Two-Stage Mesh Deep Learning for Automated Tooth Segmentation and Landmark Localization on 3D Intraoral Scans

Tai-Hsien Wu,

Division of Orthodontics, College of Dentistry, the Ohio State University, Columbus 43210, USA.

Chunfeng Lian [Member, IEEE],

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China.

Sanghee Lee,

Division of Orthodontics, College of Dentistry, the Ohio State University, Columbus 43210, USA.

Matthew Pastewait,

United States Air Force, Kadena AB, Japan.

Christian Piers,

private practice in Morganton, NC, 286555, USA.

Jie Liu,

Division of Orthodontics, College of Dentistry, the Ohio State University, Columbus 43210, USA.

Fan Wang,

Key Laboratory of Biomedical Information Engineering of Ministry of Education, Department of Biomedical Engineering, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China.

Li Wang [Senior Member, IEEE],

Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

Chiung-Ying Chiu,

SOVE Inc, Columbus 43220, USA.

Wenchi Wang,

SOVE Inc, Columbus 43220, USA.

Christina Jackson,

SOVE Inc, Columbus 43220, USA.

Wei-Lun Chao,

Department of Computer Science and Engineering, the Ohio State University, Columbus 43210, USA.

Dinggang Shen [Fellow, IEEE],

School of Biomedical Engineering, ShanghaiTech University Shanghai, China, and Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China.

Ching-Chang Ko

Division of Orthodontics, College of Dentistry, the Ohio State University, Columbus 43210, USA.

Abstract

Accurately segmenting teeth and identifying the corresponding anatomical landmarks on dental mesh models are essential in computer-aided orthodontic treatment. Manually performing these two tasks is time-consuming, tedious, and, more importantly, highly dependent on orthodontists' experiences due to the abnormality and large-scale variance of patients' teeth. Some machine learning-based methods have been designed and applied in the orthodontic field to automatically segment dental meshes (e.g., intraoral scans). In contrast, the number of studies on tooth landmark localization is still limited. This paper proposes a two-stage framework based on mesh deep learning (called TS-MDL) for joint tooth labeling and landmark identification on raw intraoral scans. Our TS-MDL first adopts an end-to-end iMeshSegNet method (i.e., a variant of the existing MeshSegNet with both improved accuracy and efficiency) to label each tooth on the downsampled scan. Guided by the segmentation outputs, our TS-MDL further selects each tooth's region of interest (ROI) on the original mesh to construct a light-weight variant of the pioneering PointNet (i.e., PointNet-Reg) for regressing the corresponding landmark heatmaps. Our TS-MDL was evaluated on a real-clinical dataset, showing promising segmentation and localization performance. Specifically, iMeshSegNet in the first stage of TS-MDL reached an averaged Dice similarity coefficient (DSC) at 0.964 ± 0.054, significantly outperforming the original MeshSegNet. In the second stage, PointNet-Reg achieved a mean absolute error (MAE) of 0.597 ± 0.761 mm in distances between the prediction and ground truth for 66 landmarks, which is superior compared with other networks for landmark detection. All these results suggest the potential usage of our TS-MDL in orthodontics.

Keywords

Tooth Segmentation; Anatomical Landmark Detection; Orthodontic Treatment Planning; 3D Deep Learning; Intraoral Scan

I. INTRODUCTION

Digital 3D dental models have been widely used in orthodontics due to their efficiency and safety. To create a patient-specific treatment plan (e.g., for the fabrication of clear aligners), orthodontists need to segment teeth and annotate the corresponding anatomical landmarks on 3D dental models to analyze and rearrange tooth positions. Manually performing these two tasks is time-consuming, tedious, and expertise-dependent, even with the assistance provided by most commercial software (integrating semi-automatic algorithms). Although there is a clinical need to develop fully automatic methods instead of manual operations, it is practically challenging, especially for tooth landmark localization, mainly due to (1) large-scale shape variance of different teeth, (2) abnormal, disarranged, and/or missing teeth

for some patients, and (3) incomplete dental models (e.g., the second/third molars) captured by intraoral scanners.

Almost 15 year ago, Zhao et al. used curvature values of mesh cells to generate feature contours for tooth segmentation in a semi-automatic manner, indicating the importance of segmentation on dental mesh models in modern dentistry [1]. Recently, some deep learning approaches [2]–[8] have been proposed for end-to-end dental surface labeling. Although these deep networks show superior segmentation accuracy than conventional (semi-)automatic methods, mainly due to task-oriented extraction and fusion of local details and semantic information, very few of them address the critical task of landmark localization. Compared with tooth segmentation, localizing anatomical landmarks is typically more sensitive to varying shape appearance of patients' teeth, as each tooth's landmarks are just small points encoding local geometric details, and the number of landmarks changes across positions.

Considering the natural correlations between the two tasks (e.g., each tooth's landmarks depend primarily on its local geometry), a two-stage framework leveraging mesh deep learning (called TS-MDL) is proposed in this paper for joint tooth segmentation and landmark localization. The schematic diagram of our TS-MDL is shown in Fig. 1.

In Stage 1, we propose an end-to-end deep neural network, called *i*MeshSegNet, to perform tooth segmentation on 3D intraoral scans. *i*MeshSegNet improves the implementation of the multi-scale graph-constrained learning module in its forerunner MeshSegNet [9], [10], a neural network for automatic tooth segmentation. In Stage 2, we extract cells that belong to individual teeth based on the segmentation results generated by Stage 1. By doing this, we narrow the entire intraoral scan down to several ROIs (i.e., individual teeth) since a tooth landmark is always on and only associated with its corresponding tooth. Inspired by the use of heatmaps to successfully detect anatomical landmarks on 2D and 3D medical images [11], we design a modified PointNet [12], called PointNet-Reg, to learn the heatmaps encoding landmark locations. The experimental results on real clinical data indicate that *i*MeshSegNet improves tooth segmentation in terms of both accuracy and efficiency, and the straightforward two-stage strategy leads to promising accuracy in anatomical landmark localization.

The rest of the paper is organized as follows. We briefly review the most related work in Section II, including deep learning on 3D dental models for automated tooth segmentation and heatmap regression for anatomical landmark localization in medical images. Section III describes the studied data and our TS-MDL method. The experimental results and the comparisons between our method and other strategies/approaches are presented in Section IV. We further discuss the effectiveness of some critical methodological designs and our current method's potential limitations in Section V. Finally, the work is concluded in Section VI.

II. RELATED WORK

A. Deep Learning-based 3D Dental Mesh Labeling

Several deep learning-based methods, leveraging convolutional neural networks (CNNs) or graph neural networks (GNNs), have been proposed for automated tooth segmentation on 3D dental meshes. For the CNN-based method, Xu et al. [2] extracted cell-level hand-crafted features to form 2D image-like inputs of a CNN, which predicts the respective semantic label of each cell. Tian et al. [3] converted the dental mesh to a sparse voxel octree model and used 3D CNNs to segment and identify individual teeth. Zhang et al. [6] mapped a 3D tooth model isomorphically to a 2D "image" encoding harmonic attributes. A CNN model was further learned to predict the segmentation mask, which was then transferred back to the original 3D space. These CNN-based methods cannot work directly on the raw dental surfaces, and they typically need to convert 3D meshes (or point clouds) to regular "images" (of hand-crafted features), inevitably losing fine-grained geometric information of teeth.

Inspired by the pioneering PointNet [12] working directly on the 3D shapes for shape-level or point/cell-level classifications, an increasing number of studies proposed to design GNN-based methods for mesh or point cloud segmentation in an end-to-end fashion. For example, Zanjani et al. [4] combined PointCNN [13] with a discriminator in an adversarial setting to automatically assign tooth labels to each point from intraoral scans. Sun et al. [5] proposed to label teeth on digital dental casts by using FeaStNet [14], which was further extended in their more recent work [15] for coupled tooth segmentation and landmark localization. Clinically, the number of teeth can vary between patients, which increases the difficulty of the segmentation task. In order to address this issue, Zanjani et al. proposed Mask-MCNet, which is analogous to Mask-RCNN [16], to conduct instance segmentation on intra-oral scans [7]. In addition, Cui et al. proposed TSegNet, which detects all the teeth using a distance-aware tooth centroid voting scheme, followed by a confidence-aware cascade segmentation module to outperform state-of-the-art approaches [8].

Our previous work, MeshSegNet [9], [10], is also based on GNN, which integrates a series of graph-constrained learning modules to hierarchically extract and integrate multi-scale local contextual features for tooth labeling on raw intraoral scans. Although it has achieved state-of-the-art segmentation accuracy, MeshSegNet has a key limitation – the heavy computational requirements due to the large-scale matrix computations of the large adjacent matrices.

B. Learning-based Landmark Localization

Landmark localization is a crucial task in both computer vision and medical imaging analysis. In 2012, Kumer et al. presented a set of specific methods to automatically identify several dental-specific features (e.g., cusps, marginal ridges, grooves, etc.) on digital dental meshes [17]. However, each method is specific to identify the corresponding feature only, which means this system cannot detect landmarks outside their well-defined domain. In deep learning, a straightforward strategy for landmark localization is to regress the coordinates directly from high-dimensional inputs (e.g., images), but learning such highly nonlinear mappings is technically challenging [18]. To deal with this challenge, various studies in

the computer vision community (e.g., Pfister et al. [18]) proposed to encode the location information of landmarks into Gaussian heatmaps, by which the point localization task is transferred as an easier image-to-image/heatmap regression task. The task of landmark localization is still a very active topic in computer vision, and several novel studies recently emerged with a focus on self-supervised methods. For example, Suwajanakorn et al. presented an end-to-end geometric reasoning framework to learn a latent set of category-specific 2D keypoints with depth information, optimized by multi-view consistency and relative pose estimation [19]. Reddy et al. also presented a graph-based framework, called Occlusion-Net, with a largely self-supervised scheme to predict the 2D and 3D locations of occluded keypoints [20].

In medical imaging analysis, the landmarking task has been widely used in diseases diagnosis, treatment planning, and surgical simulation [21]–[24]. The heatmap regression-based landmark localization has also been successfully applied in the medical imaging domain. For example, by leveraging the end-to-end image-to-image learning ability of fully convolutional networks (FCNs), Payer et al. proposed SpatialConfiguration-Net that incorporates spatial configuration of anatomical landmarks with local appearance to improve the robustness of heatmap regression even with a limited amount of training data [11], [25]. In addition, due to the association between segmentation and landmarking, Zhang et al. [26] aimed to solve the two tasks jointly and adopted heatmap regression for landmark digitization in their approach. In general, these existing studies only focus on detecting landmarks in 2D/3D medical images, without attention to more complicated data structures such as irregular dental meshes.

Recently, some studies have started working on landmark identification for point cloud data. For instance, Fernandez-Labrador et al. proposed an unsupervised learning method for category-specific keypoint identification on 3D point clouds of objects [27]. Maron et al. proposed a method to transfer the surface mesh to a flat torus (i.e., a 2D image) so that the traditional CNN would be able to apply to the corresponding flat torus [28]. However, the number of studies regarding landmark detection on dental mesh models, particularly directly working on 3D data with graph neural networks, is still limited.

III. MATERIALS AND METHOD

A. Dataset

The dataset used in this study consists of 136 patients' raw upper intraoral scans (mesh surfaces), acquired by iTero Element®, a 3D dental intraoral scanner (IOS). The Institutional Review Boards (IRBs) used in this study are IRB# 13-0924 (University of North Carolina at Chapel Hill) and 2020H0459 (The Ohio State University). With the exception of a few cases with missing teeth, each scan has 14 teeth and 66 landmarks. These 66 landmarks are commonly recognized in the orthodontic field and are useful in superimposition as well as calculating tooth movement between pre- and post-treatment scans. The full name for each landmark acronym is given in Table I. Each scan was manually annotated and checked by two experienced orthodontists, serving as the ground truth for the segmentation and localization tasks, respectively. A typical example is shown in Fig. 2, with the names of teeth and corresponding landmarks listed in Table I.

B. TS-MDL

In this paper, we propose a two-stage method for automated identification of anatomical landmarks on 3D intraoral scans. The workflow of this method is shown in Fig. 1, where the two stages correspond to tooth segmentation and heatmap regression, respectively. The idea of this two-stage method is first to segment individual teeth using *i*MeshSegNet, an improvement of our MeshSegNet [9], [10]. After that, the cells belonging to each tooth (i.e., partial mesh) are cropped as an ROI for the localization of the corresponding landmarks. That is, each ROI is fed into an individual PointNet-Reg, a variant of PointNet, to regress the heatmaps that encode the anatomical landmarks on the corresponding tooth. By doing this, we narrow the possible locations of landmarks from the entire intraoral scan down to the specific ROIs, which significantly improves localization efficiency and accuracy.

1) i*MeshSegNet* for Tooth Segmentation: The purpose of Stage 1 is to perform automatic tooth segmentation on raw intraoral scans. This stage includes three steps: pre-processing, inference, and post-processing.

In the pre-processing step, the raw intraoral scans are first downsampled from approximately 100,000 mesh cells (based on iTero Element®) to 10,000 cells. The downsampled mesh further forms an $N \times 15$ vector $\mathbf{F}^0$, which is the input of *i*MeshSegNet, where N is the number of cells. The 15 dimensions correspond to coordinates of the three vertices of each cell (9 units), the normal vector of each cell (3 units), and the relative position of each cell with respect to the whole surface (3 units), respectively. The $N \times 15$ matrix is z-score normalized.

*i*MeshSegNet inherits its forerunner, MeshSegNet [9], [10], and its architecture is shown in Fig. 3. The major difference between MeshSegNet and *i*MeshSegNet is the implementation of the graph-constrained learning module (GLM) that has been proven effective for segmentation [9], [10]. MeshSegNet utilizes symmetric average pooling (SAP) and two adjacency matrices (A_S and A_L in Refs. [9], [10]) to extract local geometric contexts. However, the two $N \times N$ adjacency matrices and the matrix multiplication cause high computational complexity and substantial memory usage when N is large. In order to overcome this drawback, *i*MeshSegNet adopts the EdgeConv operation [29] to replace SAP for local context modeling.

In the inference step, *i*MeshSegNet first consumes the input $\mathbf{F}^0$ ($N \times 15$) with a multi-layer perceptron module (MLP-1 in Fig. 3) to obtain an $N \times 64$ feature vector $\mathbf{F}^1$. A feature-transformer module (FTM) then predicts a 64×64 transformation matrix $\mathbf{T}$ based on the features learned by MLP-1, which maps the $\mathbf{F}^1$ into a canonical feature space by performing a matrix (tensor) multiplication $\widehat{\mathbf{F}}^1 = \mathbf{F}^1 \mathbf{T}$.

The first graph-constrained learning module (i.e., GLM-1) applies EdgeConv [29] to propagate the contextual information provided by neighboring cells on each centroid cell, resulting in a cell-wise feature matrix $\mathbf{F}^{E1} = \left(f_1^{E1}, f_2^{E1}, ..., f_N^{E1} \right)$ that encodes local geometric contexts, as expressed as

$$f_i^{E1} = \max_{j \in \mathfrak{N}(i)} h_\Theta(x_i - x_j, x_i), \tag{1}$$

where j and $\mathfrak{N}(i)$ denote a single adjacent cell and all adjacent cells of cell i, computed by k-nearest neighbor (k-NN) graph; and h_Θ denotes a shared-weight MLP or a 1D Conv layer. Note that the k-NN graph is computed based on the initial graph including self-loop, which is unlike the dynamic graph update in [29], and we adopt $k = 6$ for the k-NN graph in GLM-1. The EdgeConv not only bypasses the expansive matrix multiplication but also has the permutation invariance and partial translation invariance properties [29], leading to better performance in terms of both efficiency and accuracy, as verified by our experiments presented in Section IV.

The feature matrix $\mathbf{F}^{E1}$ then goes through the second MLP module (i.e., MLP-2), constituted by three shared-weight 1D Convs with 64, 128, 512 channels, respectively, becoming an $N \times 512$ feature matrix $\mathbf{F}^2$ and passing through GLM-2. Compared to GLM-1, GLM-2 further enlarges the receptive field to learn multi-scale contextual features. Specifically, $\mathbf{F}^2$ is processed by two parallel EdgeConvs defined on two different k-NN graphs ($k = 6$ and $k = 12$), resulting in two feature matrices $\mathbf{F}^{E2}$ and $\mathbf{F}^{E2'}$. They are then concatenated and fused by a 1D Conv with 512 channels.

A global max pooling (GMP) is applied on the output of GLM-2 to produce the translation-invariant holistic features that encode the semantic information of the whole dental mesh. Then, a dense fusion strategy is used to densely concatenate the local-to-global features from FTM, GLM-1, GLM-2, and upsampled GMP, followed by the third MLP module (MLP-3) to yield an $N \times 128$ feature matrix. Finally, a 1D Conv layer with softmax activation is used to predict an $N \times (C + 1)$ probability matrix where $C = 14$ (i.e., 14 teeth). The effectiveness of the GLM and dense fusion of local-to-global features have been examined systematically [9]. All 1D Convs in iMeshSegNet are followed by batch normalization (BN) and ReLU activation except the final 1D Conv.

In the post-processing step, we first refine the segmentation results predicted by deep neural networks, as they may still contain isolated false predictions or non-smooth boundaries. Similar to [2], [9], we refine the network outputs by using the multi-label graph-cut method [30] that optimizes

$$\underset{L}{\arg\min} \sum_{i=1}^{N} -log(p_i(l_i) + \epsilon) + \lambda \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} V(p_i, p_j, l_i, l_j), \tag{2}$$

where the first and second terms are the data-fitting and local-smoothness terms, respectively; cell i is labeled by the deep network with label l_i under the probability of p_i; ϵ is a small scalar (i.e., 1×10^{-4}) to ensure numerical stability; λ is a non-negative tuning parameter; j and $\mathcal{N}_i$ denote a single nearest-neighboring cell and all nearest-neighboring cells of cell i, respectively. The local smoothness term (the second term in Eq. 2) is expressed as

$$V(p_i, p_j, l_i, l_j) = \begin{cases} 0, & l_i = l_j \\ -log(\frac{\theta_{ij}}{\pi})\phi_{ij}, & l_i \neq l_j, \theta_{ij} \text{ is concave}, \\ -\beta_{ij} log(\frac{\theta_{ij}}{\pi})\phi_{ij}, & l_i \neq l_j, \theta_{ij} \text{ is convex} \end{cases} \quad (3)$$

where θ_{ij} is the dihedral angle between cell i and j; $\phi_{ij} = |c_i - c_j|$ and $\beta = 1 + |\hat{n}_i \cdot \hat{n}_j|$; c and $\hat{n}$ denote the barycenter and normal vector of a cell, respectively. The term β ($\beta = 30$ in this study) enforces the optimization to favor concave regions as the boundaries among teeth and gingiva are usually concave [2].

Finally, we project the segmentation result from the downsampled mesh (approximately 10,000 cells) back to the original intraoral scan (approximately 100,000 cells). To this end, we train a support vector machine (SVM), with the radial basis function (RBF) kernel, using downsampled cells' coordinates and corresponding labels (i.e., predicted segmentations) as training data. Then, we consider all cells in the original high-resolution intraoral scan as test data and predict their labels. By doing this, the prediction from the SVM model is the segmentation result under the original intraoral scan.

2) Point-Reg for Landmark Localization: A tooth anatomical landmark locates on and is determined by the shape of its related tooth. For example, both the mesial and distal contact points (i.e., MCP and DCP in Table I) defined in this study refer to the center of contact areas which are usually located in the upper one third of the crown on the mesial and distal sides of most teeth. Therefore, it is reasonable and more efficient to locate a landmark only using the mesh of the corresponding tooth, instead of the entire dental arch model.

Then, we modify the original PointNet [12] for cell-wise regression, called PointNet-Reg, which learns the non-linear mapping from a tooth ROI to the corresponding heatmaps that encode landmark locations. As shown in Fig. 4, we replace the softmax activation function to the sigmoid function in the final 1D Conv to output an $N' \times N^L$ Gaussian heatmap matrix that encodes the location of the landmarks, where N' and N^L represent the number of cells in the ROI and the number of anatomical landmarks, respectively. After the inference by PointNet-Reg, we choose the centroid cells having the maximum heatmap values as the predicted landmark coordinates, which naturally locate on the original intraoral scan.

C. Implementation

1) Data Augmentation: Due to the symmetry of the dental arch, all 136 samples with their annotated labels as well as 66 landmarks are flipped to double the sample size. Then, we follow the same strategy proposed in [9], [10] to augment these original and flipped data, by combining the operations of 1) random translation, 2) random rotation, and 3) random rescaling. Specifically, along each axis in the 3D space, an intraoral scan has 50% probability to be translated between [−10, 10], rotated between [−π, π], and scaled between [0.8, 1.2], respectively. The combination of these random operations generates 20 "new" cases from each original and flipped scan in this study.

2) **Heatmap Generation:** The ground-truth heatmaps in PointNet-Reg are defined by a Gaussian function, expressed as

$$h_k(\mathbf{x}_i) = H\exp(-\frac{\left(\mathbf{x}_i - \mathbf{x}_k^L\right)^2}{2\sigma_H^2}), \tag{4}$$

where $\mathbf{x}_i$ and $\mathbf{x}_k^L$ are the coordinates of centroid cell i and landmark k, respectively; h_k and H are the Gaussian height for landmark k and the maximum Gaussian height ($H = 1.0$); and σ_H is the Gaussian root mean square (RMS) width ($\sigma_H = 5.0$ mm).

3) **Training Procedure:** The model was trained by minimizing the generalized Dice loss [31] in Stage 1 (i.e., *i*MeshSegNet for tooth segmentation) and mean square error (MSE) in Stage 2 (i.e., PointNet-Reg for landmark regression), respectively, using the AMS-Grad variant of the Adam optimizer [32]. In the training step, we randomly selected 9,000 and 3,000 cells from the downsampled scan (approximate 10,000 cells) and ROI in the original resolution (approximate 3,050 cells) to form the inputs of *i*MeshSegNet (i.e., $N = 9,000$) in Stage 1 and PointNet-Reg (i.e., $N' = 3,000$) in Stage 2, respectively.

4) **Inference Procedure:** The proposed TS-MDL can process the intraoral scans of varying sizes. We downsample the unseen intraoral scan before applying the trained *i*MeshSegNet model on it, which is due to the limited GPU memory (e.g., 11GB for NIVIDIA RTX 2080 Ti). At the end of Stage 1, we transfer the downsampled segmentation result predicted by *i*MeshSegNet back to its original resolution. Then, 14 ROIs are generated in Stage 2, and each ROI is fed into the corresponding PointNet-Reg under the original resolution. In total, 66 heatmaps are predicted from 14 different PointNet-Reg models and converted to the corresponding landmark coordinates. In this study, we conducted experiments to localize 66 landmarks on 14 teeth from the upper dental model, but our TS-MDL method could be straightforwardly extended to include more landmarks and teeth as well as lower dental models (with in-house tests and data not shown) if needed.

IV. EXPERIMENTS

A. Experimental Setup

We split the 136 samples into training, validation, and test sets with a ratio of 65:15:20. Both training and validation sets were augmented using the method described in Sec. III-C1. Using the manual annotations as the ground truth, the segmentation performance in Stage 1 was quantitatively evaluated by Dice similarity coefficient (DSC), sensitiviy (SEN), postive predictive value (PPV), and Hausdorff distance (HD), and the landmark localization performance in Stage 2 was evaluated by the mean absolute error (MAE), which is defined as

$$\text{MAE}(\mathbf{x}_k^P, \mathbf{x}_k^L) = \frac{1}{N^L}\sum_{k=1}^{N^L}|\mathbf{x}_k^P - \mathbf{x}_k^L|, \tag{5}$$

where $\mathbf{x}_k^P$ and $\mathbf{x}_k^L$ denote the predicted and ground-truth coordinates of landmark k, respectively; N^L represents the number of anatomical landmarks.

B. Competing Methods

In the task of tooth segmentation, we compare *i*MeshSegNet with its forerunner MeshSegNet. Note that in our previous work [9], [10], MeshSegNet has already been compared with other state-of-the-art methods, showing superior performance.

In the task of landmark identification, our ***i*MeshSegNet+PointNet-Reg** is compared with another two-stage method (i.e., ***i*MeshSegNet+MeshSegNet-Reg**), where PointNet-Reg is replaced by MeshSegNet-Reg (a variant of MeshSegNet) to regress landmark heatmaps for each tooth based on its segmentation produced by *i*MeshSegNet. In addition, to verify if tooth segmentation can boost landmark localization, these two-stage methods are also compared with two single-stage methods (i.e., **PointNet-Reg** and *i***MeshSegNet-Reg**), which regress landmark heatmaps in an end-to-end fashion from the input 3D dental model.

C. Results

1) Tooth Segmentation Results: The quantitative segmentation results obtained by MeshSegNet and *i*MeshSegNet, both without post-processing, are compared in Table II, which show that *i*MeshSegNet significantly outperformed the original MeshSegNet in terms of all four metrics (i.e., DSC, SEN, PPV, and HD). Fig. 5 further summarizes the segmentation result (in terms of DSC and HD) for each tooth, from which we can see that, being consistent with the average segmentation accuracy in Table II, *i*MeshSegNet also led to better accuracy in segmenting each tooth. The *i*MeshSegNet segmentation results refined by the multi-label graph cut are DSC of 0.964 ± 0.054, SEN of 0.970 ± 0.061, PPV of 0.960 ± 0.054, and HD of 0.995 ± 0.722 *mm*, respectively.

A qualitative comparison among *i*MeshSegNet, MeshSegNet, and the ground truth is visualized in Fig. 6. Each row presents the segmentations of a representative example produced by the two automatic methods (with post-processing refinement) and the ground truth, respectively. We can see that *i*MeshSegNet outperformed its forerunner MeshSegNet in the challenging areas (e.g., those highlighted by the yellow dotted circles).

Besides the segmentation accuracy, we further compared the computational efficiency between MeshSegNet and *i*MeshSegNet. The training and inference time of these two methods, using the same implementation strategy and computational environment, are shown in Table III. From Table III, we can observe that the average training time of *i*MeshSegNet is 253.6 *sec* per epoch for 30 samples, roughly 20 times faster than MeshSegNet. In addition, *i*MeshSegNet only needs roughly 0.62 *sec* to conduct the segmentation of an unseen input, which is approximately 8.6 times faster than MeshSegNet.

The above results show that *i*MeshSegNet significantly outperformed the original MeshSegNet in both accuracy and efficiency, which implies the efficacy of the substitution of SAP and adjacency matrices by EdgeConv.

2) **Landmark Localization Results:** The results of the four competing methods in terms of MAE are summarized in Table IV, which lead to three observations. First, the two-stage methods (i.e, *i*MeshSegNet+PointNet-Reg and *i*MeshSegNet+*i*MeshSegNet-Reg) consistently outperformed the single-stage methods, demonstrating the efficacy of the use of ROI. Second, the single-stage *i*MeshSegNet-Reg shows a better MAE of 1.566 ± 3.711 *mm* compared to PointNet-Reg. This finding matches our expectations since *i*MeshSegNet incorporates a series of GLMs and a dense fusion strategy to learn higher-level features from the raw intraoral scans. Third, *i*MeshSegNet+PointNet-Reg obtained better results than *i*MeshSegNet+*i*MeshSegNet-Reg. Two possible explanations for this controversial finding might be that 1) the ROI implicitly provides the localized information (i.e., the learning difficulty has been significantly reduced) so the advantage of *i*MeshSegNet-Reg on extracting localized information (i.e., knowing where it is in the whole scan) is not as important as in the single-stage, and 2) due to the reduced learning difficulty in the two-stage manner and our small dataset, *i*MeshSegNet-Reg might slightly overfit and its generalization ability is worse on the test set, compared to PointNet-Reg.

Fig. 7 displays the MAE of landmarks on each tooth predicted by the four competing methods. The results from each tooth have similar trends as the overall evaluation reported in Table IV. Although there is no public dataset of intraoral scans and benchmark of mesh segmentation for fair comparison, it is worth noting that our results achieve comparable accuracies, even slightly better, compared to Ref. [15]. The MAEs for the incisor, canine, premolar, and molar are 0.51 *mm*, 0.72 *mm*, 0.51 *mm*, and 0.70 *mm*, respectively, from our *i*MeshSegNet+PointNet-Reg, whereas the MAEs for the same categories are 0.65 *mm*, 0.71 *mm*, 0.86 *mm*, and 0.96 *mm*, respectively, in Ref. [15]. Moreover, higher errors of landmark localization are observed on the first molar (i.e., UR6 and UL6), which is similar to the trend in the segmentation task that prediction accuracy of the molar (UR/L6, UR/L7) is lower than the other teeth. One possible reason is that uncompleted capture often occurs on the molars due to the difficulty in scanning the posterior areas, resulting in the lack of complete information on molars.

Furthermore, Fig. 8 illustrates the qualitative comparisons among the four competing methods and the ground truth for different teeth. The green and purple circles represent landmarks in anterior teeth (i.e., incisor and canine) and posterior teeth (i.e., premolar and molar), respectively. The qualitative observation reveals that *i*MeshSegNet+PointNet-Reg (the fourth column) has the best prediction, which is consistent with the quantitative measurement given in Fig. 7.

V. DISCUSSION

A. Static Adjacency vs. Dynamic Adjacency

Wang et al. [29] suggest using dynamic adjacency matrices defined in high-dimensional feature spaces to model the non-local dependencies between points (or cells) along the forward path of the network. However, due to the specific geometric distribution and arrangement of human teeth, we think it is more reasonable to use the static adjacency (defined in the input Euclidean space) when extracting contextual features even in deeper layers. To justify this claim, we compared the tooth segmentation results obtained by the

aformentioned strategies, as shown in Table V. From Table V, we can see that the model using static adjacency surpassed the one using dynamic adjacency in terms of DSC, SEN, and PPV in our specific task of tooth segmentation on 3D intraoral scans.

B. Clinically Acceptable Errors

According to Tanikawa et. al. [33], several methods were used to evaluate the clinically acceptable range in orthodontics. The simplest computational method is to determine whether the predicted landmarks are located in a circle within a 2 *mm* radius. In addition, the American Board of Orthodontics (ABO) objective grading system considers deviations of 0.5 *mm* as an awareness distinction [34], implying 0.5 *mm* is a high-standard criterion. The deviations over 0.5 *mm* and 1 *mm* will be penalized 1 point and 2 points, respectively, for the alignment and marginal ridge categories. Based on these definitions, we can see that the results predicted from our TS-MDL in Fig. 7 meet the criterion reported in Ref. [33] and approximate the marginal ridge and alignment criteria for the ABO. However, a "passing" case usually has an ABO score of 27 points or less, indicating that errors greater than 0.5 *mm* can still achieve clinically acceptable results.

C. Ceiling Analysis

We performed a ceiling analysis, where the ground truth of segmentation was used as input in Stage 1. The analysis is given in Table VI. Perfect segmentation resulted in 0.087 *mm* improvement in overall accuracy in terms of MAE in the proposed TS-MDL. Based on the ceiling analysis, it appears that refining the work of landmark regression in the future will yield the best overall improvement in the similar multi-stage manner, particularly if we consider 0.5 *mm* as the requirement in terms of accuracy.

D. Limitations and Future Work

Although our TS-MDL leverages the ROI to achieve state-of-the-art performance, it still has some limitations. First, the dataset only contains 136 samples, which is relatively small. In the future, we will keep collecting intraoral scans used in dental clinics. Second, intraoral scans only have surface information. In the heatmap regression-based method, we can only predict those landmarks on the dental mesh surface. However, due to areas of malocclusion, some teeth are overlapped with adjacent teeth, resulting in an incomplete mesh. If a landmark happens to be located on the overlapped area, then the heatmap regression-based method is unable to accurately predict its location. To solve this issue, one of our future works is to introduce 3D dental mesh repairing at the end of Stage 1 in order to reconstruct the incomplete areas from intraoral scans.

VI. CONCLUSION

This study has a two-fold contribution. First, we proposed an end-to-end graph-based neural network, *i*MeshSegNet, for automated tooth segmentation on dental intraoral scans, which improves upon our previous work in terms of the implementation of graph-constrained learning modules. *i*MeshSegNet shows significantly better accuracy in terms of DSC, SEN, PPV, and HD and dramatically reduces computational time in both training and prediction. Second, we proposed the TS-MDL to automatically localize tooth landmarks on intraoral

scans. In TS-MDL, we first predict the segmentation masks using *i*MeshSegNet with graph-cut refinement. We consider the segmentation masks as independent ROIs and then feed them into a series of regression network PointNet-Reg to predict the heatmaps that encode the coordinates of tooth landmarks. Our method can achieve an averaged MAE of 0.597 $\pm$ 0.761 *mm* for localizing 66 landmarks, showing that it has the potential to be utilized in orthodontic applications.

ACKNOWLEDGMENT

REFERENCES

[1]. Zhao M, Ma L, Tan W, and Nie D, "Interactive Tooth Segmentation of Dental Models," in 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, 2005, pp. 654–657.

[2]. Xu X, Liu C, and Zheng Y, "3D Tooth Segmentation and Labeling Using Deep Convolutional Neural Networks," IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 7, pp. 2336–2348, 2019. [PubMed: 29994311]

[3]. Tian S, Dai N, Zhang B, Yuan F, Yu Q, and Cheng X, "Automatic Classification and Segmentation of Teeth on 3D Dental Model Using Hierarchical Deep Learning Networks," IEEE Access, vol. 7, pp. 84817–84828, 2019.

[4]. Ghazvinian Zanjani F, Anssari Moin D, Verheij B, Claessen F, Cherici T, Tan T, and de With PHN, "Deep Learning Approach to Semantic Segmentation in 3D Point Cloud Intra-oral Scans of Teeth," in Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, ser. Proceedings of Machine Learning Research, Cardoso MJ, Feragen A, Glocker B, Konukoglu E, Oguz I, Unal G, and Vercauteren T, Eds., vol. 102. London, United Kingdom: PMLR, 2019, pp. 557–571. [Online]. Available:http://proceedings.mlr.press/v102/ghazvinian-zanjani19a.html

[5]. Sun D, Pei Y, Song G, Guo Y, Ma G, Xu T, and Zha H, "Tooth Segmentation and Labeling from Digital Dental Casts," in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 669–673.

[6]. Zhang J, Li C, Song Q, Gao L, and Lai Y-K, "Automatic 3D tooth segmentation using convolutional neural networks in harmonic parameter space," Graphical Models, vol. 109, p. 101071, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1524070320300151

[7]. Zanjani FG, Moin DA, Claessen F, Cherici T, Parinussa S, Pourtaherian A, Zinger S, and de With PHN, "Mask-MCNet: Instance Segmentation in 3D Point Cloud of Intra-oral Scans BT - Medical Image Computing and Computer Assisted Intervention MICCAI 2019," Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, and Khan A, Eds. Cham: Springer International Publishing, 2019, pp. 128–136.

[8]. Cui Z, Li C, Chen N, Wei G, Chen R, Zhou Y, Shen D, and Wang W, "TSegNet: An efficient and accurate tooth segmentation network on 3D dental model," Medical Image Analysis, vol. 69, p. 101949, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841520303133 [PubMed: 33387908]

[9]. Lian C, Wang L, Wu T, Wang F, Yap P, Ko C, and Shen D, "Deep Multi-Scale Mesh Feature Learning for Automated Labeling of Raw Dental Surfaces from 3D Intraoral Scanners," IEEE Transactions on Medical Imaging, p. 1, 2020. [PubMed: 31135355]

[10]. Lian C, Wang L, Wu T-H, Liu M, Durán F, Ko C-C, and Shen D, "MeshSNet: Deep Multi-scale Mesh Feature Learning for End-to-End Tooth Labeling on 3D Dental Surfaces BT - Medical Image Computing and Computer Assisted Intervention MICCAI 2019," in MICCAI 2019, Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, and Khan A, Eds. Cham: Springer International Publishing, 2019, pp. 837–845.

[11]. Payer C, Štern D, Bischof H, and Urschler M, "Regressing Heatmaps for Multiple Landmark Localization Using CNNs BT - Medical Image Computing and Computer-Assisted Intervention MICCAI 2016," Ourselin S, Joskowicz L, Sabuncu MR, Unal G, and Wells W, Eds. Cham: Springer International Publishing, 2016, pp. 230–238.

[12]. Qi CR, Su H, Mo K, and Guibas LJ, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," p. arXiv:1612.00593, dec 2016. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2016arXiv161200593Q

[13]. Li Y, Bu R, Sun M, Wu W, Di X, and Chen B, "PointCNN: Convolution On X-Transformed Points," p. arXiv:1801.07791, jan 2018. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2018arXiv180107791L

[14]. Verma N, Boyer E, and Verbeek J, "Feastnet: Feature-steered graph convolutions for 3d shape analysis," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2598–2606.

[15]. Sun D, Pei Y, Li P, Song G, Guo Y, Zha H, and Xu T, "Automatic Tooth Segmentation and Dense Correspondence of 3D Dental Model BT - Medical Image Computing and Computer Assisted Intervention MICCAI 2020," Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, and Joskowicz L, Eds. Cham: Springer International Publishing, 2020, pp. 703–712.

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

[17]. Kumar Y, Janardan R, and Larson B, "Automatic Feature Identification in Dental Meshes," Computer-Aided Design and Applications, vol. 9, no. 6, pp. 747–769, jan 2012. [Online]. Available: https://www.tandfonline.com/doi/abs/10.3722/cadaps.2012.747-769

[18]. Pfister T, Charles J, and Zisserman A, "Flowing convnets for human pose estimation in videos," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1913–1921.

[19]. Suwajanakorn S, Snavely N, Tompson J, and Norouzi M, "Discovery of latent 3d keypoints via end-to-end geometric reasoning," arXiv preprint arXiv:1807.03146, 2018.

[20]. Reddy ND, Vo M, and Narasimhan SG, "Occlusion-net: 2d/3d occluded keypoint localization using graph networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7326–7335.

[21]. Zhang J, Gao Y, Gao Y, Munsell BC, and Shen D, "Detecting Anatomical Landmarks for Fast Alzheimer's Disease Diagnosis," IEEE Transactions on Medical Imaging, vol. 35, no. 12, pp. 2524–2533, 2016. [PubMed: 27333602]

[22]. Meng Y, Li G, Lin W, Gilmore JH, and Shen D, "Spatial distribution and longitudinal development of deep cortical sulcal landmarks in infants," NeuroImage, vol. 100, pp. 206–218, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811914004819 [PubMed: 24945660]

[23]. Zhang J, Liu M, and Shen D, "Detecting Anatomical Landmarks From Limited Medical Imaging Data Using Two-Stage Task-Oriented Deep Neural Networks," IEEE Transactions on Image Processing, vol. 26, no. 10, pp. 4753–4764, 2017. [PubMed: 28678706]

[24]. Lian C, Liu M, Zhang J, and Shen D, "Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural MRI," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 880–893, 2020. [PubMed: 30582529]

[25]. Payer C, Štern D, Bischof H, and Urschler M, "Integrating spatial configuration into heatmap regression based CNNs for landmark localization," Medical Image Analysis, vol. 54, pp. 207–219, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841518305784 [PubMed: 30947144]

[26]. Zhang J, Liu M, Wang L, Chen S, Yuan P, Li J, Shen SG-F, Tang Z, Chen K-C, Xia JJ, and Shen D, "Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization," Medical Image Analysis, vol. 60, p. 101621, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841518302093 [PubMed: 31816592]
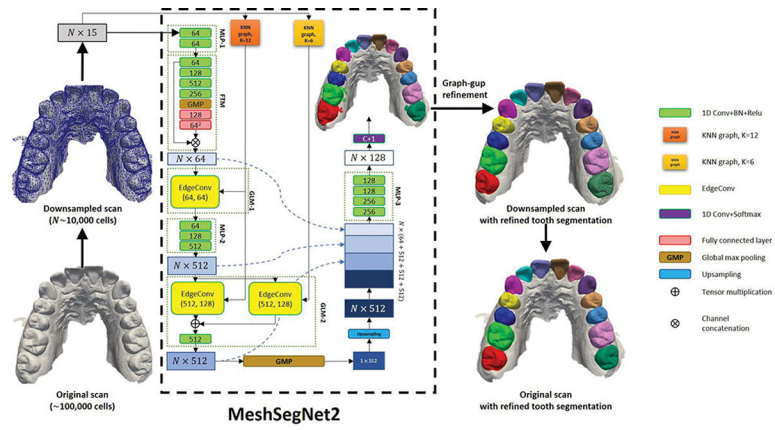
[27]. Fernandez-Labrador C, Chhatkuli A, Paudel DP, Guerrero JJ, Demonceaux C, and Gool LV, "Unsupervised learning of category-specific symmetric 3d keypoints from point sets," in Computer VisionECCV 2020: 16th European Conference, Glasgow, UK, August 2328, 2020, Proceedings, Part XXV 16. Springer, 2020, pp. 546–563.

[28]. Maron H, Galun M, Aigerman N, Trope M, Dym N, Yumer E, Kim VG, and Lipman Y, "Convolutional neural networks on surfaces via seamless toric covers." ACM Trans. Graph, vol. 36, no. 4, p. 71, 2017.

[29]. Wang Y, Sun Y, Liu Z, Sarma SE, Bronstein MM, and Solomon JM, "Dynamic Graph CNN for Learning on Point Clouds," ACM Trans. Graph, vol. 38, no. 5, oct 2019. [Online]. Available: 10.1145/3326362

[30]. Boykov Y, Veksler O, and Zabih R, "Fast approximate energy minimization via graph cuts," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 11, pp. 1222–1239, 2001.

[31]. Sudre CH, Li W, Vercauteren T, Ourselin S, and Jorge Cardoso M, "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations BT - Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support," Cardoso MJ, Arbel T, Carneiro G, Syeda-Mahmood T, Tavares JMRS, Moradi M, Bradley A, Greenspan H, Papa JP, Madabhushi A, Nascimento JC, Cardoso JS, Belagiannis V, and Lu Z, Eds. Cham: Springer International Publishing, 2017, pp. 240–248.

[32]. Reddi SJ, Kale S, and Kumar S, "On the Convergence of Adam and Beyond," p. arXiv:1904.09237, apr 2019. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019arXiv190409237R

[33]. Tanikawa C, Yagi M, and Takada K, "Automated Cephalometry: System Performance Reliability Using Landmark-Dependent Criteria," The Angle Orthodontist, vol. 79, no. 6, pp. 1037–1046, nov 2009. [Online]. Available: [PubMed: 19852592]

[34]. Casko JS, Vaden JL, Kokich VG, Damone J, James R, Cangialosi TJ, Riolo ML, Owens SE, and Bills ED, "Objective grading system for dental casts and panoramic radiographs," American Journal of Orthodontics and Dentofacial Orthopedics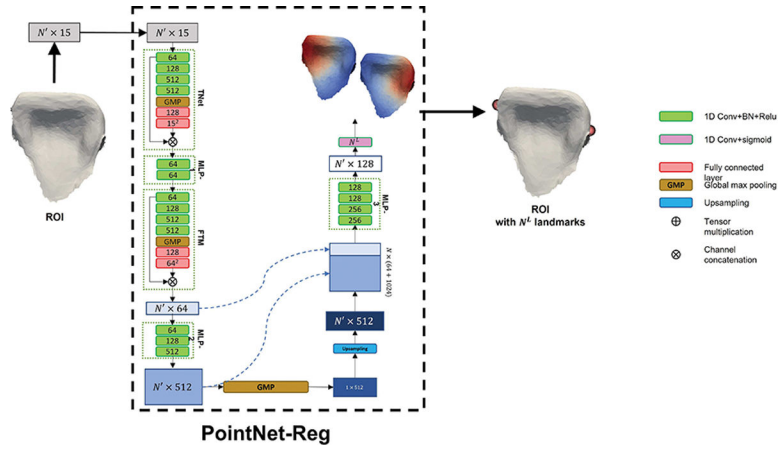, vol. 114, no. 5, pp. 589–599, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0889540698701799 [PubMed: 9810056]
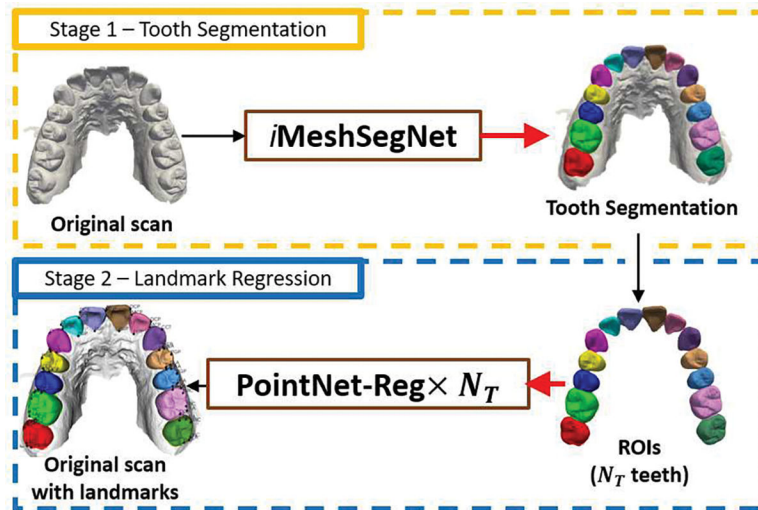
Fig. 1:

The workflow of our two-stage method for automated tooth segmentation and dental landmark localization.

Fig. 2:
The learning targets in this study are the 14 teeth and 66 anatomical landmarks on 14 teeth. The top and bottom graphics are occlusal and frontal views, respectively.
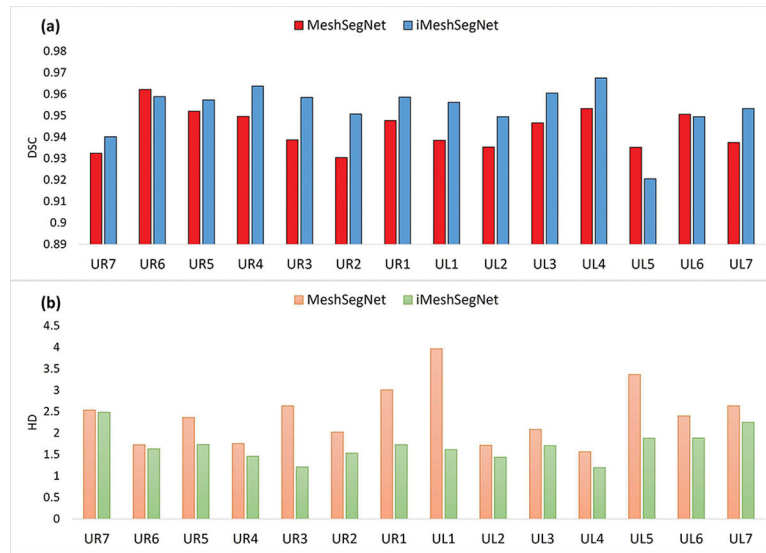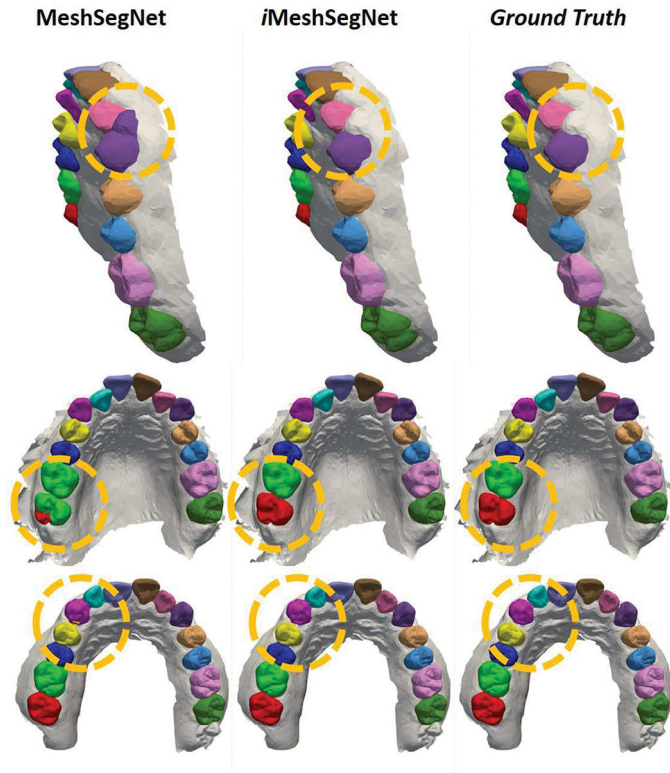
Fig. 3:
The network structure of *i*MeshSegNet for tooth segmentation in Stage 1.

Fig. 4:
The network structure of PointNet-Reg for landmark regression in Stage 2.

Fig. 5:

The results of (a) Dice similarity coefficient (DSC) and (b) Hausdorff distance (HD) of each tooth from MeshSegNet and *i*MeshSegNet.

Fig. 6:

The qualitative comparison among *i*MeshSegNet, MeshSegNet, and ground truth. Each row contains the automated segmentation results of a sample labeled by different methods (with post-processing refinement). The yellow dotted circles highlight the areas that *i*MeshSegNet outperformed its forerunner MeshSegNet, compared to the ground turth.
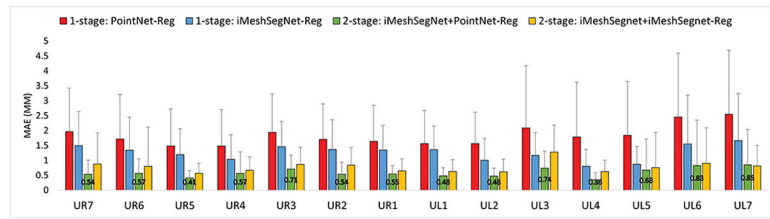
Fig. 7:

The mean absolute error (MAE) of landmarks on each tooth predicted by the four competing methods. The numbers showing on the green bars are the MAE predicted by *i*MeshSegNet+PointNet-Reg for each tooth.
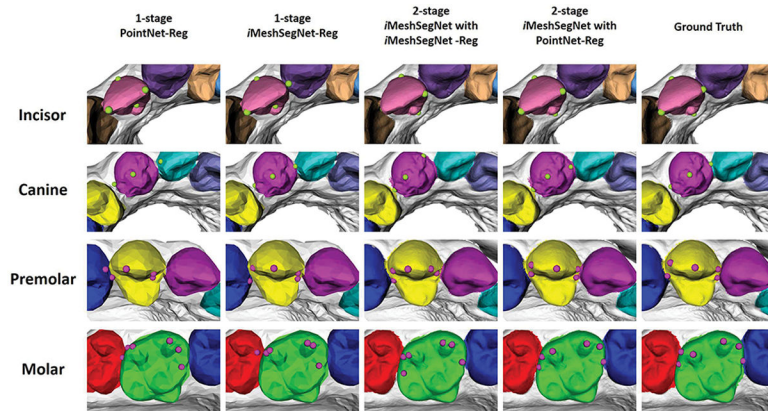
Fig. 8:

The qualitative comparison among the four competing methods and ground truth for different teeth. The green and purple circles represent landmarks in anterior teeth (i.e., incisor and canine) and posterior teeth (i.e., premolar and molar), respectively. The tooth color also indicates its segmentation result. Due to the limited GPU memory, both one-stage strategies (the first and second columns) predict results under downsampled meshes.

TABLE I:

The names of landmarks, numbers of landmarks, and their corresponding teeth.

Tooth name	Landmark name	No. of landmarks
upper right/left central incisor (UR1, UL1)	DCP, MCP, PGP, LGP	4
upper right/left lateral incisor (UR2, UL2)	DCP, MCP, PGP, LGP	4
upper right/left canine (UR3, UL3)	DCP, MCP, CCT	3
upper right/left first premolar (UR4, UL4)	MLA, DLA, PGP, MCP, DCP	5
upper right/left second premolar (UR5, UL5)	MLA, DLA, PGP, MCP, DCP	5
upper right/left first molar (UR6, UL6)	MLA, DLA, MBC, DBC, MCP, DCP	6
upper right/left second molar (UR7, UL7)	MLA, DLA, MBC, DBC, MCP, DCP	6

MCP (mesial contact point); DCP (distal contact point);

PGP (palatal gingival point); LGP (labial gingival point);

CCT (canine cusp tip);

MLA (mesial line angle); DLA (distal line angle);

MBC (mesiobuccal cusp); DBC (distobuccal cusp)

TABLE II:

The comparison of segmentation results without multi-label graph cut refinement between MeshSegNet and *i*MeshSegNet in terms of Dice similarity coefficient (DSC), sensitivity (SEN), positive predictive value (PPV), and Hausdorff distance (HD). Bold font indicates the best result.

Metric	MeshSegNet [9], [10]	*i*MeshSegNet
DSC	0.943 ± 0.044 p=0.0006	**0.953 ± 0.056**
SEN	0.946 ± 0.054 p=0.0183	**0.955 ± 0.064**
PPV	0.943 ± 0.051 p=0.0014	**0.953 ± 0.058**
HD	2.420 ± 3.140 *mm* p=0.0209	**1.696 ± 1.087** *mm*

TABLE III:

The comparison of computing time between MeshSegnet and *i*MeshSegNet. Bold font indicates the best result.

Metric	MeshSegNet [9], [10]	*i*MeshSegNet
Training (sec/epoch)	5114.4	**253.6**
Prediction (sec/scan)	5.33	**0.62**

TABLE IV:

The results of the four competing methods (two single-stage and two two-stage strategies) in terms of mean absolute error (MAE) for the landmark localization. Bold font indicates the best result.

Method	MAE (mm)
1-stage: PointNet-Reg	1.807 ± 1.558
1-stage: *i*MeshSegNet-Reg	1.250 ± 1.021
2-stage: *i*MeshSegNet+PointNet-Reg	**0.597 ± 0.761**
2-stage: *i*MeshSegNet+*i*MeshSegNet-Reg	0.773 ± 0.832

TABLE V:

The comparison of segmentation results based on static adjacency and dynamic adjacency. Bold font indicates the best result.

Metric	Dynamic adjacency	Static adjacency
DSC	0.858 ± 0.169 $p{<}0.0001$	**0.953 ± 0.056**
SEN	0.900 ± 0.175 $p{=}0.0027$	**0.955 ± 0.064**
PPV	0.834 ± 0.165 $p{<}0.0001$	**0.953 ± 0.058**

TABLE VI:

Ceiling analysis of our TS-MDL.

Stage	Accuracy (MAE, mm)	Improvement (mm)
Overall	0.597 ± 0.761	N/A
Stage 1	0.510 ± 0.477	0.087