DTR-Net: Dual-Space 3D Tooth Model Reconstruction From Panoramic X-Ray Images

Lanzhuju Mei, Yu Fang, Yue Zhao, Xiang Sean Zhou, Min Zhu, Zhiming Cui, and Dinggang Shen, *Fellow, IEEE*

***Abstract*— In digital dentistry, cone-beam computed tomography (CBCT) can provide complete 3D tooth models, yet suffers from a long concern of requiring excessive radiation dose and higher expense. Therefore, 3D tooth model reconstruction from 2D panoramic X-ray image is more cost-effective, and has attracted great interest in clinical applications. In this paper, we propose a novel dual-space framework, namely DTR-Net, to reconstruct 3D tooth model from 2D panoramic X-ray images in both image and geometric spaces. Specifically, in the image space, we apply a 2D-to-3D generative model to recover intensities of CBCT image, guided by a task-oriented tooth segmentation network in a collaborative training manner. Meanwhile, in the geometric space, we benefit from an implicit function network in the continuous space, learning using points to capture complicated tooth shapes with geometric properties. Experimental results demonstrate that our proposed DTR-Net achieves state-of-the-art performance both quantitatively and qualitatively in 3D tooth model reconstruction, indicating its potential application in dental practice.**

***Index Terms*— Tooth model reconstruction, panoramic X-ray image, CBCT image, task-oriented segmentation, implicit function.**

Manuscript received 23 May 2023; revised 18 August 2023; accepted 30 August 2023. Date of publication 26 September 2023; date of current version 2 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62131015 and Grant 62206036; in part by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 21010502600; and in part by the Key Research and Development Program of Guangdong Province, China, under Grant 2021B0101420006. *(Corresponding authors: Zhiming Cui; Dinggang Shen.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee in Shanghai Ninth People's Hospital under Approval No. SH9H-2021-T169-1.

Lanzhuju Mei and Yu Fang are with the School of Biomedical Engineering and the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: meilzhj@shanghaitech.edu.cn; fangyu@shanghaitech.edu.cn).

Yue Zhao is with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: zhaoyue@cqupt.edu.cn).

Xiang Sean Zhou is with Shanghai United Imaging Intelligence Company Ltd., Shanghai 200230, China (e-mail: sean.zhou@uii-ai.com).

Min Zhu is with Shanghai Ninth People's Hospital, Shanghai Jiao Tong University, Huangpu, Shanghai 200011, China (e-mail: ZHUM1612@sh9hospital.org.cn).

Zhiming Cui and Dinggang Shen are with the School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China, also with Shanghai United Imaging Intelligence Company Ltd., Shanghai 200230, China, and also with the Shanghai Clinical Research and Trial Center, Shanghai 201210, China (e-mail: cuizm.neu.edu@gmail.com; Dinggang.Shen@gmail.com).

Digital Object Identifier 10.1109/TMI.2023.3313795

I. Introduction

IN DIGITAL dentistry, acquiring complete tooth models with both tooth crown and tooth root is a fundamental step in dental diagnosis and treatment planning [1], [2], [3], [4]. For instance, in orthodontic treatment, tooth movement path design requires ensuring no collision between neighboring teeth during the process. Panoramic X-ray and cone-beam computed tomography (CBCT) images are the two important dental imaging modalities in clinical practice, respectively providing the 2D projection and 3D volumetric data of the oral cavity. However, due to the high radiation levels and costs associated with CBCT devices, acquiring complete tooth models for orthodontic treatment can be relatively expensive and potentially harmful, making it unaffordable in many dental clinics [5]. Therefore, reconstructing a complete 3D tooth model from 2D panoramic X-ray image is more acceptable and of great interest.

Nevertheless, 3D tooth model reconstruction from 2D panoramic X-ray images is quite challenging for two reasons. First, it is highly time-consuming and excessively radioactive to obtain paired CBCT and panoramic X-ray images from the same patient in daily dental clinics, which brings difficulty in building paired 2D and 3D data for network learning. Second, the oral cavity is extremely complicated and consists of various tissues [6], including alveolar bone and different types of teeth with similar intensity distributions. Moreover, caused by projection and distortion during the scanning, the captured panoramic X-ray images suffer from severe ambiguity in the overlapping areas to separate neighboring tissues (e.g., adjacent teeth), and also the loss of dimension to faithfully recover the actual 3D space. Thus, restoring 3D shapes of each tooth from such 2D images is extremely difficult.

To address these challenges, many previous works [7], [8], [9] exploit handcrafted geometric features for 3D tooth reconstruction from panoramic X-ray images. However, these methods are typically built on the pre-defined tooth templates, and thus lack diversity to generate patient-specific tooth models. Recently, with the advance of deep learning, many learning-based methods [10], [11], employing convolutional neural networks (CNNs), have been proposed for 3D tooth or oral cavity reconstruction from panoramic X-ray images. These methods provide feasible strategies to generate CBCT images, followed by a tooth segmentation module to reconstruct the 3D tooth model from 2D

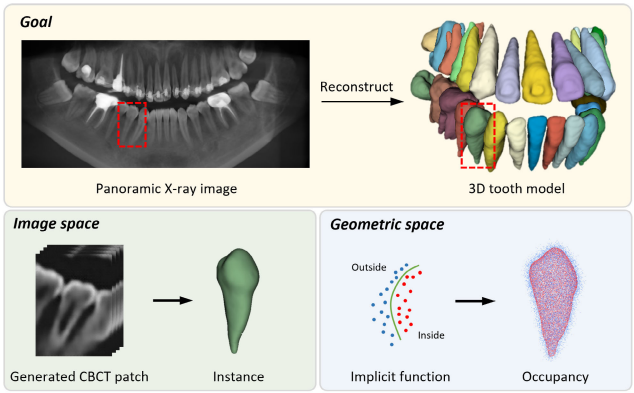


Fig. 1. Overview of our proposed *DTR-Net*, a dual-space framework to reconstruct 3D tooth models from panoramic X-ray images. In the image space, we leverage a 2D-to-3D generative model to learn intensities of CBCT image, especially on the tooth regions obtained simultaneously by a segmentation task. In the geometric space, we design an implicit function network in the continuous space to capture complicated tooth shapes with fine-grained details.

panoramic X-ray images. Unfortunately, these methods focus only on tooth reconstruction in the image space (i.e., CBCT image and tooth mask), entirely ignoring geometric properties even crucial to recover fine-grained details of 3D tooth models.

In this paper, we propose a novel dual-space framework, namely *DTR-Net*, for 3D tooth reconstruction from the 2D panoramic X-ray image. The core of our method is to reconstruct 3D tooth models in dual spaces, including 1) image space (i.e., CBCT and tooth mask) and 2) geometric space (i.e., implicit surface function representation). In the image space, we exploit a 2D-to-3D generative model to recover a 3D CBCT image from an input panoramic X-ray image. In particular, to reduce the impact of other tissues with similar intensity values (e.g., neighboring teeth or alveolar bone) during the generation process, we further leverage a tooth segmentation module to guide more attention to the foreground tooth area. In this way, the 3D tooth model can be reconstructed from the segmentation mask accordingly. However, geometric properties are not considered in the image space, despite their importance in producing fine-grained tooth shape details. Hence, we further introduce an implicit function network to recover 3D tooth models in the geometric space. Note that it is a coordinate network defined in a continuous space where a large number of query points are sampled to determine surface occupancy (i.e., inside and outside), thus being able faithfully to represent tooth shapes at arbitrary resolutions. Eventually, we merge dual-space outputs and generate final 3D tooth models with detailed geometric properties. Our main contributions are summarized below:

- We propose a novel dual-space framework to reconstruct a 3D tooth model from the 2D panoramic X-ray image in the image space and the geometric space, respectively.
- In the image space, we leverage a 2D-to-3D generative model to predict intensities of the 3D CBCT image, which is further enhanced by a segmentation module to focus on the foreground tooth region. In the geometric space, an implicit function network is introduced in the continuous space to capture complicated tooth shapes with fine-grained details.
- Extensive experiments and ablation studies show that our proposed framework can generate accurate 3D tooth models from 2D panoramic X-ray images. Compared with state-of-the-art methods, our method can achieve superior results both qualitatively and quantitatively, demonstrating effectiveness of our proposed dual-space framework.

II. Related Work

A. 3D Tooth Segmentation and Generation

3D tooth model reconstruction from dental imaging (e.g., intra-oral scanning data, CBCT images, and panoramic X-ray images) is fundamental in digital dentistry, especially for orthodontic treatments. Existing approaches [12], [13], [14], [15] have achieved promising results in tooth instance segmentation on intra-oral scanning data or CBCT image. Typically, traditional methods usually apply region growing, level-set algorithms, or their variants to extract tooth models [16], [17], [18], [19], [20]. With the development of computer vision techniques, many learning-based methods have been proposed to surpass these traditional methods with better effectiveness and efficiency [4], [14], [15], [21], [22]. However, the intra-oral scanning data only contains tooth crowns without any root information, and CBCT image is not affordable in many dental clinics due to massive radiation and high cost.

Instead of segmenting individual teeth from the intra-oral scanning data or CBCT image, Laura et al. [7] propose to reconstruct 3D tooth models only from 2D panoramic X-ray image with B-spline interpolation and free-form surface. Further, inspired by the generative model in computer vision, Song et al. [10] propose to exploit a generative model to reconstruct 3D flattened images from a panoramic X-ray image, which is further deformed into a reconstructed CBCT image with a predicted dental arch curve. Unfortunately, this method can only recover intensities of the 3D CBCT image, whereas the 3D tooth model is not considered. Recently, Liang et al. [11] adopt a two-stage framework to first localize each tooth in a panoramic X-ray image, and then directly reconstruct 3D tooth models from respective X-ray patches. Although it is a feasible strategy for individual tooth model reconstruction, it is difficult to directly obtain 3D tooth models from 2D X-ray images, which also often produce artifacts around tooth models. Moreover, this method utilizes only image-level supervision to generate 3D tooth models, where many important tooth shape properties and details (e.g., tooth root number) are usually inaccurate. To effectively reconstruct 3D tooth models, we design to utilize the image space (i.e., intensity and segmentation masks) and geometry space (i.e., implicit surface function representation) as dual-space supervision in our framework.

B. 3D Reconstruction From 2D Input

In the computer vision and graphics community, 3D reconstruction from 2D inputs usually aims to restore 3D information (e.g., 3D image intensities or 3D shapes) from the

2D representation of the target. These tasks can be generally concluded as two categories based on the number of 2D views, i.e., 1) multi-view reconstruction and 2) single-view reconstruction.

For multi-view reconstruction, given a dense sampling of views, the typical methods [23], [24] design to reconstruct the scene with these sampled views through camera pose estimation in a unified coordinate system. Further, many works [25], [26] are proposed to synthesize a 3D target by optimizing a continuous implicit function with 2D views from different angles. Furthermore, in medical imaging applications, Liu et al. [27] propose a multi-view learning for disease diagnosis. Kasten et al. [28] propose a 2D-to-3D convolutional neural network to reconstruct the 3D knee bone from bi-planar X-ray images. Although these methods have been proven effective in different tasks, taking multi-view images as input is impossible in our specific task, since the panoramic X-ray image used in dental clinics is only a single-view 2D projection captured with a moving camera.

Recently, many works have been explored to reconstruct 3D images or shapes from a single-view image. For example, Henzler et al. [29] utilize a convolutional neural network to extend a single 2D image to a 3D image. Shen et al. [30] propose a deep learning algorithm to map a single-view radiograph to a corresponding 3D image. SCSCN [31] leverages a separated channel-spatial convolutional network to reconstruct 3D shape from a 2D image at any viewpoint. However, these methods are designed only to consider voxel-wise supervision in the image space (e.g., intensity distribution or segmentation masks), while ignoring geometric attributes of the target (e.g., shapes), which usually leads to many artifacts in the reconstructed results.

III. Method

This section presents a novel dual-space framework for 3D tooth model reconstruction from 2D panoramic X-ray images. An overview of our framework is shown in Fig. 3. We first briefly introduce the paired X-ray and CBCT data building, and then carefully explain the respective details of our dual-space architecture designed in the image space and the geometric space.

A. Data Building

Our goal is to obtain 3D tooth models accurately from a 2D panoramic X-ray image. We formulate this task as a tooth model reconstruction problem in the image space (i.e., CBCT image and 3D tooth masks) and the geometric space (i.e., tooth surface). To build this framework, we need paired panoramic X-ray images and CBCT images for supervised learning. However, the paired data with both CBCT and panoramic X-ray images is often unavailable in dental clinics, as collecting both data is time-consuming and also excessively radioactive. To this end, we alternatively design to synthesize pairwise X-ray images from CBCT images for the purpose of network learning.

In a typical panoramic X-ray imaging setup, a rotating arm holds both the X-ray source and the receptor. As the arm rotates, the X-ray beam focuses on a narrow area, capturing a continuous projected image of the dental arch. Leveraging this principle, we first identify the dental arch curve on CBCT images and project the CBCT image to generate multi-planar images based on the computed arch curve. Finally, we employ the ray-sum algorithm to simulate paired panoramic X-ray images. This simulation algorithm, widely utilized in panoramic X-ray image generation from CBCT images, has been extensively discussed in prior studies [10], [11], [32].

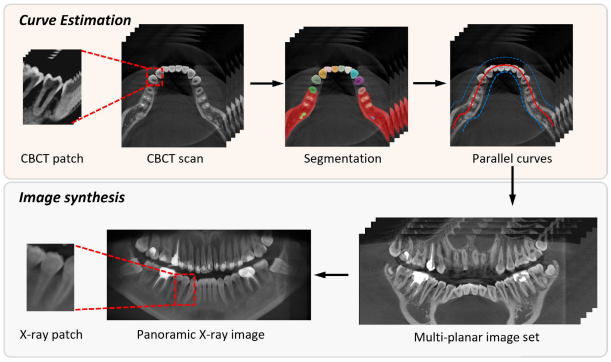


Fig. 2. Overview of panoramic X-ray patch generation. We first estimate the oral curve (i.e., red curve) and its two corresponding parallel curves (i.e., two blue curves) of the dental arch on the CBCT image. Afterward, we unwarp the image slices between the parallel curves to generate a multi-planar image set. Furthermore, we synthesize the pairwise 2D panoramic X-ray image by applying the ray-sum algorithm to the multi-planar image set. Moreover, the pairwise CBCT image patch and the X-ray image patch for each tooth can be cropped using the tooth center and the normal vector that is perpendicular to the oral curve.

Specifically, we first collect a set of CBCT scans $X = \{X_1, X_2, \ldots, X_N\}$ in the dental clinics and annotate the corresponding 3D tooth masks $Y = \{Y_1, Y_2, \ldots, Y_N\}$, where N denotes the total number of CBCT scans. Each voxel at a CBCT image $X_i \in \mathbb{R}^{H\times W\times D}$ has an intensity value, and its corresponding location in the label map $Y_i \in \mathbb{R}^{H\times W\times D}$ owns a value indicating property of belonging to tooth or background. Then, as shown in Fig. 2, to synthesize the pairwise 2D panoramic X-ray images, we first estimate the parallel oral curves by applying the thinning algorithm [33] and also cubic spline interpolation on mandible segmentation. In Fig. 2, the red curve denotes the estimated oral curve, while two blue curves denote two parallel oral curves covering the whole tooth region. After estimating oral curves, we unwarp image slices between parallel oral curves to generate a multi-planar image set. Finally, we synthesize the pairwise 2D panoramic X-ray image by applying the ray-sum algorithm onto the multi-planar image set [32]. Notably, we can also generate the pairwise CBCT patch (i.e., image patch x_i^t and mask patch y_i^t for the t-th tooth of the i-th patient) and X-ray image patch (i.e., z_i^t for the t-th tooth of i-th patient).

B. Tooth Reconstruction in the Image Space

Taking a 2D panoramic X-ray image, we first detect each tooth, and feed the cropped X-ray image patch into the network. To predict the 3D tooth model from a 2D X-ray image patch, the most intuitive way is to reconstruct the

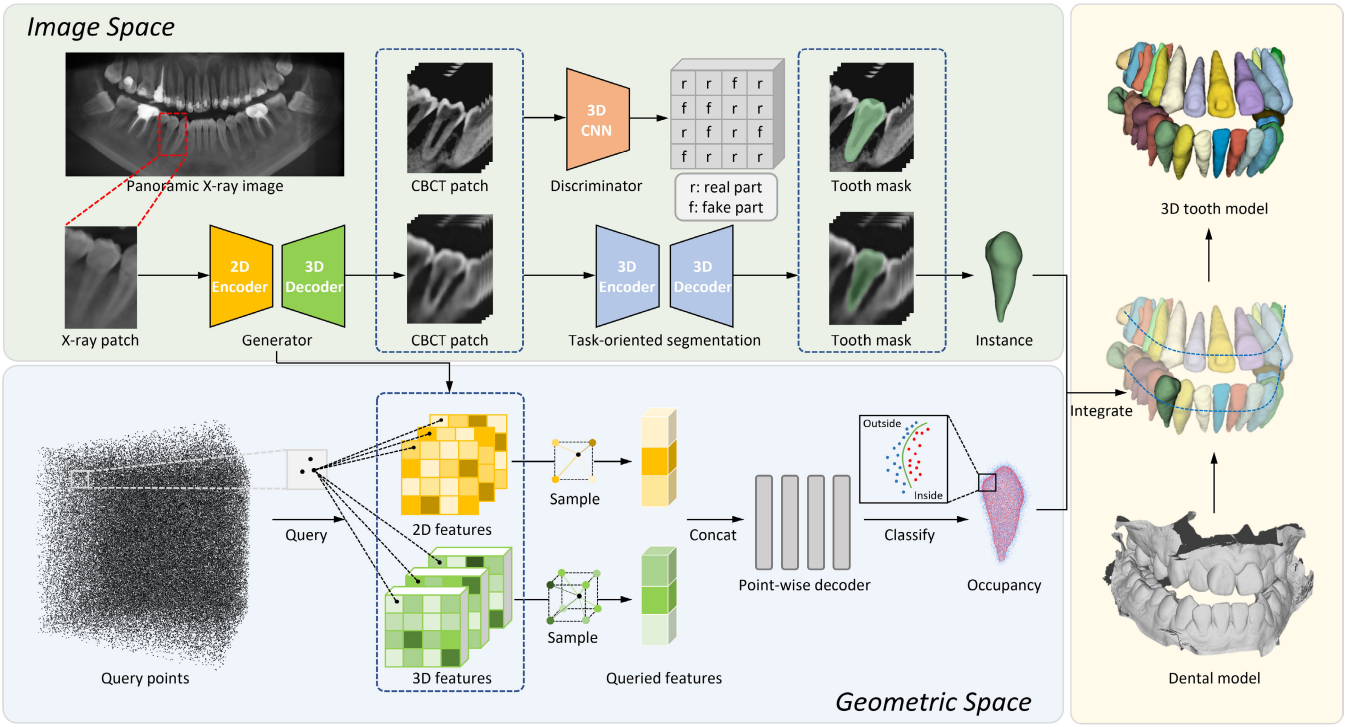


Fig. 3. Overview of DTR-Net. We design a dual-space framework to reconstruct a 3D tooth model from a 2D panoramic X-ray image. In the image space, we utilize the auto-encoder with a discriminator to generate corresponding CBCT image patches from 2D X-ray image patches, enhanced by a task-oriented segmentation module (for obtaining the tooth mask) to focus on the foreground tooth area. While, in the geometry space, we design an implicit function network to predict the occupancy of the query points (i.e., outside or inside the tooth model) around the tooth surface within the corresponding CBCT image patch. Eventually, we obtain a 3D tooth model by integrating the results from dual spaces, which are placed in their corresponding positions by the paired intra-oral scanned data.

corresponding 3D CBCT image patch, and then perform the tooth segmentation. By following this idea, we design three modules in the image space, including 1) a 2D-to-3D generator, 2) a discriminator, and 3) a tooth segmentation module. Nevertheless, our method differs from previous works since our proposed segmentation module is task-oriented, i.e., differentiable to the 2D-to-3D generation process, where the foreground teeth in each patch can be generated more accurately.

1) Generator: Since each pixel of the 2D X-ray image is represented as the accumulation of the CBCT voxels along the ray direction, we first apply a 2D-to-3D auto-encoder in our framework to learn the inverse mapping $\mathbf{G}$, expecting to recover 3D image intensities (i.e., a CBCT image patch) with the input of a 2D X-ray image patch, as defined below:

$$\mathbf{G}(z) : \mathbb{R}^{h\times w} \mapsto \mathbb{R}^{h\times w\times d}, \tag{1}$$

where z denotes a 2D X-ray image patch, and h, w, d refers to three dimension of the cropped patch size. The network is composed of a 2D encoder and a 3D decoder, with a feature transformation module to bridge them. Details of the generator architecture can be found in IV-C. To train the generator, an objective function $\mathcal{L}_{\mathbf{G}}(\hat{x}, x)$ of mean square error (MSE) is used to measure the error between the generated CBCT image patch $\hat{x}_i^t$ and the ground truth CBCT image patch x_i^t.

2) Discriminator: As the CBCT image reconstruction with voxel-wise supervision (i.e., using MSE) usually leads to over-smooth results with less high-frequency information, especially in the tooth boundary areas, we further adopt a commonly-used image-level discriminator to obtain a more realistic intensity distribution for the generated CBCT image, particularly in local tooth regions:

$$\mathbf{D}(x) : \mathbb{R}^{h\times w\times d} \mapsto [0,1]^{\frac{h}{s}\times\frac{w}{s}\times\frac{d}{s}}, \tag{2}$$

where x is the predicted CBCT image patch or the cropped CBCT image patch, and s is the downsampling factor. Each voxel in $\frac{h}{s}\times\frac{w}{s}\times\frac{d}{s}$ indicates whether the input is a real CBCT image patch or a generated CBCT image patch. During the network training, the generator and the discriminator are updated accordingly based on the minimax loss $\mathcal{L}_{\mathbf{D}}(\hat{x}, x)$.

3) Task-Oriented Tooth Segmentation: With the reconstructed CBCT image patch, we first utilize a segmentation network [34], which is pre-trained on real CBCT image patches to obtain tooth masks:

$$\mathbf{S}(x) : \mathbb{R}^{h\times w\times d} \mapsto [0,1]^{h\times w\times d}. \tag{3}$$

The output of the segmentation network is a probability map $\hat{y}$, with values ranging from 0 to 1, indicating the probability of belonging to the background or the target tooth. To train the segmentation network, we employ the objective function by combining the cross-entropy loss and the Dice based loss.

Although it is feasible to generate 3D tooth masks from 2D panoramic X-ray image patches with the 2D-to-3D generative network and the pretrained segmentation network, the intensity distribution gap between the generated CBCT image patches and the real CBCT image patches is ignored. This could lead

to unsatisfactory segmentation results for 3D tooth model reconstruction, especially at tooth boundaries with blurred signals or at the molars with significant shape variations. Recently, many works, especially in the medical imaging community, leverage one highly-related task to guide the targeted task, such as integrating segmentation task into classification or registration task [35], [36]. Inspired by this, instead of simply fixing the parameters of the pretrained segmentation network, we further train the 2D-to-3D generative network and the segmentation network in a collaborative way, where the segmentation network is trained by both generated and real CBCT image patches.

Moreover, to further guide the generator to pay more attention to the foreground tooth regions, we update the optimization strategy for the CBCT image reconstruction process. Specifically, the generator $\mathbf{G}$ and the discriminator $\mathbf{D}$ are first optimized with the initial objective functions $\mathcal{L}_{\mathbf{G}}$ and $\mathcal{L}_{\mathbf{D}}$ in the first 50 epochs. For the rest epochs, with the probability map $\hat{y}$ predicted by the segmentation network, we transfer it as a weighted map to supervise the generator, as defined below:

$$\mathcal{L}'_*(\hat{x}, x) = \mathcal{L}_*(\hat{x}, x) \cdot (\lambda_f \cdot \hat{y} + \lambda_b \cdot (1 - \hat{y})), \tag{4}$$

where $\mathcal{L}'_*(\hat{x}, x)$ refers to the updated loss function $\mathcal{L}'_{\mathbf{G}}(\hat{x}, x)$ or $\mathcal{L}'_{\mathbf{D}}(\hat{x}, x)$. And we set $\lambda_f = 0.7$ for the foreground voxels, and $\lambda_b = 0.3$ for the background voxels, respectively.

C. Tooth Reconstruction in the Geometric Space

Although tooth reconstruction in the image space can recover the intensity distribution of the CBCT image and segment foreground teeth, many geometric properties of the tooth shape are not considered, especially at the tooth roots with large variation. To address this problem, we adopt an implicit function network to further represent and reconstruct each tooth model in the geometric space. The network is defined in a continuous field, such as the signed distance function or occupancy field, which can represent each tooth surface concisely at arbitrary spatial resolutions, and faithfully capture the fine-grained geometric properties. Note that, the tooth surface is subsequently reconstructed from the segmented tooth mask using the widely-used Marching-Cube algorithm. The details of the implicit function network for tooth reconstruction are given below, including 1) multi-scale and long-range encoding, 2) point-wise occupancy decoding, and 3) confidence-aware dual-space integration.

1) Multi-Scale and Long-Range Feature Encoding: Generally, in the implicit function network, given a CBCT image patch x_i^t, we sample a number of query points $\mathcal{P}$ in the 3D space within x_i^t. We aim to predict the occupancy $\hat{o}_i^t \in [0, 1]$ of the query points, i.e., outside or inside the 3D tooth surface. To obtain representative query points in the 3D space, we first randomly sample 10^5 points on the tooth surface, and move these points with random displacements based on Gaussian distribution. In this manner, these sampled points can cover the whole space of the CBCT image patch. More importantly, most of these query points are located around the tooth surface for effectively representing the complicated tooth shapes.

After sampling these query points, the next step is to obtain point-wise deep features and predict the tooth occupancy field. We design a point-wise encoding based on the 2D-to-3D generator, which produces multi-scale feature maps for point-wise querying in different dimensions, i.e., 2D for the encoder, and 3D for the decoder. Specifically, as shown in Fig. 3, for each 3D query point p, we get its features on the 2D and 3D multi-scale feature maps, respectively. The feature of point p on a 3D feature map can be easily interpolated according to Euclidean distances on neighboring grids. For the 2D feature maps, we first project the 3D query point p onto 2D space along the X-ray direction, and then employ the same feature interpolation in the 2D space. Finally, we concatenate these feature vectors together as the final feature vector $\mathcal{F}_p$ of point p.

We further notice that the point-wise features are queried entirely from the convolutional feature maps. However, the fixed-size convolutional kernel and the limited receptive field usually inhibit the network's ability to capture long-term dependencies [37], [38], particularly when these dependencies are distantly located. On the other hand, self-attention is inherently better at capturing these long-term dependencies, making them particularly suited for the global tooth shape information in our task. Accordingly, we design a respective method to capture both global and local tooth information. Specifically, given the input 2D or 3D feature map before the point-wise feature querying, we first apply three fully connected layers to obtain the query features, key features, and value features, respectively. We then compute the correlation matrix between the query features and key features, effectively modeling spatial and channel-wise relationships across the feature map irrespective of distance. Finally, the value features are reweighted by this correlation matrix and added to the original feature maps to generate the final feature maps. In this way, the concatenated point-wise features would thus include both global and local tooth information, and contribute to more reliable occupancy results.

2) Point-Wise Occupancy Decoding: To predict the occupancy of a query point p with its feature vector $\mathcal{F}_p$, we employ a multi-layer perceptron (MLP) that takes $\mathcal{F}_p$ as input, and conduct a binary classification to indicate whether the point is outside or inside the tooth surface, as defined below:

$$\mathbf{MLP}(\mathcal{F}_p) : \mathbb{R}^C \mapsto [0, 1], \tag{5}$$

where C is the length of the concatenated feature vector, 0 and 1 indicate outside and inside, respectively. To train the implicit function network, we use the cross entropy loss $\mathcal{L}_{\mathbf{MLP}}(\hat{o}, o)$ to measure error between the predicted occupancy $\hat{o}$ and the ground-truth occupancy o.

3) Confidence-Aware Dual-Space Integration: In the continuous geometric space, the learned implicit function can predict accurate point-wise occupancy $\hat{o}_i^t$ to represent detailed tooth shapes. Along with the voxel-wise tooth segmentation in the image space, the overall framework can provide more reliable results by integrating the dual spaces. Specifically, our proposed segmentation module outputs a probability map $\hat{y}_i^t$, indicating the probability of each voxel belonging to the foreground tooth. However, due to complicated tooth shapes and limited intensity contrast of reconstructed CBCT images, the regions near tooth boundaries usually have uncertain

segmentations. To eventually utilize the results from both image space and geometric space, we integrate the segmentation probability map and the occupancy classification probability. Specifically, integration of predictions from these two spaces is only carried out during the inference stage. This is achieved by calculating the confidence score (represented by the entropy) for each voxel's prediction, defined as:

$$f(h,w,d) = \begin{cases} \hat{y}_i^t(h,w,d) \text{ if } H_y(h,w,d) < H_o(h,w,d) \\ \hat{o}_i^t(h,w,d) \text{ if } H_y(h,w,d) \geq H_o(h,w,d), \end{cases} \tag{6}$$

where $H_*(h,w,d)$ represents the entropy at position (h,w,d) (i.e., the entropy of occupancy is $H_o(h,w,d) = -\hat{o}_i^t(h,w,d) \cdot \ln \hat{o}_i^t(h,w,d)$ and the entropy of segmentation is $H_y(h,w,d) = -\hat{y}_i^t(h,w,d) \cdot \ln \hat{y}_i^t(h,w,d)$). Note that a prediction with lower entropy is indicative of higher confidence. Consequently, this prediction with higher confidence is selected as the final prediction f for each specific voxel.

To place the individual tooth models into the accurate dental cavity position, we employ intra-oral scanning data as guidance. Since the tooth crowns in our reconstructed tooth model and the intra-oral scanning data should be identical, as they are obtained from the same patient, we employ a rigid registration algorithm such as Iterative Closest Point (ICP) to integrate the 3D tooth model with the corresponding intra-oral scan. This process aligns the tooth model to the intra-oral scan by minimizing the distance between corresponding points, thus ensuring precise matching within the 3D dental cavity model.

D. Overall Objective Function

The overall objective function for our proposed DTR-Net is defined as

$$\mathcal{L} = \lambda_{\mathbf{G}}\mathcal{L}_{\mathbf{G}} + \lambda_{\mathbf{D}}\mathcal{L}_{\mathbf{D}} + \lambda_{\mathbf{MLP}}\mathcal{L}_{\mathbf{MLP}} + \lambda_{\mathbf{S}}\mathcal{L}_{\mathbf{S}}, \tag{7}$$

where $\lambda_{\mathbf{G}}$, $\lambda_{\mathbf{D}}$, $\lambda_{\mathbf{MLP}}$, and $\lambda_{\mathbf{S}}$ are the hyperparameters to balance different modules during the network training.

IV. Experiments

In this section, we first introduce how we build our dataset, i.e., the paired X-ray image patches and CBCT image patches. We then present the metrics that quantitatively measure the performance in both image space and geometric space. Finally, we provide implementation details for training our framework, and also the pipeline to reconstruct 3D tooth model from a 2D panoramic X-ray image in the inference stage.

A. Dataset

To train the network, we first collect 40 CBCT scans with the spacing of $0.16 \times 0.16 \times 0.16$ from a dental clinic, where 70% is used for training, 10% for validation, and 20% for testing. These CBCT images are collected from Shanghai NinthPeople's Hospital. The dataset is approved by the ResearchEthics Committee, and the reference number is SH9H-2021-T169-1. Most of these patients were seeking orthodontic and implant treatments, indicating that there are teeth crowding, missing, and misalignment problems in this dataset. As for age information, they range from 11 to 57 years old. Moreover, as described in Fig. 2, since there are no paired 2D panoramic X-ray images in dental clinics, we utilize real-world CBCT images to synthesize paired X-ray image patches and CBCT image patches for each tooth. In this way, we produce 900 tooth-wise pairs of 2D panoramic X-ray image patches (with a size of 96×128) and CBCT image patches (with a size of $96 \times 96 \times 128$), where each CBCT image patch has a corresponding annotated mask (with a size of $96 \times 96 \times 128$). To learn the implicit representation of each tooth shape, we randomly sample and train with a total of 10^5 query points on the surface reconstructed from the tooth mask. Then, random displacements with Gaussian distributions are added to these query points, and their occupancies (i.e., outside or inside the surface) are used for supervision. More specifically, we use two Gaussian distributions with deviations of $\sigma_1 = 0.03$ and $\sigma_2 = 0.1$, for half of the points, respectively. Note that, for each iteration during the training, we only randomly sample 50000 query points from this pool to train the implicit function network, which is relatively small compared to the 3D CNNs operated on volumetric data. In conclusion, in the training stage, each tooth sample consists of a group of data, with a 2D panoramic X-ray image patch, a ground-truth 3D CBCT image patch, an annotated tooth mask of the CBCT image patch, and 10^5 query points with occupancy. In the inference stage, the total number of query points is based on the resolution of the 3D CBCT images (i.e., voxel grids of $96 \times 96 \times 128$). This ensures the output of the image space and geometric space to be of equal size, allowing them to be integrated into the final stage.

Notice that 3D tooth shapes vary greatly between tooth categories, as each type owns distinctive shapes according to their respective roles for chewing. To this end, we discuss and analyze the results for each specific tooth category in this paper. As shown in Fig. 4, we denote the tooth categories as T1-T7 for illustration purposes, based on the dental notation system [39].

B. Evaluation Metrics

To quantitatively evaluate the performance of our method, we introduce four metrics in both image and geometric space. We use the structural similarity (SSIM) metric to measure the quality of CBCT image reconstruction. For the 3D tooth model reconstruction task, we utilize three metrics to validate the performance at both image and surface level, including the Dice coefficient (Dice), Hausdorff Distance (HD), and Average Surface Distance (ASD). More specifically, the Dice metric is performed on the segmented tooth mask, while the HD and ASD metrics are measured on the generated tooth surface, respectively, to indicate the maximum and average surface distances.

C. Implementation Details

Our framework was implemented on the PyTorch platform with Pytorch Lightning library, and trained on an NVIDIA A100 GPU with the Adam optimizer. The learning rate is initially set to $3e^{-4}$, and decays by 0.1 for every 50 epochs.

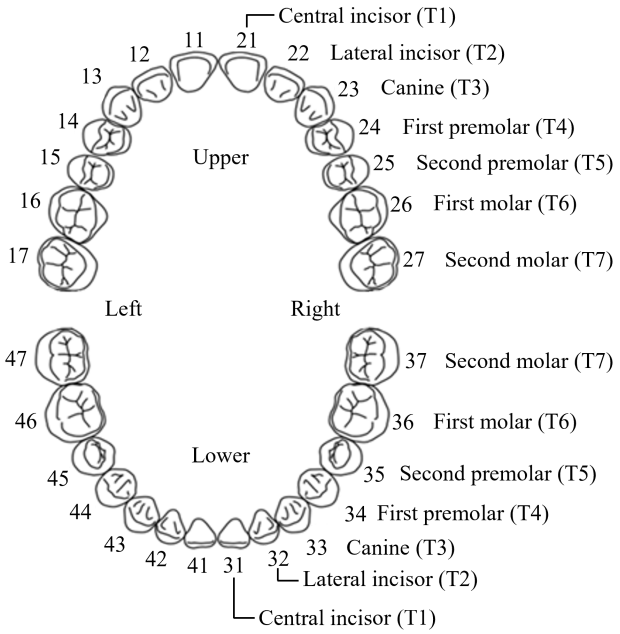


Fig. 4. Tooth categories in the dental notation system. The oral cavity is split into four quadrants: upper left, upper right, lower right, and lower left. Each quadrant has seven teeth of different types. The two digits of each tooth respectively represent its quadrant and its number from the midline of the face. Specifically, wisdom teeth are excluded from this study due to their limited samples. In this paper, we denote T1-T7 as different categories of teeth for convenience (e.g., T1 represents the set of teeth with ID 11, 21, 31, 41).

Notably, for the discriminator, it is updated only when the loss $\mathcal{L}_{\mathbf{PD}}$ is greater than a threshold of 0.1. As for the overall loss function, the hyperparameters for different loss terms set as $\lambda_{\mathbf{G}} = 1$, $\lambda_{\mathbf{D}} = 0.1$, $\lambda_{\mathbf{MLP}} = 0.1$, and $\lambda_{\mathbf{S}} = 0.2$ in this paper. In regard to the network architecture, our DTR-Net framework is composed of dual space modules. In the image space, we adapt the network architecture from DRR2CT's 2D-to-3D autoencoder [30], which consists of four convolutional blocks and four transposed convolutional blocks. Our 3D discriminator comprises five 3D convolution blocks. Each of these blocks contains a single convolution layer, a batch normalization layer, and a LeakyReLU layer. For the task-oriented segmentation network, we employ the V-Net architecture, a widely-used method in this domain [34]. In the geometric space, the feature point sampling strategy and the architecture of the MLP layers are directly adapted from IF-Net [40].

During the inference time, we first locate the center point of each tooth on the 2D panoramic X-ray image with a center-based detection network. Like the training stage, we crop the X-ray image into several small patches centered at each tooth, with a fixed size of $96 \times 96 \times 128$. Then, we take the cropped X-ray patches as input to our dual-space framework for reconstructing the 3D tooth models. Finally, with the curve of the dental arch, estimated from the panoramic X-ray image or optionally provided by the paired intra-oral scanning data (if available), we put the reconstructed 3D tooth models back to their corresponding positions in the 3D oral space.

V. Results

To demonstrate the advantage of our method, we first compare our framework with several state-of-the-art methods. Then, we conduct extensive ablation experiments to validate the effectiveness of our dual-space design, including task-oriented segmentation in the image space and also the implicit function network in the geometric space.

A. Comparison

For comparison, we implement several recent state-of-the-art methods, which are highly related to 3D tooth reconstruction from 2D panoramic X-ray image, including:

- DRR2CT [30]: It is a 2D-to-3D auto-encoder network for 3D CT image generation from a 2D panoramic X-ray image, basically achieved by a transformation module that converts 2D feature maps to 3D feature maps.
- Oral-3D [10]: It is also a 2D-to-3D auto-encoder network for 3D oral cavity reconstruction from a 2D panoramic X-ray image. This method utilizes a deformation module based on convolution structures with dense connections to unwarp the generated 3D flatten images into a 3D oral cavity.
- X2Teeth [11]: Notice that it is specifically designed for 3D tooth model reconstruction from 2D panoramic X-ray image, which conducts the same task as in this paper. It first exploits a 2D localization network to detect the bounding box of each tooth from 2D panoramic X-ray image, and then involves a 3D tooth model reconstruction network to estimate 3D tooth model directly from localized 2D features.

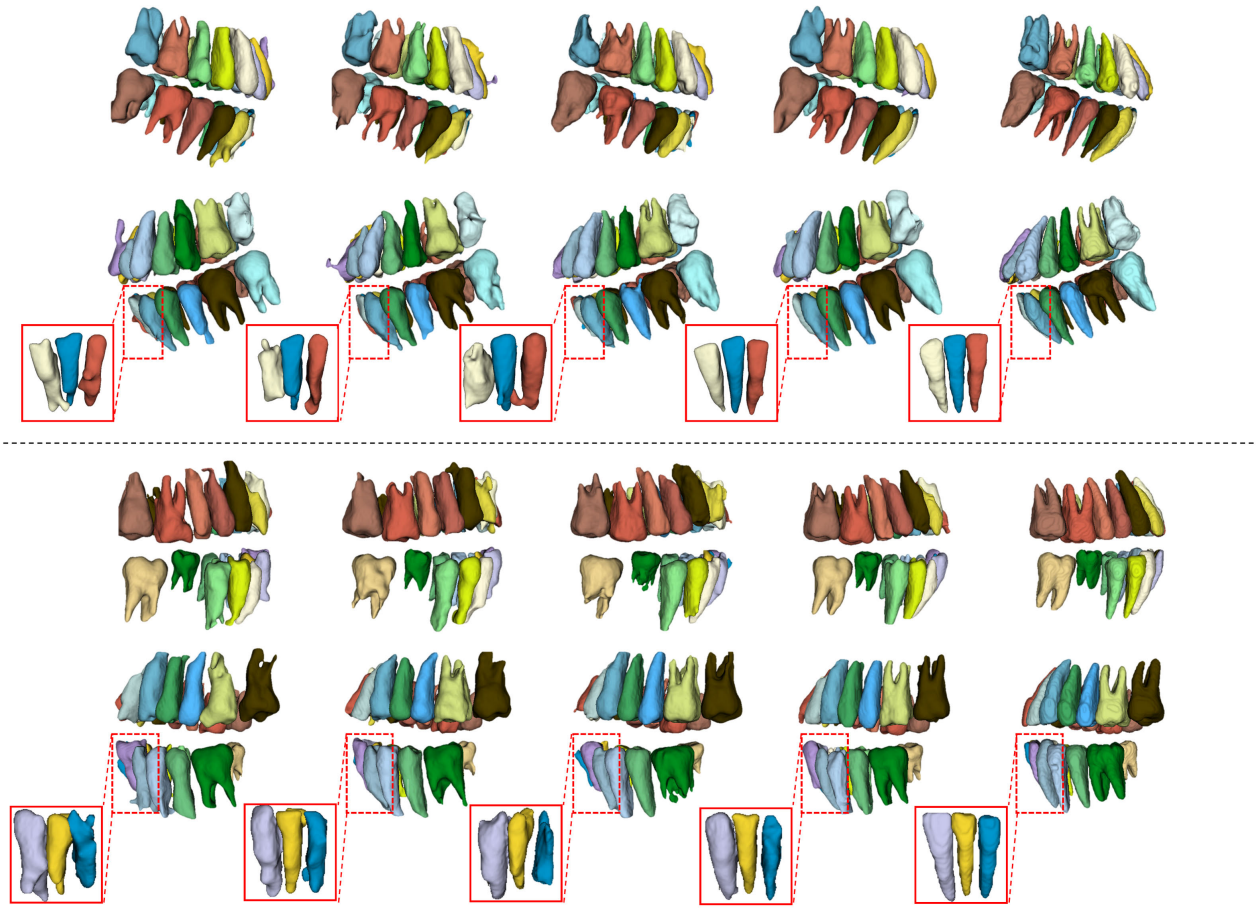
The overall 3D tooth model reconstruction results are presented in Table I, where our method significantly outperforms other state-of-the-art methods in terms of all metrics. Concretely, since original DRR2CT and Oral-3D only generate CBCT images and do not acquire 3D tooth models, we adopt the same pre-trained segmentation network used in our method to segment tooth instances from the cropped CBCT patches. X2Teeth directly reconstructs 3D tooth models without intermediate CBCT image reconstruction. Thus the image reconstruction metric (i.e., SSIM) is not available for quantitative comparison.

Compared with DRR2CT, which simply performs a 2D-to-3D auto-encoder to reconstruct CBCT image from X-ray image, our proposed method achieves remarkable improvement with respect to SSIM score (2.25%), Dice score (2.14%), and HD error (2.22mm), demonstrating the advantages of our proposed dual-space network architecture. Moreover, compared to Oral-3D that only uses a patch discriminator to reconstruct 3D CBCT images, our method leads to 4.01% improvement in Dice score and 2.45mm reduction in HD error, respectively. Most notably, compared to the state-of-the-art performance produced by X2Teeth, our method further boosts the Dice score from 84.75% to 86.11%, and reduces the HD error from 1.37mm to 1.20mm. We also quantify results based on each tooth type in Table I. Note that the ASD errors of molars (i.e., first molar T6 and second molar T7) with the most complicated shapes have been significantly reduced.

TABLE I

STATISTICAL COMPARISON OF 3D TOOTH MODEL RECONSTRUCTION BY OUR DTR-NET AND THREE STATE-OF-THE-ART METHODS

	Method	T1	T2	T3	T4	T5	T6	T7	Mean
SSIM (%) ↑	DRR2CT	87.37	87.83	84.99	88.04	86.14	78.36	77.70	84.08±0.05
	Oral-3D	86.34	85.84	83.95	86.43	85.32	78.54	76.24	83.12±0.05
	X2Teeth	-	-	-	-	-	-	-	-
	DTR-Net	**89.81**	**90.74**	**87.59**	**89.93**	**88.49**	**79.96**	**79.11**	**86.32 ± 0.05**
Dice (%) ↑	DRR2CT	76.51	80.91	86.45	89.45	86.52	83.59	86.53	83.97±0.09
	Oral-3D	75.08	76.45	84.37	87.38	85.15	83.53	83.11	82.10±0.09
	X2Teeth	75.73	80.64	87.40	90.45	87.79	83.63	86.03	84.75±0.09
	DTR-Net	**79.61**	**84.13**	**88.14**	**90.97**	**88.21**	**85.88**	**88.29**	**86.11±0.08**
HD (mm) ↓	DRR2CT	3.74	4.22	3.67	3.91	3.70	2.25	2.13	3.43±1.73
	Oral-3D	4.16	4.42	3.83	4.08	4.04	2.39	2.42	3.65±1.52
	X2Teeth	2.27	1.64	1.12	0.75	0.96	1.53	1.39	1.37±0.82
	DTR-Net	**1.78**	**1.47**	**0.92**	**0.68**	**0.89**	**1.26**	**1.06**	**1.20±0.91**
ASD (mm) ↓	DRR2CT	0.76	0.68	0.51	0.40	0.50	0.47	0.44	0.56±0.29
	Oral-3D	0.79	0.81	0.58	0.51	0.60	0.48	0.55	0.63±0.28
	X2Teeth	0.59	0.44	0.32	0.22	0.29	0.41	0.41	0.39±0.20
	DTR-Net	**0.49**	**0.35**	**0.28**	**0.21**	**0.28**	**0.35**	**0.33**	**0.34±0.22**



(a) DRR2CT (b) Oral-3D (c) X2Teeth (d) DTR-Net (e) GT

Fig. 5. Comparison with three state-of-the-art methods. We visualize two typical cases of reconstructed 3D tooth models (two row) by different methods (column). We demonstrate that DTR-Net (1) succeeds in capturing significant shape variance for different teeth, especially for the molar teeth, and (2) shows effectiveness in local-detail restoration compared with other three methods.

All these results indicate effectiveness of using both the image space (i.e., CBCT image and tooth mask) and the geometric space (i.e., implicit surface function representation), where the fine-grained tooth shape details can be produced accurately in the reconstructing 3D tooth models.

To further illustrate the advantage of our proposed method, the visual comparison of two typical examples is shown in Fig. 5. It can be observed that the 3D tooth models produced by our method match better with the ground truth, especially at tooth roots with significant shape variations. In addition, the competing methods without tooth segmentation modules (i.e., DRR2CT and Oral-3D) fail to recover complex tooth shapes, where many artifacts and discontinuous surfaces are introduced. This manifests that, with the

TABLE II
ABLATION STUDY OF OUR METHOD WITH DIFFERENT VARIANTS, INCLUDING THE TASK-ORIENTED SEGMENTATION, SELF-ATTENTION, AND DIFFERENT SPACES (I.E., IMAGE-, GEOMETRIC- AND DUAL-SPACE)

Method	Space	Dice (%) ↑	SSIM (%) ↑	HD (mm) ↓	ASD (mm) ↓
B-Net	Image-space	81.99±0.55	83.47±0.10	3.17±1.66	0.60±0.31
TS-Net		85.42±0.12	86.13±0.08	1.27±0.93	0.36±0.24
G-Net	Geometric-space	84.89±0.10	85.97±0.09	1.30±0.87	0.38±0.27
SE-Net		85.45±0.10	86.27±0.07	1.23±0.75	0.35±0.22
DTR-Net	Dual-space	**86.11±0.08**	**86.33±0.05**	**1.20±0.68**	**0.34±0.22**

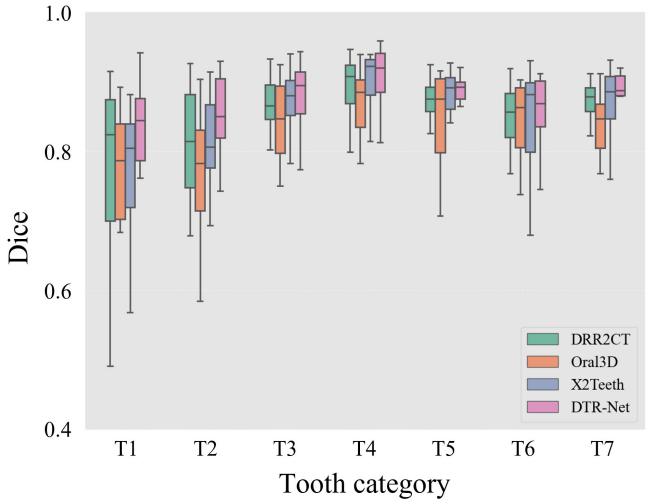


Fig. 6. Quantitative comparisons of different methods on different tooth categories (T1-T7). The Dice scores of our proposed method and state-of-the-art methods are compared through boxplots.

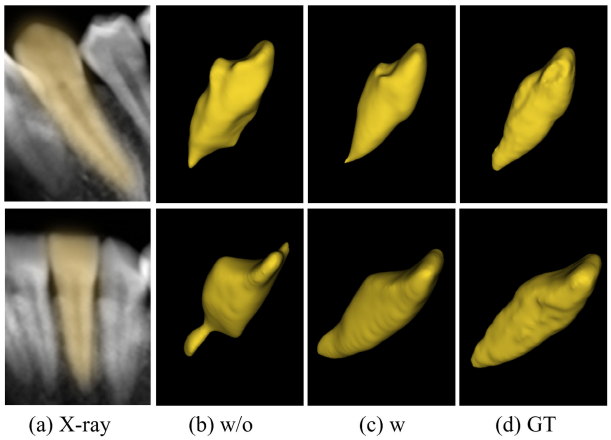


Fig. 7. Comparison of 3D tooth model reconstruction with or without using the task-oriented segmentation module: (a) two 2D panoramic X-ray image patches; (b) 3D tooth model reconstruction without using the task-oriented segmentation; (c) 3D tooth model reconstruction using the task-oriented segmentation module; (d) ground truth.

involvement of the segmentation module, our framework can capture more effective semantic information for tooth shape modeling. Furthermore, although X2Teeth also introduces a tooth segmentation module, as shown in the 3rd column of Fig. 5, the tooth geometric details still cannot be accurately preserved. This is because it directly segments 3D tooth instances from 2D X-ray images, ignoring the core supervision from CBCT image reconstruction in the image space and tooth surface reconstruction in the geometric space. Generally, the qualitative results shown in Fig. 5 are consistent with the quantitative results given in Table I, which further demonstrates the effectiveness of our proposed method for 3D tooth model reconstruction from 2D X-ray panoramic image.

B. Ablation Studies

We conduct extensive experiments to demonstrate the effectiveness of each of our proposed modules, including the task-oriented segmentation module in the image space and the implicit function network in the geometric space, respectively. As shown in Table II, we present our proposed method in five configurations: (1) The baseline network, denoted as B-Net, directly reconstructs CBCT patches from X-ray patches with a discriminator, followed by a pretrained segmentation network to segment individual teeth from reconstructed CBCT image patches. (2) Instead of simply using a pretrained segmentation network on the CBCT image patches, we train the CBCT image reconstruction and the tooth segmentation in a collaborative way as introduced in Sec. III-B.3, denoted as TS-Net. (3) We perform an implicit function network only in the geometric space, denoted as G-Net. (4) Based on G-Net, the self-attention module is employed in the feature map. (5) The full configuration, namely DTR-Net, aims to reconstruct 3D tooth models in both image space and geometric space.

1) Effectiveness of Task-Oriented Segmentation Module: Compared with the baseline B-Net, TS-Net (i.e., B-Net with task-oriented segmentation) involves both generated and real CBCT image patches for tooth instance segmentation in a collaborative way, by effectively bridging these two domains. Moreover, the predicted tooth mask of the generated CBCT patch can also benefit the CBCT image reconstruction, and promote network attention to the foreground tooth regions. As shown in Table II, TS-Net significantly improves tooth reconstruction accuracy in terms of all metrics (e.g., improving 3.43% for Dice score and reducing 1.9mm for HD). Moreover, we present two typical visual results in Fig. 7. For these two cases of incisors, we notice that the baseline B-Net completely fails to recover the tooth crown and root shapes, while the TS-Net can produce results similar to the ground truth. These results show that the task-oriented segmentation module can provide more effective semantic information to capture the global tooth shape.

2) Effectiveness of Self-Attention Module: In comparison to the implicit function network in the geometric space (i.e., G-Net), our approach employs a self-attention module. This feature enhances the capture of long-term dependencies,

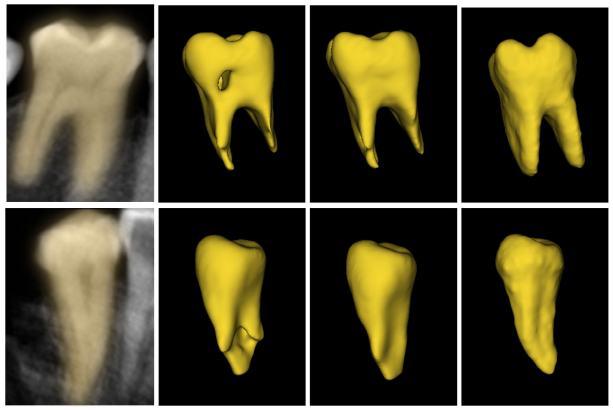


Fig. 8. Comparison of 3D tooth model reconstruction with or without using the implicit function network in the geometric space: (a) two 2D panoramic X-ray image patches; (b) 3D tooth model reconstruction without using the implicit function network; (c) 3D tooth model reconstruction using the implicit function network; (d) ground truth.

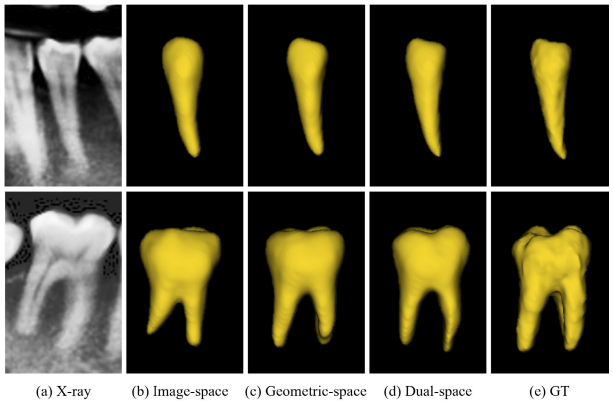


Fig. 9. The qualitative results of the tooth models reconstructed from different spaces, including the image space, geometric space, and dual space.

optimizing the framework for processing global tooth shape information. To validate the effectiveness of this approach, we have observed that incorporating the self-attention module (implemented in SE-Net) consistently improves outcomes across all metrics, such as an improvement of 0.07mm in terms of Hausdorff Distance (HD).

3) Effectiveness of the Implicit Function Network: In our full framework, we augment TS-Net with an implicit function network in the geometric space. This encourages the network to recover more geometric details for the 3D tooth models, especially at the tooth roots with significant shape variation. As illustrated in Table II, compared to the networks designed only in the image space (i.e., TS-Net) or the geometric-space (i.e., SE-Net), our full method based on the dual-space, DTR-Net, achieves the best performance. Also, qualitatively, as shown in Fig. 8 and Fig. 9, the tooth models match better with the ground truth, by carrying more fine-grained tooth details. All these results indicate that our full method built in the dual spaces can generate accurate 3D tooth models from 2D panoramic images.

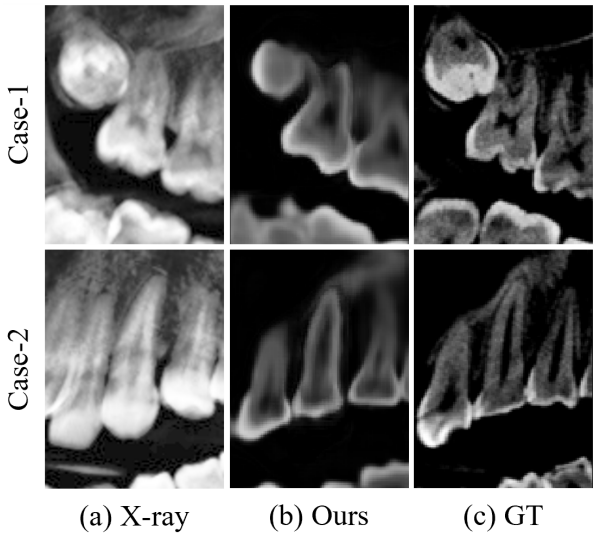


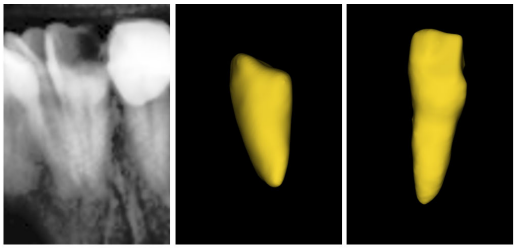
Fig. 10. Two typical reconstructed CBCT patches of our framework.

VI. Discussion

In digital dentistry, tooth model reconstruction is the prerequisite for dental diagnosis and treatment planning, for which panoramic X-ray images would be the most cost-effective modality. Unfortunately, despite great significance, tooth model reconstruction from panoramic X-ray images has not been well studied. This task involves major challenges in reconstructing 3D tooth models, including the ill-posed problem of recovering actual 3D space from 2D images, and the complexity of tooth shapes. Current methods fail to provide reliable tooth shapes, as they simply study the intensity distribution in the image space, without learning the foremost geometric properties. To address this problem, we present a novel dual-space framework by considering both perspectives in the image space and the geometric space.

Our method jointly learns through both image space and geometric space, to supervise the learning of image intensity and 3D shapes, respectively. Specifically, the image space provides the basic intensity distribution of the target tooth, which is further enhanced by the task-oriented segmentation to focus on the foreground tooth area. Notably, the geometric space is learned gracefully with a coordinate network, which enjoys continuous space to study the geometric properties of tooth shapes and contours. Extensive experiments have demonstrated that our dual-space method surpasses state-of-the-art methods quantitatively and qualitatively. More specifically, image generation with task-oriented segmentation in the image space, and the implicit function network in the geometric space have both shown their effectiveness in our ablation studies.

Additionally, our framework not only generates solid tooth models, but also 3D CBCT patches from X-Ray images, offering more comprehensive texture and oral information, as shown in Fig.10. However, this is merely an intermediate step in image processing. The detailed anatomical structures, such as the root canal and alveolar bone, are not reliably reproduced in this step (with a reliability measure of 19.45 dB),



(a) X-ray (b) Ours (c) GT

Fig. 11. The failure case of reconstructed models with overlapped teeth in the 2D panoramic X-ray image.

making this approach unsuitable for diagnosing these diseases, especially in their early stages. Therefore, our framework is more beneficial for addressing dental issues that rely on solid tooth models, such as orthodontic planning. It is not suitable for diagnosing dental diseases that require precise image density information.

While our method has demonstrated impressive performance, it still has some limitations. 1) As the paired panoramic X-ray and CBCT images are not available in real clinical scenarios, we alternatively utilize DRR to simulate a paired panoramic X-ray image from each CBCT image, and supervise the network by training with these data pairs. This yet leaves a domain gap between authentic panoramic X-ray images and our simulated ones. 2) Also, panoramic X-ray images often suffer from severe ambiguity in the overlapping areas to separate adjacent teeth. In this paper, we manage to alleviate this impact by focusing on the foreground tooth with the help of task-oriented segmentation. However, the CBCT patch generated from the X-ray patch often contains information from adjacent teeth, which causes ambiguity to likely mislead the reconstruction, as shown in Fig.11. 3) Learning-based methods commonly yield smoother results, as also observed in the computer vision community when reconstructing 3D facial or human body models from 2D images. Furthermore, the use of linear interpolation to derive query point features from feature maps leads to a loss of high-frequency details and thus a smoother output. To capture more shape details in the reconstruction, a promising strategy is to incorporate 2D hints, such as root landmarks. In future work, we also plan to incorporate information from 3D intra-oral scanning data, a non-radiology data source with detailed tooth crown information, to guide and recover tooth surface details from 2D X-ray images.

VII. Conclusion

In this paper, we have presented a novel dual-space method, namely DTR-Net, to reconstruct 3D tooth models from 2D panoramic X-ray images. Considering no large set of pairwise panoramic X-ray images and CBCT images in real dental clinics, we propose to simulate X-ray image patches from CBCT images to build data pairs for network training. We further solve our problem in both image space and geometric space. Specifically, in the image space, we first apply a 2D-to-3D generative model to recover intensities of the CBCT image, guided by a task-oriented segmentation module. In the geometric space, we leverage the implicit function network to learn important geometric details of tooth shapes. Extensive experiments demonstrate that our proposed DTR-Net can effectively reconstruct 3D tooth models (even with complicated tooth shapes) by restoring local geometric details.

References

[1] W. E. Harrell Jr., "3D diagnosis and treatment planning in orthodontics," *Seminars Orthodontics*, vol. 15, no. 1, pp. 35–41, 2009.

[2] N. Shah, N. Bansal, and A. Logani, "Recent advances in imaging technologies in dentistry," *World J. Radiol.*, vol. 6, no. 10, p. 794, 2014.

[3] E. Corbet, D. Ho, and S. Lai, "Radiographs in periodontal disease diagnosis and management," *Austral. Dental J.*, vol. 54, pp. S27–S43, Sep. 2009.

[4] Y. Chen et al., "Automatic segmentation of individual tooth in dental CBCT images from tooth surface map by a multi-task FCN," *IEEE Access*, vol. 8, pp. 97296–97309, 2020.

[5] G. Magat, E. Oncu, S. Ozcan, and K. Orhan, "Comparison of cone-beam computed tomography and digital panoramic radiography for detecting peri-implant alveolar bone changes using trabecular micro-structure analysis," *J. Korean Assoc. Oral Maxillofacial Surgeons*, vol. 48, no. 1, pp. 41–49, Feb. 2022.

[6] P. Kamrani and N. M. Sadiq, *Anatomy, Head and Neck, Oral Cavity (Mouth)*. Treasure Island, FL, USA: StatPearls, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK545271/

[7] L. Mazzotta, M. Cozzani, A. V. Razionale, S. Mutinelli, A. Castaldo, and A. Silvestrini-Biavati, "From 2D to 3D: Construction of a 3D parametric model for detection of dental roots shape and position from a panoramic radiograph—A preliminary report," *Int. J. Dentistry*, vol. 2013, Mar. 2013, Art. no. 964631.

[8] S. Barone, A. Paoli, and A. V. Razionale, "Geometrical modeling of complete dental shapes by using panoramic X-ray, digital mouth data and anatomical templates," *Comput. Med. Imag. Graph.*, vol. 43, pp. 112–121, Jul. 2015.

[9] F. Wu, V. Matov, and J. Chen, "Three-dimensional tooth modeling using a two-dimensional X-ray image," U.S. Patent 10 076 389, Sep. 18, 2018.

[10] W. Song, Y. Liang, J. Yang, K. Wang, and L. He, "Oral-3D: Reconstructing the 3D bone structure of oral cavity from 2D panoramic X-ray," 2020, *arXiv:2003.08413*.

[11] Y. Liang, W. Song, J. Yang, L. Qiu, K. Wang, and L. He, "X2Teeth: 3D teeth reconstruction from a single panoramic radiograph," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2020, pp. 400–409.

[12] X. Xu, C. Liu, and Y. Zheng, "3D tooth segmentation and labeling using deep convolutional neural networks," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 7, pp. 2336–2348, Jul. 2019.

[13] M. Chung et al., "Pose-aware instance segmentation framework from cone beam CT images for tooth segmentation," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103720.

[14] Z. Cui et al., "TSegNet: An efficient and accurate tooth segmentation network on 3D dental model," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101949.

[15] 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. (CVPR)*, Jun. 2019, pp. 6368–6377.

[16] S. Keyhaninejad, R. A. Zoroofi, S. K. Setarehdan, and G. Shirani, "Automated segmentation of teeth in multi-scale CT images," in *Proc. Int. Conf. Vis. Inf. Eng.* London, U.K.: IET, 2006, pp. 339–344.

[17] H. Akhoondali, R. A. Zoroofi, and G. Shirani, "Rapid automatic segmentation and visualization of teeth in CT-scan data," *J. Appl. Sci.*, vol. 9, no. 11, pp. 2031–2044, May 2009.

[18] 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," *Med. Phys.*, vol. 42, no. 1, pp. 14–27, Dec. 2014.

[19] D. X. Ji, S. H. Ong, and K. W. C. Foong, "A level-set based approach for anterior teeth segmentation in cone beam computed tomography images," *Comput. Biol. Med.*, vol. 50, pp. 116–128, Jul. 2014.

[20] H.-T. Yau, T.-J. Yang, and Y.-C. Chen, "Tooth model reconstruction based upon data fusion for orthodontic treatment simulation," *Comput. Biol. Med.*, vol. 48, pp. 8–16, May 2014.

[21] X. Ren et al., "Interleaved 3D-CNNs for joint segmentation of small-volume structures in head and neck CT images," *Med. Phys.*, vol. 45, no. 5, pp. 2063–2075, May 2018.

[22] X. Wu, H. Chen, Y. Huang, H. Guo, T. Qiu, and L. Wang, "Center-sensitive and boundary-aware tooth instance segmentation and classification from cone-beam CT," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 939–942.

[23] D. Martinec and T. Pajdla, "Robust rotation and translation estimation in multiview reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[24] M. Jancosek and T. Pajdla, "Multi-view reconstruction preserving weakly-supported surfaces," in *Proc. CVPR*, Jun. 2011, pp. 3121–3128.

[25] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[26] M. Oechsle, S. Peng, and A. Geiger, "UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5589–5599.

[27] M. Liu, D. Zhang, E. Adeli, and D. Shen, "Inherent structure-based multiview learning with multitemplate feature representation for Alzheimer's disease diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1473–1482, Jul. 2016.

[28] Y. Kasten, D. Doktofsky, and I. Kovler, "End-to-end convolutional neural network for 3D reconstruction of knee bones from bi-planar X-ray images," in *Proc. Int. Workshop Mach. Learn. Med. Image Reconstruct.* Berlin, Germany: Springer, 2020, pp. 123–133.

[29] P. Henzler, V. Rasche, T. Ropinski, and T. Ritschel, "Single-image tomography: 3D volumes from 2D cranial X-rays," *Comput. Graph. Forum*, vol. 37, no. 2, pp. 377–388, 2018.

[30] L. Shen, W. Zhao, and L. Xing, "Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning," *Nature Biomed. Eng.*, vol. 3, no. 11, pp. 880–888, Oct. 2019.

[31] J. Ma, H. Zhang, P. Yi, and Z. Wang, "SCSCN: A separated channel-spatial convolution net with attention for single-view reconstruction," *IEEE Trans. Ind. Electron.*, vol. 67, no. 10, pp. 8649–8658, Oct. 2020.

[32] Z. Yun, S. Yang, E. Huang, L. Zhao, W. Yang, and Q. Feng, "Automatic reconstruction method for high-contrast panoramic image from dental cone-beam CT data," *Comput. Methods Programs Biomed.*, vol. 175, pp. 205–214, Jul. 2019.

[33] L. Ben Boudaoud, A. Sider, and A. Tari, "A new thinning algorithm for binary images," in *Proc. 3rd Int. Conf. Control, Eng. Inf. Technol. (CEIT)*, May 2015, pp. 1–6.

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

[35] Z. Al-Milaji, I. Ersoy, A. Hafiane, K. Palaniappan, and F. Bunyak, "Integrating segmentation with deep learning for enhanced classification of epithelial and stromal tissues in H&E images," *Pattern Recognit. Lett.*, vol. 119, pp. 214–221, Mar. 2019.

[36] M. S. Elmahdy et al., "Joint registration and segmentation via multi-task learning for adaptive radiotherapy of prostate cancer," *IEEE Access*, vol. 9, pp. 95551–95568, 2021.

[37] A. Araujo, W. Norris, and J. Sim, "Computing receptive fields of convolutional neural networks," *Distill*, vol. 4, no. 11, p. e21, Nov. 2019. [Online]. Available: https://distill.pub/2019/computing-receptive-fields

[38] N. Zou, Z. Xiang, Y. Chen, S. Chen, and C. Qiao, "Boundary-aware CNN for semantic segmentation," *IEEE Access*, vol. 7, pp. 114520–114528, 2019.

[39] E. F. Harris, "Tooth-coding systems in the clinical dental setting," *Dental Anthropol. J.*, vol. 18, no. 2, pp. 43–49, Sep. 2018.

[40] J. Chibane, T. Alldieck, and G. Pons-Moll, "Implicit functions in feature space for 3D shape reconstruction and completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6970–6981.