Automatic 3D Teeth Reconstruction From Five Intra-Oral Photos Using Parametric Teeth Model

Yizhou Chen, Shuojie Gao, Puxun Tu, and Xiaojun Chen, *Member, IEEE*

***Abstract*—Orthodontic treatment is a lengthy process that requires regular in-person dental monitoring, making remote dental monitoring a viable alternative when face-to-face consultation is not possible. In this study, we propose an improved 3D teeth reconstruction framework that automatically restores the shape, arrangement, and dental occlusion of upper and lower teeth from five intra-oral photographs to aid orthodontists in visualizing the condition of patients in virtual consultations. The framework comprises a parametric model that leverages statistical shape modeling to describe the shape and arrangement of teeth, a modified U-net that extracts teeth contours from intra-oral images, and an iterative process that alternates between finding point correspondences and optimizing a compound loss function to fit the parametric teeth model to predicted teeth contours. We perform a five-fold cross-validation on a dataset of 95 orthodontic cases and report an average Chamfer distance of 1.0121 mm^2 and an average Dice similarity coefficient of 0.7672 on all the test samples in the cross-validation, demonstrating a significant improvement compared with the previous work. Our teeth reconstruction framework provides a feasible solution for visualizing 3D teeth models in remote orthodontic consultations.**

***Index Terms*—Statistical shape modeling, parametric teeth model, teeth contour extraction, teeth reconstruction.**

I. Introduction

ORTHODONTIC treatment is a process aimed at correcting misaligned teeth and jaws, which can take up to several years. Regular dental monitoring is crucial to ensure the progress of the treatment. Intra-oral photos are commonly used by orthodontists to monitor the treatment progress and maintain orthodontic records [1], [2]. A typical set of orthodontic intra-oral photographs, as shown in Fig. 1, includes five specific views: anterior, left buccal, right buccal, maxillary, and mandibular.

Fig. 1. The typical set of the orthodontic intra-oral photographs from different views: (a) anterior view, (b) left buccal view, (c) right buccal view, (d) maxillary view, and (e) mandibular view.

Although these photographs provide some geometric information about the teeth, they are two-dimensional and do not fully capture the spatial arrangement of the teeth and jaws, making it difficult for orthodontists to fully understand the patients' conditions.

In the past decade, intra-oral scanners have become increasingly popular among dentists for dental monitoring due to their ability to efficiently generate high-quality 3D digital teeth models [3]. In fact, a recent report by [4] indicates that approximately 53% of dentists in the United States use intra-oral scanners. However, despite their accuracy and convenience, the high cost of investment and the need for specialized knowledge to operate them limit their wider adoption. Furthermore, they are unable to meet the requirements of remote diagnosis caused by the pandemic or other unforeseen circumstances.

Reconstructing the three-dimensional (3D) shape of human teeth from a set of orthodontic intra-oral photographs is an effective solution for remote diagnosis and dental monitoring in orthodontic treatment. This approach facilitates efficient communication between orthodontists and patients, as well as improves productivity for orthodontists in remote situations by presenting the patient's teeth in a digital 3D model format. Additionally, digital 3D teeth models are measurable, which enables orthodontists to measure the displacement of each tooth, observe the dental occlusion of the patient, and evaluate whether the recent progress of orthodontic treatment is in line with the treatment plan. Given that these intra-oral photographs can be taken at home with the guidance of an orthodontist, the restored 3D teeth models can further promote virtual orthodontic consultation. With 3D teeth models, the method in [5] allows users to preview the post-treatment teeth alignment.

In this paper, we present an improved template-based framework for reconstructing digital 3D models of upper and lower teeth from the typical five orthodontic intra-oral photographs, utilizing prior knowledge of human teeth shape. The proposed framework comprises three phases: first, the parameterization of teeth arrangement and shape through statistical shape modeling;

Manuscript received 5 October 2022; revised 30 April 2023; accepted 15 May 2023. Date of publication 19 May 2023; date of current version 1 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 81971709, M-0019, and 82011530141, in part by the Foundation of Science and Technology Commission of Shanghai Municipality under Grant 20490740700, in part by Shanghai Pudong Science and Technology Development Fund under Grant PKX2021-R04, in part by Shanghai Jiao Tong University Foundation on Medical and Technological Joint Science Research under Grants YG2019ZDA06, YG2021ZD21, YG2021QN72, and YG2022QN056. Recommended for acceptance by Y. He. *(Corresponding author: Xiaojun Chen.)*

Yizhou Chen, Shuojie Gao, and Puxun Tu are with the Institute of Biomedical Manufacturing and Life Quality Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yizhou.chen@sjtu.edu.cn; gaoshuojie@sjtu.edu.cn; puxuntu@sjtu.edu.cn).

Xiaojun Chen is with the School of Mechanical Engineering, Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xiaojunchen@sjtu.edu.cn).

Digital Object Identifier 10.1109/TVCG.2023.3277914

second, teeth boundary extraction from intra-oral images using a U-net based method; and finally, 3D teeth reconstruction based on the prior parametric teeth model. Our contribution is two-fold:

- We propose enhancements to an existing 3D teeth reconstruction framework by introducing a robust initialization approach for camera pose estimation and utilizing deep neural networks for teeth boundary extraction. We then apply the improved framework to accurately restore the shape and arrangement of teeth from five orthodontic intra-oral photographs.
- We propose a robust non-rigid registration algorithm to align point sets and establish point correspondences in statistical shape modeling.

The source code is released and updated at: github.com/SJTUzhou/3D-Teeth-Reconstruction-from-Five-Intra-oral-Images.

II. Related Work

Reconstructing the 3D human teeth from multiple images is an ill-posed problem in the field of image-based object reconstruction. The difficulty arises from the fact that the surface of human teeth is typically textureless in photographs, and the spatial arrangement of teeth is not directly predictable from the images. Traditional reconstruction methods, such as Structure from Motion [6], establish and match feature correspondences across different views and reconstruct scenes through epipolar geometry and triangulation. However, this approach becomes highly challenging when the camera positions of the images are far apart or when occlusions are present [7]. To solve this multiple occluded-object 3D reconstruction problem, we draw attention to a broad range of relevant work, the methods of which can be categorized into two groups: methods based on shape modeling and those using deep learning techniques.

A. Reconstruction Methods Based on Shape Modeling

Combining the shape-from-shading (SFS) algorithm with statistical shape priors is a feasible approach for teeth reconstruction. Abdelrahim et al. utilized the SFS algorithm along with the Oren-Nayar diffuse reflection model and the statistical shape priors of teeth to reconstruct the surface of a single tooth [8]. They further enhanced their SFS single tooth surface reconstruction algorithm with a more sophisticated reflection model and a 2D-PCA shape model, which is created by converting 3D tooth shapes to 2D height maps [9]. This type of reconstruction framework can be extended to the reconstruction of the anterior teeth in a dental arch. Farag et al. proposed a model-based SFS approach to reconstruct the visible surface of human jaws from a single optical image by modeling the effect of illumination with spherical harmonics and relating the photometric information to the 3D shape using a statistical model [10]. Mostafa et al. further developed this model-based SFS approach for teeth restoration by introducing a coupled statistical model that links coefficients of the 2D texture model, 3D shape model, and spherical harmonics projection images model [11]. Despite the promising results obtained from leave-one-out tests, the statistical SFS framework of teeth reconstruction requires high-quality photos captured from the same direction and cannot handle cases with missing teeth.

Deforming a 3D morphable model to fit 2D contours or landmarks is a popular framework for 3D shape reconstruction, with extensive research being conducted on 3D face reconstruction under this framework [12]. However, this approach has received limited attention in restoring the 3D shape of teeth. Wu et al. proposed a model-based teeth reconstruction method from a sparse set of extra-oral photographs [13]. They built a parametric model of the entire tooth row and fit it to the 2D teeth boundary extracted from images, modeling the problem as maximum a posteriori estimation. Their method requires minimal manual intervention during initialization and achieves an average euclidean error of 0.86 mm over the non-root teeth vertices on an unspecified test dataset. Wirtz et al. [14] reconstructed the anterior twelve teeth of each dental arch from five orthodontic documentary photos using a model-based 3D reconstruction framework that incorporates 2D contours. They trained and used deep neural networks to do binary segmentation of the related teeth regions in photos. Five view-specific 2D coupled shape models were built and fitted to extract the teeth contours to initialize their 3D coupled shape models. The 2D contours projected from the 3D teeth models were used to drive the optimization process by minimizing a silhouette-based loss. The model-based 3D reconstruction framework incorporating 2D contours shows robust and plausible results in the teeth reconstruction task.

B. Reconstruction Methods via Deep Learning

Deep neural networks (DNN) have demonstrated impressive capabilities in restoring 3D objects and human shapes from images. For instance, Jackson et al. utilized volumetric regression networks to reconstruct a 3D human body from a single image [15], while Natsume et al. introduced an implicit representation using 2D silhouettes and 3D joints of a body pose to restore a textured 3D human body model with deep generative models [16]. Similarly, Mustafa et al. proposed a multi-person reconstruction network that restores the 3D shapes and positions of multiple individuals coherently from a single image [17]. Alldieck et al. presented a novel deep neural network methodology that recovers both geometric and color information of the human body from single monocular images [18]. In addition, Han et al. provided a comprehensive review of the methods that employ deep learning in the field of image-based 3D object reconstruction [19]. Nonetheless, restoring the 3D shapes of teeth with high precision from optical RGB images via deep learning without prior knowledge of their shapes is challenging due to the high level of occlusion among them in optical images.

Recently, there has been research on recovering the 3D shape of teeth from certain types of medical imaging. Cui et al. proposed a DNN-based method to automatically segment and identify tooth instances from cone beam CT (CBCT) images, which produces promising results and does not require any user intervention or post-processing step [20]. In addition, Liang et al. decomposed the task of reconstructing teeth into teeth localization and single-tooth shape estimation, and proposed

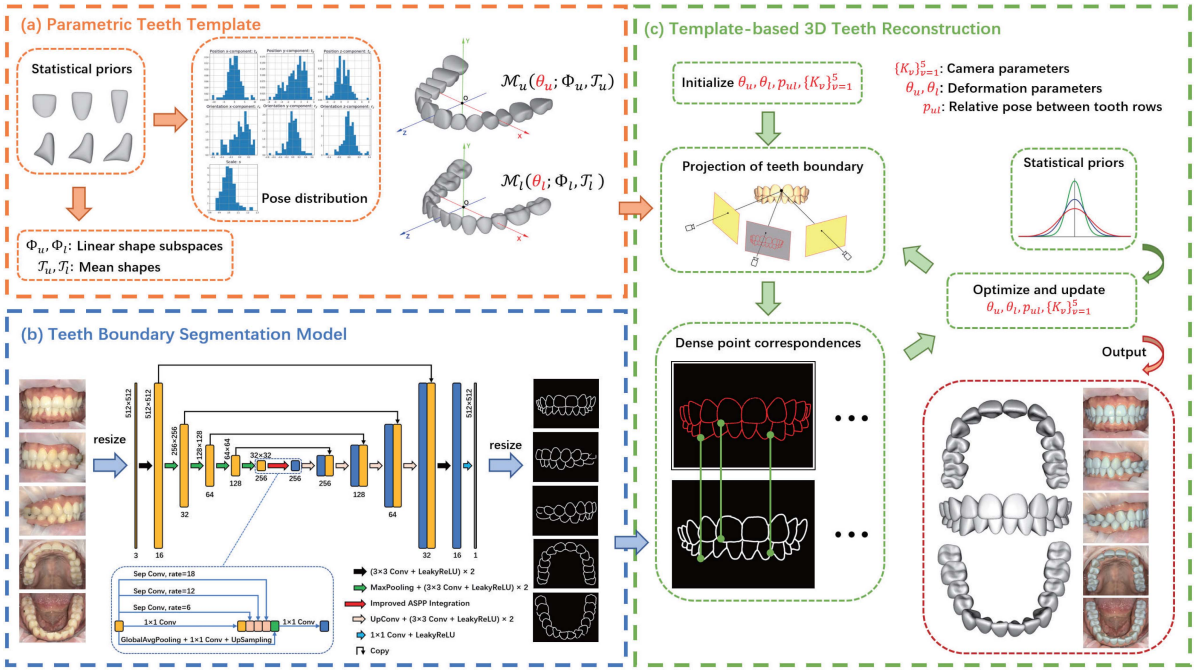


Fig. 2. The template-based reconstruction framework restoring the 3D shape and arrangement of the teeth from five orthodontic intra-oral photographs with the prior knowledge of the shapes and poses of teeth.

a DNN-based framework to restore 3D teeth from a single panoramic radiograph [21]. Moreover, Song et al. reconstructed the 3D oral cavity from a single panoramic X-ray image and prior information of the dental arch, using a generative model to learn the cross-dimension transformation from 2D to 3D and a deformation module with the obtained dental arch curve [22]. Although reconstructing teeth from CBCT images or a single panoramic radiograph is relatively accurate, these methods expose patients to radiation and are not suitable for teleconsultation.

III. Methods

We present a framework for reconstructing 3D teeth models, specifically crowns, from five orthodontic intra-oral photographs. Our approach utilizes prior knowledge of tooth arrangement and shape, as illustrated in Fig. 2. Building upon the methods proposed in [13] and [14], our framework consists of three main components. First, we construct statistical shape models for each tooth in the dental arches, and derive their mean shapes with spatial information to serve as teeth templates. Second, we apply a simplified U-net with atrous spatial pyramid pooling (ASPP) module to extract the boundaries of specific teeth in the intra-oral images. Lastly, we fit the teeth templates to the extracted teeth boundaries in order to restore person-specific 3D teeth models, accomplished through an iterative process of finding 3D-to-2D point correspondences and minimizing a compound loss function.

A. Parametric Teeth Model

The parametric model for the upper or lower teeth characterizes the pose, size, and shape of the teeth in a given tooth row.

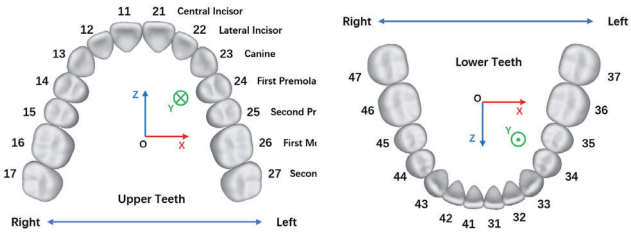


Fig. 3. Two-digit tooth numbers in ISO Tooth Numbering System where the former digit stands for the quadrant of the tooth and the latter digit stands for the number of the tooth from the face midline (wisdom teeth excluded).

Specifically, we focus on tooth rows containing N teeth, where $N = 14$ if there is no tooth missing. The parameterization of both upper and lower teeth is identical. The model for a tooth row can be formulated as $\mathcal{M}(\boldsymbol{\theta}; \boldsymbol{\Phi}, \mathcal{T})$, where $\boldsymbol{\theta}$ represents the set of parameters that captures the variations in pose, size, and shape of the N teeth relative to the teeth template, $\boldsymbol{\Phi}$ represents the set of linear shape subspaces and corresponding eigenvalues of the N teeth, which is obtained through statistical shape modeling, and $\mathcal{T}$ denotes the standard teeth template that encodes the standard poses, mean sizes, and mean shapes of the N teeth. To simplify the notation, we adopt the tooth numbering system suggested by the International Standards Organization (ISO), which assigns two-digit numbers to each tooth, as shown in Fig. 3.

1) Cartesian Coordinates for Teeth: Establishing an appropriate coordinate system for the tooth row is crucial to accurately describe the poses of individual teeth. To this end, we compute the centroid of the vertices of each tooth mesh and represent a tooth row using the N resulting tooth centroids. For the upper tooth rows, we align these tooth centroids rigidly based on point

correspondence and calculate the N average tooth centroids. We follow a similar procedure for the lower tooth rows. The tooth-row coordinate frame, illustrated in Fig. 3, is created as follows:

i) Let origin O the center of the N average tooth centroids;
ii) Let A denote the midpoint of the line connecting the centroids of left and right central incisors, and take the direction of $\overrightarrow{OA}$ as positive Z axis $\overrightarrow{OZ}$;
iii) Let B and C denote the mean centroids of left and right second molars, respectively, and take the direction of $\overrightarrow{OA} \times \overrightarrow{CB}$ as positive Y axis $\overrightarrow{OY}$;
iv) Positive X axis $\overrightarrow{OX}$ is defined by $\overrightarrow{OX} = \overrightarrow{OY} \times \overrightarrow{OZ}$.

Furthermore, we establish a tooth-specific coordinate system for each tooth in the row. The origin of this coordinate system is located at the corresponding average tooth centroid, and the orientation of the X, Y, and Z axes is aligned with that of the tooth-row coordinate system.

2) Point Distribution Models of Teeth: To parameterize the shape of each tooth (a total of $2N$ teeth), we construct point distribution models (PDM). However, constructing a PDM requires the same number of points and dense point correspondence among different samples. Therefore, we develop a robust non-rigid registration algorithm that aligns a group of point sets and identifies the corresponding points. Our algorithm alternates between computing the mean shape as the reference point-set and finding the point correspondences between each point-set and the reference one through non-rigid registration. We use also $M = 1500$ points to describe the surface shape of each tooth as [14] did. Algorithm 1 provides a detailed description of our method for aligning the point-sets of the teeth, determining the tooth poses, and identifying the corresponding points among all samples.

To illustrate our method, we consider the example of the maxillary left central incisor. First, we randomly select a tooth sample and use farthest point sampling (FPS) to downsample it to M points, generating the initial reference point-set Y_{ref}. We then perform similarity registration between each point-set sample X and the reference point-set Y_{ref} using the Coherent Point Drift (CPD) algorithm [23]. To ensure accurate point correspondences, we employ Gaussian process (GP) to model non-rigid deformation that fits the reference point-set Y_{ref} to each registered tooth sample X_t using a user-specified kernel $\boldsymbol{k}$ [24]. For each point $\nu \in Y_{ref}$, we identify its corresponding point in X_t as the closest point. We re-order these points and obtain a group of point-sets with point correspondences $[Y_1, Y_2, \ldots, Y_n]$, where $Y_i \in \mathbb{R}^{M\times 3}$. We take the mean point-set $\overline{Y}$ as the reference point-set in the next iteration and use the mean distance between corresponding points ϵ to measure the variation between two consecutive reference point-sets Y_{ref}^{prev} and Y_{ref} for assessing convergence.

In the non-rigid deformation process, we use the Chamfer distance d_{CD} between two point-sets S_1 and S_2 as the loss function, which is expressed as

$$d_{CD} = \frac{1}{|S_1|}\sum_{x\in S_1}\min_{y\in S_2}\|x-y\|_2^2 + \frac{1}{|S_2|}\sum_{y\in S_2}\min_{x\in S_1}\|y-x\|_2^2. \tag{1}$$

The Gaussian kernel $\boldsymbol{k}(x, x')$ that we choose is a matrix-valued one [25], which is expressed as

$$\boldsymbol{k}(x,x') = s\begin{pmatrix} g(x,x') & 0 & 0 \\ 0 & g(x,x') & 0 \\ 0 & 0 & g(x,x') \end{pmatrix}, \tag{2}$$

where $g(x, x') = exp(-\frac{\|x-x'\|^2}{l^2})$. We get the low-rank approximation of the Gaussian process using its Karhunen–Loève expansion [26]. The parameters s and l in (2) need to be determined in fine tuning.

We use the classical method of statistical shape modeling [27] to build PDMs. The group of registered point-sets with

Algorithm 1: CPD-GP-Based Point-Set Registration.

Input: Initial 3D tooth point-sets $[X_1, X_2, \ldots, X_n]$, a convergence threshold ϵ_{thre}, a matrix-valued kernel $\boldsymbol{k} \in \mathbb{R}^{3\times 3}$.
Output: Registered 3D tooth point-sets with point correspondence $[Y_1, Y_2, \ldots, Y_n]$ where $Y_i \in \mathbb{R}^{M\times 3}$, relative sizes and poses of tooth $[(s_1, \boldsymbol{p_1}), (s_2, \boldsymbol{p_2}), \ldots, (s_n, \boldsymbol{p_n})]$.

procedure point_set_registration
  $(\overline{x_c}, \overline{y_c}, \overline{z_c}) \leftarrow$ Mean of the centroids of $[X_1, \ldots, X_n]$
  $\boldsymbol{t_c} \leftarrow [-\overline{x_c}, -\overline{y_c}, -\overline{z_c}]$
  for $i \leftarrow 1, 2, \ldots, n$ **do**
    $X_i \leftarrow X_i \oplus \boldsymbol{t_c}$[1]
  end for
  $Y_{ref} \leftarrow \mathrm{FPS}(X_1, M)$, $Y_{ref}^{prev} \leftarrow \varnothing$, $\epsilon \leftarrow \infty$
  while $\epsilon > \epsilon_{thre}$ **do**
    $\widehat{\mathrm{GP}}(\mathrm{Y_{ref}}, \boldsymbol{k}) \leftarrow$ Low-rank approx. of $\mathrm{GP}(Y_{ref}, \boldsymbol{k})$
    for $i \leftarrow 1, 2, \ldots, n$ **do**
      $X \leftarrow X_i$
      $(X_t, (\widetilde{s}, \widetilde{\boldsymbol{p}})) \leftarrow \mathrm{SimRegis}(X, Y_{ref}) \triangleright$ CPD
      $s_i \leftarrow 1/\widetilde{s}$, $\boldsymbol{p_i} \leftarrow -\widetilde{\boldsymbol{p}}$
      $\widehat{Y_d} \leftarrow \arg\min_{Y_d} d_{CD}(X_t, \widehat{\mathrm{GP}}(Y_{ref}, \boldsymbol{k}))$
      for $j \leftarrow 1, 2, \ldots, M$ **do**
        $Y_i[j] \leftarrow \arg\min_{\nu\in X_t} \|\widehat{Y_d}[j] - \nu\|_2$
      end for
    end for
    $Y_{ref}^{prev} \leftarrow Y_{ref}$
    $\overline{s} \leftarrow \frac{1}{n}\sum_{i=1}^n s_i$; $\overline{\boldsymbol{p}} \leftarrow \frac{1}{n}\sum_{i=1}^n \boldsymbol{p_i}$; $\overline{Y} \leftarrow \frac{1}{n}\sum_{i=1}^n Y_i$
    $Y_{ref} \leftarrow \mathrm{SimTrans}(\overline{Y}, (\overline{s}, \overline{\boldsymbol{p}}))$
    $\epsilon \leftarrow \frac{1}{M}\sum_{j=1}^M \|Y_{ref}^{prev}[j] - Y_{ref}[j]\|_2$
  end while
end procedure

[1] $\oplus$ denotes adding a row vector to each row in a matrix.

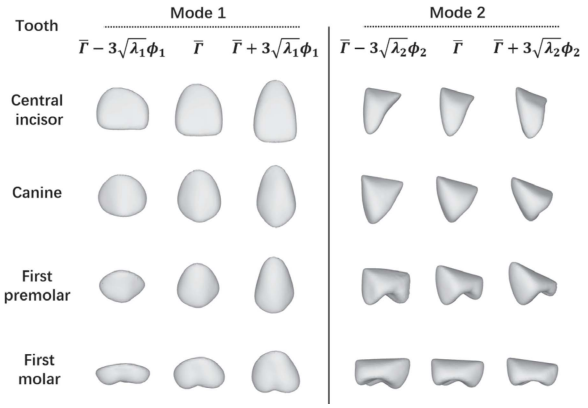


Fig. 4. Visualization of the first two modes of shape variation for the selected teeth in the ride side of the upper jaw where $\bar{\Gamma}$ denotes the mean shape, λ denotes the eigenvalue of this mode, and ϕ denotes its corresponding eigenvector.

point correspondence $[Y_1, Y_2, \ldots, Y_n]$ are flattened into a group of vectors $[\Gamma_1, \Gamma_2, \ldots, \Gamma_n]$ with $\Gamma_i \in \mathbb{R}^{3M\times 1}$ representing the shape of tooth sample i. We denote the reverse reshape transform from $\Gamma \in \mathbb{R}^{3M\times 1}$ to $Y \in \mathbb{R}^{M\times 3}$ as vec^{-1}. A commonly assumed probabilistic model for tooth shape is a normal distribution, i.e., $\Gamma \sim \mathcal{N}(\bar{\Gamma}, \Sigma)$, where $\bar{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i$ represents the mean shape of the teeth and $\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(\Gamma_i - \bar{\Gamma})(\Gamma_i - \bar{\Gamma})^T$ is the covariance matrix. By performing a Principal Component Analysis (PCA) and retaining the first C components ($C = 10$ in this work), a compact probabilistic representation of the tooth shape can be obtained. Specifically, the shape of the tooth can be expressed as

$$\Gamma = \bar{\Gamma} + \sum_{j=1}^{C} b_j\sqrt{\lambda_j}\phi_j, \quad \text{where } b_j \sim \mathcal{N}(0,1), \tag{3}$$

where λ_j and ϕ_j are the jth eigenvalue and eigenvector of the covariance matrix, respectively [24].

In order to obtain a direct interpretation of the shape parameters in the parametric teeth model, we plot the first two modes of shape variation, namely b_1 and b_2, for the right upper teeth in Fig. 4. It is evident that the first mode of shape variation for each tooth is associated with its width length ratio, while the second mode is correlated with the relative size of the facial and lingual tooth surfaces.

3) Parameterization of a Tooth Row: The standard teeth template $\mathcal{T}$ for a tooth row can be represented as $\{\bar{\Gamma}^{(i)}, \boldsymbol{t_c}^{(i)}\}_{i\in\mathcal{I}_u/\mathcal{I}_l}$, where $\mathcal{I}u$ and $\mathcal{I}_l$ denote the numbering sets of upper and lower teeth, respectively. Here, $\bar{\Gamma}^{(i)}$ encodes the mean shape, size, orientation, and position of tooth (i) in its tooth-specific coordinate frame, while $\boldsymbol{t_c}^{(i)}$ represents the translation vector from the tooth-specific coordinate frame to the tooth-row coordinate frame. The parametric N-teeth tooth-row model is expressed as

$$\mathcal{M}(\boldsymbol{\theta};\boldsymbol{\Phi},\mathcal{T}) = \left\{s^{(i)}\ \text{vec}^{-1}\left(\Gamma^{(i)}\right)R\left(\boldsymbol{r}^{(i)}\right)\oplus\left(\boldsymbol{t_c}^{(i)}+\boldsymbol{t}^{(i)}\right)\right\}_{i\in\mathcal{I}_u/\mathcal{I}_l}, \tag{4}$$

where $\boldsymbol{\theta}$ is related to the size s, pose $\boldsymbol{p}$ (position $\boldsymbol{t}$ and orientation $\boldsymbol{r}$), and shape vector $\boldsymbol{b}$ of each tooth, $\boldsymbol{\Phi}$ is the set of linear shape subspaces and the corresponding eigenvalues, and $R(\boldsymbol{r})$ is the rotation matrix related to $\boldsymbol{r}$.

By introducing a relative pose $\boldsymbol{p}_{ul}$ between the upper teeth and the lower ones, the arrangement and shape of the $2N$ teeth can be parameterized by $\{\mathcal{M}_u(\boldsymbol{\theta}_u;\boldsymbol{\Phi}_u,\mathcal{T}_u),\ \mathcal{M}_l(\boldsymbol{\theta}_l;\boldsymbol{\Phi}_l,\mathcal{T}_l),\ \boldsymbol{p}_{ul}\}$.

TABLE I
TOOTH NUMBERS OF THE SELECTED TEETH TO EXTRACT CONTOURS IN DIFFERENT VIEWS

View type ν	Tooth numbers of the outlined teeth $\mathcal{I}_\nu$
1 Anterior view	11-16, 21-26, 31-36, 41-46
2 Left view	11, 21-26, 31-36, 41
3 Right view	11-16, 21, 31, 41-46
4 Maxillary view	11-17, 21-27
5 Mandibular view	31-37, 41-47

B. Teeth Boundary Extraction

We employ a simplified U-net [28] with integrated improved atrous spatial pyramid pooling (ASPP) [29] to extract the boundaries of specific teeth based on their view types and corresponding tooth numbers, as specified in Table I. The improved ASPP integration reduces computation cost by adopting depthwise separable convolution while maintaining performance similar to that of traditional ASPP [30]. It is incorporated after the bottleneck of the encoder part of the U-net, as illustrated in Fig. 2(b). Our teeth boundary segmentation network takes $512 \times 512 \times 3$ RGB images as input and generates 512×512 binary teeth contours.

The choice of loss function affects the performance of boundary segmentation. Ma et al. [31] have shown that using a Dice-related compound loss is often a good choice in medical image segmentation. To further reduce blurriness in the output, we incorporate the Structural Similarity (SSIM) Index, which measures the similarity between images in aspect of structural information by comparing the luminance, contrast, and structure [32]. Therefore, we take the weighted sum of the structural similarity loss L_{SSIM} [29] and the soft Dice loss L_{Dice} [33] as the final loss L to train our neural network, as shown in (5).

$$L = L_{Dice} + \lambda_{ssim}\cdot L_{SSIM}. \tag{5}$$

C. Template-based 3D Teeth Reconstruction

1) Problem Formulation: We adopt a methodology similar to that proposed in [13] for restoring 3D teeth. The approach estimates the camera parameters in each of the five views and determining the parameters of the teeth model, using an iterative process that alternates between identifying corresponding points and minimizing a compound loss function. The ensemble of parameters for estimation denoted by $\mathcal{X}$ is expressed as

$$\mathcal{X} = \{\boldsymbol{T}_\nu, \boldsymbol{K}_\nu\}_{\nu=1}^{5} \cup \{\boldsymbol{p}_{ul}\} \cup \left\{\boldsymbol{t}^{(i)}, \boldsymbol{r}^{(i)}, s^{(i)}, \boldsymbol{b}^{(i)}\right\}_{i\in\mathcal{I}_u\cup\mathcal{I}_l}. \tag{6}$$

Here, $\boldsymbol{T}_\nu \in \mathbb{R}^{4\times 4}$ and $\boldsymbol{K}_\nu \in \mathbb{R}^{3\times 3}$ represent the camera's extrinsic and intrinsic matrices in view ν, respectively. Additionally, $\boldsymbol{p}_{ul}$ denotes the relative pose of the lower tooth row with respect to the upper row, while the position, orientation, size, and

shape parameters of tooth (i) are denoted by $\boldsymbol{t}^{(i)}, \boldsymbol{r}^{(i)}, s^{(i)}, \boldsymbol{b}^{(i)}$, where $i \in \mathcal{I}_u \cup \mathcal{I}_l$.

The ensemble of known parameters $\mathcal{Y}$ is

$$\begin{aligned}\mathcal{Y} = &\{\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\} \cup \{\boldsymbol{\mu}_p^{(i)}, \boldsymbol{\Sigma}_p^{(i)}\}_{i\in\mathcal{I}_u\cup\mathcal{I}_l} \\ &\cup \{\boldsymbol{\Phi}_u, \boldsymbol{\Phi}_l, \mathcal{T}_u, \mathcal{T}_l, \mathcal{I}_u, \mathcal{I}_l\} \cup \{\mathcal{I}_\nu\}_{\nu=1}^{5}, \end{aligned} \tag{7}$$

where $\boldsymbol{\mu}_s$ and $\boldsymbol{\Sigma}_s$ are the mean vector and covariance matrix of the sizes of teeth, $\boldsymbol{\mu}_p^{(i)}$ and $\boldsymbol{\Sigma}_p^{(i)}$ are the mean vector and covariance matrix of the pose of tooth (i), $\boldsymbol{\Phi}_u$ and $\boldsymbol{\Phi}_l$ are the sets of shape subspaces of the upper and lower teeth, $\mathcal{T}_u$ and $\mathcal{T}_l$ are the standard templates of the upper and lower teeth, $\mathcal{I}_u$ and $\mathcal{I}_l$ are the numbering sets of the patient's teeth (usually $\mathcal{I}_u = \{11, 12, \ldots, 17, 21, 22, \ldots, 27\}$ and $\mathcal{I}_l = \{31, 32, \ldots, 37, 41, 42, \ldots, 47\}$ but there could be missing teeth), and $\mathcal{I}_\nu$ is the numbering set of the outlined teeth in view v, defined in Table I.

We apply the method of [13], which is expressed as

$$\widehat{\boldsymbol{c}_i} = \underset{\widehat{\boldsymbol{c}_j}}{\arg\min} \|\boldsymbol{c}_i - \widehat{\boldsymbol{c}_j}\|_2^2 \cdot exp\left(-\frac{\langle \boldsymbol{n}_i, \widehat{\boldsymbol{n}_j}\rangle^2}{\sigma_{angle}^2}\right), \tag{8}$$

to find the correspondences between the points $\{\boldsymbol{c}_i\}$ in the predicted teeth contours and the visible boundary points $\{\widehat{\boldsymbol{c}_j}\}$ of the relevant teeth projected from the parametric teeth model. $\{\boldsymbol{n}_i\}$ and $\{\widehat{\boldsymbol{n}_j}\}$ are the corresponding normal vectors of predicted and projected contour points in the image coordinate, $\langle\cdot,\cdot\rangle$ represents the inner product, and $\sigma_{angle} = 0.3$ [13]. To filter out the bad correspondences, we define the correspondence loss of a predicted contour point $\boldsymbol{c}_i$ as $l(\boldsymbol{c}_i)$, which is expressed as

$$l(\boldsymbol{c}_i) = \|\boldsymbol{c}_i - \widehat{\boldsymbol{c}_i}\|_2^2 \cdot exp\left(-\frac{\langle \boldsymbol{n}_i, \widehat{\boldsymbol{n}_i}\rangle^2}{\sigma_{angle}^2}\right), \tag{9}$$

and mask the point $\boldsymbol{c}_i$ as outlier if $l(\boldsymbol{c}_i) > P_{99}$ where P_{99} is the 99th percentile of l calculated from all predicted contour points in the same view.

The optimal parameters $\mathcal{X}^*$ is obtained by minimizing the sum of the compound loss in different views.

$$\mathcal{X}^* = \underset{\mathcal{X}}{\arg\min} \sum_{\nu=1}^{5} \mathcal{L}(\mathcal{X}, \boldsymbol{B}_\nu; \mathcal{Y}). \tag{10}$$

The compound loss function of one view is composed of three parts: $\mathcal{L}_p$ the point-to-point loss, $\mathcal{L}_n$ the point-to-plane loss, and $\mathcal{L}_{prior}$ the penalization based on the teeth priors.

$$\mathcal{L}(\mathcal{X}, \boldsymbol{B}_\nu; \mathcal{Y}) = w_p \cdot \mathcal{L}_p + w_n \cdot \mathcal{L}_n + \mathcal{L}_{prior}, \tag{11}$$

w_p and w_n are weighting factors, and $\boldsymbol{B}_\nu$ is the predicted teeth contours in view ν. The point-to-point loss $\mathcal{L}_p$ is the average squared euclidean distance between the corresponding points.

$$\mathcal{L}_p = \frac{1}{|\boldsymbol{B}_\nu|} \sum_{i=1}^{|\boldsymbol{B}_\nu|} \|\boldsymbol{c}_i - \widehat{\boldsymbol{c}_i}\|_2^2. \tag{12}$$

The point-to-plane loss $\mathcal{L}_n$ is the mean squared inner product between the vectors of corresponding points $\{\boldsymbol{c}_i - \widehat{\boldsymbol{c}_i}\}$ and the image-plane normals of projected points $\{\widehat{\boldsymbol{n}_i}\}$.

$$\mathcal{L}_n = \frac{1}{|\boldsymbol{B}_\nu|} \sum_{i=1}^{|\boldsymbol{B}_\nu|} \langle \boldsymbol{c}_i - \widehat{\boldsymbol{c}_i}, \widehat{\boldsymbol{n}_i}\rangle^2. \tag{13}$$

The penalization based on the teeth priors $\mathcal{L}_{prior}$ is the sum of squared Mahalanobis distance [34] of the sizes, poses, and shapes of the outlined teeth, which is expressed as

$$\mathcal{L}_{prior} = d_M^{size^2}(\boldsymbol{s}(\mathcal{I}'_\nu)) + \sum_{i\in\mathcal{I}'_\nu} \left[d_M^{pose^2}(\boldsymbol{p}^{(i)}) + d_M^{shape^2}(\boldsymbol{b}^{(i)})\right], \tag{14}$$

where the numbering set of the outlined teeth is defined as $\mathcal{I}'_\nu = (\mathcal{I}_u \cup \mathcal{I}_l) \cap \mathcal{I}_\nu$.

2) Robust Initialization: Adequate initialization of the estimation parameters is paramount for solving the non-linear problem, as it is highly susceptible to local minima. We propose a robust strategy for initializing the camera parameters $\boldsymbol{T}_\nu$ and $\boldsymbol{K}_\nu$ for each view ν, as well as the relative pose $\boldsymbol{p}_{ul}$ between the tooth rows, while the remaining parameters are initialized using the mean values obtained from statistical shape modeling. Given that the size of tooth rows is not negligible in comparison to the distance between teeth and the lens, we utilize perspective camera models and assume that pixels possess uniform aspect ratio. Specifically, $\boldsymbol{T}_\nu$ has six degrees of freedom (DoFs) and $\boldsymbol{K}_\nu$ has three DoFs for each view ν.

The view directions of intra-oral photos exhibit substantial similarity, albeit with noticeable variations attributed to manual operations by the dentists. To account for this, we derive diverse initial camera settings listed in Table II. For each intra-oral image, we perform in parallel similarity registrations between predicted contour points and their projections under various initial settings. We then establish dense point correspondences between the predicted contour points and each transformed projected contour points using (8). The mean correspondence loss, obtained using (9), is computed, and the transformed projected contour point set that yields the minimum mean correspondence loss is retained. Subsequently, we refine the initial camera setting for each view using the Direct Linear Transform algorithm [35] to solve the Perspective-n-Point problem between the transformed projected contour points and their corresponding 3D points in the teeth model.

We take the upper tooth row coordinates as the world coordinates and estimate the relative pose $\boldsymbol{p}_{ul}$ of the lower tooth row with respect to the upper one. The search space of the initial positions of the lower tooth row is listed in Table III and the orientation is initialized by an identity matrix. We adopt a similar approach as that used to identify the best initial camera setting to determine the most appropriate initial value of $\boldsymbol{p}_{ul}$. In our experiments, we first determine the type of dental occlusion, categorized as Class I (slight overbite), Class II (severe overbite), or Class III (severe underbite), which is reflected by the translation along the z-axis in tooth-row coordinates, denoted by z_d. After acquiring z_d, we employ the default $\boldsymbol{p}_{ul}$ (highlighted in bold in Table III) and search for the best initial camera setting. We then fix the camera setting and identify the optimal initial $\boldsymbol{p}_{ul}$. To evaluate the robust initialization strategy, we

TABLE II
INITIAL CAMERA SETTINGS FOR PHOTOS OF DIFFERENT VIEWS WHERE r_x, r_y, r_z IS THE SPACE-FIXED EULER ANGLE REPRESENTATION OF THE INITIAL ORIENTATION IN THE WORLD COORDINATE, t IS THE POSITION VECTOR, AND f_{pix} IS THE FOCAL LENGTH IN PIXELS. THE PRINCIPAL POINT IN THE INTRINSIC PARAMETERS IS INITIALIZED BY THE CENTER OF THE IMAGE

View type	Extrinsic param.				Intrinsic param.
	r_x	r_y	r_z	$t(mm)$	f_{pix}
Anterior view	$0.98\pi, \mathbf{1.0\pi}, 1.02\pi$	0	0	$[0, -2, 120]^T$	1666.67
Left view	1.0π	$0.1\pi, 0.15\pi, \mathbf{0.2\pi}, 0.25\pi, 0.3\pi$	0	$[-5, 0, 120]^T$	1666.67
Right view	1.0π	$-0.1\pi, -0.15\pi, \mathbf{-0.2\pi}, -0.25\pi, -0.3\pi$	0	$[5, 0, 120]^T$	1666.67
Maxillary view	$0.6\pi, 0.65\pi, \mathbf{0.7\pi}, 0.75\pi, 0.8\pi$	0	0	$[0, 0, 70]^T$	1000
Mandibular view	$-0.6\pi, -0.65\pi, \mathbf{-0.7\pi}, -0.75\pi, -0.8\pi$	0	0	$[0, 0, 70]^T$	1000

TABLE III
INITIAL RELATIVE POSITION SETTINGS OF THE LOWER TOOTH ROW WHERE $z_d \in \{0, \mathbf{3}, 6\}$

Param.	$t_x(mm)$	$t_y(mm)$	$t_z(mm)$
Value	$-1, \mathbf{0}, 1$	$-7, -6, \mathbf{-5}, -4$	$\boldsymbol{z_d} \pm \{0, 1, 2, 3\}$

conduct experiments and compare the reconstruction accuracy obtained using it and a fixed initialization (highlighted in bold in Tables II and III).

3) Optimization: We employ a coarse-to-fine optimization strategy to refine the parametric teeth model. Specifically, we apply the sequential least square programming (SLSQP) method with the derived gradient for optimization, while ensuring that the variables in $\{\boldsymbol{T}_\nu\}_{\nu=1}^5$, $\{\boldsymbol{K}_\nu\}_{\nu=1}^5$, and $\boldsymbol{p}_{ul}$ satisfy empirical bounds. Additionally, we restrict $\{\boldsymbol{t}^{(i)}, \boldsymbol{r}^{(i)}, s^{(i)}, \boldsymbol{b}^{(i)}\}_{i\in\mathcal{I}_u\cup\mathcal{I}_l}$ to their scaled standard deviations obtained from statistical shape modeling. To expedite the convergence, we introduce anisotropic scale variables for the tooth row, which can be decomposed into the size and position of each tooth, as they affect the scaling of every vertex in the parametric teeth model. We can express the size $s^{(i)}$ and position $\boldsymbol{t}^{(i)}$ of tooth (i) explicitly in terms of the tooth row's anisotropic scales in the tooth row's coordinate system $\boldsymbol{s}^{row} = [s_x^{row}, \boldsymbol{s}_y^{row}, \boldsymbol{s}^{row}z] \in \mathbb{R}^3$ using element-wise product $\odot$, as shown in (15) and (16). We assume that a person's upper and lower tooth rows share the same anisotropic scales.

$$s^{(i)} = \left(\boldsymbol{s}_x^{row} \cdot \boldsymbol{s}_y^{row} \cdot \boldsymbol{s}_z^{row}\right)^{\frac{1}{3}} \quad \forall\, i \in \mathcal{I}_u \cup \mathcal{I}_l \tag{15}$$

$$\boldsymbol{t}^{(i)} = (\boldsymbol{s}^{row} - \mathbf{1}_{1\times 3}) \odot \boldsymbol{t}_{\boldsymbol{c}}^{(i)} \quad \forall\, i \in \mathcal{I}_u \cup \mathcal{I}_l. \tag{16}$$

Our optimization procedure is as follows:

a) Robust initialization.
b) Take the set of parameters to be optimized: $\widetilde{\mathcal{X}} = \{\boldsymbol{T}_\nu, \boldsymbol{K}_\nu\}_{\nu=1}^5 \cup \{\boldsymbol{p}_{ul}\}$.
c) Project the parametric model $\{\mathcal{M}_u, \mathcal{M}_l, \boldsymbol{p}_{ul}\}$.
d) Find the point correspondences in different views.
e) Optimize using SLSQP and update parameters: $\widetilde{\mathcal{X}}^{k+1} = \arg\min_{\mathcal{X}} \sum_{\nu=1}^5 \mathcal{L}(\widetilde{\mathcal{X}}^k, \boldsymbol{B}_\nu; \mathcal{Y})$.
f) Repeat Step (c) — (e) for 10 iterations to optimize $\widetilde{\mathcal{X}}$.
g) Set $\widetilde{\mathcal{X}} = \{\boldsymbol{T}_\nu, \boldsymbol{K}_\nu\}_{\nu=1}^5 \cup \{\boldsymbol{p}_{ul}, \boldsymbol{s}^{row}\}$ and repeat Step (c) — (e) for 5 iterations.
h) Decompose $\boldsymbol{s}^{row}$ into $\{s^{(i)}, \boldsymbol{t}^{(i)}\}_{i\in\mathcal{I}_u\cup\mathcal{I}_l}$.
i) Set $\widetilde{\mathcal{X}} = \{\boldsymbol{T}_\nu, \boldsymbol{K}_\nu\}_{\nu=1}^5 \cup \{\boldsymbol{p}_{ul}\} \cup \{\boldsymbol{x}^{(i)}\}_{i\in\mathcal{I}_u\cup\mathcal{I}_l}$ and do Step (c) — (e) once.
j) Let $\boldsymbol{x}$ in Step (i) take $\boldsymbol{t}$, $\boldsymbol{r}$, s, and $\boldsymbol{b}$ alternatively and do Step (i) for 40 iterations.
k) Return the optimal parameters: $\mathcal{X}^* = \{\boldsymbol{T}_\nu, \boldsymbol{K}_\nu\}_{\nu=1}^5 \cup \{\boldsymbol{p}_{ul}\} \cup \{\boldsymbol{t}^{(i)}, \boldsymbol{r}^{(i)}, s^{(i)}, \boldsymbol{b}^{(i)}\}_{i\in\mathcal{I}_u\cup\mathcal{I}_l}$.

4) Poisson Surface Reconstruction: The reconstructed tooth models are represented as surface point clouds, as illustrated in Fig. 5, and are subsequently transformed into triangle meshes using the Poisson surface reconstruction method [36]. In order to perform Poisson reconstruction, the point cloud's oriented normals are required. To compute the normal of each point, we find its nearest 30 points and calculate the principal axis of the adjacent points using covariance analysis. Next, we compute the centroid of each tooth surface point cloud. For each point, we calculate the angle between its normal and the vector originating from the centroid to itself. If the angle is greater than 90 degrees, we orient the normal in the opposite direction to ensure that all the normals point outward. Finally, the normals are further oriented in accordance with their consistent tangent planes to facilitate the Poisson surface reconstruction process. The obtained point normals of the point clouds for different teeth are illustrated in Fig. 6.

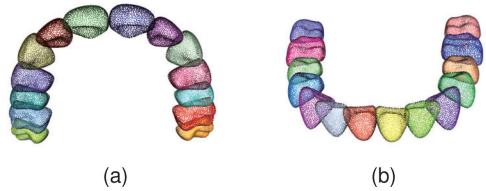


Fig. 5. Point clouds of reconstructed (a) upper and (b) lower tooth rows.

IV. EXPERIMENTS AND RESULTS

A. Data Sets and Implementation Details

We prepare two distinct groups of data for different purposes. To construct statistical shape models, we collect a total of 128 digital 3D models of upper and lower teeth, which are presented in surface mesh format. These teeth meshes are derived from digital dental scans created using optical intra-oral scanners at various orthodontic clinics. Subsequently, these meshes are

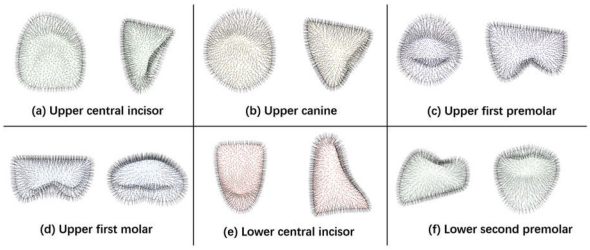


Fig. 6. The oriented point normals of different types of teeth point clouds prepared for Poisson surface reconstruction.

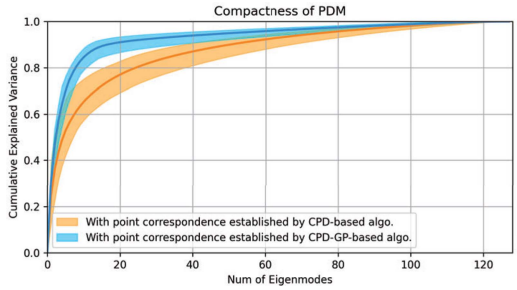


Fig. 7. Compactness of the constructed PDMs with point correspondence established by different algorithms. The dark blue/orange line indicates the average cumulative explained variance of that method.

manually segmented, labelled, and repaired by experts. To extract the teeth boundaries and restore the 3D shape and arrangement of teeth, we gather an additional 95 orthodontic cases, each comprising five intra-oral photographs and the corresponding teeth mesh obtained using the same method. The images in our dataset possess a resolution of 1440×1080.

Our teeth reconstruction framework involves fine-tuning various parameters and implementing techniques in each phase. In the statistical shape modeling phase, the only parameters requiring determination are s and l of the matrix-valued kernel in (2). After performing fine-tuning, we set $s = 2$ and $l = 10$ when executing the non-rigid deformation between two point sets of teeth. For training the teeth boundary extraction model, we randomly flip and rotate training images by an angle within $[-30°, 30°]$ to facilitate data augmentation. During the training process, we use a learning rate of 0.0005, a batch size of 8, and $\lambda_{ssim} = 1$. We train the model using the Adam optimizer for 50 epochs and save the model that performs best on the test data set. In our experiments, we set $w_p = 0.04$ and $w_n = 2$ in (11) after conducting a grid search.

B. Evaluation of the CPD-GP-Based Registration Algorithm and the Shape Correspondence

The quality of shape correspondence can be evaluated by the specificity, generalization, and compactness of the point distribution model (PDM) [37]. Specificity is typically computed as the average root mean squared error (RMSE) between randomly generated samples in the feature shape space and their nearest counterparts in real data [24]. Generalization is measured by computing a shape space spanned by the eigenmodes on all training cases except for one and then calculating the average RMSE between the remaining shape and its projection onto this shape space, commonly called the leave-one-out test [24]. Compactness refers to the tightness of a probability distribution and can be represented by the accumulative explained variation as a function of the number of used eigenmodes [24].

Three experiments are performed to validate the performance of the non-rigid registration algorithm in aligning point-sets and finding the point correspondences. The proposed CPD-GP-based algorithm is compared with the CPD-based algorithm. The difference between the two algorithms lies in the deformation stage, where the proposed algorithm uses Gaussian Process while the CPD-based algorithm uses the CPD deformation [23] for registration. Point correspondences are established in the same way for both algorithms. To evaluate the specificity, each constructed PDM is tested with 1,000 randomly generated samples from the distribution of its feature space. To evaluate the generalization, a leave-one-out test is performed for each PDM spanned with the number of eigenmodes that keeps 95% explained variance. To evaluate the compactness, the cumulative explained variance is computed as a function of the number of eigenmodes for each PDM.

The results of PDMs' specificity and generalization are categorized according to the tooth type and listed in Table IV. The PDMs with point correspondence established by the CPD-GP-based algorithm perform well in generalization but worse in specificity, compared to that using the CPD-based algorithm. Since the reconstruction framework focuses on unseen cases, the ability of generalization is preferred in the trade-off between generalization and specificity. Fig. 7 shows the compactness of the constructed PDMs with the cumulative explained variance. With fewer modes of variation, the PDMs with the CPD-GP-based algorithm have a higher explained variance, which indicates they are more compact than the ones built with the CPD-based algorithm. Therefore, the proposed CPD-GP-based algorithm outperforms the CPD-based algorithm in terms of point-correspondence performance.

C. Evaluation of the Teeth Reconstruction Method

1) Precision of the Parametric Teeth Model: The statistical shape models are built and evaluated using 128 digital intra-oral scans. Various metrics are utilized to provide a comprehensive assessment of the accuracy of the parametric teeth model, namely the root mean squared surface distance (RMSD), average symmetric surface distance (ASSD), Hausdorff distance (HD), Chamfer distance (CD), and Dice similarity coefficient (DSC). These same metrics are also used to evaluate the proposed teeth reconstruction framework.

We evaluate the precision of our teeth model by directly aligning it with the corresponding 3D ground-truth teeth mesh, the result summarized in Table V. Our compact parametric teeth model, which employs the first $C = 10$ modes of shape variation for each tooth, can provide an accurate representation of the original teeth mesh, as indicated by an average Dice similarity coefficient (DSC) value of 0.9615. It is worth noting that the precision of the parametric teeth model sets a theoretical upper limit on the accuracy achievable through the proposed reconstruction framework. However, due to the limited number

TABLE IV
SPECIFICITY AND GENERALIZATION OF THE CONSTRUCTED PDMS MEASURED BY THE MEAN AND STANDARD DEVIATION OF ROOT MEAN SQUARE ERROR (MM) BETWEEN CORRESPONDING POINTS AND CATEGORIZED BY THE TOOTH TYPE. EACH PDM IS EVALUATED BY 1000 RANDOMLY GENERATED SAMPLES FOR SPECIFICITY AND BY A LEAVE-ONE-OUT TEST FOR GENERALIZATION

Tooth type of PDM		Incisor	Canine	Premolar	Molar
Specificity	CPD-based algo.	$1.2713_{\pm 0.6538}$	$1.4572_{\pm 0.8365}$	$1.2803_{\pm 0.7613}$	$1.5420_{\pm 0.9079}$
	CPD-GP-based algo.	$1.4634_{\pm 0.7888}$	$1.9779_{\pm 1.2614}$	$1.4167_{\pm 0.8017}$	$2.1256_{\pm 1.3872}$
Generalization	CPD-based algo.	$0.1801_{\pm 0.0405}$	$0.1891_{\pm 0.0408}$	$0.1877_{\pm 0.0447}$	$0.1985_{\pm 0.0530}$
	CPD-GP-based algo.	$0.1230_{\pm 0.0332}$	$0.1406_{\pm 0.0337}$	$0.1241_{\pm 0.0240}$	$0.1415_{\pm 0.0361}$

TABLE V
QUANTITATIVE EVALUATION RESULTS OF DIFFERENT METRICS ON THE TEST DATA USING GROUND TRUTH ALIGNMENT, THE METHOD OF [14], NEAREST RETRIEVAL, SCALED NEAREST RETRIEVAL, AND OURS. (RMSD: ROOT MEAN SQUARED SURFACE DISTANCE, ASSD: AVERAGE SYMMETRIC SURFACE DISTANCE, HD: HAUSDORFF DISTANCE, CD: CHAMFER DISTANCE, DSC: DICE SIMILARITY COEFFICIENT)

Metric (avg. ± std.)	RMSD $(mm)\downarrow$	ASSD $(mm)\downarrow$	HD $(mm)\downarrow$	CD $(mm^2)\downarrow$	DSC/F1-Score $\uparrow$
Ground truth alignment	$0.1712_{\pm 0.0448}$	$0.1588_{\pm 0.0396}$	$0.4377_{\pm 0.1592}$	$0.0626_{\pm 0.0564}$	$0.9615_{\pm 0.0413}$
Method of [14]	$1.0343_{\pm 0.4453}$ [14]	$0.8479_{\pm 0.3788}$ [14]	$2.6268_{\pm 0.9151}$ [14]	—	$0.6588_{\pm 0.1397}$ [14]
Nearest retrieval	$0.9422_{\pm 0.4270}$	$0.8017_{\pm 0.3551}$	$2.2126_{\pm 0.8914}$	$2.1401_{\pm 2.2194}$	$0.6526_{\pm 0.1581}$
Scaled nearest retrieval	$0.7902_{\pm 0.3415}$	$0.6758_{\pm 0.2835}$	$1.9359_{\pm 0.7569}$	$1.4821_{\pm 1.5633}$	$0.7057_{\pm 0.1398}$
Ours (annotated teeth boundary)	$0.6334_{\pm 0.3015}$	$0.5470_{\pm 0.2515}$	$1.5558_{\pm 0.6837}$	$0.9849_{\pm 1.2880}$	$0.7693_{\pm 0.1230}$
Ours (predicted teeth boundary)	$0.6392_{\pm 0.3123}$	$0.5522_{\pm 0.2615}$	$1.5633_{\pm 0.6942}$	$1.0121_{\pm 1.6772}$	$0.7672_{\pm 0.1217}$

of modes used in the reconstruction process, fine details such as sharp edges may not be fully restored.

2) Five-Fold Cross Validation: The methods of teeth boundary extraction and teeth restoration are trained and evaluated on 95 orthodontic cases (475 intra-oral images and 95 corresponding teeth meshes) by a five-fold cross-validation. In each fold, 80% of the data is used to train the teeth boundary segmentation model, while the remaining 20% is reserved for testing purposes. Subsequently, the teeth boundary is predicted, and the 3D teeth models of the test samples are reconstructed. The Poisson surface reconstruction method [36] is employed to obtain triangle meshes. The reconstructed teeth meshes are aligned with the ground-truth meshes by utilizing a global similarity transformation.

The proposed teeth reconstruction framework is evaluated quantitatively using different metrics on the test data in all folds (95 samples). Table V presents the evaluation results, where our method demonstrates better reconstruction accuracy in comparison with different methods. Specifically, nearest retrieval involves selecting the most similar teeth mesh in the dataset for each test sample, building statistical shape models, and rigidly aligning it with the test sample to compute the metrics. Scaled nearest retrieval, on the other hand, replaces global rigid registration in nearest retrieval with similarity registration. It should be noted that the nearest retrieval dataset comprises 128 samples, yielding a relatively accurate result. Additionally, Fig. 8 offers an intuitive visualization of the reconstruction results for several orthodontic cases, including those involving misaligned teeth, missing teeth, protruding teeth, and malocclusion. These results demonstrate that our method accurately reconstructs the spatial arrangement and 3D shape of teeth for such cases.

To verify the accuracy of teeth boundary segmentation, we perform a direct reconstruction of the teeth from the annotated teeth boundary. The reconstruction accuracy using the predicted

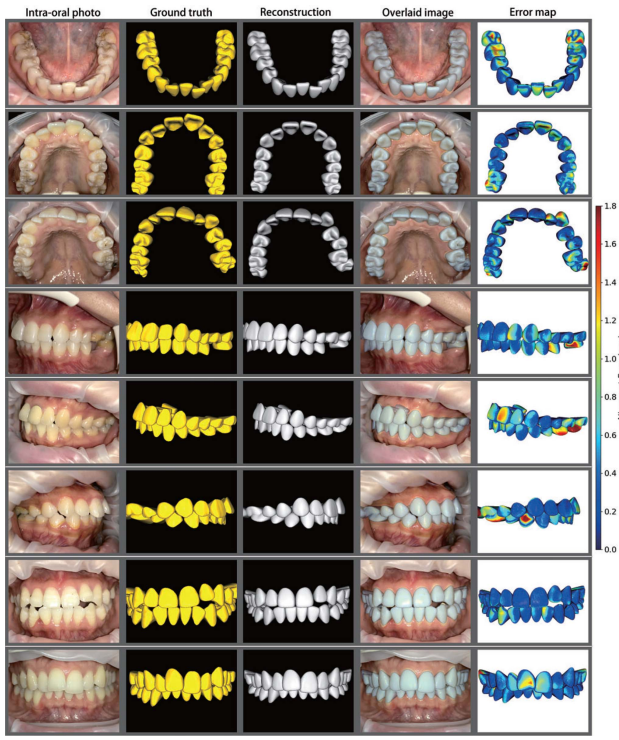


Fig. 8. Visualization of the original intra-oral photos, the ground truth of the teeth mesh, the reconstructed teeth mesh, the intra-oral images overlaid with the semi-transparent projection of the reconstructed teeth mesh, and the alignment error maps for different orthodontic cases.

teeth boundary shows no significant decrease, compared with the annotated teeth boundary (see Table V). The result validates the accuracy of the teeth boundary segmentation and indicates that the reconstruction error mainly comes from the optimization stage.

To analyze the spatial distribution of reconstruction error, we computed the error for every tooth individually. Fig. 9 displays

TABLE VI
RESULTS OF ABLATION EXPERIMENTS WITH DIFFERENT SETTINGS OF THE RECONSTRUCTION METHOD, AVERAGED OVER ALL THE TEETH OF THE 95 CASES (RMSD: ROOT MEAN SQUARED SURFACE DISTANCE, ASSD: AVERAGE SYMMETRIC SURFACE DISTANCE, HD: HAUSDORFF DISTANCE, CD: CHAMFER DISTANCE, DSC: DICE SIMILARITY COEFFICIENT)

Settings	RMSD $(mm)\downarrow$	ASSD $(mm)\downarrow$	HD $(mm)\downarrow$	CD $(mm^2)\downarrow$	DSC/F1-Score $\uparrow$
Default	$0.6392_{\pm 0.3123}$	$0.5522_{\pm 0.2615}$	$1.5633_{\pm 0.6942}$	$1.0121_{\pm 1.6772}$	$0.7672_{\pm 0.1217}$
(w/o) Robust Init.	$0.7538_{\pm 0.3781}$	$0.6481_{\pm 0.3190}$	$1.8071_{\pm 0.7901}$	$1.4223_{\pm 2.4799}$	$0.7224_{\pm 0.1416}$
(w/o) $\mathcal{L}_n$	$0.6700_{\pm 0.3193}$	$0.5780_{\pm 0.2700}$	$1.6303_{\pm 0.7001}$	$1.1015_{\pm 2.1639}$	$0.7556_{\pm 0.1224}$
(w/o) $\mathcal{L}_{prior}$	$0.7139_{\pm 0.3613}$	$0.6131_{\pm 0.3028}$	$1.7632_{\pm 0.8134}$	$1.2804_{\pm 1.9748}$	$0.7379_{\pm 0.1365}$
(w/o) $\mathcal{L}_n + \mathcal{L}_{prior}$	$0.7323_{\pm 0.3492}$	$0.6281_{\pm 0.2942}$	$1.8047_{\pm 0.7987}$	$1.3164_{\pm 2.3581}$	$0.7336_{\pm 0.1291}$

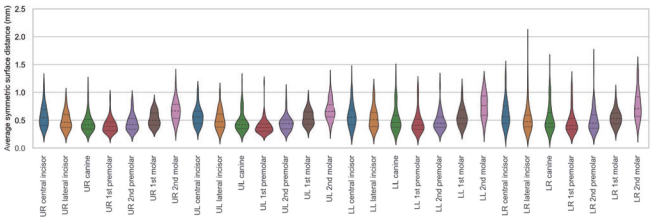


Fig. 9. Reconstruction error measured by average symmetric surface distance for each individual tooth. The same color is used for the teeth of the same type in different sides of jaws (UR: upper jaw's right side, UL: upper jaw's left side, LL: lower jaw's left side, LR: lower jaw's right side).

the reconstruction error measured by ASSD for each tooth in the test data. The analysis revealed a relatively large reconstruction error for the second molars in the upper and lower jaws. This is primarily due to the occlusion caused by the tongue, lips and other teeth obstructs the visibility of these teeth in the photos.

3) Ablation Study: Ablation experiments are performed to validate different components of the teeth reconstruction method. Specifically, the effectiveness of the robust initialization, the point-to-plane loss $\mathcal{L}_n$, and the penalization based on the prior distribution $\mathcal{L}_{prior}$ are evaluated. The default setting includes all the aforementioned components and the point-to-point loss $\mathcal{L}_p$ that is commonly used as a cost function in bundle adjustment [38]. Each test is performed over all the teeth of the 95 cases during the test phase of the five-fold validation. The results in Table VI indicate that the proposed robust initialization significantly improves the reconstruction outcome by providing reasonable initial camera settings and pose of the lower tooth row. Furthermore, the prior penalty is found to contribute more to the enhancement of reconstruction accuracy than the point-to-plane loss.

4) Robustness and Failed Cases: The proposed method exhibits robustness in challenging scenarios, including crowded teeth, missing teeth, severe overbite, teeth with impurities, and teeth with fillings, shown in Fig. 10. Furthermore, it performs well with intra-oral images where some teeth are out of focus. The illustration of each case comprises the original intra-oral image, the predicted teeth boundaries, the intra-oral images overlaid with a semi-transparent projection of the reconstructed teeth mesh, and the alignment error maps.

The proposed method fails to reconstruct severely misaligned teeth, broken teeth, or cases with severe underbite. Fig. 11 illustrates that although the teeth boundary prediction is accurate, the reconstruction fails, resulting in an alignment error exceeding

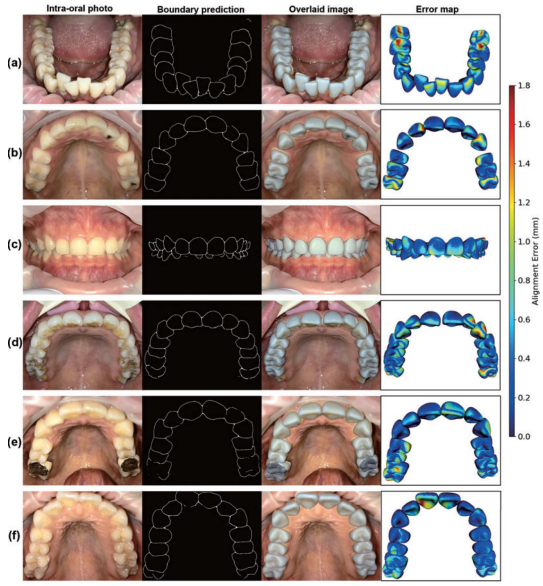


Fig. 10. Robust performance of the proposed method against some difficult cases: (a) crowded teeth, (b) missing teeth, (c) severe overbite, (d) teeth with impurities, (e) teeth with fillings, and (f) out-of-focus teeth.

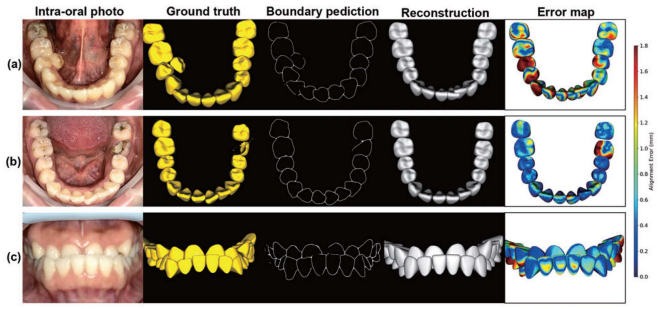


Fig. 11. Failed cases of the proposed method: (a) severely misaligned teeth, (b) broken teeth, and (c) severe underbite.

1.8 mm (regions marked in dark red). For severely misaligned teeth, bad point correspondence and the prior penalization in the loss function are the main contributing factors to such failure. Fig. 11(a) shows a patient with a severely mispositioned lower right second premolar, which poses challenges for determining point correspondence without specifying the source tooth of contour points. The rare occurrence of this case in optimization leads to a large penalty by the prior penalization. A multi-class boundary segmentation with an explicit label for each contour point indicating its source tooth could potentially address this

problem. For broken teeth, the constructed PDM is unable to simulate such significant deformation, which is beyond its generalization ability. While free-form deformation techniques can be applied, determining point correspondence around the broken part can be challenging. In cases of severe underbite, the difficulty lies in predicting the relative position of the upper and lower teeth, particularly premolars and molars. Although the proposed method yields reasonable results, the alignment error remains large.

V. Discussion

The proposed framework for 3D teeth reconstruction can automatically reconstruct the 3D shape and arrangement of teeth using five intra-oral photos. It outperforms the work of Wirtz et al. [14] in terms of reconstruction accuracy for the same task. This improvement can be attributed to several factors. First, the assumption made by Wirtz et al. that similar photos share the same view direction is not practical, as orthodontists may have different shooting habits. In contrast, the proposed method has a robust initialization of camera settings and relative pose between tooth rows. Second, Wirtz et al. simply placed their 2D coupled shape models on the predicted teeth segmentation to generate teeth boundaries, which cannot handle occlusion problems, resulting in poor performance for overbite cases, where lower teeth may be incompletely reconstructed. The proposed method can generate more reasonable reconstruction results by leveraging a prior parametric teeth template. Third, the parameters in the proposed teeth template, which include scale, pose, and shape vectors, are more explicit and suitable for step-by-step optimization. In contrast, the feature vector of the coupled shape model in Wirtz et al. entangles the original parameters, and optimizing all parameters simultaneously can lead to local optima. In summary, the proposed method is robust to variations in view directions of intra-oral photos and can handle malocclusion and missing teeth without human intervention. Additionally, predicting the relative pose between upper and lower teeth allows the proposed method to provide a rough estimation of the occlusion status of the patient, which can facilitate diagnosis by orthodontists.

Although the proposed method has shown promising results, there is still room for improvement. The reconstruction error for second molars is larger compared to other teeth due to occlusion from lips, tongue, and other teeth. Additionally, orthodontic cases with severe malocclusion are still difficult to predict automatically and require human intervention. Moreover, as we use only teeth silhouettes for reconstruction, fine details such as teeth ridges cannot be recovered. However, this is not a problem for the application of our method in orthodontics, which aims to correct teeth arrangement. To improve the accuracy of our teeth reconstruction framework, several additional techniques can be applied. Providing camera parameters can enhance the performance of our method. Collision detection between teeth can also be introduced as a supplementary constraint in optimization. However, we argue that this constraint is slightly redundant as the geometric teeth boundary and loss related to teeth priors penalize such collided situations.

VI. Conclusion

In this article, we present an improved template-based 3D teeth reconstruction framework that utilizes a robust initialization strategy and a modified loss function. We apply this framework to reconstruct 3D teeth models (excluding wisdom teeth) using five orthodontic intra-oral photographs. To describe the shape and arrangement of teeth accurately and compactly, we construct a parametric teeth model by leveraging prior statistical knowledge of orthodontic cases. Quantitative evaluation results confirm the accuracy of our method. Our 3D teeth reconstruction method provides a promising solution for remote diagnosis and dental monitoring of orthodontic treatment by enabling orthodontists to intuitively visualize the spatial arrangement and dental occlusion of their patients' teeth.

Acknowledgments

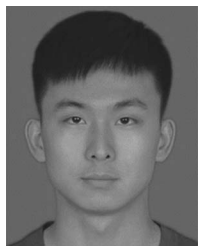
We would like to thank Smartee Denti-Technology Corporation Limited for segmenting, labelling, and repairing the digital dental scans and preparing the entire data set.

References

[1] J. Sandler and A. Murray, "Clinical photographs–The gold standard," *J. Orthodontics*, vol. 29, no. 2, pp. 158–161, 2002.
[2] V. Desai and D. Bumb, "Digital dental photography: A contemporary revolution," *Int. J. Clin. Pediatr. Dent.*, vol. 6, no. 3, pp. 193–196, 2013.
[3] P. Hong-Seok and S. Chintal, "Development of high speed and high accuracy 3d dental intra oral scanner," *Procedia Eng.*, vol. 100, pp. 1174–1181, 2015.
[4] M. Revilla-Leon et al., "Intraoral scanners: An american dental association clinical evaluators panel survey," *J. Amer. Dent. Assoc.*, vol. 152, no. 8, pp. 669–670, 2021.
[5] L. Yang et al., "iOrthoPredictor: Model-guided deep prediction of teeth alignment," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–15, 2020.
[6] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion*," *Acta Numerica*, vol. 26, pp. 305–364, 2017.
[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[8] A. S. Abdelrahim, M. T. El-Melegy, and A. A. Farag, "Realistic 3D reconstruction of the human teeth using shape from shading with shape priors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 64–69.
[9] A. S. Abdelrehim, A. A. Farag, A. M. Shalaby, and M. T. El-Melegy, "2D-PCA shape models: Application to 3D reconstruction of the human teeth from a single image," in *Proc. Int. MICCAI Workshop Med. Comput. Vis.*, Springer, 2013, pp. 44–52.
[10] A. Farag, S. Elhabian, A. Abdelrehim, W. Aboelmaaty, A. Farman, and D. Tasman, "Model-based human teeth shape recovery from a single optical image with unknown illumination," in *Proc. Int. MICCAI Workshop Med. Comput. Vis.*, 2012, pp. 263–272.
[11] E. Mostafa, S. Elhabian, A. Abdelrahim, S. Elshazly, and A. Farag, "Statistical morphable model for human teeth restoration," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 4285–4288.
[12] A. Morales, G. Piella, and F. M. Sukno, "Survey on 3D face reconstruction from uncalibrated images," *Comput. Sci. Rev.*, vol. 40, 2021, Art. no. 100400.
[13] C. Wu et al., "Model-based teeth reconstruction," *ACM Trans. Graph.*, vol. 35, no. 6, 2016, Art. no. 220.
[14] A. Wirtz, F. Jung, M. Noll, A. Wang, and S. Wesarg, "Automatic model-based 3-D reconstruction of the teeth from five photographs with predefined viewing directions," in *Proc. Med. Imag. Image Process.*, SPIE, 2021, pp. 198–212.
[15] A. S. Jackson, C. Manafas, and G. Tzimiropoulos, "3D human body reconstruction from a single image via volumetric regression," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 64–77.
[16] R. Natsume et al., "SiCloPe: Silhouette-based clothed people," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4480–4490.

[17] A. Mustafa, A. Caliskan, L. Agapito, and A. Hilton, "Multi-person implicit reconstruction from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14474–14483.
[18] T. Alldieck, M. Zanfir, and C. Sminchisescu, "Photorealistic monocular 3D reconstruction of humans wearing clothing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1506–1515.
[19] X.-F. Han, H. Laga, and M. Bennamoun, "Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1578–1604, May 2021.
[20] Z. Cui, C. Li, and W. Wang, "ToothNet: Automatic tooth instance segmentation and identification from cone beam CT images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6368–6377.
[21] Y. Liang, W. Song, J. Yang, L. Qiu, K. Wang, and L. He, "X2Teeth: 3D teeth reconstruction from a single panoramic radiograph," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, Springer, 2020, pp. 400–409.
[22] W. Song, Y. Liang, J. Yang, K. Wang, and L. He, "Oral-3D: Reconstructing the 3D structure of oral cavity from panoramic X-ray," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 566–573.
[23] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2262–2275, Dec. 2010.
[24] G. Zheng, S. Li, and G. Szekely, *Statistical Shape and Deformation Analysis: Methods, Implementation and Applications*. Cambridge, MA, USA: Academic Press, 2017.
[25] M. Lüthi, T. Gerig, C. Jud, and T. Vetter, "Gaussian process morphable models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1860–1873, Aug. 2018.
[26] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Berlin, Germany: Springer Science & Business Media, 2011.
[27] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interactive Techn.*, 1999, pp. 187–194.
[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, Springer, 2015, pp. 234–241.
[29] H. He, D. Yang, S. Wang, S. Wang, and Y. Li, "Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss," *Remote Sens.*, vol. 11, no. 9, 2019, Art. no. 1015, doi: 10.3390/rs11091015.
[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[31] J. Ma et al., "Loss odyssey in medical image segmentation," *Med. Image Anal.*, vol. 71, 2021, Art. no. 102035.
[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[33] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.
[34] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," *Chemometrics Intell. Lab. Syst.*, vol. 50, no. 1, pp. 1–18, 2000.
[35] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[36] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proc. 4th Eurograph. Symp. Geometry Process.*, 2006, pp. 61–70.
[37] M. A. Styner et al., "Evaluation of 3D correspondence methods for model building," in *Proc. Biennial Int. Conf. Inf. Process. Med. Imag.*, 2003, pp. 63–75.
[38] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment–A modern synthesis," in *Proc. Int. Workshop Vis. Algorithms*, Springer, 1999, pp. 298–372.

Yizhou Chen received the BEng degree in mechanical engineering from Shanghai Jiao Tong University, in 2021. He is currently working toward the master's degree with the Institute of Biomedical Manufacturing and Life Quality Engineering in Shanghai Jiao Tong University. His research interests lie in computer vision, including semantic segmentation and multi-view 3D reconstruction.



Shuojie Gao is currently working toward the bachelor's degree with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include computer vision, deep learning and statistical shape analysis in various applications.

Puxun Tu received the BEng degree in mechanical engineering from Chongqing University, in 2019. He is currently working toward the PhD degree with the Institute of Biomedical Manufacturing and Life Quality Engineering, Shanghai Jiao Tong University. His research interests include image guided surgery and augmented reality.

Xiaojun Chen (Member, IEEE) received the PhD degree in mechanical engineering from Shanghai Jiao Tong University, China, in 2006. He is currently a professor of the School of Mechanical Engineering, Shanghai Jiao Tong University. His research focuses on computer-assisted surgery, including medical image analysis, computer graphics, surgical navigation, VR/AR technology in medicine, surgical robotics, medical 3D printing, etc.