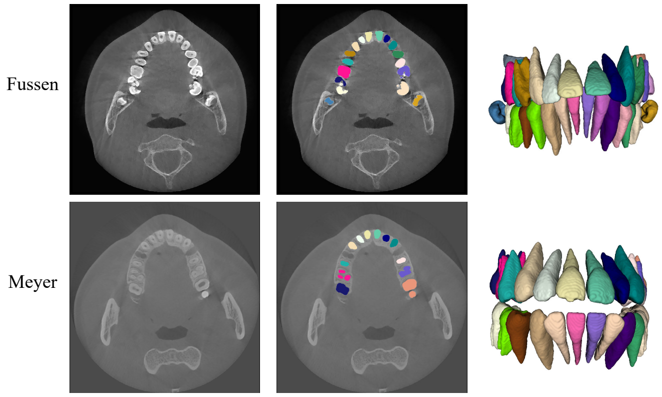

(a) Original image (b) Ground Truth (c) 3D UNet (d) Ours

Fig. 9. The segmentation results under metal artifacts and missing teeth case (denote by red boxes). The first and second row shows the effects of more or fewer artifacts, and the third row shows the effects of missing teeth.

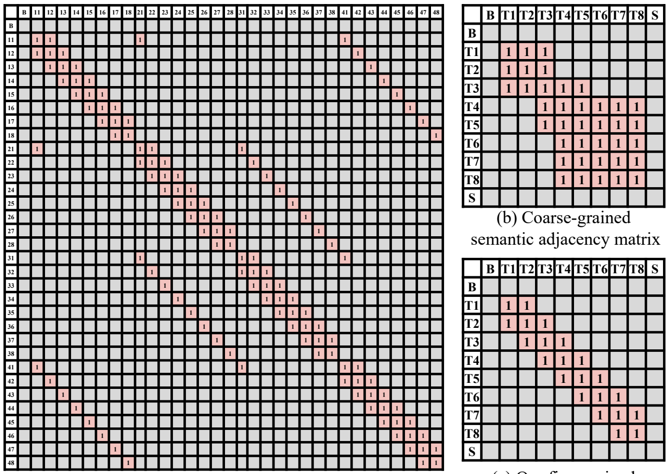
4) Generalization on different imaging protocols: The results presented in Table I show the state-of-the-art performance of our method in segmenting teeth with varying conditions on this dataset. To further check the generalizability of our method across different imaging protocols, we conduct an additional experiment to apply our method to a set of 37 dental CBCT images acquired by other manufacturers, i.e., Fussen Technology Co., Ltd. (voxel resolution: 0.25 mm; intensity range: -3200 to 4200) and Hefei Meyer Optoelectronic Technology Inc. (voxel resolution: 0.275 mm; intensity range: -1100 to 2600). Fig. 10 shows the qualitative segmentation results, we can see that the proposed method also led to competitive results in this newly dataset. It suggests the promising generalizability of our method in CBCT tooth segmentation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2022.3179128, IEEE Transactions on Medical Imaging



(a) Original image (b) 2D segmentation (c) 3D Reconstruction

Fig. 10. External image segmentation results from two different image protocols, the first and second rows show the segmentation results on the Fussen manufacture and the Meyer manufacture, respectively.



	B	11	12	13	14	15	16	17	18	21	22	23	24	25	26	27	28	31	32	33	34	35	36	37	38	41	42	43	44	45	46	47	48
B																																	
11		1	1							1																1							
12		1	1	1																							1						
13			1	1	1																							1					
14				1	1	1																							1				
15					1	1	1																							1			
16						1	1	1																							1		
17							1	1	1																							1	
18								1	1																								1
21		1								1	1							1															
22										1	1	1							1														
23											1	1	1							1													
24												1	1	1							1												
25													1	1	1							1											
26														1	1	1							1										
27															1	1	1							1									
28																1	1								1								
31										1								1	1							1							
32											1							1	1	1													
33												1							1	1	1												
34													1							1	1	1											
35														1							1	1	1										
36															1							1	1	1									
37																1							1	1	1								
38																	1							1	1								
41		1																1								1	1						
42			1																							1	1	1					
43				1																							1	1	1				
44					1																							1	1	1			
45						1																							1	1	1		
46							1																							1	1	1	
47								1																							1	1	1
48									1																							1	1

(a) Tooth instance adjacency matrix

	B	T1	T2	T3	T4	T5	T6	T7	T8	S
B										
T1		1	1	1						
T2		1	1	1						
T3		1	1	1	1	1				
T4				1	1	1	1	1	1	
T5				1	1	1	1	1	1	
T6					1	1	1	1	1	
T7					1	1	1	1	1	
T8					1	1	1	1	1	
S										

(b) Coarse-grained semantic adjacency matrix

	B	T1	T2	T3	T4	T5	T6	T7	T8	S
B										
T1		1	1							
T2		1	1	1						
T3			1	1	1					
T4				1	1	1				
T5					1	1	1			
T6						1	1	1		
T7							1	1	1	
T8								1	1	
S										

(c) Our fine-grained semantic adjacency matrix

Fig. 11. The different constructed adjacency matrix based on the ISO tooth numbering system.

5) Analysis of different adjacency matrix settings: We further compare the segmentation performance in terms of different adjacency matrix settings. Specifically, three different conditions shown in Fig. 11 have been considered in our newly added experiment: i) The adjacency matrix based on instance labels, i.e., for all 32 teeth defined according to the ISO tooth numbering system order. In this adjacency matrix, except for the third molars which have only 3 adjacent teeth, all remaining teeth have four adjacent teeth; ii) The coarse-grained adjacency matrix based on 8 semantic labels, i.e., 2 incisor categories (central incisor T1, and lateral incisor T2), 1 canine category (canine T3), and five molar categories (premolar and molar T4, T5, T6, T7, T8); iii) The fine-grained adjacency matrix based on 8 semantic labels, i.e., that adopted in our all previous experiments. From Table VIII, we can see the constructed tooth instance adjacency matrix with four adjacent teeth and fine-grained semantic adjacency matrix with three teeth are better than coarse-grained semantic adjacency matrix with six adjacent teeth. In future work, we will try to find more adaptive ways to define the tooth spatial relationships.

TABLE VIII
QUANTITATIVE RESULTS COMPARISON ON DIFFERENT ADJACENCY MATRIX SETTINGS.

Method	DSC	Jaccard	Precision	Recall	HD (mm)
Instance adjacency	90.44	83.28	90.56	91.18	2.83
Coarse-grained semantic adjacency	89.33	81.70	92.74	87.11	2.70
Ours fine-grained semantic adjacency	**91.13**	**84.80**	**92.13**	**91.23**	**1.00**

VI. CONCLUSION

This paper proposes an accurate tooth segmentation method based on a semantic graph attention mechanism. It contains a graph-based attention module and a global-context attention module to explicitly model the anatomical topology of the teeth in each quadrant, based on which voxel-wise discriminative feature embeddings are learned for the accurate delineation of teeth boundaries. Experiments on the clinic dental CBCT dataset show the superiority of the proposed method compared with state-of-the-art methods, which allow the proposed methods to improve the intelligence level of dental CAD systems. In future work, we will further improve the current method to better deal with challenging cases with metal braces and metal implants.

REFERENCES

[1] L. Wang, Y. Gao, F. Shi, G. Li, K.-C. Chen, Z. Tang, J. J. Xia, and D. Shen, "Automated segmentation of dental cbct image with prior-guided sequential random forests," *Medical physics*, vol. 43, no. 1, pp. 336–346, 2016.

[2] A. Z. Arifin, E. Tanuwijaya, B. Nugroho, A. M. Priyatno, R. Indraswari, E. R. Astuti, and D. A. Navastara, "Automatic image slice marking propagation on segmentation of dental cbct," *TELKOMNIKA*, vol. 17, no. 6, pp. 3218–3225, 2019.

[3] R. Pauwels, R. Jacobs, H. Bosmans, P. Pittayapat, P. Kosalagood, O. Silkosessak, and S. Panmekiate, "Automated implant segmentation in cone-beam ct using edge detection and particle counting," *International journal of computer assisted radiology and surgery*, vol. 9, no. 4, pp. 733–743, 2014.

[4] I.-B. Pavaloiu, N. Goga, A. Vasilateanu, I. Marin, A. Ungar, I. Patrascu, and C. Ilie, "Neural network based edge detection for cbct segmentation," in *2015 E-Health and Bioengineering Conference (EHB)*, pp. 1–4, IEEE, 2015.

[5] J. Michetti, A. Basarab, F. Diemer, and D. Kouame, "Comparison of an adaptive local thresholding method on cbct and μct endodontic images," *Physics in Medicine & Biology*, vol. 63, no. 1, p. 015020, 2017.

[6] Y. Fan, R. Beare, H. Matthews, P. Schneider, N. Kilpatrick, J. Clement, P. Claes, A. Penington, and C. Adamson, "Marker-based watershed transform method for fully automatic mandibular segmentation from cbct images," *Dentomaxillofacial Radiology*, vol. 48, no. 2, p. 20180261, 2019.

[7] Y. Jiang, J. Qian, S. Lu, Y. Tao, J. Lin, and H. Lin, "Lrvrg: a local region-based variational region growing algorithm for fast mandible segmentation from cbct images," *Oral Radiology*, pp. 1–10, 2021.

[8] M. Hosntalab, R. A. Zoroofi, A. A. Tehrani-Fard, and G. Shirani, "Segmentation of teeth in ct volumetric dataset by panoramic projection and variational level set," *International Journal of Computer Assisted Radiology and Surgery*, vol. 3, no. 3-4, pp. 257–265, 2008.

[9] H. Gao and O. Chae, "Individual tooth segmentation from ct images using level set method with shape and intensity prior," *Pattern Recognition*, vol. 43, no. 7, pp. 2406–2417, 2010.

[10] Y. Gan, Z. Xia, J. Xiong, Q. Zhao, Y. Hu, and J. Zhang, "Toward accurate tooth segmentation from computed tomography images using a hybrid level set model," *Medical physics*, vol. 42, no. 1, pp. 14–27, 2015.

[11] Z. Xia, Y. Gan, L. Chang, J. Xiong, and Q. Zhao, "Individual tooth segmentation from ct images scanned with contacts of maxillary and mandible teeth," *Computer methods and programs in biomedicine*, vol. 138, pp. 1–12, 2017.

[12] Y. Wang, S. Liu, G. Wang, and Y. Liu, "Accurate tooth segmentation with improved hybrid active contour model," *Physics in Medicine & Biology*, vol. 64, no. 1, p. 015012, 2018.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[14] Z. Cui, C. Li, and W. Wang, "Toothnet: Automatic tooth instance segmentation and identification from cone beam ct images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6368–6377, 2019.

[15] 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 *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 939–942, IEEE, 2020.

[16] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2021. DOI: 10.1109/TPAMI.2021.3086072.

[17] M. Ezhov, A. Zakirov, and M. Gusarev, "Coarse-to-fine volumetric segmentation of teeth in cone-beam ct," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 52–56, IEEE, 2019.

[18] G. Li, M. Muller, A. Thabet, and B. Ghanem, "Deepgcns: Can gcns go as deep as cnns?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9267–9276, 2019.

[19] H. Chang, S. Zhao, H. Zheng, Y. Chen, and S. Li, "Multi-vertebrae segmentation from arbitrary spine mr images under global view," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 702–711, Springer, 2020.

[20] S. Pang, C. Pang, L. Zhao, Y. Chen, Z. Su, Y. Zhou, M. Huang, W. Yang, H. Lu, and Q. Feng, "Spineparsenet: Spine parsing for volumetric mr image by a two-stage segmentation framework with semantic image representation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 262–273, 2020.

[21] H. C. Kang, C. Choi, J. Shin, J. Lee, and Y.-G. Shin, "Fast and accurate semiautomatic segmentation of individual teeth from dental ct images," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[22] H. Akhoondali, R. Zoroofi, and G. Shirani, "Rapid automatic segmentation and visualization of teeth in ct-scan data," *Journal of Applied Sciences*, vol. 9, no. 11, pp. 2031–2044, 2009.

[23] B. Jiang, Y. Zhang, X. Tang, and H. Shi, "Region growing model with edge restrictions for multiple roots tooth segmentation," in *Proceedings of the Third International Symposium on Image Computing and Digital Medicine*, pp. 171–174, 2019.

[24] L. Hiew, S. Ong, K. W. Foong, and C. Weng, "Tooth segmentation from cone-beam ct using graph cut," in *Proceedings of the Second APSIPA Annual Summit and Conference*, pp. 272–275, ASC, Singapore, 2010.

[25] Y. Gan, Z. Xia, J. Xiong, G. Li, and Q. Zhao, "Tooth and alveolar bone segmentation from dental computed tomography images," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 196–204, 2017.

[26] Y. Zhao, P. Li, C. Gao, Y. Liu, Q. Chen, F. Yang, and D. Meng, "Tsasnet: Tooth segmentation on dental panoramic x-ray images by two-stage attention segmentation network," *Knowledge-Based Systems*, vol. 206, p. 106338, 2020.

[27] Q. Chen, Y. Zhao, Y. Liu, Y. Sun, C. Yang, P. Li, L. Zhang, and C. Gao, "Mslpnet: multi-scale location perception network for dental panoramic x-ray image segmentation," *Neural Computing and Applications*, pp. 1–15, 2021.

[28] Z. Wang, C. Liu, D. Cheng, L. Wang, X. Yang, and K.-T. Cheng, "Automated detection of clinically significant prostate cancer in mp-mri images based on an end-to-end deep neural network," *IEEE transactions on medical imaging*, vol. 37, no. 5, pp. 1127–1139, 2018.

[29] Y. Zhang, H. Li, J. Du, J. Qin, T. Wang, Y. Chen, B. Liu, W. Gao, G. Ma, and B. Lei, "3d multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification," *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1618–1631, 2021.

[30] J. He, Y. Yang, Y. Wang, D. Zeng, Z. Bian, H. Zhang, J. Sun, Z. Xu, and J. Ma, "Optimizing a parameterized plug-and-play admm for iterative low-dose ct reconstruction," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 371–382, 2018.

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[32] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[33] Y. Liu, F. Zhang, Q. Zhang, S. Wang, Y. Wang, and Y. Yu, "Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3812–3822, 2020.

[34] Z. Tian, X. Li, Y. Zheng, Z. Chen, Z. Shi, L. Liu, and B. Fei, "Graph-convolutional-network-based interactive prostate segmentation in mr images," *Medical physics*, vol. 47, no. 9, pp. 4164–4176, 2020.

[35] D. Sun, Y. Pei, P. Li, G. Song, Y. Guo, H. Zha, and T. Xu, "Automatic tooth segmentation and dense correspondence of 3d dental model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 703–712, Springer, 2020.

[36] Q. Ma, G. Wei, Y. Zhou, X. Pan, S. Xin, and W. Wang, "Srf-net: Spatial relationship feature network for tooth point cloud classification," in *Computer Graphics Forum*, vol. 39, pp. 267–277, Wiley Online Library, 2020.

[37] C. Lian, L. Wang, T.-H. Wu, F. Wang, P.-T. Yap, C.-C. Ko, and D. Shen, "Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3d intraoral scanners," *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2440–2450, 2020.

[38] L. Zhang, Y. Zhao, D. Meng, Z. Cui, C. Gao, X. Gao, C. Lian, and D. Shen, "Tsgcnet: Discriminative geometric feature learning with two-stream graph convolutional network for 3d dental model segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6699–6708, 2021.

[39] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, pp. 424–432, Springer, 2016.

[40] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[42] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[43] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2018.

[44] 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.

[45] J. Li, P. Jin, J. Zhu, H. Zou, X. Xu, M. Tang, M. Zhou, Y. Gan, J. He, Y. Ling, *et al.*, "Multi-scale gcn-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary oct images," *Biomedical Optics Express*, vol. 12, no. 4, pp. 2204–2220, 2021.

[46] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, IEEE, 2016.

[47] J.-H. Lee, S.-S. Han, Y. H. Kim, C. Lee, and I. Kim, "Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs," *Oral surgery, oral medicine, oral pathology and oral radiology*, vol. 129, no. 6, pp. 635–642, 2020.

[48] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

[49] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.