NKUT: Dataset and Benchmark for Pediatric Mandibular Wisdom Teeth Segmentation

Zhenhuan Zhou, Yuzhu Chen, Along He, Xitao Que, Kai Wang, Rui Yao, and Tao Li

***Abstract*—Germectomy is a common surgery in pediatric dentistry to prevent the potential dangers caused by impacted mandibular wisdom teeth. Segmentation of mandibular wisdom teeth is a crucial step in surgery planning. However, manually segmenting teeth and bones from 3D volumes is time-consuming and may cause delays in treatment. Deep learning based medical image segmentation methods have demonstrated the potential to reduce the burden of manual annotations, but they still require a lot of well-annotated data for training. In this paper, we initially curated a Cone Beam Computed Tomography (CBCT) dataset, NKUT, for the segmentation of pediatric mandibular wisdom teeth. This marks the first publicly available dataset in this domain. Second, we propose a semantic separation scale-specific feature fusion network named WTNet, which introduces two branches to address the teeth and bones segmentation tasks. In WTNet, We design a Input Enhancement (IE) block and a Teeth-Bones Feature Separation (TBFS) block to solve the feature confusions and semantic-blur problems in our task. Experimental results suggest that WTNet performs better on NKUT compared to previous state-of-the-art segmentation methods (such as TransUnet), with a maximum DSC lead of nearly 16%.**

***Index Terms*—CBCT dataset, pediatric wisdom teeth segmentation, pediatric germectomy, multi-scale feature fusion.**

I. Introduction

MANDIBULAR wisdom teeth (MWT) typically erupt in the mouth between the ages of 17 and 24. Due to limited growing space, they have a high probability of developing into horizontally impacted teeth. Such teeth are often associated with oral pathological changes such as anterior crowding, periodontal diseases, and even damage to adjacent teeth [1]. Removal is generally considered the most effective solution for impacted MWT, which may cause pain or lead to pathological changes. However, the removal of mature MWT may be associated with some short-term postoperative complications, such as alveolar osteitis and pain [2], which may significantly affect a patient's daily life. Germectomy is a frequently performed surgery in pediatric dentistry to avoid complications caused by MWT. It involves the extraction of incipient MWT germs when their crowns and roots have not yet completed development [3]. However, the implementation of germectomy depends heavily on the experience of dentists and can be highly subjective. Therefore, it is promising to adopt computer-aided diagnosis (CAD) methods to develop a more objective surgical plan with minimal trauma. Pediatric MWT segmentation is a prerequisite for building such a system.

Recently, AI methods based on traditional machine learning and deep learning have achieved certain success in the field of dental computer-aided diagnosis. Wu et al. [4] introduced a four-stage method based on dental panoramic radiographs to help dentists obtain a reliable assessment of parameters for orthodontic evaluation. Reference [5] proposed a novel framework called Dental Diagnosis System for dental diagnosis based on the hybrid approach of segmentation, classification and decision making. This work was tested under a real dental case of Hanoi Medical University and achieved good performance. Deep learning [6] based methods have also made considerable progress in teeth analysis and segmentation tasks. For example, Tian et al. [7] developed a computer-aided Deep Adversarial-driven dental Inlay reStoration (DAIS) framework for the automatic reconstruction of a realistic surface for a damaged tooth, which providing higher clinical applicability. Lai et al. [8] introduced an approach employing deep convolutional neural networks to aid in human identification through the automatic and precise matching of 2-D panoramic dental X-ray images. Rajee et al. [9] proposed an algorithm to estimate the gender of human from dental x-ray image (DXI). The proposed method was trained and tested on a dataset consisting of 1000 DXI images, and the results indicate that the performance of the proposed method has achieved better outcomes. Tian et al. [10] introduced a gingival margin line reconstruction (GMLR) framework driven by a deep adversarial network to automatically acquire the personalized gingival contour for a partially edentulous patient. Qiu et al. [11] presented a low-cost annotation way and a dental

Manuscript received 20 November 2023; revised 29 February 2024; accepted 22 March 2024. Date of publication 1 April 2024; date of current version 6 June 2024. This work was supported in part by the National Natural Science Foundation under Grant 62272248, in part by the Open Project Fund of State Key Laboratory of Computer Architecture, in part by the Institute of Computing Technology, Chinese Academy of Sciences under Grant CARCHA202108, in part by the Natural Science Foundation of Tianjin of China under Grant 21JCZDJC00740 and Grant 21JCYBJC00760. *(Corresponding authors: Tao Li; Rui Yao.)*

Zhenhuan Zhou, Along He, and Kai Wang are with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: bdor25@163.com; healong2020@163.com; wangk@nankai.edu.cn).

Yuzhu Chen is with the School of Medicine, Nankai University, Tianjin 300350, China (e-mail: 2110214@mail.nankai.edu.cn).

Xitao Que and Rui Yao are with the Department of Pediatric Dentistry, Tianjin Stomatological Hospital, Tianjin 300041, China (e-mail: quexitao@163.com; yaorui73@163.com).

Tao Li is with the College of Computer Science, Nankai University, Tianjin 300350, China, and also with the Haihe Lab of ITAI, Tianjin 300459, China (e-mail: litao@nankai.edu.cn).

Dataset and codes will be released at https://github.com/nkicsl/NKUT.

Digital Object Identifier 10.1109/JBHI.2024.3383222

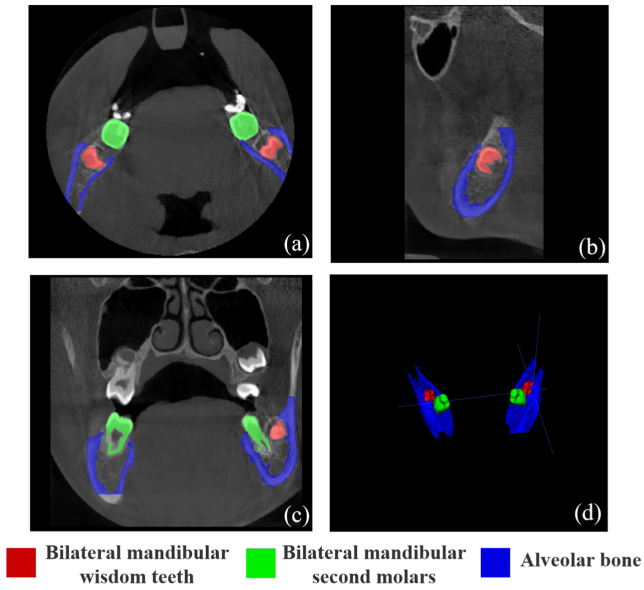

Fig. 1. Images and annotations from the proposed NKUT dataset. (a), (b) and (c) denote the axial, coronal, and sagittal views respectively, and (d) represents the 3D rendering results of annotations via ITK-Snap.

arch prior-assisted method for 3D tooth instance segmentation, Zhang et al. [12] explored a two-stream graph convolutional network to learn multi-view geometric information in 3D dental models. Jang et al. [13] proposed a hierarchical multi-step model to identify and segment 3D individual teeth from dental CBCT images. These methods have achieved good results, further demonstrated the significant potential of using deep learning methods in pediatric MWT segmentation tasks. However, there are still several limitations when applying these methods to pediatric MWT segmentation task. Firstly, most previous works were trained and evaluated on in-house datasets and mainly focused on the analysis and segmentation of adult teeth rather than children. Secondly, the majority of previous works did not specifically focus on the analysis of wisdom teeth, and almost no research dedicated to addressing the significant multi-scale problems faced by segmenting pediatric MWT germs, second molars and alveolar bone. Therefore, a publicly available dataset with high-quality expert annotations and a effective model for the multi-scale challenges in pediatric MWT segmentation tasks are urgently needed for the researches of CAD on pediatric germectomy.

To address these issues, in this paper we first constructed a CBCT dataset called NKUT, specifically for the pediatric MWT segmentation task. The dataset contains 133 CBCT volumes encompassing more than 53,000 slices. The patients range in age from 7 to 22 years old, with an average age of 13.2 years. All scans in NKUT were manually labeled in great detail by pediatric experts and covering three different pixel-level annotations: bilateral MWT germs, bilateral mandibular second molars (SM) and partial bilateral mandibular alveolar bones (AB), as shown in Fig. 1. The eruption trend of MWT and their positional distribution relative to the surrounding AB are very important for the clinical diagnosis of pediatric germectomy. By segmenting and reconstructing the annotated area mentioned above, pediatric dentists can objectively determine the risk of MWT germs developing into horizontally impacted wisdom teeth, while assess the necessity, optimal timing and precise locations for pediatric germectomy procedures. We will provide a detailed introduction to our NKUT dataset in Section III.

The pediatric MWT segmentation task presents serious semantic confusion and multi-scale challenges. On the one hand, children in the mixed dentition stage may have not only MWT germs but also some unerupted permanent teeth germs in their alveolar bone. These germs are very similar in shape, which can easily cause semantic confusion. At the same time, for the clinical analysis of pediatric germectomy, we only need to segment a small part of AB that surrounding the MWT and SM, instead of the entire alveolar bones. Therefore, it is also crucial to guide the model to segment the AB in the correct locations. On the other hand, because the tooth germs are soft tissue without calcified, they usually terms to have a smaller volume compared to SM, and SM also have a more smaller volume than the surrounding AB. These factors imply that if a model wants to simultaneously segment a patient's MWT germs, SM and AB, it must be capable of addressing the multi-scale problems. To address these issues, we proposed a semantic separation scale-specific feature fusion network named WTNet. It contains of two main components: the Input Enhancement (IE) block, which enhances the original input to provide stronger ROI feature representations and avoid confusions. The Teeth-Bones Feature Separation (TBFS) block, which utilizes a shared encoder and two independent decoders to predict teeth and bones respectively, so as to improve the segmentation performance. In the IE block, we designed a Regional Feature Enhancement (RFE) block that uses a binary ROI mask to let the share encoder pay more attention to the target segmentation areas. In TBFS block, we developed a Scale-specific Feature Fusion (SFF) block that enables the teeth and bones branches to adaptively select suitable features from the encoder. The experimental results show that WTNet can outperform the previous SOTA segmentation methods on NKUT dataset.

The key contributions in this paper can be summarized as follows:

- We collected and annotated a 3D high-quality pediatric CBCT dataset called NKUT, which is the first dataset specifically designed for pediatric MWT segmentation. We believe the NKUT will be valuable for the researches of CAD on paediatric dentistry germectomy.
- We proposed a semantic separation scale-specific feature fusion network called WTNet to segment teeth and bones separately and address the problem of semantic-blur and feature confusions. WTNet consists of two main components i.e., IE block and TBFS block. In the IE block, we designed an RFE block to enhance the target features of original input slices,while in the TBFS block, we designed an SFF block to let the teeth and bones branches choose suitable encoder features.

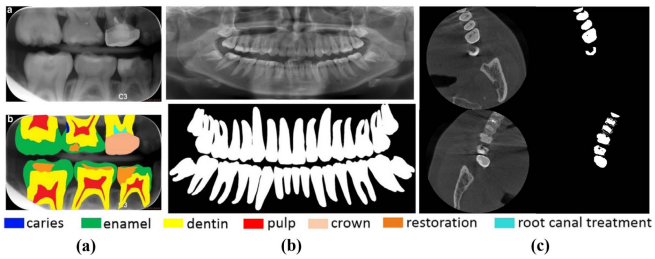

Fig. 2. Examples of some previous public oral datasets, (a) Dental X-ray, (b) LNDb, (c) CTooth.

- Extensive experiments were conducted on the proposed NKUT dataset to establish benchmarks for pediatric MWT segmentation. The results show that our WTNet outperforms previous SOTA networks and can serve as a strong baseline for the pediatric MWT segmentation task.

II. Related Works

In this section, we briefly review previous open-source dental datasets as well as the recent works on CBCT teeth segmentation and multi-stage or multi-scale approaches.

A. Dental Datasets

Dental X-ray dataset [14] as shown in Fig. 2(a) was collected in 2015, it contains 400 cephalometric radiographs and 120 bitewing radiographys. 19 landmarks were manually annotated in each cephalometric radiographs and 7 locations including caries, enamel, crown, dentin, pulp, root canal treatment and restoration were manually annotated in each bitewing radiographys. Another X-ray dataset LNDb was collected in 2016 by Jader et al. [15]. Compared to Dental X-ray, LNDb contains 1,500 annotated panoramic X-ray images. Images in this dataset were manually categorized among 10 categories and all images were cropped to 1991 $\times$ 1127 pixels. As shown in Fig. 2(b), this dataset provided binary annotations between background and full set of teeth. CTooth [16] and CTooth+[17] were the first public 3D CBCT datasets proposed by Cui et al. in 2022. The datasets consist of 5504 annotated CBCT slices of 22 patients and 25876 unlabeled CBCT slices of 146 patients. As shown in Fig. 2(c), CTooth only provided binary annotations and ignored alveolar bones. Although CTooth has more than 168 volumes and 30,000 slices, most of them are without any annotations. However, these datasets are randomly selected and not specifically obtained from pediatric dentistry, and none of them have provided the annotations of the alveolar bones. These factors leads to significant limitations to directly use them for pediatric MWT segmentation task. Therefore, building a public available high-quality dataset is crucial for advancing researches in pediatric MWT segmentation.

B. Teeth Segmentation in CBCT

CBCT images show substantial advantages clinically, such as accurate measurements and excellent resolutions, thus CBCT has become the preferred imaging procedure for comprehensive orthodontic treatments [18]. In recent years, with the development of deep learning, some methods have been proposed to solve the teeth segmentation tasks for CBCT images [19]. Cui et al. proposed a two-stage deep convolutional neural network (ToothNet) [20] for automatic and accurate tooth instance segmentation and identification from CBCT images. It is the first neural network used for CBCT teeth segmentation task, achieving the highest performance on a in-house CBCT dataset. Cui et al. also presented a strong deep-learning-based AI system with an ROI generation network and a specific two-stage deep network to localize the foreground and explicitly leverage the comprehensive geometric information [21], [22]. This framework was evaluated on the largest in-house CBCT dataset, including 4938 CBCT scans of 4215 patients. Zheng et al. proposed a novel anatomically constrained Dense U-Net [23] for integrating oral-anatomical knowledge, benefiting from the integration with anatomical domain knowledge, it achieved good results on a small dataset. In 2023, Liu et al. proposed a metal artifacts and blurring robust CBCT tooth segmentation method ToothSegNet [24], it generates degraded images and do channel-wise cross fusion between the information of both high and low quality images to reduce the semantic-blur problem between encoder and decoder. They also constructed a private CBCT dataset with sparse annotations to test their network and achieved good performance. However, all these methods aim to segment fully developed teeth rather than children's teeth germs. Consequently, they lack the ability to effectively address the serious multi-scale and semantic-blur challenges in pediatric MWT segmentation tasks. In conclusion, it is necessary to develop a model with robust multi-scale feature learning capabilities for pediatric MWT segmentation.

C. Multi-Stage Methods & Multi-Scale Feature Fusion

The multi-stage approaches divide a major problem into several sub-problems to be solved step by step, thereby improving efficiency and effectiveness. In recent years, researchers have proposed some multi-stage methods for various tasks. Pandey et al. [25] introduced a two-stage generative adversarial network to artificially increase the number of training image-mask pairs. The approach used two GAN to generate the synthesized masks and images, it was evaluated using the cell nuclei image segmentation task and demonstrated the superior performance. Lei et al. [26] proposed an unsupervised domain adaptation method based image synthesis and feature alignment method to segment optic disc and cup on fundus images. The method can be divided into two main parts: the GAN-based image synthesis system to generate high-quality target-like query images, and content and style feature alignment to ensure the feature consistency. Experimental results have demonstrated the effectiveness of their method. Inspired by them, we employ the multi-stage strategy to accomplish feature enhancement and segmentation of pediatric MWT.

Multi-scale feature fusion is crucial for medical image segmentation, especially when dealing with lesions which have small volumes and low contrast. Recently, the researchers have proposed some multi-scale fusion methods for medical image

TABLE I
STATISTICS OF THE EXISTING TEETH DATASETS AND OUR NKUT.ML, WT AND PD REPRESENT MULTI-LABEL, WISDOM TEETH AND PEDIATRIC DENTISTRY, RESPECTIVELY

Dataset	Modality	Year	Slices	Task	ML	WT	PD
Dental X-ray [14]	2D X-ray	2015	400 + 120	caries detecting and X-ray analysing	✓	×	×
LNDb [15]	2D Panoramic X-ray	2016	1500	X-ray binary teeth segmentation	×	✓	×
CTooth [16]	3D CBCT	2022	22 volumes (7368 slices)	Tooth volume binary segmentation	×	✓	×
CTooth+ [17]	3D CBCT(unlabeled)	2022	146 volumes (25876 slices)	Tooth volume binary segmentation	×	×	×
Our NKUT	3D CBCT	2023	133 volumes (53000+ slices)	mandibular wisdom teeth segmentation	✓	✓	✓

segmentation. For example, Wu et al. [27] introduced a efficient adaptive dual attention module and a spatial information weighting method for automated skin lesion segmentation. He et al. [28] designed a progressively multi-scale fusion network, which improves the multi-class fundus lesion segmentation accuracy by integrating features from the current encoder layer and adjacent encoder layers. Wang et al. [29] proposed a cross-scale boundary-aware transformer named XBound-Former to address the boundary variation problem of the skin lesion segmentation. Different from previous works, our TBFS block use the SFF blocks to ensure that each layer of the decoder receives contextual information from each level of the encoder, thereby increasing the class-specific segmentation capability of different branches.

III. NKUT DATASET

The lack of high-quality datasets is the major factor hindering the CAD of pediatric dentistry germectomy. We aim to introduce a valuable dataset with high-quality annotations for the relevant researches. This study has undergone review by the Medical Ethics Committee of Tianjin Stomatological Hospital and has been confirmed to fully comply with the Helsinki Declaration and relevant regulations on biomedical human experiments in the People's Republic of China. The approval number is PH2021-B-024_001, and the approval date is March 30, 2021.

A. Collection

Several 2D and 3D open-source dental datasets are listed in Table I. Different from them, our NKUT focus on the pediatric MWT segmentation task. NKUT contains 133 CBCT volumes, which were scanned using NewTom VGi scanners without any appearance enhancements. We collected all original DICOM files from the Department of Pediatric Dentistry, Tianjin Stomatological Hospital (grade-A tertiary hospital). All files have been desensitized for public use. During the data collecting stage, radiologists initially examined the patients' volumes in the database and selected the clear volumes with appropriate ages and MWT development status. After the first round primary selection, the chosen volumes will be reviewed by two paediatric dentistry experts to ensure their quality. Finally, a total of 133 CBCT volumes were selected to establish the NKUT dataset. The average age of these patients is 13.2 years, with 81 males and 52 females. The detailed distribution of the dataset is illustrated in Fig. 3. It is worth to note that we collected 34 cases aged 18 to 22 years in the NKUT dataset for two main reasons: Firstly, we aim to enhance the diversity of NKUT, encompassing various developmental stages of MWT, thereby improving the robustness of NKUT and facilitating subsequent researches; Secondly, we intended to enhance the generalization performance of the models. We aspire that the models trained on NKUT can not only identify MWT germs but also recognize MWT at various stages of development, from germ to complete calcification. When we release NKUT, we will clearly indicate which data belong to individuals aged 18 and above. Researchers can then decide whether to incorporate this subset into their studies.

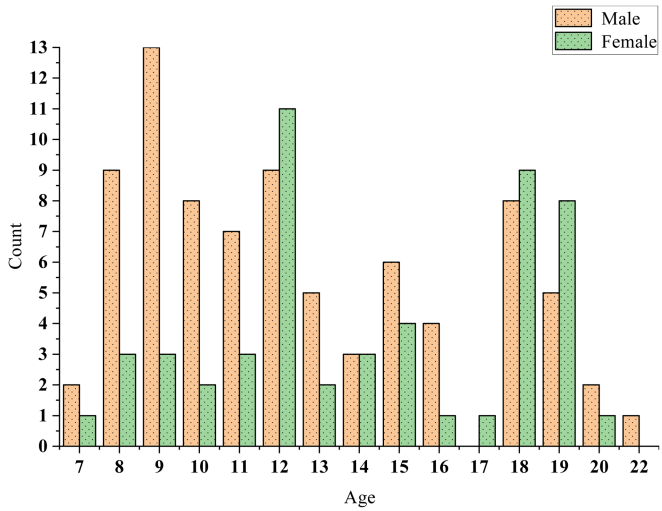

Fig. 3. Distribution of age and gender in the NKUT dataset. The horizontal axis of the table corresponds to age and the left vertical axis shows the count.

B. Annotation

NKUT provides three pixel-level labels, including bilateral MWT germs, SM and a portion of surrounding AB. All scans in NKUT were manually annotated in detail, the annotation tool we used is ITK-SNAP, and the major annotation work was done by two senior experts in paediatric dentistry with at least ten years of experience from Tianjin Stomatological Hospital. The annotation can be divided into two stages: annotating and reviewing. During the annotating stage, we split the data into

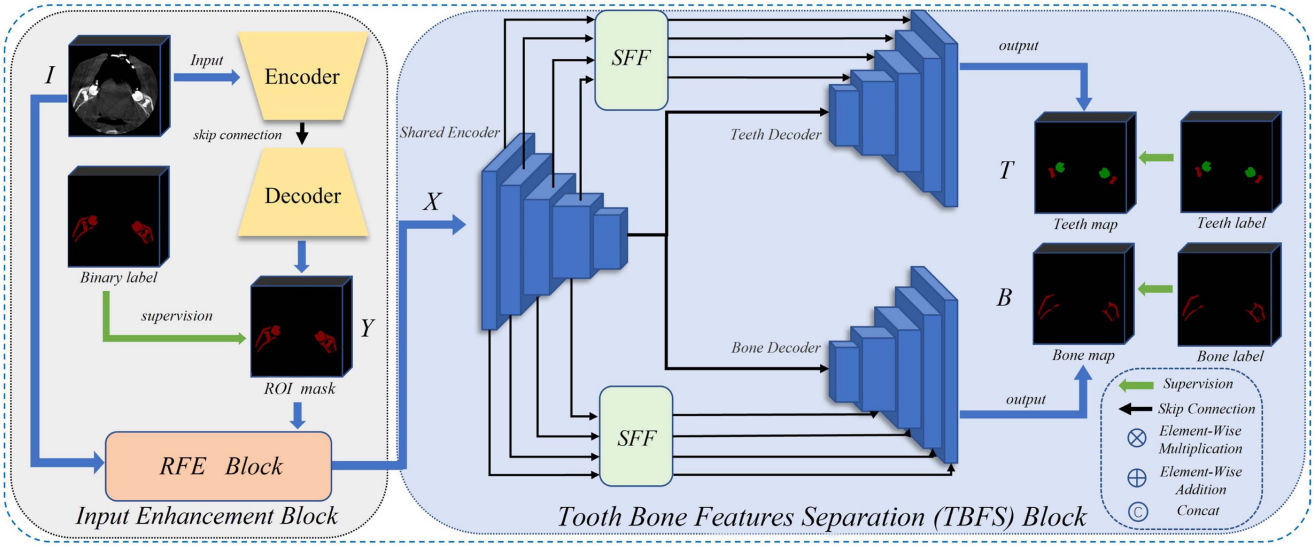

Fig. 4. Overall structure of the proposed WTNet. It consists of two basic components: Input Enhancement (IE) block and Teeth-Bones Feature Separation (TBFS) block. Labels represent the corresponding ground truths and maps denote the prediction of input slices.

two non-overlapping subsets and assign each subset to different experts for annotation, two experts used automatic threshold method to obtain the different rough masks of bilateral MWT, SM and AB. Then they fine-tuned all rough masks slice-by-slice in the axial, coronal and sagittal views to get the high-quality pixel-level labels. Finally, we obtained the annotation results for all data. It takes us about 10 hours to get and refine the rough masks of each volume. In the reviewing stage, after shuffling all the data, we invite the experts to review and discuss all 133 volumes one by one. For data on which a consensus is reached, we organize and store it in the database. For data on which the two experts cannot reach a consensus, they will adjust the annotations until an agreement is reached. Finally, we cropped out the redundant (without annotations) regions along the horizontal plane in the scans to complete the establishment of the NKUT dataset. The whole process of collecting, annotating, reviewing and adjusting the entire dataset took us more than 12 months.

IV. Methodology

In this section, we first provide a brief overview of the proposed network and then detail each component in the following subsections.

A. Overview

The overall pipeline of WTNet is illustrated in Fig. 4, it can be divided into two main components: (1) Input Enhancement (IE) block aims to enhance the feature representation of the target segmentation areas. Additionally, a RFE block is designed to induce the module and let it pay more attention to the regions that need to be segmented. (2) Teeth-Bones Feature Separation (TBFS) block is responsible for predicting the segmentation masks of teeth and bones separately. In order to enable the independent branches in TBFS block to learn appropriate multi-scale features, we propose Scale-specific Feature Fusion (SFF) blocks for semantic dissociated multi-scale feature fusion. The SFF blocks allow adaptive feature selections for the segmentation of teeth and bones, which enhance the ability of multi-scale segmentation in WTNet.

Specifically, given a image $I \in \mathbb{R}^{H\times W\times C\times D^*}$, here H, W and C denote the height, width and the channel numbers of the image. D^* represents an optional dimension: when the input is 3D data patches, this term indicates the depth of input data; In the case of 2D slices, this term is omitted. I will be sent to IE block to get a binary ROI mask $Y \in \mathbb{R}^{H\times W\times D^*\times 1}$, then I and Y will be sent to the RFE block to get the enhanced input $X \in \mathbb{R}^{H\times W\times D^*\times C}$. In the TBFS block, X will be first sent to a shared encoder, which has four encoder layers and a bottleneck layer follows the way of VGG16 [30] (for 3D training, the share encoder is same as the 3D-Unet [31] encoder) to produce the hierarchical features of X. We denote the output features of encoder-layers as $F_i \in \mathbb{R}^{\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}\times\frac{D^*}{2^{i-1}}\times C_i}$, where $i \in \{1, 2, 3, 4\}$. Then all F_i will be sent to the SFF blocks of the teeth branch and the bones branch respectively, so that different branches can adaptively select multi-scale features and local spatial information for scale-specific feature fusion. Finally, the outputs include a teeth segmentation mask $T \in \mathbb{R}^{H\times W\times D^*\times 3}$ and a bones segmentation mask $B \in \mathbb{R}^{H\times W\times D^*\times 2}$. Next, we detail each component of the proposed method.

B. Input Enhancement Block

1) Motivation: In the pediatric MWT segmentation task, it is crucial to ensure that the network can accurately localize the target areas. Taking inspiration from previous works like [20], we design the IE block to enhance the target features of the original input slices. We further designed an RFE block, which is the crucial and major component of the IE block. By enhancing the original slices X using the generated ROI mask in the RFE

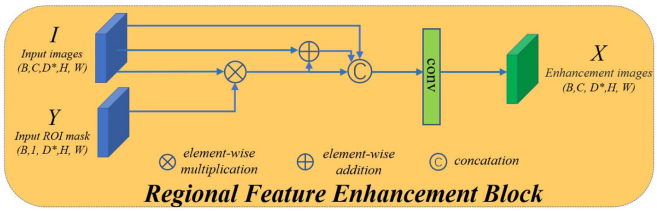

Fig. 5. Schematic diagram of the RFE Block in the IE Block.

block, we can strengthen the feature representations and induce the network to pay more attention to the target areas. This plays a vital role in distinguishing MWT, SM, and AB from other regions.

2) Structure Details: Specifically, in IE block, we first employ a Unet [32] structure with VGG16 [30] backbone to get the Binary ROI mask of the input data (for 3D training, we utilized a 3D-Unet [31] directly). In this process, we employ a binary label generated from the original label for supervision. Given original label with annotations i, where $i \in \{0, 1, 2, 3\}$ and they denote backgrounds, MWT, SM and AB, respectively. The binary label used in IE block can be produced by setting the pixel values of SM and AB in original label to 1. After that, the output Binary ROI mask Y and original input slices I will be sent to the RFE block to get the enhanced output X, as shown in Fig. 5. We can obtain X using the following formula:

$$X = \text{Conv}_{1\times1(\times1)}(\text{Concat}(((I \otimes Y) \oplus I), I \otimes Y, I)) \quad (1)$$

where Concat denotes concatenation along the channel dimension and we utilize this step to effectively fuse the features of the original and baited slices, while also avoiding the excessive influence of ROI mask on the final segmentation results. $\text{Conv}_{1\times1(\times1)}$ refers to a $1 \times 1(\times1)$ convolutional layer with $padding = 0$, which can further fuse features and perform channel dimension reduction. $\otimes$, $\oplus$ donate the element-wise multiplication and element-wise addition.

C. Teeth-Bones Feature Separation Block

1) Motivation: Children's MWT germs, SM and surrounding AB differ significantly in volume and morphology, leading to serious multi-scale challenges. In order to allow the model to learn specific features related to teeth, germs and bones, we utilize the divide-and-conquer strategy to generate segmentation masks in different scales. Inspired by [33] we propose Teeth-Bones Feature Separation (TBFS) Block to use a structure with shared encoder and specific decoders to specifically enhance the model's ability to address the multi-scale problems.

2) Structure Details: We utilize a shared encoder to extract hierarchical features from the enhanced slices. We employ two independent specific decoders to generate segmentation maps for teeth and bones. Specifically, the teeth branch is only responsible for segmentation of MWT and SM, the ground truth of teeth branch can be produced by setting the label of AB in the original labels to 0. Another branch is only used to segment AB, the ground truth of bones branch can be produced by setting the label of MWT and SM in the original labels to 0 and the AB label to 1. The output feature maps of the shared encoder can be represented as F_i, where $i \in \{1, 2, 3, 4, 5\}$, please note that F_5 corresponds to the output of bottleneck layer. In contrast to traditional skip connections, which only fuse features with the same spatial resolution on a single scale, $\{F_1, F_2, F_3, F_4\}$ in WTNet will be sent to the SFF blocks for scale-specific integration. We will introduce the SFF blocks in the following subsection.

D. Scale-Specific Feature Fusion Block

1) Motivation: To enhance the feature learning capabilities of the specific branches in TBFS, we introduce the Scale-specific Feature Fusion (SFF) Block. It is widely recognized that maintaining low-level features is crucial for segmenting small objects [34] such as the tiny MWT germs and SM. Previous works [35] and [28] have demonstrated the importance of multi-scale feature fusion. Taking inspiration from them, we designed the SFF blocks, which take the full-scale output features of the shared encoder as input and then send them to the specific decoders for feature fusion. The SFF blocks enable each level of the decoders to integrate rich full-scale features, thereby enhancing the scale-specific features fusion ability of the TBFS block.

2) Structure Details: As shown in Fig. 6, within the SFF blocks, we utilize the full-scale features $\{F_1, F_2, F_3, F_4\}$ obtained from the shared encoder as input. In each SFF block, the four outputs are first adjusted to the same spatial resolution and subsequently concatenated along the channel dimension. More specifically, in the teeth SFF block, $\{F_1, F_2, F_3, F_4\}$ are first adjusted the resolutions and concatenated to $Z_T \in \mathbb{R}^{B\times C_f\times\frac{D^*}{2}\times\frac{H}{2}\times\frac{W}{2}}$ using the following formula:

$$Z_T = \text{Concat}(\text{AP}_{k=2}(F_1), F_2, \text{UP}_{k=2}(F_3), \text{UP}_{k=4}(F_4)) \quad (2)$$

where AP, UP and k represent adaptive average pool layers, bilinear upsampling layers and the scale factor, respectively. Here $C_f = C_1 + C_2 + C_3 + C_4$ and B represents the batch size. Then the output Z_T will pass through a Channel Attention (CA) block following the approach of [36] to adaptively adjust the attention scores across different feature scales. Let the output of CA be denoted as $\hat{Z}_T$ with the same size of Z_T. The final output $\hat{F}_i^t \in \mathbb{R}^{\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}\times\frac{D^*}{2^{i-1}}\times C_i}$ of the teeth branch SFF can be calculated using the following formula:

$$\hat{F}_i^t = \text{SA}(\text{Restore}(\hat{Z}_T) \oplus F_i), i = (1, 2, 3, 4) \quad (3)$$

Here SA refers to the Spatial Attention block [37], and the symbol $\oplus$ donates element-wise addition. The Restore layer in SFF serves to convert $\hat{Z}_T$ to four different X_i, each of which matches the spatial resolution and channel number of F_i. More specifically, in the Restore layer, four $1 \times 1(\times1)$ convolutional layers are employed to adjust the channel number of $\hat{Z}_T$ to align with F_i. Then we use various resample layers to restore the spatial resolution of $\hat{Z}_T$ to match that of F_i. The overall process of the Restore layer of teeth branch SFF can be considered the inverse operation of (2). Consequently, $\hat{Z}_T$ will be transformed

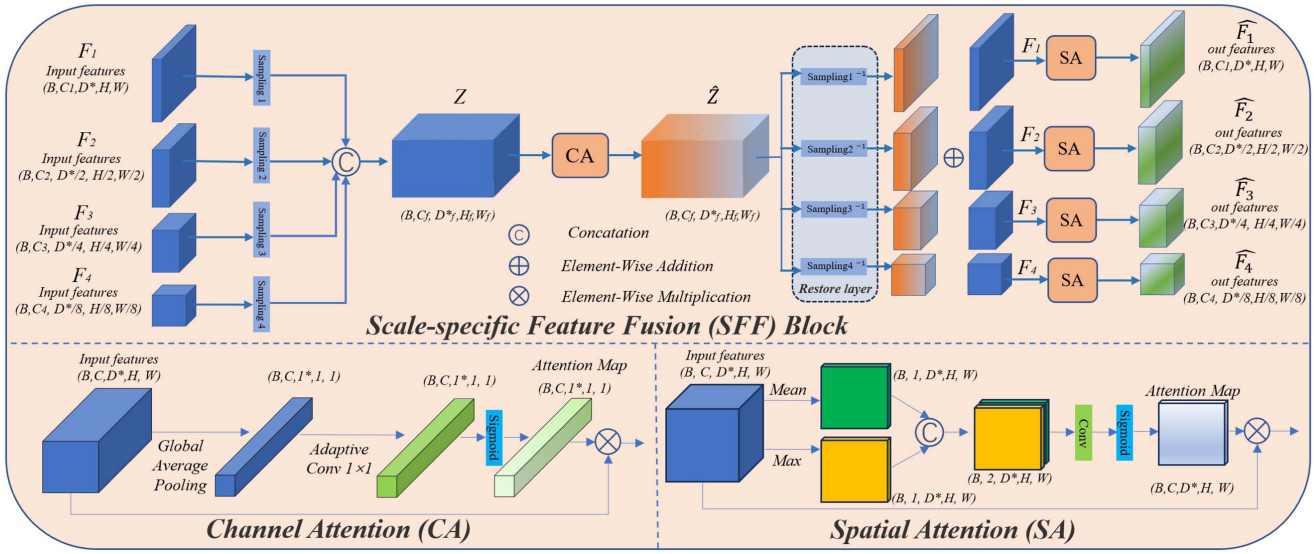

Fig. 6. Overall structure of the proposed SFF blocks in both Tooth branch and Bone branch. The CA and SA denote channel attention and spatial attention, respectively. Samplingx^{-1} represents the inverse process of the corresponding sampling operation. For example, if Sampling1 is downsampling 2×, then Sampling1^{-1} is upsampling 2×. If Sampling is Identity, then Sampling^{-1} is also Identity.

into four different X_i, where $i \in \{1, 2, 3, 4\}$ and the size of X_i is consistent with F_i. Then each X_i undergoes element-wise addition with its corresponding F_i. The results will be then passed through four independent SA layers to extract spatial features and get the final output $\hat{F}_i^t$. The $\hat{F}_i^t$ are used in the skip connections at corresponding positions within the decoder.

Similarly, in bones branch, the $Z_B \in \mathbb{R}^{B\times C_f\times\frac{D^*}{4}\times\frac{H}{4}\times\frac{W}{4}}$ can be calculated using the following formula:

$$Z_B = \text{Concat}(\text{AP}_{k=2}(F_1), \text{AP}_{k=4}(F_2), F_3, \text{UP}_{k=2}(F_4)) \quad (4)$$

please note that in the teeth branch, we adjust the resolutions of feature maps in all scales to half of the original input resolutions, while here we adjust them to a quarter of the original input resolutions. The reason is that we hope more low-level features can be retained in the teeth branch, and more valuable information can be mined from the high-level features in the bones branch. Thanks to the divide-and-conquer approach of the TBFS, we can enable the teeth and bones branch to better learn scale-specific features more effectively. The output of the CA in the bones branch is denoted as $\hat{Z}_B$, and the final output $\hat{F}_i^b$ of the bones SFF branch can be calculated using (5). Note that the Restore here represents the Restore layers in bones SFF, which can be considered as the inverse operation of (4), differing from the teeth branch.

$$\hat{F}_i^b = \text{SA}(\text{Restore}(\hat{Z}_B) \oplus F_i), i = (1, 2, 3, 4) \quad (5)$$

E. Decoders

In each specific decoder of the 2D training stage, we employ linear interpolation for up-sampling, followed by two $3 \times 3(\times 3)$ convolutional layers with ReLU activation for feature learning. To be specific, the output of each decoder in both the teeth branch and bones branch can be calculated using the following formula:

$$D_i = \text{Concat}(\text{Conv}_{3\times3(\times3)}(\text{UP}(D_{i+1})), \hat{F}_i), i \in \{1, 2, 3, 4\} \quad (6)$$

Where D_5 represents the output of the bottleneck layer F_5 in the shared encoder. In the final output layer, the feature maps in each branch will be resized to the same resolution with the original inputs X. Then a 1×1 convolution layer is utilized to predict the segmentation results, yielding $\text{T} \in \mathbb{R}^{H\times W\times D^*\times 3}$ and $\text{B} \in \mathbb{R}^{H\times W\times D^*\times 2}$. Here T and B represent the final segmentation results of teeth branch and bones branch, respectively. In the 3D training stage, we directly use the 3D-Unet [31] decoder as our specific decoders.

F. Loss Function

During the training stage, we employ a combination of the cross-entropy loss $\mathcal{L}_{ce}$ and the Dice loss [38] $\mathcal{L}_{dice}$ as the loss function of each output layers. The definitions of $\mathcal{L}_{ce}$ and $\mathcal{L}_{dice}$ are as follows:

$$\mathcal{L}_{ce}(P, G) = -\sum_{i=1}^{H\times W\times D^*} G_i \cdot log(P_i) \quad (7)$$

$$\mathcal{L}_{dice}(P, G) = 1 - \frac{2 \cdot \sum_{i=1}^{H\times W\times D^*} P_i G_i + \tau}{\sum_{i=1}^{H\times W\times D^*} P_i^2 + \sum_{i=1}^{H\times W\times D^*} G_i^2 + \tau} \quad (8)$$

where P and G represent the predicted maps and ground truth of the input CBCT slices, respectively, and $H \times W$ donates the pixel numbers of output masks. τ is a smooth term. The entire architecture can be trained using the following loss function:

$$\mathcal{L} = \mathcal{L}_{ce}^B + \mathcal{L}_{dice}^B + \mathcal{L}_{ce}^t + \mathcal{L}_{dice}^t + \mathcal{L}_{ce}^b + \mathcal{L}_{dice}^b \quad (9)$$

where $\mathcal{L}^B$, $\mathcal{L}^t$ and $\mathcal{L}^b$ represent the loss function for the binary mask, teeth mask and bones mask, respectively.

TABLE II
QUANTITATIVE COMPARISONS WITH SOME PREVIOUS STATE-OF-THE-ART METHODS ON THE PROPOSED NKUT DATASET.

Methods	mIOU% ↑	ACC%↑	HD95↓	ASSD↓
3D-Unet [31]	59.12	66.88	71.47	12.12
Vnet [38]	49.68	60.14	89.57	14.15
3D-DenseNet [45]	46.41	50.68	121.04	21.68
UneTR [46]	49.54	61.26	88.79	14.55
VT-Unet [47]	52.18	61.65	94.30	15.33
WTNet-3D (ours)	**59.58**	**67.82**	**70.06**	**11.60**
Unet [32]	67.12	78.12	19.18	4.63
PSPNet [48]	49.28	58.82	21.82	6.63
DeepLabV3+ [49]	65.11	75.88	22.60	6.26
TransUnet [50]	65.64	76.53	23.14	6.49
SegFormer [51]	61.02	72.22	21.03	5.60
ToothSegNet [24]	63.76	78.95	19.39	5.27
WTNet-2D (ours)	**68.72**	**79.70**	**18.76**	**4.38**

We conuct 5-FOLD cross validation and report their average results. The bold value represents the best result of a certain indicator.

TABLE III
DSC COMPARISONS WITH SOME PREVIOUS SOTA METHODS ON THE PROPOSED NKUT DATASET.

Methods (3D)	DSC↑(%)				Paired T-test	
	MWT	SM	AB	Avg	|T| (2.132)	P
3D-Unet [31]	66.72	72.69	66.77	68.73	1.042	3.561e-1
Vnet [38]	61.70	64.59	61.81	62.70	3.113	3.577e-2
3D-DenseNet [45]	57.75	62.29	50.54	56.86	5.102	6.971e-3
UneTR [46]	58.29	65.11	61.86	61.75	4.190	1.380e-2
VT-Unet [47]	62.90	66.48	62.73	64.04	10.152	5.301e-4
WTNet (ours)	**66.96**	**73.16**	**67.39**	**69.17**	-	-
Methods (2D)	**DSC↑(%)**				**Paired T-test**	
	MWT	SM	AB	Avg	|T| (2.353)	P
Unet [32]	80.87	80.91	75.52	79.10	5.615	3.117e-2
PSPNet [48]	66.40	67.71	62.58	65.56	7.255	5.404e-3
DeepLabV3+ [49]	79.84	82.00	73.15	78.33	2.389	9.685e-2
TransUnet [50]	80.98	81.03	74.69	78.90	2.698	7.389e-2
SegFormer [51]	76.52	77.67	72.30	75.50	4.732	1.788e-2
ToothSegNet [24]	78.09	77.47	**77.36**	77.64	4.819	1.703e-2
WTNet (ours)	**83.41**	**83.39**	76.40	**81.07**	-	-

We conuct 5-FOLD cross validation and report their average results. Additionally, we conducted paired t-tests with a significance level of 0.1 using the average results corresponding to each fold, and reported the t-statistic and p-value in the table. The value next to |T| represents the t critical value for the current test.
The bold value represents the best result of a certain indicator.

V. EXPERIMENTS

In this section, we will first introduce the data processing methods, implementation details and evaluation metrics. Next, we will compare WTNet with some SOTA segmentation methods and present the qualitative and quantitative experimental results. Finally, we will conduct the ablation studies to verify the effectiveness of each component in our method.

A. Data Processing

We performed the following processing on proposed NKUT for network training and testing. The window-level and window-width of all CBCT images were adjusted to 800 and 2500, respectively. The HU values of all CBCT images were normalized to the range of [0,255]. For the 2D training and testing stage, we extracted images and labels slice by slice along the horizontal plane from the original CBCT images. The input for 2D networks is the 3-channel RGB images with a size of 256×256. For the 3D training stage, due to the limitation of GPU memory, we randomly cut 150 patches with a size of $64 \times 64 \times 64$ around the labeled areas in each training set CBCT scan, so, the input for 3D networks is the single channel gery level CBCT image blocks with a size of $64 \times 64 \times 64$. For the testing stage of 3D networks, we used a sliding window with non-overlap to sequentially predict $64 \times 64 \times 64$ patches. In both 2D and 3D testing, for pixels where there are ambiguities between the teeth and bone branches, we directly set those pixel values of the segmentation masks to 0 to generate the final segmentation masks.

B. Implementation Details and Evaluation Metrics

1) Implementation Details: In the experiments, we applied data augmentations of horizontal flips, vertical flips, and random rotations to reduce overfitting. We utilized Adam optimizer [39] to train all models. The total number of epochs was set to 200 with an initial learning rate of 0.0001. To ensure fairness, we employed the same learning rate decay strategy to train the all models, i.e., if the training loss does not decrease for more than 3 epochs, the learning rate will be decreased by multiplying λ (where $\lambda = 0.8$ in our work). Additionally, if the validation loss does not decrease for more than 10 epochs, the training process will be terminated. The numbers of feature channels for the four encoder-layers in both IE and TBFS blocks are $[64, 128, 256, 512]$ (the same settings are used for 2D and 3D.). The 2D encoder backbones in IE and TBFS blocks were VGG16 pre-trained on the ImageNet [40]. For SegFormer we used its b_0 version module pre-trained on VOC2012 [41]. Since the 3D pretrained backbone is hard to find, the parameters of all 3D methods were randomly initialized. All training were based on Pytorch [42], torchio [43] and monai [44]. We use a Nvidia RTX 3090 GPU for 3D experiments and a Nvidia RTX 4090 GPU for 2D experiments.

2) Evaluation Metrics: The following metrics were adopted for systematic performance evaluation, including mean Intersection Over Union (mIOU), pixel-wise Accuracy (Acc), Dice Similariy Coefficient (DSC), Hausdorff distance (HD95) and Average Symmetric Surface Distance (ASSD).

C. Comparisons With Other Methods

To demonstrate the effectiveness of our method, we compared its 2D and 3D versions with some state-of-the-art 2D and 3D image segmentation methods using the proposed NKUT dataset. We utilized five-fold cross-validation to conduct comparative experiments and reported the average results. Tables II and III presents the quantitative results of these networks. We can observe that the 2D version of WTNet achieves optimal segmentation results. While the ToothSegNet [24] slightly outperforms WTNet in bone segmentation, WTNet outperforms it in MWT, SM, and overall performance.

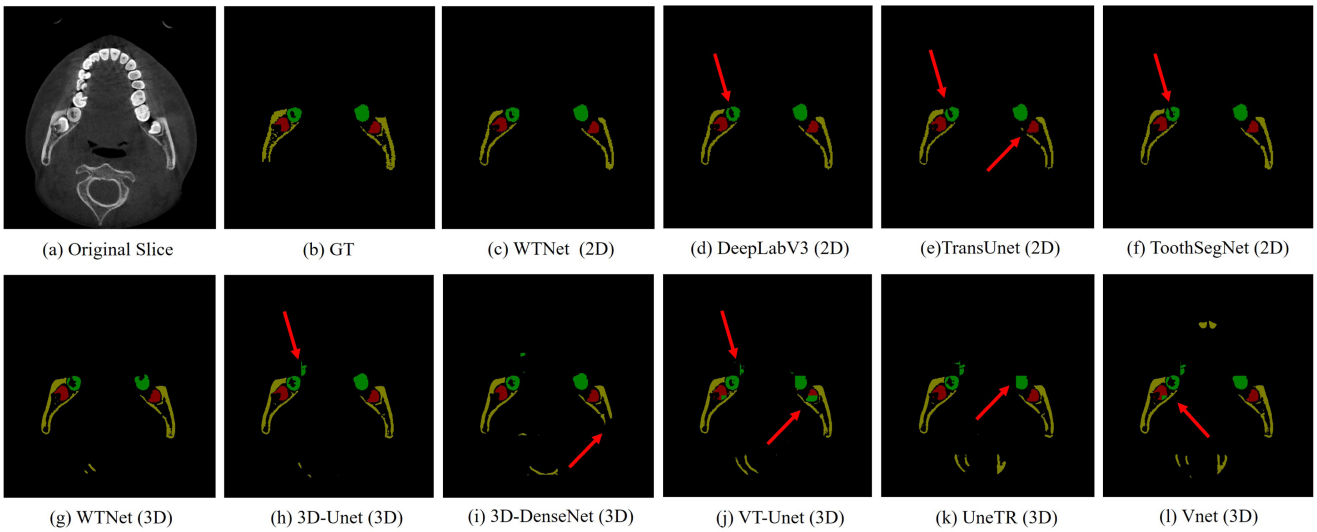

Fig. 7. Qualitative results of our proposed WTNet and other state-of-the-art 2D and 3D segmentation methods, including (d) DeepLab V3+, (e) TransUnet, (f) ToothSegNet, (h) 3D-Unet, (i) 3D-DenseNet, (j) VT-Unet, (k) UneTR and (l) Vnet. (a) and (b) donate original CBCT slice and ground truth, respectively.

The visual comparison results are displayed in Fig. 7. WTNet-2D exhibits outstanding performance on the NKUT dataset. In comparison to other methods, it can segment intact teeth and bones with well-defined boundaries, and the segmentation results are the most similar to the manual annotations of pediatric dentists. Thanks to the divide-and-conquer strategy and the robust scale-specific feature fusion capability within the TBFS block, WTNet can accurately segment teeth, germs and bones at various scales simultaneously. In contrast to the skip connection in Unet and its variants, SFF ensures scale-specific feature fusion, preserving the semantic integrity of the decoders. This is essential for achieving higher segmentation accuracy and structural completeness, particularly when identifying tiny teeth germs. The inclusion of the IE block enables WTNet to accurately delineate target areas without confusion with other neighboring teeth or bones.

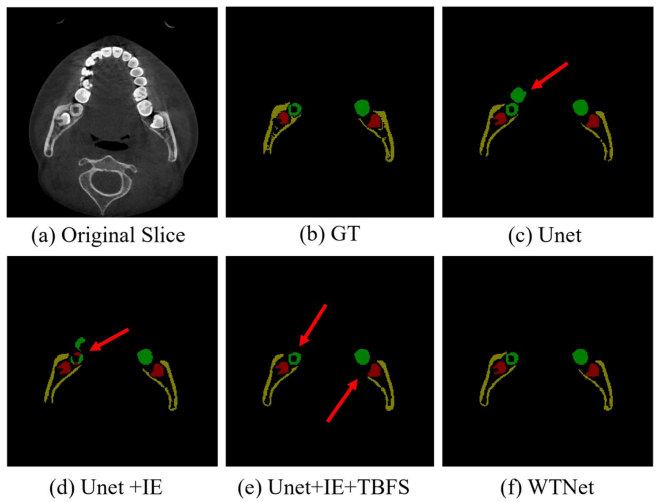

Fig. 8. Visual comparisons of the segmentation results between different configurations.

D. Ablation Studies

To verify the effectiveness of the IE, TBFS and SFF blocks, we conducted ablation studies based on the baseline Unet to better understand the impacts of each component. We give the qualitative and quantitative segmentation results and analyze each block.

1) Analysis of IE Block: Comparing ② with the baseline Unet ①, we can clearly observe the effectiveness of the IE block. From a quantitative perspective, after incorporating the IE block, the DSC of MWT, SM, and AB has improved by 1.09%, 1.30%, and 0.46%, respectively, compared to the baseline UNet. These results indicates that enhancing the target features in input images using the IE block can effectively improve the network's segmentation performance to some extent. Furthermore, when comparing Fig. 8(c) and (d), it is not difficult to see that the baseline Unet confuses the first molar with the second molar due to their highly similar structures. After adopting the IE block, the network provides a more accurate assessment of the target areas but still displays some confusion in the distinguishing between MWT and SM. This suggests that while the IE block enhances performance, it is still unable to effectively address specific multi-scale feature learning challenges.

2) Analysis of TBFS Block: Expanding upon Tabel IV ②, we replaced the encoder-decoder structure in unet with the TBFS block. We investigated the effectiveness of the TBFS block in Table IV ③, revealing that the TBFS block delivers significant improvements when compared to the baseline UNet. After adopting the TBFS block, the segmentation performance of MWT, SM and AB has shown substantial enhancements, with respective enhancements of 0.51%, 0.43% and 0.14% compared to ②. The visualization results, as shown in Fig. 8(e),

TABLE IV
THE CONTRIBUTIONS OF MAIN COMPONENTS IN WTNET-2D ON THE TEST SET OF OUR NKUT DATASET.

Method	MWT	SM	AB	Avg	Δ%	#Param
① Unet	80.87	80.91	75.52	79.10	-	24.89M
② Unet+IE	81.96	82.21	75.98	80.05	↑0.95	35.07M
③ Unet+IE+TBFS	82.47	82.64	76.12	80.41	↑1.31	59.96M
④ WTNet	**83.41**	**83.39**	**76.40**	**81.07**	↑1.97	61.81M

Choosing the DSC as the reference metric.
The bold value represents the best result of a certain indicator.

demonstrate that, with the combination of IE and TBFS, the network can more accurately localize the target regions and provide more accurate and comprehensive segmentation results. In (e), both bilateral MWT germs and SM are segmented well, representing a significant enhancement compared to (d). The continuity and completeness of bone segmentation results have also been further enhanced. This demonstrates that the specificity-independent decoders within the TBFS contribute to further improving the network performance.

3) Analysis of SFF Blocks: The standard skip connections used in TBFS can only fuse features with the same resolution at a single scale, limiting the network's capacity to solve multi-scale segmentation problems. SFF blocks utilize features from all scales provided by the shared encoder as inputs. Simultaneously, the different SFF blocks in the teeth and bones branches enable the model to dynamically select the specific features suitable for each branch. As shown in Table IV ④, we incorporated the SFF blocks into the skip connections of various branches in TBFS, forming the final structure of our proposed WTNet. In comparison with the previous stage ③, the addition of SFF has led to a notable improvement in segmentation performance for all targets.

From the visual results in Fig. 8, it becomes more apparent that the segmentation results displayed in (f) exhibit more intricate details and sharper boundaries. The network no longer confuses the MWT germs and SM, and the segmentation results for the bones are more comprehensive. The segmentation results of Unet+IE+TBFS+SFF (④ WTNet) closely resemble the GT shown in Fig. 8(b), underscoring the indispensable role played by each module within WTNet.

VI. CONCLUSION AND FUTURE WORK

In this paper, we first collected and annotated a CBCT dataset called NKUT, which is the first 3D CBCT dataset for the researches of pediatric MWT germs segmentation and CAD of pediatric dentistry germectomy. Subsequently, we introduced WTNet, composed of two key components: IE block with the RFE block and TBFS block with the SFF blocks. WTNet effectively address the challenges of semantic confusion and multi-scale issues in pediatric MWT segmentation tasks. When compared to other state-of-the-art 2D and 3D image segmentation methods, WTNet achieves the best results on the NKUT dataset. Furthermore, we also proved the effectiveness of the three main components by ablation studies and visualization results. We believe that our work will prove invaluable to the ediatric dentistry germectomy research community.

However, our work still has some shortcomings. *From the perspective of establishing a public dataset:* NKUT consists of only 133 examples, which is still insufficient in scale and difficult to meet the training requirements of complex and large models; *From the perspective of the network model:* There is still some room for optimization in the parameter quantity of WTNet. To further improve these problems, we will continue to expand the scale of the NKUT dataset and utilize methods such as model quantization compression to reduce the parameter count of WTNet. Additionally, we will explore efficient semi-supervised methods to fully leverage labeled and unlabeled data.

In future, we will develop a pediatric intelligent diagnostic system based on segmentation results and deploy it for clinical trials. By analyzing the reconstructed segmentation results, we can obtain the developmental trends of MWT and their shortest distance from the surrounding AB. These findings can assist doctors in making more objective assessments of the necessity for surgery and determining the optimal surgical plans.

REFERENCES

[1] T. D. G. Mettes, H. Ghaeminia, M. E. Nienhuijs, J. Perry, W. J. van der Sanden, and A. Plasschaert, "Surgical removal versus retention for the management of asymptomatic impacted wisdom teeth," *Cochrane Database Systematic Rev.*, vol. 5, 2012, Art. no. CD003879.
[2] E. Bailey, W. Kashbour, N. Shah, H. V. Worthington, T. F. Renton, and P. Coulthard, "Surgical techniques for the removal of mandibular wisdom teeth," *Cochrane Database Systematic Rev.*, vol. 7, 2020, Art. no. CD004345.
[3] G. D'Angeli, F. Zara, I. Vozza, F. M. D'Angeli, and G. L. Sfasciotti, "The evaluation of further complications after the extraction of the third molar germ: A pilot study in paediatric dentistry," *Healthcare*, vol. 9, 2021, Art. no. 121.
[4] C.-H. Wu, W.-H. Tsai, Y.-H. Chen, J.-K. Liu, and Y.-N. Sun, "Model-based orthodontic assessments for dental panoramic radiographs," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 545–551, Mar. 2018.
[5] T. M. Tuan et al., "Dental diagnosis from X-ray images: An expert system based on fuzzy computing," *Biomed. Signal Process. Control*, vol. 39, pp. 64–73, 2018.
[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[7] S. Tian et al., "Efficient computer-aided design of dental inlay restoration: A deep adversarial framework," *IEEE Trans. Med. Imag.*, vol. 40, no. 9, pp. 2415–2427, Sep. 2021.
[8] Y. Lai et al., "LCANet: learnable connected attention network for human identification using dental images," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 905–915, Mar. 2021.
[9] M. Rajee and C. Mythili, "Gender classification on digital dental x-ray images using deep convolutional neural network," *Biomed. Signal Process. Control*, vol. 69, 2021, Art. no. 102939.
[10] S. Tian et al., "Efficient tooth gingival margin line reconstruction via adversarial learning," *Biomed. Signal Process. Control*, vol. 78, 2022, Art. no. 103954.
[11] L. Qiu, C. Ye, P. Chen, Y. Liu, X. Han, and S. Cui, "DArch: Dental arch prior-assisted 3D tooth instance segmentation with weak annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20720–20729.
[12] L. Zhang et al., "TSGCNet: Discriminative geometric feature learning with two-stream graph convolutional network for 3D dental model segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6695–6704.
[13] 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.

[14] C.-W. Wang et al., "A benchmark for comparison of dental radiography analysis algorithms," *Med. Image Anal.*, vol. 31, pp. 63–76, 2016.

[15] G. Silva, L. Oliveira, and M. Pithon, "Automatic segmenting teeth in x-ray images: Trends, a novel data set, benchmarking and future perspectives," *Expert Syst. Appl.*, vol. 107, pp. 15–31, 2018.

[16] W. Cui et al., "CTooth: A fully annotated 3D dataset and benchmark for tooth volume segmentation on cone beam computed tomography images," in *Proc. Intell. Robot. Appl.: 15th Int. Conf.*, 2022, pp. 191–200.

[17] W. Cui et al., "CTooth : A large-scale dental cone beam computed tomography dataset and benchmark for tooth volume segmentation," in *Proc. MICCAI Workshop Data Augmentation, Labelling, Imperfections*, Singapore, 2022, pp. 64–73.

[18] G. L. Machado, "CBCT imaging–A boon to orthodontics," *Saudi Dent. J.*, vol. 27, no. 1, pp. 12–21, 2015.

[19] A. Polizzi et al., "Tooth automatic segmentation from CBCT images: A systematic review," *Clin. Oral Investigations*, vol. 27, pp. 3363–3378, 2023.

[20] 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.*, 2019, pp. 6361–6370.

[21] Z. Cui et al., "A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 2096.

[22] Z. Cui et al., "Hierarchical morphology-guided tooth instance segmentation from CBCT images," in *Proc. Inf. Process. Med. Imag.: 27th Int. Conf.*, 2021, pp. 150–162.

[23] Z. Zheng, H. Yan, F. C. Setzer, K. J. Shi, M. Mupparapu, and J. Li, "Anatomically constrained deep learning for automating dental CBCT segmentation and lesion detection," *IEEE Trans. Automat. Sci. Eng.*, vol. 18, no. 2, pp. 603–614, Apr. 2021.

[24] J. Liu, T. Hu, Y. Feng, W. Ding, and Z. Liu, "ToothsegNet: Image degradation meets tooth segmentation in CBCT images," in *Proc. IEEE 20th Int. Symp. Biomed. Imag.*, 2023, pp. 1–5.

[25] S. Pandey, P. R. Singh, and J. Tian, "An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation," *Biomed. Signal Process. Control*, vol. 57, 2020, Art. no. 101782.

[26] H. Lei, W. Liu, H. Xie, B. Zhao, G. Yue, and B. Lei, "Unsupervised domain adaptation based image synthesis and feature alignment for joint optic disc and cup segmentation," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 90–102, Jan. 2022.

[27] H. Wu, J. Pan, Z. Li, Z. Wen, and J. Qin, "Automated skin lesion segmentation via an adaptive dual attention module," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 357–370, Jan. 2021.

[28] A. He, K. Wang, T. Li, W. Bo, H. Kang, and H. Fu, "Progressive multiscale consistent network for multiclass fundus lesion segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3146–3157, Nov. 2022.

[29] J. Wang et al., "XBound-former: Toward cross-scale boundary modeling in transformers," *IEEE Trans. Med. Imag.*, vol. 42, no. 6, pp. 1735–1745, Jun. 2023.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[31] 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. Interv.*, 2016, pp. 424–432.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[33] J. Shi et al., "Semantic decomposition network with contrastive and structural constraints for dental plaque segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 4, pp. 935–946, Apr. 2023.

[34] Z. Liu, G. Gao, L. Sun, and Z. Fang, "HRDNet: High-resolution detection network for small objects," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[36] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11534–11542.

[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[38] 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.*, 2016, pp. 565–571.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds., May 7–9, 2015. [Online]. Available: https://iclr.cc/archive/www/doku.php\%3Fid=iclr2015:accepted-main.html

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[41] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, pp. 98–136, 2015.

[42] A. Paszke et al., "Automatic differentiation in PyTorch," 2017. [Online]. Available: https://pytorch.org/

[43] F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Comput. Methods Prog. Biomed.*, vol. 208, 2021, Art. no. 106236.

[44] M. J. Cardoso et al., "MONAI: An open-source framework for deep learning in healthcare," 2022, *arXiv:2211.02701*. [Online]. Available: https://github.com/Project-MONAI/MONAI

[45] T. D. Bui, J. Shin, and T. Moon, "3D densely convolutional networks for volumetric segmentation," 2017, *arXiv:1709.03199*.

[46] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 574–584.

[47] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A robust volumetric transformer for accurate 3D tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2022, pp. 162–172.

[48] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[49] 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.*, 2018, pp. 801–818.

[50] J. Chen et al., "TransuNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[51] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.