Continuous Bijection Supervised Pyramid Diffeomorphic Deformation for Learning Tooth Meshes from CBCT Images

Zechu Zhang, Weilong Peng, Jinyu Wen, Keke Tang, Meie Fang, David Dagan Feng, *Life Fellow, IEEE*, and Ping Li, *Member, IEEE*

***Abstract*—Accurate and high-quality tooth mesh generation from cone-beam computerized tomography (CBCT) is an essential computer-aided technology for digital dentistry. However, existing segmentation-based methods require complicated post-processing and significant manual correction to generate regular tooth meshes. In this paper, we propose a method of continuous bijection supervised pyramid diffeomorphic deformation (PDD) for learning tooth meshes, which could be used to directly generate high-quality tooth meshes from CBCT Images. Overall, we adopt a classic two-stage framework. In the first stage, we devise an enhanced detector to accurately locate and crop every tooth. In the second stage, a PDD network is designed to deform a sphere mesh from low resolution to high one according to pyramid flows based on diffeomorphic mesh deformations, so that the generated mesh approximates the ground truth infinitely and efficiently. To achieve that, a novel continuous bijection distance loss on the diffeomorphic sphere is also designed to supervise the deformation learning, which overcomes the shortcoming of loss based on nearest-neighbour mapping and improves the fitting precision. Experiments show that our method outperforms the state-of-the-art methods in terms of both different evaluation metrics and the geometry quality of reconstructed tooth surfaces.**

***Index Terms*—Shape generation, tooth segmentation, diffeomorphic deformation, CBCT, continuous mapping.**

I. Introduction

Manuscript received 25 April 2024; revised 19 November 2024; accepted 20 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62394330 and 62072126, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012064, in part by the Fundamental Research Projects Jointly Funded by Guangzhou Council and Municipal Universities under Grant 2024A03J0394, in part by the Fundamental Research Projects Funded by Liwan Institute under Grant LWYJ202418, and in part by The Hong Kong Polytechnic University (PolyU) under Grants P0048387, P0042740, P0044520, P0043906, P0049586, and P0050657, and in part by the PolyU Research Institute for Sports Science and Technology under Grant P0044571. *(Zechu Zhang and Weilong Peng contributed equally to this work.) (Corresponding authors: Meie Fang; Ping Li.)*

Zechu Zhang, Weilong Peng, Jinyu Wen, and Meie Fang are with the Metaverse Research Institute, School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 511400, China (e-mail: zhangzechu314@gmail.com; wlpeng@gzhu.edu.cn; wjy1361120721@163.com; fme@gzhu.edu.cn).

Keke Tang is with the Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China (e-mail: tangbohutbh@gmail.com).

David Dagan Feng is with the Biomedical and Multimedia Information Technology Research Group, School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: dagan.feng@sydney.edu.au).

Ping Li is with the Department of Computing, the School of Design, and the Research Institute for Sports Science and Technology, The Hong Kong Polytechnic University, Hong Kong (e-mail: p.li@polyu.edu.hk).

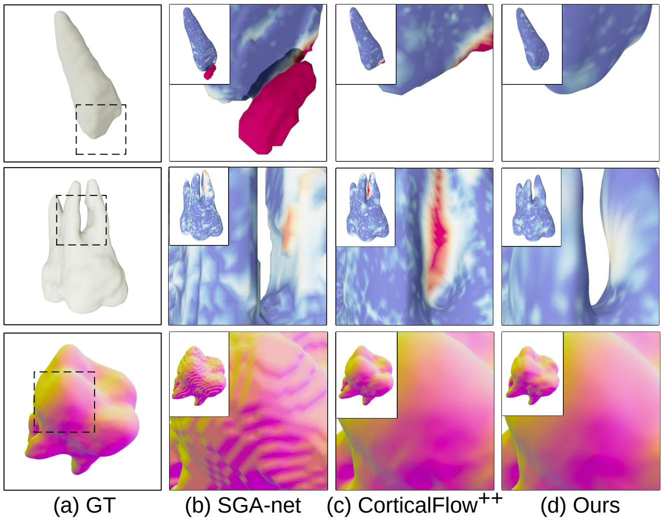


Fig. 1. Visualization of tooth reconstruction results: (a) the ground truth of the target teeth, (b) the isosurface extracted from the segmentation result of SGA-net [1], (c) the mesh generated by CorticalFlow^{++} [2], and (d) the mesh generated by our work. The surfaces of (b) are limited by the resolution of segmentation. (c) shows a large error for the multi-root case with the nearest-neighbour mapping. And, (d) is closer to the target by using a continuous bijection mapping.

GIVEN cone-beam computerized tomography (CBCT) images, and tooth models could be further reconstructed to assist doctors in making detailed treatments. Actually, tooth segmentation [1] is the first step in the clinical task of digital dentistry, and the subsequent diagnosis and treatment processes include biomechanical analysis, teeth alignment [3], [4], and 3D printing, etc. Before being applied to the subsequent processes, the segmentation results usually need to be translated to the form of regular meshes, as shown in Fig. 1(b). Therefore, if the teeth meshes could be obtained directly from CBCT images instead of segmentation, the unnecessary conversion process and errors in the subsequent processes can be avoided, improving the efficiency and accuracy of dental clinical diagnosis and treatment. So it's significant to directly build mesh models from CBCT images.

Currently, various methods have been proposed to develop instance segmentation technology on CBCT images for obtaining tooth models, which can be classified into traditional-knowledge-based (e.g., level set [5]–[7], region growing [8], statistical shape models [9]) and learning-based methods [1],

[10]–[16]. Overall, learning-based methods are superior to traditional conventional methods in automation and robustness, and existing deep-learning methods represent 3D shapes as discrete voxels. Cui et al. [10] propose a deep-learning network framework for individual tooth segmentation and identification based on Mask R-CNN. Wu et al. [11] propose a two-level hierarchical deep neural network, with one heatmap for locating tooth centers and DENSEASPP-Unet for ROI-based segmentation. Lee et al. [12] implement heatmap regression and box regression for each tooth and propose a novel Gaussian disentanglement penalty for all adjacent tooth pairs to more precisely position. Cui et al. [15] use tooth morphological features including tooth centroid, landmarks at root apices, skeleton, and boundary surface for tooth instance segmentation. Li et al. [1] explicitly model the spatial associations between different teeth for precise delineation in a coarse-to-fine fashion. The semantic graph attention mechanism is designed to model the anatomical topology of the teeth in each quadrant.

Existing deep learning methods focus on improving segmentation accuracy, but ignore the mesh quality requirements in subsequent applications including finite element analysis (FEA), computer aided geometric design (CAGD), simulations, and 3D printing. High-quality tooth meshes can be used for quantitatively analysing tooth collision and occlusion, and to help making detailed treatment plans. However, tooth meshes extracted from segmentation results may contain non-manifold and disconnected component errors, and thus require expensive post-processing topology correction [17]. Unlike previous segmentation methods, we aim to learn the genus-0 surfaces of teeth directly. Intuitively, which can be achieved by deforming a mesh template according to extracted information from medical volumetric image, e.g., [2], [18], [19]. However, multiple tooth individuals in the same CBCT image make deformation computing computationally intensive, and the loss based on the discontinuous nearest-neighbour mapping between meshes cannot handle the complex shape of the tooth well, especially for multi-root individuals, as shown in Fig. 1(c). Our motivation is to extract top-to-bottom flows from CBCT and use them to deform a sphere mesh to the target tooth shape efficiently. And most importantly, we require the mapping to be continuously bijective, so that it could approach the best reconstruction accuracy even for the worst multi-root case, as Fig. 1(d).

To achieve this purpose, we propose a pyramid diffeomorphic deformation (PDD) network and use continuous bijection distance (CBD) loss to supervise it for learning genus-0 tooth mesh from CBCT images. Overall, we adopt a widely-used two-stage schema in tooth reconstruction. Firstly, an improved anchor-free detector is designed to locate each tooth accurately. Secondly, PDD network is designed to deform a sphere mesh to a high-resolution tooth mesh progressively according to the extracted pyramidal flows. To avoid the local minimum brought by the discontinuous nearest-neighbour mapping, we propose a continuous bijection distance to measure the difference between two meshes and apply CBD loss to supervising the learning of PDD network. Since the continuous bijection guarantees that the deformation is conducted continuously along the manifold surface, our method greatly improves the reconstruction accuracy. Finally, we validate the effectiveness and efficiency of our method by comparing it with previous segmentation methods and other explicit surface learning methods. In a nutshell, our contributions are summarized as:

- We are the first to formulate the 3D tooth reconstruction task as mesh surface learning, abandoning the commonly-used segmentation process, to produce tooth surface from CBCT images directly.
- We devise a PDD network to guide the tooth shape generation efficiently by deforming from a low-detailed level to a high-quality one.
- The CBD loss is proposed to force the nearest-neighbour-based mapping between the prediction mesh and the ground truth to be continuously bijective, promoting the precision of tooth shape generation.
- Our solution shows superior performance on both segmentation precision and geometric quality of teeth to the state-of-art methods by experiments.

II. Related Work

A. Tooth Detection and Identification

To improve the performance of tooth instance segmentation, many researchers adopt a two-stage schema and locate the tooth instances in the first stage ([10]–[13], [15]). This approach, which involves initially locating teeth, has found application not only in CBCT imaging but also extends its usage widely to other modalities encompassing mesh models ([20], [21]) and panoramic radiographs ([22], [23]). Most two-stage frameworks extend anchor-based object detectors for tooth localization ([10], [13]). ToothNet [10] proposes a learned similarity matrix to remove redundant proposals. Chung et al. [13] design a 2D convolutional neural network to regress CBCT image pose, then use Faster R-CNN [24] to propose ROIs. Anchor-based methods predict a large number of ROIs which need further post-processing (e.g., non-maximum suppression algorithm), and it is tricky to get the best settings for anchors. In contrast, anchor-free methods have fewer hyperparameters. Wu et al. [11] use a global stage heatmap to accurately locate tooth centers, then crop individual tooth volumes of fixed size $64 \times 48 \times 48$. Recently, [12] proposes a point-based tooth localization network based on anchor-free detector CenterNet [25] and produces impressive average precision of detection, and proposes a new Gaussian disentanglement loss to penalize the overlap of Gaussian distribution regression for each tooth. However, the recall of [12] is relatively low, because of the similar topology and proximate nature of teeth. Considering the characteristics of the tooth, we revise the branches of CenterNet to obtain higher average precision and recall.

B. 3D Tooth Representation

For traditional methods, a tooth is represented as a set of contours [5]–[7], statistical shape models [9], or voxels [8]. Existing tooth segmentation networks [10]–[15] use voxel-based representations, which can be naturally processed with varied convolution operations. Indirect mesh prediction

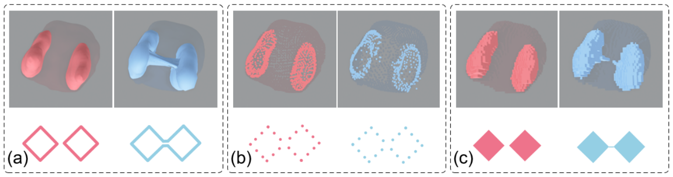


Fig. 2. The predictions (blue) to the tooth ground truths (red) with two close roots: the cases of (a) mesh, (b) point cloud, and (c) voxels. Metrics including surface distance in (b) and overlap ratio in (c) can't measure the discontinuous nearest neighbour mappings between predictions and ground truths. Note that without considering the mesh topology, the loss may bring the wrong predictions.

methods require expensive post-processing, such as Marching Cubes or Marching Tetrahedral to extract explicit surfaces. Moreover, post-processing is not differentiable and usually leads to non-manifold errors [2], [18]. Although Deep Marching Cubes [26] and Deep Marching Tetrahedra [27] make the surface extraction process differentiable, they still lack constraints on the topological structure. Direct mesh reconstruction aims to fit the target shape with a pre-defined topology (e.g., template [2], [18], [28]–[31], or a union of primitives [32], part-level geometries [19], [33], [34]). To obtain high-quality tooth meshes, we adopt the scheme based on a deforming template, which will only change vertices' positions and conserve partial topology properties, e.g., non-manifold edges and vertices.

Recently, the deforming template framework [2], [18], [30], [35], [36] has been highly successful in generating 3D models from medical volumetric images. Voxel2Mesh [18] uses a graph-based convolutional network (GCN) to predict the deformation of each vertex on a template sphere and has been tested on MRI brain, hippocampus, and liver datasets. Meanwhile, [30] represents the entire heart as alternative template meshes which represent a subset of the geometries, and uses GCN to predict the transformation of control points enclosing the template heart to cater to different modelling requirements. In contrast, CorticalFlow^{++} [2] learns a diffeomorphic deformation from a genus zero smooth template to the target cortical surface, the template tightly wraps all training cortical surfaces. Teeth with different numbers of roots are structurally dissimilar, especially molars, whose root count is variable. Considering that tooth shapes and orientations vary considerably while sharing the same 0-genus, we use a template of spherical mesh instead of template generation [35] or template selection [36].

C. Diffeomorphic Deformation

Diffeomorphic deformation is an invertible function that could map a differentiable manifold to another smoothly. Ordinary Differential Equation (ODE) is a widely used tool to define invertible deformations. Furthermore, when the velocity field of ODE is globally Lipschitz continuous, any two ODE trajectories will not intersect [37]. This guarantees diffeomorphic flow with strong implicit regularizations to prevent self-intersection and non-manifold faces [31]. The Flow ODE is utilized in shape generation networks [2], [31], [38], [39], while the stationary velocity field (SVF) framework, where the velocity field is time-independent, has been successful in medical registration problems [40], [41]. The work closest to ours is CorticalFlow^{++} [2], which expands the SVF approach to points in real coordinates and introduces a diffeomorphic mesh deformation module (DMD). The method uses a multi-scale approach with three sequential deformations. Since the deformations need to be trained separately, the training process is time-consuming. Differently, we develop a novel coarse-to-fine deformation strategy based on pyramid flows extraction, with fewer parameters, less training time and inferencing time.

D. Losses in Mesh Generation

In the area of mesh learning and generation, the losses are very important for generating high-quality results. Generally, surface distance [42]–[44] and overlap ratio [32] are used as metrics to measure the difference between two shapes. The distance between two surfaces is naively defined as the distance between each point in the surface and its closest neighbour in the other surface, CD loss [42] calculates the average of the distances between meshes. Hausdorff distance (HD) [43] finds the point on one mesh that is farthest from the closest point on the other mesh, and is sensitive to large geometric differences. Neither CD nor HD is sensitive to the mismatched point density. In contrast, Earth mover's distance (EMD) [44] finds the bijective between two point clouds by solving an optimal transmission problem, which results in the computation cost of EMD being much higher than CD and HD. Paschalidou et al. [32] propose occupancy loss which converts the implicit surface to an indicator function by inverse homeomorphic mapping. However, the occupancy loss has poor adaptability to other frameworks. Losses forced on only point clouds bring unnecessary sharp corners and faces, thus Laplacian loss and edge length loss are proposed by Pixel2Mesh [28] to smooth the surface. Laplacian loss avoids mesh self-intersection, and edge length loss is further applied to balance the edge length between faces. However, for complex shapes like multi-root teeth, surface distance and overlap ratio losses may fall into a local minimum, resulting in convergence to an incorrect shape, as demonstrated in Fig. 2. In human body reconstruction [45], the definition of correspondences between scanned models has a crucial impact on the accuracy of their alignment. The mesh generation loss should encourage a correct mapping between the predicted and target shapes. Essentially, these metrics neglect mesh topology, and cannot penalize discontinuous nearest neighbour mappings between predicted and target meshes. To address this, we propose a loss based on continuous bijection distance to enforce a continuous bijective mapping between prediction and target.

III. Problem Formulation

A. Diffeomorphic Deformation of Tooth

By assuming that the tooth surface can be reconstructed by deforming from an initial sphere to ground truth continuously, the deformation trajectory and velocity could be described according to the following ODE [46]:

$$\frac{\partial \mathbf{s}(t)}{\partial t} = \mathbf{v}(\mathbf{s}(t), t), \text{ with } t \in [0, T], \tag{1}$$

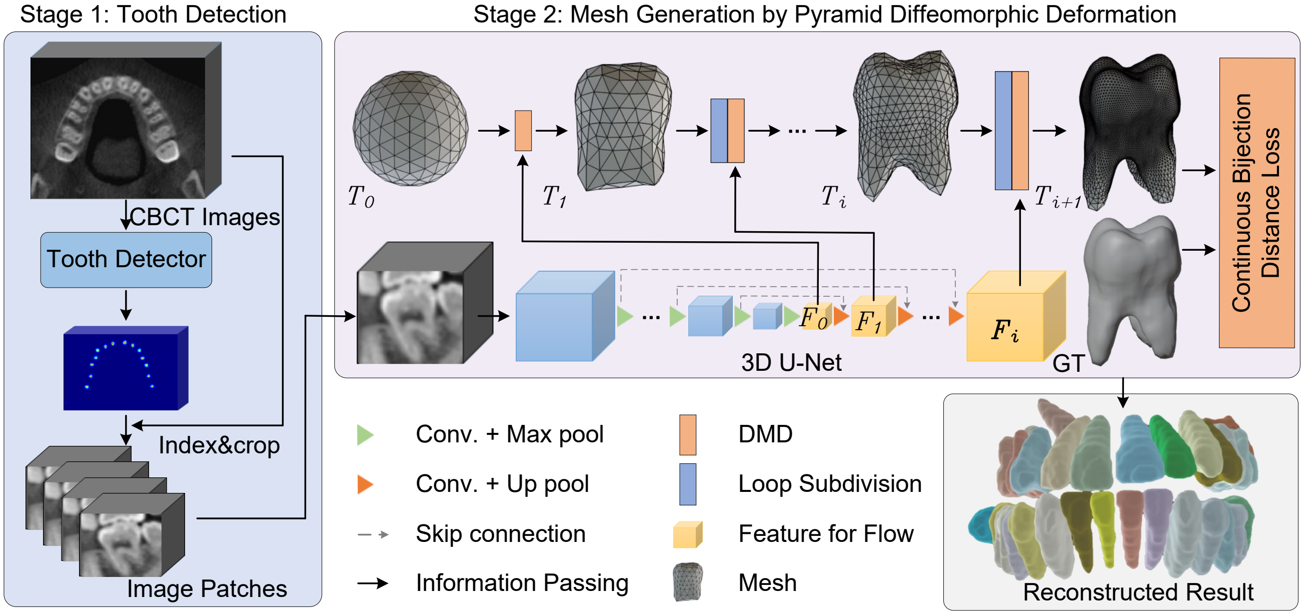


Fig. 3. Two-stage framework for pyramid diffeomorphic deformation. In the first stage, an improved tooth detector is used to locate teeth and crop them. In the second stage, tooth meshes are generated by a pyramid diffeomorphic deformation (PDD) network based on a 3D U-net backbone. In particular, the features $\{F_i\}$ are extracted from different scales of the decoder layers in 3D U-Net and fed into pyramid flows based on DMDs. And the learning of mesh generation is supervised by continuous bijection distance loss.

where $\mathbf{s} : [0, T] \to \mathbb{R}$ is the continuous 3D trajectory of a point on the surface during the deformation, and $\mathbf{v} : \mathbb{R}^3 \times [0, T] \to \mathbb{R}^3$ is the continuous velocity field in space and time.

Since the point $\mathbf{s}(0)$ is given on the initial sphere surface, the following spatial positions could be propagated step-by-step using the Euler method according to velocity flow by the equation:

$$\mathbf{s}(t + \Delta t) = \mathbf{s}(t) + \mathbf{v}(\mathbf{s}(t))\Delta t. \tag{2}$$

B. Motivation

It is time-consuming to compute the velocity flow of every point in each step, especially for high-quality mesh generation. Therefore, we consider obtaining the low-resolution flows from top-level CBCT features incipiently, then elevate the resolution gradually from the bottom ones in the subsequent steps, so that computation time can be reduced significantly. Note that the top-level information also can be propagated to the bottom efficiently in the pyramid processing. In addition, the loss function is also key for high-quality results. Classic losses, e.g., CD, HD, and EMD, perform unsatisfactorily in guaranteeing the predicted shape closely matches the target tooth shape. Therefore, we propose the CBD loss that promotes the nearest neighbour mapping between the predicted mesh and the target tooth to be a continuous bijection, which is crucial for achieving the reconstruction with high precision. Based on the above two aspects, the tooth shapes are hoped to be generated effectively and efficiently.

IV. Methods

In section, we will describe how our method is designed to generate the tooth surfaces from CBCT images. The adopted framework is a two-stage network, as shown in Fig. 3. In the first stage, an enhanced detector is designed to identify the center point, size, and classification of each tooth, and then crop the CBCT image patches according to the detection result. In the second stage, the CBCT patch is fed into a pyramid diffeomorphic deformation network for mesh reconstruction to generate the tooth mesh efficiently. To implement a high-quality reconstruction, we further designed the CBD loss based on continuous bijection between prediction and target to supervise the learning of the deformation network.

A. Enhanced Detector for Cropping Tooth

Following the ISO standard tooth numbering system, each tooth needs to be identified with one of 32 labels. CenterNet uses a 32-channel heatmap to regress bounding box centers. However, a tooth is prone to be repeatedly classified with different labels, while its adjacent teeth may be ignored. Moreover, it is difficult to regress the Gaussian distributions of two adjacent teeth that are similar, causing Gaussian disentanglement issue [12]. Therefore, we redesign the detection head of CenterNet. We perform downsampling 4 times in the encoder and upsampling 2 times in the decoder with stride 2. At last, a 1-channel heatmap is used to predict all bounding box centers, while the 32-channel heatmap is utilized for classification. In particular, each channel of the 32-channel heatmap is related to the target bounding box center with a certain label following [47]. And the 1-channel heatmap predicts the pixel-wise maximum of the 32-channel heatmap. Finally, the loss function of the 1st-stage network is

$$L_{1st} = \lambda_k L_k + \lambda_c L_c + \lambda_{off} L_{off} + \lambda_{size} L_{size}, \tag{3}$$

where focal losses [48] L_c and L_k are used to train predicted 1-heatmap and predicted 32-channel heatmap, and smooth L1

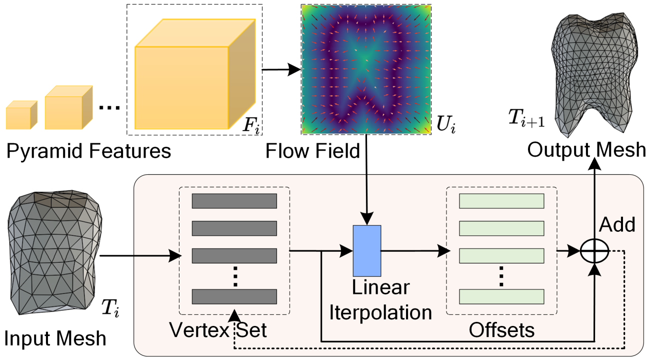


Fig. 4. The DMD takes feature map F_i as input and generates diffeomorphic mapping from input mesh T_i to refined mesh T_{i+1}.

loss L_{off} and L_{size} are used to supervise the local offset and boxes size at the center of every bounding box as in the work [24]. Based on this design, our detector focuses on detecting all teeth and achieves a higher recall than the original CenterNet.

B. Surface Generating with Pyramid Diffeomorphic Deformation

We design a PDD network for tooth surface generation in the 2nd-stage. It is composed of a 3D Unet and a progressive deformation module, as shown in Fig. 3. The deformation head consists of multiple DMD modules. Drawing inspiration from the feature pyramid network [49], [50] to detect objects with various scales, we learn the pyramid feature maps from the Unet decoder, and put them into the DMD modules so that a template deforms progressively. During the process, the resolution and precision of the mesh are boosted step by step until it matches the target tooth mesh.

Pyramid Flows based DMDs. In the architecture, the early-stage DMD accepts the low-resolution feature map from the top-level feature, while the later-stage ones accept the high-resolution feature map from the bottom. Thus, the formulation of progressive deformation could be defined as follow:

$$T_{i+1} = DMD_i(F_i, T_i), i = 0, 1, 2, 3, \tag{4}$$

where F_i is the feature map after the i-th upsampling in the decoder, and T_0 is a low-resolution initial template. The DMD module takes F_i and T_i from the previous stage as input, and outputs the deformed shape T_{i+1}, as shown in Fig. 4. In particular, all $\{T_i\}$ have the same genus as the initial template.

In the DMD module, the feature map F_i is translated to flow vector field U_i, and the input mesh is converted into a point cloud. Therefore, the diffeomorphic mapping $\Phi : [0, 1] \times \mathbb{R}^3 \rightarrow \mathbb{R}^3$, from T_i to T_{i+1}, could be generated based on flow ODE in Eq. (2). Specifically, it is given by:

$$\hat{\Phi}(h, \mathbf{x}) = \mathbf{x} + hU_i(\mathbf{x}), \tag{5}$$

where h is step size, U_i is the 3D flow field, and $U_i(\mathbf{x})$ means linear interpolation of U_i at vertex position $\mathbf{x}$. $\hat{\Phi}$ is a numerical approximation of Φ in consideration of discretization.

To make the progressive deformation adapt to the pyramid deformation, the loop subdivision algorithm [51] is followed after the DMD to increase the resolution of predicted meshes. Thus, the ground truth with high resolution could be approximated immensely as needed.

C. Continuous Bijection Distance Loss

During the surface learning, the 2nd-stage network needs to be supervised by loss between the predictions $\{T_i\}$ and ground truth T_g. Therefore, a unified loss, referred to as CBD loss, is designed for each prediction. Specifically, we measure the surface distance between two genus-0 shapes based on continuous bijective mapping instead of the nearest-neighbour mapping, ensuring the predicted shape deforms towards the target shape during the training process. We assume that all teeth are genus-0, and note the meshes of prediction and ground truth as the T_p and T_g respectively. They are conformal to parametric spheres S_p and S_g under the given restrictions of topology connections via the following mappings:

$$S_p = \varphi_p \circ T_p \quad \text{and} \quad S_g = \varphi_g \circ T_g, \tag{6}$$

where φ_p and φ_g are their corresponding spherical conformal mapping operations.

Nearest Neighbour Metric. A general surface distance metric between two shapes T_p and T_g is based nearest-neighbour mapping algorithm like:

$$D_{nn}(T_p, T_g) = \frac{1}{|T_p|} \sum_{v_p \in T_p} \min_{v_g \in T_g} ||f(v_p) - f(v_g)||_2^2, \tag{7}$$

where $f(v)$ indicates the geometric feature (or just coordinate) extracted at vertex v. Briefly, we re-formulate Eq. (7) as below:

$$D_{nn}(T_p, T_g) = \frac{1}{|T_p|} ||T_p - \tau_\varepsilon \circ T_g||_2^2, \tag{8}$$

where τ_ε is the operation that find the closest point in T_g for each point in T_p in Euclidean space,

$$\tau_\varepsilon = \text{NN}(T_p, T_g). \tag{9}$$

The nearest-neighbour metric (NNM) has been widely used in the losses, e.g., CD and HD losses, for the task of geometric reconstruction. Since the metric is only applied to local vertices, it ignores the inner topology of shapes and doesn't satisfy the continuous mapping between two meshes. So NNM works badly for accurate mesh fitting.

Continuous Bijection Metric. To overcome the drawback of NNM, we adjust the nearest-neighbour mapping τ_ε to be continuous bijective. In particular, we define a novel continuous bijection metric (CBM):

$$D_{cb}(T_p, T_g) = \frac{1}{|T_p|} ||T_p - \tau_\mathcal{M} \circ T_g||_2^2. \tag{10}$$

Here $\tau_\mathcal{M}$ aims to find the closest point on the manifold surface, instead of that in Euclidean space. To make it continuous bijective, we obtain $\tau_\mathcal{M}$ by adjusting τ_ε on parametric sphere S_g,

$$\tau_\mathcal{M} = \text{NN}(\Delta_p(\tau_\varepsilon \circ S_g), S_g), \tag{11}$$

where Δ_p is a local adjustment operation to remove the overlapping area on $\tau_\varepsilon \circ S_g$. Firstly, nearest-neighbour mapping τ_ε is used to locate the corresponding parametric points in sphere S_g for prediction. But it may occur self-intersection of

Algorithm 1: Continuous Bijection Distance Loss

```
Input: Genus-0 meshes $T_p$ and $T_g$
Output: CBD loss $L_{cbd}$
begin
    Calculate spherical conformal mapping $\varphi_g$ for $T_g$;
    Parametric sphere $S_g \leftarrow \varphi_g \circ T_g$;
    Compute neighbour mapping $\tau_\mathcal{E}$ from $T_p$ to $T_g$;
    $S_t = copy(T_p)$;
    for $v_i \in S_t$ do
        Find $u_j \in S_g$ corresponding to $v_i$ based on $\varphi_g$;
        $v_i \leftarrow u_j$;
    for $1$ to $t$ do
        for $v_i \in S_t$ do
            $\mathbf{n}_p[v_i] \leftarrow$ the nomral of $v_i$ in surface $T_p$;
            $\mathbf{n}_g[v_i] \leftarrow$ the nomral of $v_i$ in surface $T_g$;
            if $\boldsymbol{n}_p[v_i] \cdot \boldsymbol{n}_g[v_i] < \theta$ then
                $v_i \longleftarrow v_i + h \cdot \Delta_p v_i$;
    Compute neighbour mapping $\tau_\mathcal{M}$ from $S_t$ to $S_g$;
    $L_{cbd} \leftarrow 0$;
    for $v_i \in T_p$ do
        Find $u_j \in T_g$ corresponding to $v_i$ based on $\tau_\mathcal{M}$;
        $L_{cbd} \leftarrow L_{cbd} + ||v_i - u_j||_2^2$;
    $L_{cbd} \leftarrow \frac{1}{|T_p|} \times L_{cbd}$;
    return $L_{cbd}$;
```

the manifold in the cases of non-convex tooth shapes. Thus, we use laplacian Δ_p to adjust $\tau_\mathcal{E} \circ S_g$ locally so that the self-intersection could be eliminated,

$$v_i \longleftarrow v_i + h \cdot \Delta_p v_i, \quad \text{for} \quad v_i \in S_{op}. \tag{12}$$

Here S_{op} indicates the operation area defined as below:

$$S_{op} = \{v | \mathbf{n}_p[v] \cdot \mathbf{n}_g[v] < \theta \text{ for } v \in \tau_\mathcal{E} \circ S_g\}, \tag{13}$$

where $\mathbf{n}_p[v]$ and $\mathbf{n}_g[v]$ are normal pair in surface T_p and T_g located with v in parametric sphere.

Geometric Interpretation. When all the overlapping areas are removed, NN in Eq. (11) will find a continuous bijection $\tau_\mathcal{M}$ from $\Delta_p(\tau_\mathcal{E} \circ S_g)$ to S_g. In other words, the shapes T_p and T_g, between which the nearest-neighbour mapping is not continuously bijective originally, could be translated to the same sphere $\Delta_p(\tau_\mathcal{E} \circ S_g)$ consistently. The distance minimization based on CBM aims to deform T_p to the target T_g along the manifold surface. In contrast to NNM, CBM forces the prediction to approach the target with higher precision, as shown in Fig. 5. Finally, the learning of mesh deformation is to optimize the output T_p of PDD network to match $\tau_\mathcal{M} \circ T_g$. CBD loss is as below:

$$L_{cbd} = D_{cb}(T_p, T_g). \tag{14}$$

The CBD loss is presented as pseudocode in Algorithm 1.

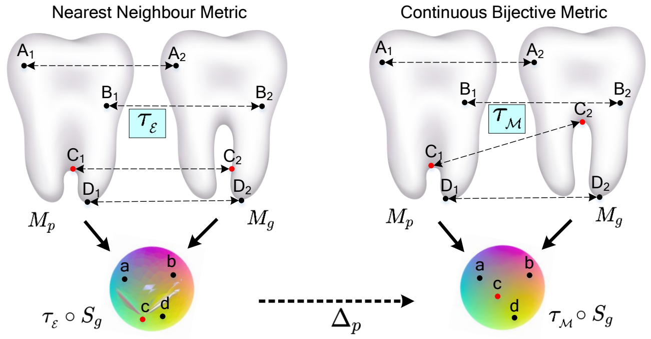


Fig. 5. Nearest neighbour metric vs. continuous bijection metric. The nearest neighbour mapping $\tau_\mathcal{E}$ ignores the topology of T_p and T_g, so T_p cannot be continuously mapped to T_g, e.g., point C_1 at the saddle of the tooth root is mapped to C_2 close to the root apex. Self-intersection occurs in the spherical mesh $\tau_\mathcal{E} \circ S_g$ when mapping T_p to the spherical parametric domain same to T_g. In contrast, the continuous bijection mapping $\tau_\mathcal{M}$ maps T_p to T_g continuously with the help of the adjustment operation Δ_p, avoiding any self-intersection. The $\tau_\mathcal{M}$ preserves the topological relationships between neighbouring regions, inducing correct match even between the meshes with large distance, e.g., the root saddle region where point C_1 is mapped to the root saddle region of another mesh.

D. Loss of Second Stage Network

The PDD network will output several predictions $\{T_i\}$. We design the loss function for each T_i as follow:

$$L_{2nd} = \lambda_{cd} L_{cd} + \lambda_{edge} L_{edge} + \lambda_{cbd} L_{cbd} + \lambda_{lap} L_{lap}. \tag{15}$$

Besides CBD loss, we also minimize CD loss L_{cd} to make the predicted mesh match ground-truth surfaces. Mesh regularizers, i.e., Laplacian L_{lap} and edge lengths L_{edge} [18], are used to penalize non-manifold errors. For the early two meshes, we set the hyperparameters $\lambda_{edge} = 1, \lambda_{cd} = 1, \lambda_{cbd} = 0, \lambda_{lap} = 0$. For the latter two ones, we set $\lambda_{edge} = 1, \lambda_{cd} = 1, \lambda_{cbd} = 1, \lambda_{lap} = 0.1$. The losses L_{2nd} of each T_i are summed as the training loss.

V. Experimental Results

A. Dataset

We adopt two CBCT datasets [15], [16] with crowding, missing, or oblique teeth. The voxel resolution of these scans is $0.4mm$. The first dataset in [16] is a publicly available dataset consisting of 150 instances. The first 100 instances in it have annotations of high degree of quality, while the remaining 50 instances have annotations with slightly coarse precision. The second dataset in [15] encompasses 100 instances, all of which possess annotations of superior quality. In the experiments, We utilize the first 100 instances in dataset [16] and all CBCT scans in dataset [15]. The CBCT images are cropped and resized to $256 \times 256 \times 256$, and intensity values are normalized to $[0, 1]$. Pixel-level segmentation labels are manually annotated. The explicit surfaces of teeth are extracted by the Marching Cubes method. Meshes with non-manifold, non-watertight, or non-genus-zero problems are automatically processed using blender [52] and manually fixed. The CBCT images are randomly cropped into $160 \times 160 \times 160$ to increase the batch size for the 1st stage network, and teeth are cropped

TABLE I
PERFORMANCE COMPARISON OF TOOTH DETECTION AND IDENTIFICATION METHODS

Model	AP50	OIR	FA	CD (mm)
Faster R-CNN [24]	98.50	97.34	93.20	—
Deformable DETR [54]	82.10	54.67	87.77	2.3293
nnDetection [53]	98.71	97.85	99.14	1.0293
CenterNet [25]	98.95	97.76	93.42	0.9220
Ours	**99.35**	**98.02**	**99.24**	**0.6768**

and resized to 64 × 64 × 64 for the 2nd stage network. All experiments were conducted using a single Nvidia GeForce RTX$^{\text{TM}}$ 3090 Ti graphics card, and the network architectures were constructed using the PyTorch library.

B. *Performance of Tooth Detector*

Baselines. We compare the proposed tooth detector with state-of-the-art detection networks, including Faster R-CNN [24], nnDetection [53], Deformable DETR [54], [55], and CenterNet [25]. Faster R-CNN is used as the backbone and extended to a two-stage framework for CBCT segmentation [10], [13], while CenterNet is also used as a tooth detector in [12]. nnDetection is designed to automate the configuration of medical object detection models and supports various 3D medical image inputs. Deformable DETR is an end-to-end object detection method that combines deformable attention modules with Transformers to directly predict object bounding boxes and categories, achieving excellent results in 2D object detection.

Evaluation Metrics. For the two-stage tooth segmentation network, the detection accuracy of the detector directly affects the performance of instance segmentation. The average precision (AP) is a critical metric for evaluating the effectiveness of tooth detection, quantifying the area under the precision-recall (PR) curve. When the threshold value for intersection over union is 0.5, the AP is known as AP50. The 2nd-stage network only segments tooth shape after cropping, which is suitable for measuring the ratio of the target tooth region inside the detected bounding box. The Object Include Ratio (OIR) [13] is also adopted as the metric which is the ratio (%) of the region of the detected object in a detected bounding box to the complete ground-truth object. We use identification accuracy (FA) [10] to measure the accuracy of identification. The similar topology and proximate nature of adjacent teeth are considered to be one of the main difficulties for classification. [12] figures that this difficulty also causes a deviation in positioning the detected bounding box center. Our detection network considers teeth as a single class when locating tooth instances, which improves the localization accuracy and recall while suppressing the number of detected bounding boxes. The CD is used to measure the distance from detected bounding box centers to its corresponding ground truth.

Comparison. Table I shows the evaluation results for tooth detection and identification. The proposed method outperforms all baseline methods. Among all the baselines, nnDetection performs the best. However, compared to nnDetection, our method shows improvements in the performance metrics of CD, AP50, and OIR, with values approximately $0.3525mm$, 0.64%, and 0.17% better, respectively. Compared to CenterNet, the proposed method demonstrates a significant improvement in FA by 5.82% and outperforms it by approximately $0.245mm$ in CD, indicating that our tooth center localization is more accurate. Fig. 6 is a visualization of the detection results. In the first row, the images include the pixel-wise maximum of CenterNet's classification branch and the output of our network heatmap branch. As shown in Fig. 6(a), both CenterNet and our network output similar images to the pixel-wise maximum of multiple Gaussian distributions. But our output heatmap has smaller standard deviations and higher values at the peaks, which means more accurately detected bounding box centers and higher recall of predicted results. Fig. 6(b)–(d) are adjacent channels of the 32-channel heatmap. Our pixel-wise classification is based on heatmap regression. Fig. 6(e) shows that our method can effectively avoid misclassification.

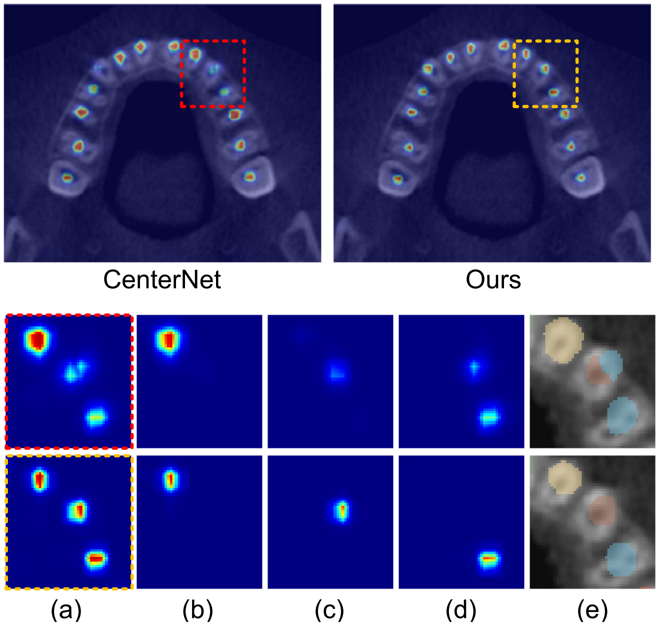


Fig. 6. Visualization of detection results for CenterNet and our network. Our heatmap has smaller standard deviations and higher peak values. As a result, our method yields more accurate predicted centers and higher recall of predicted results. Additionally, as demonstrated in (e), our approach is effective in avoiding misclassification.

C. *Performance of Shape Generation*

Baselines. We compare the proposed method with state-of-the-art voxel-based methods and mesh-based methods. The voxel-based methods include not only ToothNet [10] and SGA-net [1], which are specifically designed for tooth segmentation in dental CBCT, but also one-stage 3D medical image segmentation networks such as UNETR [56], Swin UNETR [57], and 3D UX-Net [58], which have demonstrated excellent performance across multiple datasets. The mesh-based methods involve Voxel2mesh [18] and CorticalFlow^{++} [2] which are designed for a single target.

TABLE II
QUANTITATIVE RESULTS COMPARISON WITH BOTH VOXEL-BASED (FIRST FIVE ROWS) AND MESH-BASED SEGMENTATION METHODS

Models	DSC	Precision	Recall	CD (mm)	HD (mm)	HD95 (mm)
ToothNet [10]	0.916	—	—	0.30	2.82	—
UNETR [56]	0.924	0.933	0.926	0.486	9.360	2.688
Swin UNETR [57]	0.934	0.938	0.934	0.332	1.398	0.704
3D UX-Net [58]	0.932	0.939	0.930	0.399	5.493	1.179
SGA-net [1]	0.936	0.942	0.932	0.330	1.400	0.843
Voxel2mesh [18]	0.924	0.924	0.931	0.219	1.405	0.682
CorticalFlow^{++} [2]	0.937	**0.945**	0.930	0.207	1.004	0.492
Ours	**0.945**	0.942	**0.949**	**0.168**	**0.862**	**0.389**

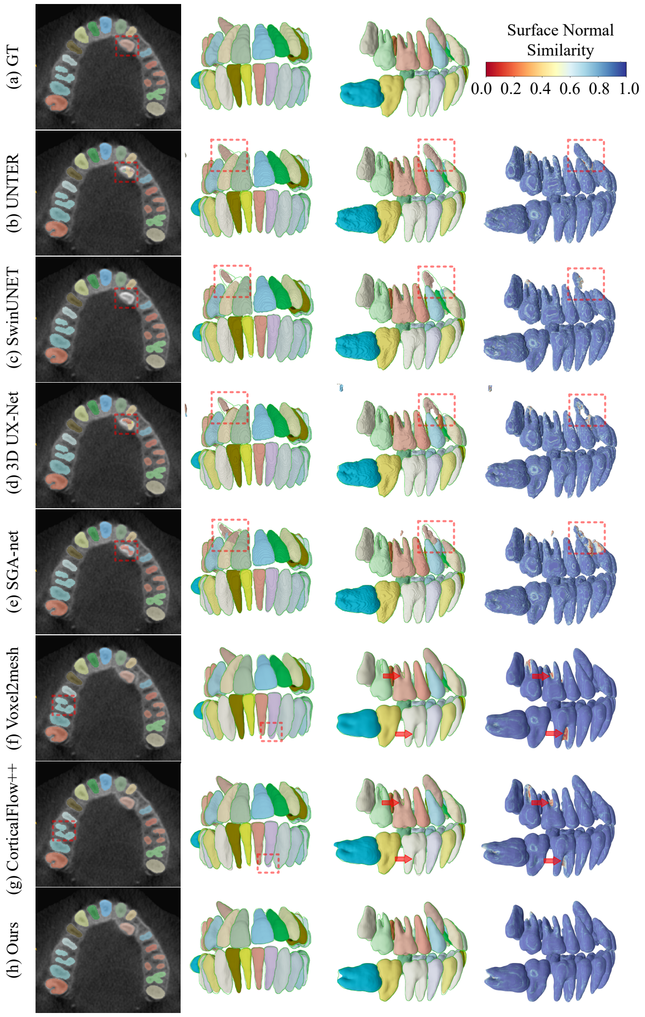


Fig. 7. Visualization of segmentation results, the predicted meshes, and surface normal similarity. The first column shows the segmentation results on CBCT, while the second and third columns visualize the predicted tooth model from the front and side views, with the ground-truth tooth model's edge contours highlighted in green. The fourth column illustrates the consistency of surface normals between the predicted and ground-truth meshes.

Evaluation Metrics. We evaluate our framework using various metrics, including segmentation performance measured by dice similarity coefficient (DSC), Precision, and Recall, geometric accuracy measured by CD and HD, and surface regularity. The metrics of geometric accuracy are computed with the surface sampling points for both mesh-based and voxel-based methods. Furthermore, we compare the surface regularity of different CBCT networks based on non-manifold vertices(NM Vert.), non-manifold edges(NM Edg.), and the ratio of self-intersection faces(SIF) [31]. In addition, we calculate the percentage of meshes that contain disconnected components(DC).

TABLE III
QUANTITATIVE RESULTS COMPARISON WITH MESH-BASED SEGMENTATION METHODS IN PUBLICLY AVAILABLE DATASET

Model	DSC	CD (mm)	HD (mm)	HD95 (mm)
Voxel2mesh [18]	0.936	0.190	1.156	0.498
CorticalFlow^{++} [2]	0.939	0.191	1.156	0.549
Ours	**0.944**	**0.178**	**1.042**	**0.454**

Comparison of Segmentation Performance. Table II shows the quantitative comparison of segmentation performance with both state-of-the-art voxel-based methods and mesh-based methods. Among all voxel-based networks, SGA-net achieves the best segmentation performance, our method outperforms SGA-net 1.7% in Recall, and 0.9% in DSC. The voxel-based network with the best surface distance performance is Swin UNETR, and our network surpasses it by approximately 0.164mm in CD, 0.536mm in HD, and 0.315mm in HD95. In summary, compared to voxel-based methods, our method achieves significant improvement, especially in reducing the surface distance to a very small value. Voxel2mesh has to make a trade-off between geometric accuracy and mesh regularity. Our method performs better according to all metrics and outperforms CorticalFlow^{++} in terms of segmentation accuracy and surface distance. The CBD loss makes the generated mesh more accurate, especially at the tooth root. In order to facilitate subsequent comparative analyses, our network and two mesh-based segmentation methods are individually trained and tested on the publicly available Dataset [16]. The results, as presented in Table III, unequivocally demonstrate the superior performance of our approach in comparison to the other two methods.

Visualization of Segmentation Results. In Fig. 7, the first column presents CBCT segmentation results, the second and third columns show front and side views of the predicted tooth model with ground-truth contours in green. For voxel-based methods without detecting tooth location, the segmentation

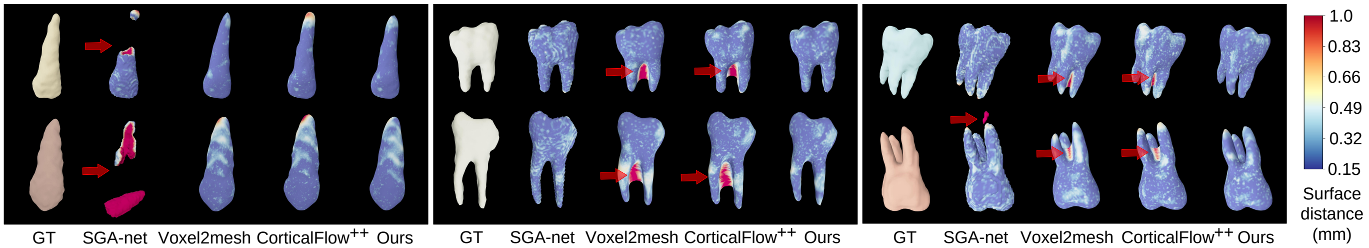


Fig. 8. The visual comparison of reconstructions between different methods on the teeth with root numbers 1 (left), 2 (middle), and 3 (right).

TABLE IV
COMPARISON OF TOOTH MESH GENERATION METHODS UPON SURFACE REGULARITY AND TOPOLOGY DEFECT

Metrics	SGA-net [1]	Voxel2mesh [18]	CorticalFlow-3 [2]	Our network
NM Vert.	×	✓	✓	✓
NM Edg.	×	✓	✓	✓
DC (%)	3.9	0.0	0.0	0.0
SIF (1e-4)	0.0	1.583	0.427	0.132

TABLE V
COMPARISON OF GEOMETRIC ACCURACY AND SEGMENTATION PERFORMANCE BETWEEN THE METHODS WITHOUT PYRAMID FLOWS AND OURS (OUR NETWORK-BASE) WITH PYRAMID FLOWS

Model	DSC	Precision	Recall	CD (mm)	HD (mm)	HD95 (mm)
CorticalFlow-1 [2]	0.932	0.924	0.942	0.222	1.237	0.629
CorticalFlow-3	0.937	**0.945**	0.930	0.207	1.004	0.492
Our network-base	**0.943**	0.941	**0.947**	**0.178**	**0.997**	**0.476**

results are often incomplete, especially for the dislocated and grown oblique tooth, as shown in Fig. 7(b)–(e). In Fig. 7(f)–(h), the tooth mesh generated by other mesh-based methods satisfies the requirement of connectivity but has serious artifacts between the branches of the tooth root. The fourth column shows the consistency of the surface normal between the predicted mesh and the ground-truth mesh. The consistency is measured by the dot product of the surface normal vectors. We observe that the predicted meshes generated by voxel-based methods exhibit a uniform distribution of surface normal similarity, but they do not achieve the same quality as mesh-based methods due to the limitations in mesh fineness imposed by image resolution. In Fig. 8, the teeth are arranged from 1 to 3 roots, of which shapes also become more and more complex. Either disconnected components or rough surface shapes emerge in all cases for SGA-net. Mesh-based methods perform poorly on multi-root teeth. Serious adhesion emerges between roots. Obviously, our method solves all these problems and outputs high-quality meshes.

Comparison of Surface Quality. We compare our method to the current state-of-the-art mesh-based and voxel-based methods using surface regularity and topological defects metrics, as shown in Table IV. Compared to the deformation-based methods Voxel2Mesh and CorticalFlow-3, our method produces tooth meshes with significantly higher geometric accuracy and a lower percentage of self-intersecting faces. Mesh-based methods only alter the positions of vertices, preserving the topology properties of the original meshes and avoiding to emerge non-manifold edges or vertices. In contrast, the voxel-based SGA-net methods produce mesh with non-manifold vertices and edges due to reconstruction by the Marching Cubes algorithm. The meshes generated by the SGA-net method may be separated into disconnected components, and there may be genus errors in the form of handles and holes. After expensive post-processing topology correction, the voxel-based SGA-net method can produce meshes without self-intersecting faces. However, compared with manual correction, those post-processing topology corrections lack the interpretation of CBCT images and generate non-plausible corrections, as described in [17].

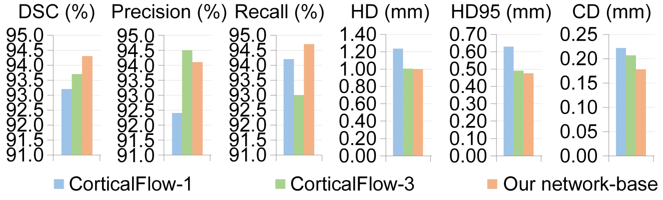


Fig. 9. Histograms Comparison on DSC, Precision, Recall, HD, HD95 and CD respectively between the methods without pyramid flows and ours (Our network-base) with pyramid flows.

D. Ablation Study

We conduct the ablation study to demonstrate the effectiveness of our tooth shape generation network which benefits from multiple novel components. In this study, we present the results of four configurations: (1) CorticalFlow-1 is the CorticalFlow^{++} network with only one deformation. (2) CorticalFlow-3 follows CorticalFlow^{++} [2] which uses three successive deformations. (3) Our network-base represents considering pyramid flows with CD loss. (4) Our network-occ is trained with both CD loss and occupancy loss [32] to compare the effectiveness of CBD loss with overlap ratio loss. (5) Our network-cbd combines the CD and CBD losses. All configurations were trained for 1200 epochs, except for the 2nd and 3rd deformations of CorticalFlow-3, which were trained for 600 epochs. Note that all four configurations use the same detector as the 1st-stage network.

Effects of Pyramid Flows. To validate the effectiveness of our PDD module, CorticalFlow-1 is adopted as the baseline network. Both have only one layer of deformation. Compared with CorticalFlow-1, PDD module significantly improves the geometric accuracy and segmentation performance. As shown

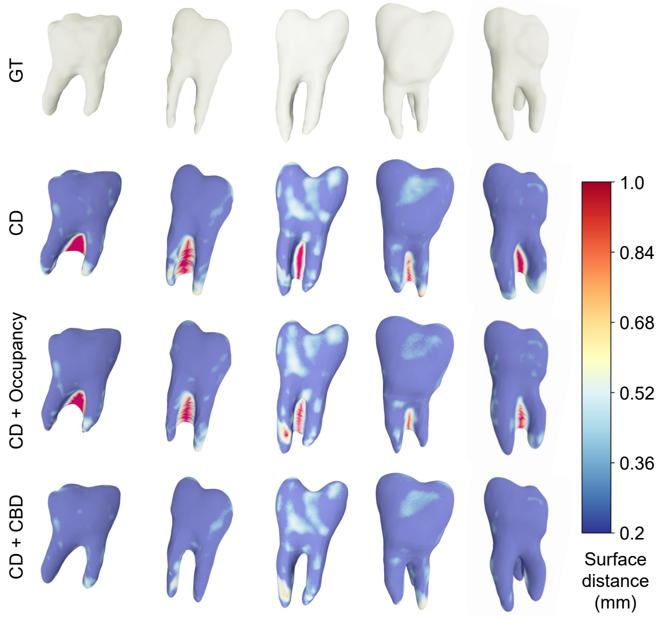


Fig. 10. Comparison of the surface distance between the loss without continuous mapping and ours (CD+CBD) with continuous mapping. Neither the CD loss based on surface distance nor CD+occupancy loss [32] takes mesh topology into account. Therefore, they are unable to penalize instances where there are discontinuous nearest-neighbour mappings between the predicted and ground truth meshes.

TABLE VI
COMPARISON ON INFERENCE TIME AND TRAINING TIME

Model	Inference runtime (s)	Training time (h)
CorticalFlow-3 [2]	0.741	56.06
CorticalFlow-1	0.324	22.82
Our network-cbd	**0.229**	20.20
Our network-base	**0.229**	**18.87**

in Table V and Fig. 9, the average percentages of surface distance (HD95, HD, CD) decrease by 24.3%, 19.4%, 19.8%, and the average increases of segmentation accuracy (DSC, Precision, Recall) are 1.1%, 1.7%, 0.5%. When the original deformation and PDD have the same integration steps and output meshes with the same number of vertices, PDD requires fewer vertices to be calculated in the initial layers. As a result, compared with CorticalFlow-1 and CorticalFlow-3, our network-base significantly decreases the time of reconstruction and training (see Table VI). And, it is easier to train PDD, and our method gets slightly better results than CorticalFlow-3 in geometric accuracy and segmentation performance.

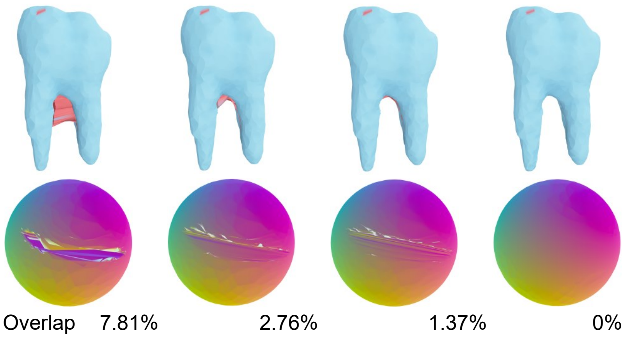


Fig. 11. Visualization of reconstruction quality under the mappings with different overlap ratios. Note that the overlap ratio could be modulated according to the laplacian smoothing in CBM.

TABLE VII
EFFECTS OF CONTINUOUS BIJECTION DISTANCE LOSS ON OUR NETWORK: ABLATION STUDY

Model	DSC	Precision	Recall	CD (mm)	HD (mm)	HD95 (mm)
Our network-base	0.943	0.941	0.947	0.178	0.997	0.476
Our network-occ	0.943	**0.948**	0.939	0.182	1.010	0.409
Our network-cbd	**0.945**	0.942	**0.949**	**0.168**	**0.862**	**0.389**

Effects of CBD Loss. We use the novel CBD loss to supervise the training of our network-cbd to validate its effectiveness. The quantitative results are shown in Table VII. With the CBD loss, the geometric accuracy of the predicted mesh is significantly improved, and the average decrease of surface distance (HD95, HD) is 18.3% and 13.5%. Furthermore, we also employed the CBD loss for training Voxel2mesh on publicly available datasets [16], denoted as Voxel2mesh-w. When compared to the training results of Voxel2mesh-w/o, which did not utilize the CBD loss, the precision of tooth mesh generation by Voxel2mesh-w exhibited a notably higher level of accuracy, as illustrated in Table VIII. Compared with other losses used to penalize surface distance and overlap ratio, only the CBD loss can penalize discontinuous errors of nearest-neighbour mapping between the predicted shape and target shape, then ensure the predicted shape converges to the target shape, as shown in Fig. 10. The deformation direction guided by CD loss falls into a wrong local minimal, while CBD loss avoids such errors, especially for the molars with multiple roots. We also visualized the target meshes of CD loss and CBD loss in Fig. 11. The CD loss guides predicted meshes significantly different from the ground truth shape due to the presence of discontinuous errors in nearest-neighbour mapping. In contrast, because CBD loss makes the nearest-neighbour mapping become continuous and bijective, so the predicted mesh is closer to the ground truth mesh. To further analyze the effectiveness, we visualized the backpropagation of CBD loss and CD loss at vertices position in Fig. 12. Using the CBD loss to train our network resulted in an approximately 7% increase in training time, with GPU memory usage remaining unchanged at 22966MB.

Step Size of Euler Method. To assess the impact of step size on the accuracy of diffeomorphic deformation solutions, we conducted ablation experiments with different step sizes. Following the approach in medical image registration [59] for evaluating the computational accuracy of diffeomorphic registration, the displacement error is defined as the distance between the initial and final positions of any point on the image after applying the deformation fields and their inverse sequentially. We use displacement error to evaluate the accuracy of our diffeomorphic deformation. The deformation from the template to the target shape is denoted as Φ, and its inverse deformation as Φ^{-1}. Both are solved using the Euler method, with reverse accumulation based on the flow vector field. For

TABLE VIII
EFFECTS OF CONTINUOUS BIJECTION DISTANCE LOSS ON MESH GENERATION METHODS IN PUBLICLY AVAILABLE DATASET: ABLATION STUDY

Model	DSC	CD (mm)	HD (mm)	HD95 (mm)
Voxel2mesh w/o [18]	0.936	0.190	1.156	0.498
Voxel2mesh w	0.940	0.183	1.100	0.466

TABLE IX
EFFECTS OF STEP SIZE ON THE ACCURACY OF DIFFEOMORPHIC DEFORMATION SOLUTIONS: ABLATION STUDY

Step Size	$\Delta\boldsymbol{u}(\Phi, \Phi^{-1})$	$\Delta\boldsymbol{u}(\Phi^{-1}, \Phi)$	Cost Time	SIF
(N, h)	Mean, Max (mm)	Mean, Max (mm)	(ms)	($1e-4$)
3, $10^{-0.48}$	1.088, 12.11	0.522, 7.263	3.07	5.526
10, $10^{-1.0}$	0.343, 4.843	0.181, 3.146	7.597	0.229
31, $10^{-1.49}$	0.112, 1.714	0.061, 1.160	22.47	0.142
100, $10^{-2.0}$	0.035, 0.545	0.019, 0.372	60.77	0.136
316, $10^{-2.5}$	0.011, 0.174	0.006, 0.119	168.6	0.132
1000, $10^{-3.0}$	0.004, 0.055	0.002, 0.038	515.3	0.144
3162, $10^{-3.5}$	0.001, 0.018	0.0006, 0.012	1589	0.145

all 3D pixel coordinates within the field, denoted as I, the displacement error resulting from the sequential application of Φ and Φ^{-1} is given by $\Delta\boldsymbol{u}(\Phi, \Phi^{-1}) = \left|I - \Phi \circ \Phi^{-1}\right|$. Similarly, the displacement error generated by the reversed sequence of deformations, Φ^{-1} and Φ, is expressed as $\Delta\boldsymbol{u}(\Phi^{-1}, \Phi)$. The mean and maximum displacement errors are used as metrics for evaluating the accuracy of diffeomorphic deformation. Additionally, we also compute the ratio of self-intersection faces (SIF) under different step sizes h, and measure the computation time for solving Φ using the Euler method. The step size is denoted by h, and the number of steps is given by $N = 1/h$. The result is shown in Table IX. As the step size h decreases exponentially from top to bottom, the number of steps N required for the solution increases rapidly. The number of steps N is approximately linearly correlated with the computational time. When $N = 316$ and $h = 10^{-2.5}$, the maximum displacement errors $\Delta\boldsymbol{u}(\Phi, \Phi^{-1})$ and $\Delta\boldsymbol{u}(\Phi^{-1}, \Phi)$ are both smaller than $0.2mm$, which is half the length of a pixel, indicating that the error in the diffeomorphic deformation will not affect the CBCT segmentation. As the step size continues to decrease, the displacement error keeps reducing, but the SIF does not decrease significantly. Therefore, we set the step size $h = 10^{-2.5}$.

VI. CONCLUSION

In this paper, we propose a pyramid diffeomorphic deformation network under a two-stage framework for reconstructing the tooth meshes from CBCT images. It is an efficient method that directly generates genus-0 tooth mesh from volume data without post-processing and with better accuracy than current segmentation methods. Its success owns to two keys: one is the pyramid diffeomorphic deformation that makes progressive deformation from low resolution and high resolution, and another is the continuous bijection distance loss that enforces the mesh to deform along the continuous manifold surface. Experiments validate that our approach demonstrates a significant improvement compared to the state-of-the-art methods in terms of segmentation accuracy and surface distance. The only lackness is that our subdivision in PDD is not adaptive to input, and it may bring unnecessary computation because of density heterogeneity. In the future, we will develop a mesh deformation mechanism with a subdivision that is adaptive to both the input and the flows, to improve accuracy and efficiency further.

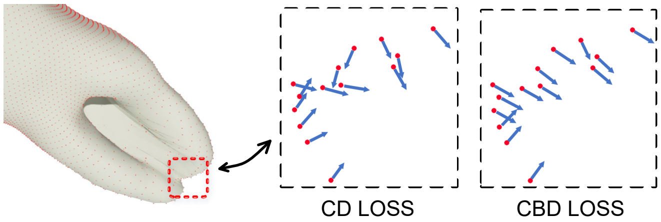


Fig. 12. Visualization of deformation via CD loss vs. CBD loss: red points mark the vertices, and blue lines mark the directions brought by backpropagation of the loss. Note that the deformation directions guided by CD loss are rather haphazard at the molars with multiple roots, while our CBD loss consistently guides deformation directions towards the correct orientation.

REFERENCES

[1] P. Li, Y. Liu, Z. Cui, F. Yang, Y. Zhao, C. Lian, and C. Gao, "Semantic graph attention with explicit anatomical association modeling for tooth segmentation from CBCT images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3116–3127, 2022.

[2] R. Santa Cruz, L. Lebrat, D. Fu, P. Bourgeat, J. Fripp, C. Fookes, and O. Salvado, "CorticalFlow^{++}: Boosting cortical surface reconstruction accuracy, regularity, and interoperability," in *Medical Image Computing and Computer Assisted Intervention*, 2022, pp. 496–505.

[3] G. Wei, Z. Cui, Y. Liu, N. Chen, R. Chen, G. Li, and W. Wang, "TANet: Towards fully automatic tooth arrangement," in *European Conference on Computer Vision*, 2020, pp. 481–497.

[4] L. Yang, Z. Shi, Y. Wu, X. Li, K. Zhou, H. Fu, and Y. Zheng, "iOrthoPredictor: Model-guided deep prediction of teeth alignment," *ACM Transactions on Graphics*, vol. 39, no. 6, pp. 216:1–216:15, 2020.

[5] Y. Gan, Z. Xia, J. Xiong, G. Li, and Q. Zhao, "Tooth and alveolar bone segmentation from dental computed tomography images," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 196–204, 2018.

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

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

[8] H. Akhoondali, R. A. Zoroofi, and G. Shirani, "Rapid automatic segmentation and visualization of teeth in CT-scan data," *Journal of Applied Sciences*, vol. 9, pp. 2031–2044, 2009.

[9] S. Barone, A. Paoli, and A. V. Razionale, "CT segmentation of dental shapes by anatomy-driven reformation imaging and B-spline modelling," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 32, no. 6, pp. e02 747:1–e02 747:22, 2016.

[10] Z. Cui, C. Li, and W. Wang, "ToothNet: Automatic tooth instance segmentation and identification from cone beam CT images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6361–6370.

[11] 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 *IEEE International Symposium on Biomedical Imaging*, 2020.

[12] J. Lee, M. Chung, M. Lee, and Y.-G. Shin, "Tooth instance segmentation from cone-beam CT images through point-based detection and Gaussian disentanglement," *Multimedia Tools and Applications*, vol. 81, no. 13, pp. 18 327–18 342, 2022.

[13] M. Chung, M. Lee, J. Hong, S. Park, J. Lee, J. Lee, I.-H. Yang, J. Lee, and Y.-G. Shin, "Pose-aware instance segmentation framework from cone beam CT images for tooth segmentation," *Computers in Biology and Medicine*, vol. 120, pp. 103 720:1–103 720:11, 2020.

[14] T. J. Jang, K. C. Kim, H. C. Cho, and J. K. Seo, "A fully automated method for 3D individual tooth identification and segmentation in dental CBCT," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6562–6568, 2022.

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

[16] Z. Cui, Y. Fang, L. Mei, B. Zhang, B. Yu, J. Liu, C. Jiang, Y. Sun, L. Ma, J. Huang *et al.*, "A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images," *Nature Communications*, vol. 13, no. 1, pp. 2096:1–2096:11, 2022.

[17] F. Ségonne, J. Pacheco, and B. Fischl, "Geometrically accurate topology-correction of cortical surfaces using nonseparating loops," *IEEE Transactions on Medical Imaging*, vol. 26, no. 4, pp. 518–529, 2007.

[18] U. Wickramasinghe, E. Remelli, G. Knott, and P. Fua, "Voxel2Mesh: 3D mesh model generation from volumetric data," in *Medical Image Computing and Computer Assisted Intervention*, 2020, pp. 299–308.

[19] D. Paschalidou, L. Van Gool, and A. Geiger, "Learning unsupervised hierarchical part decomposition of 3D objects from a single RGB image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1057–1067.

[20] Z. Cui, C. Li, N. Chen, G. Wei, R. Chen, Y. Zhou, D. Shen, and W. Wang, "TSegNet: An efficient and accurate tooth segmentation network on 3D dental model," *Medical Image Analysis*, vol. 69, pp. 101 949:1–101 949:12, 2021.

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

[22] P. Li, C. Gao, C. Lian, and D. Meng, "Spatial prior-guided bi-directional cross-attention transformers for tooth instance segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 11, pp. 3936–3948, 2024.

[23] X. Wang, L. Wang, Z. Yang, J. Zhou, Y. Zheng, F. Chen, R. Hong, J. Yu, and F. Yang, "DSIS-DPR: Structured instance segmentation and diffusion prior refinement for dental anatomy learning," *IEEE Transactions on Multimedia*, vol. 26, pp. 9464–9476, 2024.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[25] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, pp. 1–12, 2019.

[26] Y. Liao, S. Donné, and A. Geiger, "Deep Marching Cubes: Learning explicit surface representations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2916–2925.

[27] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler, "Deep Marching Tetrahedra: A hybrid representation for high-resolution 3D shape synthesis," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 6087–6101.

[28] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2Mesh: Generating 3D mesh models from single RGB images," in *European Conference on Computer Vision*, 2018, pp. 55–71.

[29] Z. Xu, W. Chang, Y. Zhu, L. Dong, H. Zhou, and Q. Zhang, "Building high-fidelity human body models from user-generated data," *IEEE Transactions on Multimedia*, vol. 23, pp. 1542–1556, 2021.

[30] F. Kong and S. C. Shadden, "Learning whole heart mesh generation from patient images for computational simulations," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 533–545, 2022.

[31] K. Gupta and M. Chandraker, "Neural Mesh Flow: 3D manifold mesh generation via diffeomorphic flows," in *Advances in Neural Information Processing Systems*, 2020, pp. 1747–1758.

[32] D. Paschalidou, A. Katharopoulos, A. Geiger, and S. Fidler, "Neural Parts: Learning expressive 3D shape abstractions with invertible neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3203–3214.

[33] B. Deng, K. Genova, S. Yazdani, S. Bouaziz, G. Hinton, and A. Tagliasacchi, "CvxNet: Learnable convex decomposition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 31–41.

[34] Z. Hao, H. Averbuch-Elor, N. Snavely, and S. Belongie, "DualSDF: Semantic shape manipulation using a two-level representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7628–7638.

[35] S. Sun, K. Han, D. Kong, H. Tang, X. Yan, and X. Xie, "Topology-preserving shape reconstruction and registration via neural diffeomorphic flow," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 813–20 823.

[36] J. Yang, U. Wickramasinghe, B. Ni, and P. Fua, "ImplicitAtlas: Learning deformable shape templates in medical imaging," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 840–15 850.

[37] E. Dupont, A. Doucet, and Y. W. Teh, "Augmented neural ODEs," in *Advances in Neural Information Processing Systems*, 2019, pp. 3140–3150.

[38] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Occupancy Flow: 4D reconstruction by learning particle dynamics," in *IEEE International Conference on Computer Vision*, 2019, pp. 5378–5388.

[39] Q. Ma, L. Li, E. C. Robinson, B. Kainz, D. Rueckert, and A. Alansary, "CortexODE: Learning cortical surface reconstruction by neural ODEs," *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 430–443, 2022.

[40] J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi, "Learning a probabilistic model for diffeomorphic registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2165–2176, 2019.

[41] M. Wilms, J. J. Bannister, P. Mouches, M. E. MacDonald, D. Rajashekar, S. Langner, and N. D. Forkert, "Invertible modeling of bidirectional relationships in neuroimaging with normalizing flows: Application to brain aging," *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2331–2347, 2022.

[42] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2463–2471.

[43] J. Zhang, J. Pang, J. Yu, and P. Wang, "An efficient assembly retrieval method based on Hausdorff distance," *Robotics and Computer-Integrated Manufacturing*, vol. 51, pp. 103–111, 2018.

[44] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[45] X. Zuo, S. Wang, J. Zheng, W. Yu, M. Gong, R. Yang, and L. Cheng, "SparseFusion: Dynamic human avatar modeling from sparse RGBD images," *IEEE Transactions on Multimedia*, vol. 23, pp. 1617–1629, 2021.

[46] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems*, 2018, p. 6572–6583.

[47] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 642–656, 2020.

[48] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[49] J. Xie, Y. Pang, J. Nie, J. Cao, and J. Han, "Latent feature pyramid network for object detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 2153–2163, 2023.

[50] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 1968–1979, 2022.

[51] J. Stam, "Evaluation of loop subdivision surfaces," in *ACM SIGGRAPH*, 1998, pp. 1–15.

[52] *Blender: A 3D modelling and rendering package*, Blender Foundation, Amsterdam, the Netherlands, 2021, release 2.92.

[53] M. Baumgartner, P. F. Jäger, F. Isensee, and K. H. Maier-Hein, "nnDetection: A self-configuring method for medical object detection," in *Medical Image Computing and Computer Assisted Intervention*, 2021, pp. 530–539.

[54] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021, pp. 1–16.

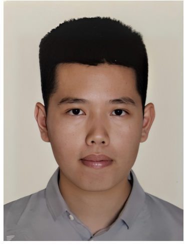
[55] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, 2020, pp. 213–229.

[56] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 1748–1758.

[57] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Medical Image Computing and Computer Assisted Intervention Brainlesion Workshop*, 2021, pp. 272–284.

[58] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman, "3D UX-Net: A large kernel volumetric convNet modernizing hierarchical transformer for medical image segmentation," in *International Conference on Learning Representations*, 2023, pp. 1–15.

[59] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Medical Image Analysis*, vol. 57, pp. 226–236, 2019.



Zechu Zhang received the B.Eng. degree in computer science from the Dongguan University of Technology, Dongguan, China, in 2019. He is currently pursuing the Ph.D. degree in computer science with the School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. His current research interests include medical image analysis and intelligent graphics.

Weilong Peng received the Ph.D. degree in computer application technology from the Tianjin University, Tianjin, China, in 2017. He is currently a Lecturer with the Guangzhou University, Guangzhou, China. He was a Senior Researcher with the Tencent Youtu Lab, Tencent Technology (Shenzhen) Co., Ltd. His current research interests include image processing, computer vision, deep learning, and intelligent graphics.

Jinyu Wen received the M.Sc. degree in engineering from the Guangxi University for Nationalities, Nanning, China, in 2019. She is currently pursuing the Ph.D. degree in computer science with the School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. Her current research interests include machine learning, deep learning, and medical image analysis.

Keke Tang received the B.Eng. degree in computer science from the Jilin University, Changchun, China, in 2012, and the Ph.D. degree in computer science from the University of Science and Technology of China, Hefei, China, in 2017. He is currently an Associate Professor with the Guangzhou University, Guangzhou, China. He was a Post-Doctoral Fellow with The University of Hong Kong, Hong Kong. His current research interests include robotics, computer vision, computer graphics, and cyberspace security.

Meie Fang received the Ph.D. degree in applied mathematics from Zhejiang University, Hangzhou, China. She is currently a Full Professor with the School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. She worked in the Institute of Computer Graphics and Image, Hangzhou Dianzi University from June 2007 to June 2017, and was transferred to Guangzhou University in June 2017. She has served as a Postdoctoral Fellow in the State Key Lab of CAD & CG, Zhejiang University and the Postdoctoral Station of Computer Application Technology, Shanghai Jiao Tong University. She visited City University of Hong Kong and Purdue University of the United States for the purpose of academic exchange several times in recent years. Her current research interests include computer graphics, medical imaging analysis and AI security.

David Dagan Feng (Life Fellow, IEEE) received the M.Eng. degree in electrical engineering and computer science (EECS) from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.Sc. degree in biocybernetics and the Ph.D. degree in computer science from the University of California, Los Angeles (UCLA), Los Angeles, CA, in 1985 and 1988, respectively, where he received the Crump Prize for Excellence in Medical Engineering. He is currently the head in the School of Information Technologies, the director in the Biomedical & Multimedia Information Technology Research Group, and the research director in the Institute of Biomedical Engineering and Technology at the University of Sydney, Sydney, Australia. He has published over 700 scholarly research papers, pioneered several new research directions, and made a number of landmark contributions in his field. More importantly, however, is that many of his research results have been translated into solutions to real-life problems and have made tremendous improvements to the quality of life for those concerned. He has served as the chair in the International Federation of Automatic Control (IFAC) Technical Committee on Biological and Medical Systems, has organized/chaired over 100 major international conferences/symposia/workshops, and has been invited to give over 100 keynote presentations in 23 countries and regions. He is a fellow of the IEEE and Australian Academy of Technological Sciences and Engineering.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently an Assistant Professor with the Department of Computing and an Assistant Professor with the School of Design, The Hong Kong Polytechnic University, Hong Kong. He has published over 250 top-tier scholarly research articles, pioneered several new research directions, and made a series of landmark contributions in his areas. He has an excellent research project reported by the *ACM TechNews*, which only reports the top breakthrough news in computer science worldwide. More importantly, however, many of his research outcomes have strong impacts to research fields, addressing societal needs and contributed tremendously to the people concerned. His current research interests include image/video stylization, colorization, artistic rendering and synthesis, and creative media.