Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Accurate estimation of 6-DoF tooth pose in 3D intraoral scans for dental applications using deep learning*#

Wanghui DING†‡[1], Kaiwei SUN[2], Mengfei YU[1], Hangzheng LIN[2], Yang FENG[3], Jianhua LI[4], Zuozhu LIU†‡[1,2]

[1]*Stomatology Hospital, School of Stomatology, Zhejiang University School of Medicine, Zhejiang Provincial Clinical Research Center for Oral Diseases, Key Laboratory of Oral Biomedical Research of Zhejiang Province, Cancer Center of Zhejiang University, Engineering Research Center of Oral Biomaterials and Devices of Zhejiang Province, Hangzhou 310000, China*

[2]*Zhejiang University–University of Illinois at Urbana-Champaign Institute, Zhejiang University, Haining 314400, China*

[3]*Angel Align Inc., Shanghai, Shanghai 200433, China*

[4]*Hangzhou Dental Hospital, Hangzhou 310006, China*

†E-mail: godson888@zju.edu.cn; zuozhuliu@intl.zju.edu.cn

Received Sept. 1, 2023; Revision accepted Sept. 11, 2023; Crosschecked Aug. 28, 2024

Abstract: A critical step in digital dentistry is to accurately and automatically characterize the orientation and position of individual teeth, which can subsequently be used for treatment planning and simulation in orthodontic tooth alignment. This problem remains challenging because the geometric features of different teeth are complicated and vary significantly, while a reliable large-scale dataset is yet to be constructed. In this paper we propose a novel method for automatic tooth orientation estimation by formulating it as a six-degree-of-freedom (6-DoF) tooth pose estimation task. Regarding each tooth as a three-dimensional (3D) point cloud, we design a deep neural network with a feature extractor backbone and a two-branch estimation head for tooth pose estimation. Our model, trained with a novel loss function on the newly collected large-scale dataset (10 393 patients with 280 611 intraoral tooth scans), achieves an average Euler angle error of only 4.780°–5.979° and a translation L1 error of 0.663 mm on a hold-out set of 2598 patients (77 870 teeth). Comprehensive experiments show that 98.29% of the estimations produce a mean angle error of less than 15°, which is acceptable for many clinical and industrial applications.

Key words: Artificial intelligence; Digital dentistry; Deep learning; Orthodontics; Tooth pose; Neural network
https://doi.org/10.1631/FITEE.2300596 **CLC number:** TP391; R783.5

1 Introduction

With the development of artificial intelligence and deep learning techniques, computer-aided dentistry has become prevalent in both dental research and the industry. Highly efficient and accurate treatment planning is of intense research interest to the field of digital dentistry, and many efforts are being made to achieve these goals, such as those devoted to model preparation, tooth segmentation, and tooth alignment. However, it is still challenging to precisely model the coordinates or pose of individual teeth in digital models, especially with high degrees of freedom. Accurate tooth pose

‡ Corresponding authors
* Project supported by the National Natural Science Foundation of China (Nos. 11932012 and 62106222), the Zhejiang University–Angelaling Inc. R&D Center for Intelligent Healthcare, the Zhejiang Provincial Public Welfare Research Program (No. LTGY23H140007), the Clinical Research Project of the Chinese Orthodontic Society (No. COS-C2021-09), and the Research and Development Project of Stomatology Hospital, Zhejiang University School of Medicine (No. RD2022YFZD01)
# Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2300596) contains supplementary materials, which are available to authorized users
ORCID: Wanghui DING, https://orcid.org/0000-0003-0645-4647; Zuozhu LIU, https://orcid.org/0000-0002-7816-502X
© Zhejiang University Press 2024

enables the precise simulation of three-dimensional (3D) tooth movement and can help elucidate the optimal force magnitudes and directions in orthodontics and implants. Current solutions used in clinical software require heavy manual work, which is time-consuming. Therefore, an accurate and efficient system that can automatically characterize individual tooth pose in digital dental models is needed to fully automate the treatment planning process.

The tooth pose task can be considered a six-degree-of-freedom (6-DoF) object pose estimation problem. In 3D digital dental models, each tooth is a rigid body with 3D translation and 3D rotation (Wei et al., 2020; Chen et al., 2023; Zhu et al., 2023). The task could be resolved by predicting the rotation and translation of each tooth, which is initially located in a world coordinate as acquired from 3D intraoral scans (IOSs). However, the task of tooth pose estimation is not yet well-defined. Little research has been conducted towards the goal of an automatic and accurate system for tooth pose estimation. In computer science, much previous research focused on estimating the 6-DoF pose of objects in two-dimensional (2D) images via holistic (Gu and Ren, 2010; Hinterstoisser et al., 2012; Su et al., 2015; Xiang et al., 2018), key point based (Kendall et al., 2015; Newell et al., 2016; Oberweger et al., 2018; Peng et al., 2019; Wang C et al., 2019), or dense correspondence (Liebelt et al., 2008; Sun et al., 2010; Li et al., 2019; Park et al., 2019; Wang H et al., 2019; Cai and Reid, 2020) methods. Recently, some research has been carried out on pose estimation for 3D objects in RGB-D images or point clouds (Zhou and Tuzel, 2018; Qi et al., 2019). However, none of these works is designated for 3D medical data, which differ from images of nature scenes both geometrically and anatomically. In addition, it is difficult to propose a universal framework to deal with all 3D dental data from different patients, which are usually composed of different geometric features. The 3D IOS data contain different initial positions and orientations in the world coordinate system (upper and lower jaws), different tooth shapes (incisors, canines, premolars, and molars), and different tooth sizes. Even the same tooth across different patients may differ in orientation, location, and shape. However, a clinically applicable model is expected to handle all these variations and provide a robust transformation with good generalization ability.

In this work, we propose a novel, automatic, and efficient model for accurate tooth pose estimation using deep learning to address this challenge. We have built a large dataset with 12 991 patients (358 481 tooth scans), and each tooth is associated with an annotated rigid transformation provided by human experts. In addition, we have designed a novel neural network based on Edge-Conv blocks (Wang Y et al., 2019) and resolved the tooth pose estimation problem with two jointly optimized tasks by estimating the corresponding translation and rotation transformations. Experimental results showed that our method can achieve excellent performance for tooth pose estimation, with a mean Euler angle error of less than 15° for 98.29% of the tooth scans and a mean Euler angle error of less than 10° for 99.54% of patients in the test cohort. To the best of our knowledge, our method presents the first attempt at accurate tooth pose estimation based on deep learning, and it has been integrated into clinical software for orthodontics in China.

2 Materials and methods

2.1 Data preprocessing and statistics

We constructed a large dataset that contains 358 481 3D IOS tooth meshes from 12 991 patients, and each tooth mesh is composed of thousands of triangular mesh faces. The dataset includes 32 permanent teeth, i.e., teeth with notations {11–18, 21–28, 31–38, 41–48} following the Federation Dentaire Internationale (FDI) standard (Herrmann, 1967). Our model was trained with 288 864 meshes (10 393 patients) and evaluated on the remaining 69617 meshes (2598 patients). All data were randomly collected from different hospitals from 2018 to 2020 in China, without overlap among them. More dataset statistics are available in Table S1 in the supplementary materials.

In our dataset, each tooth's pose is characterized by a quaternion and a translation term annotated by human experts that can transform each tooth from the world coordinate to its local anatomical coordinate, where the local Cartesian axes denote the mesiodistal, buccolingual, and root crown axis in the anatomy, separately, as illustrated in Fig. 1a. In our experiments, the quaternion and translation terms are converted to

rigid transformations $\boldsymbol{T}=[\boldsymbol{R}, \boldsymbol{t}]$, which consist of a rotation $\boldsymbol{R}\in\mathrm{SO}(3)$ and a translation $\boldsymbol{t}\in\mathbb{R}^3$, where $\mathrm{SO}(3)$ denotes the 3-DoF rotation group in geometry.

For data preprocessing, we transform each tooth mesh to a point cloud with 3000 points sampled from triangle mesh face centers, empirically filtering out teeth with fewer than 3000 mesh faces. We extract six-dimensional (6D) predefined features for each point, including the 3D coordinates and normal vectors. The 3D coordinates of each point correspond to the location of the gravity center of each face, which is denoted as $\boldsymbol{h}_{\mathrm{c}}=[x_{\mathrm{c}}, y_{\mathrm{c}}, z_{\mathrm{c}}]\in\mathbb{R}^3$. We further extract more geometrical features from the original mesh by computing the normal vector $\boldsymbol{h}_{\mathrm{n}}\in\mathbb{R}^3$ for each face. Hence, each tooth can be represented as a point cloud $\boldsymbol{P}\in\mathbb{R}^{n\times 6}$, where n=3000 is the number of points, and 6 denotes the dimension of features $\boldsymbol{h}=[\boldsymbol{h}_{\mathrm{c}}, \boldsymbol{h}_{\mathrm{n}}]\in\mathbb{R}^6$ associated with each point.

2.2 Deep learning based tooth pose estimation

The deep neural network we proposed includes two main components: a feature extractor backbone $\boldsymbol{\Theta}$ and a two-branch pose estimation head $\Phi=\{\boldsymbol{\Phi}_{\mathrm{t}}, \boldsymbol{\Phi}_{\mathrm{r}}\}$, as illustrated in Fig. 1b. We call our model the tooth pose estimation network (TP-Net). The feature extractor backbone first transforms the tooth point cloud to a canonical space. Then, it employs a stack of EdgeConv blocks to extract complex geometric features for each tooth, as inspired by the dynamic graph convolutional neural network (DGCNN) (Wang Y et al., 2019). Each EdgeConv block contains a k-nearest-neighbor (kNN) graph feature extractor, which determines the size of local graphs, and a convolution layer that captures the topological structures and features of these detailed local graphs. The neighbors are computed based on latent representations, leading to different local and global proximities from the bottom to the top blocks. Hence, the stacked EdgeConv blocks can learn local geometric features in the bottom layers and semantic characteristics and global shape properties in the top layers. Lastly, we concatenate the features from the top EdgeConv block with permutation invariant mean pooling and max pooling aggregation operations to obtain a global tooth representation for subsequent pose estimation, i.e., $\boldsymbol{z}=\boldsymbol{\Theta}(\boldsymbol{P})\in\mathbb{R}^{1024}$.

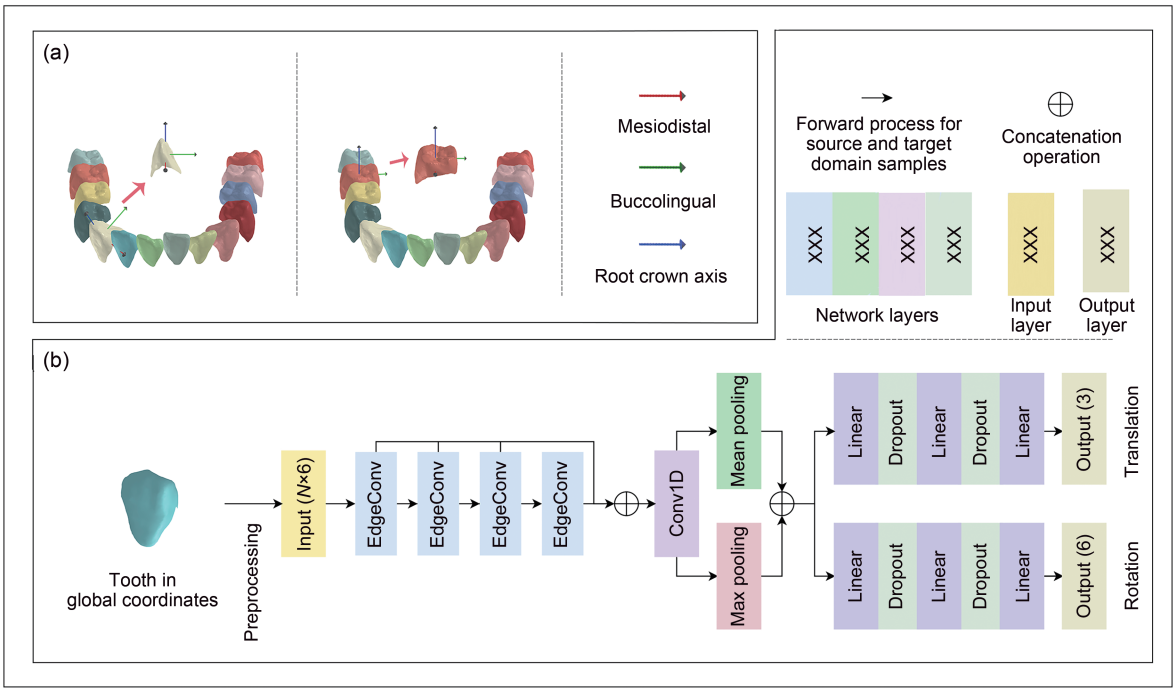

Fig. 1 Problem formulation and overview of TP-Net: (a) visualization of tooth pose estimation by predicting its orientation and location; (b) detailed architecture design of TP-Net, which consists of a feature extractor backbone and a two-branch estimation head (References to color refer to the online version of this figure)

The detailed settings of the encoder architecture are as follows: We use n_e EdgeConv blocks to learn the local geometrical relationship and global semantic relationship information. Each of the EdgeConv blocks θ_i ($i\in[1, n_e]$) is composed as kNN feature extractor-> Conv2D[c_i]->max pooling. Conv1D and Conv2D denote the 1D and 2D convolution layers, respectively. The number inside the bracket denotes the number of filters for Conv1D and Conv2D. The filter number for each EdgeConv is denoted as a set $C=\{c_1, c_2, ..., c_{n_e}\}$, where c_i (i=1, 2, ..., n_e) is the number of filters for the ith EdgeConv block. Unless otherwise indicated, the Conv2D layers use a kernel size of [1, 1] and a stride size of [1, 1] with LeakyReLU activation. The kNN feature extractor will first be applied to each sampled point in the point cloud and will calculate the nearest k neighbors. These k neighbors will then be concatenated with their center points and fed to the following 2D convolution layer. These n_e EdgeConv blocks are stacked together. We denote the output of each EdgeConv block θ_i as h_i, and $H=\{h_1, h_2, ..., h_{n_e}\}$ is the concatenated form of all the EdgeConv block outputs. H will be fed to the last 1D convolution layer, Conv1D[1024]. Then, we apply both max pooling and mean pooling over H to capture hierarchical features and obtain a good global tooth representation.

In practice, we stacked n_e=4 EdgeConv blocks and set the number of filters inside each EdgeConv block as $C=\{c_1, c_2, c_3, c_4\}=\{64, 64, 128, 256\}$. We set the value of k in the kNN algorithm as 20, since empirically it would not bring a significant improvement when we set k larger than 25 on the tooth point cloud. All the LeakyReLU in the architecture used a slope rate of 0.2, and the dropout rate in all the dropout layers was 0.6. We used the Adam-based Look-Ahead optimizer with five synchronization periods, and the slow weight learning rate α was 0.5. The Adam optimizer had the learning rate lr=1e-5. We used PyTorch to implement the neural network. The model was trained for 50 epochs on NVIDIA GeForce RTX 3090 24 GB GPU with a batch size of 32.

We used a two-branch head to estimate the orientation and location of each tooth separately. The rotation and translation transformations have different physical meanings, and note that they are usually of different scales for tooth pose estimation. Hence, we used two separate heads to concurrently estimate the rotation and translation transformations, while they share the underlying feature extractor backbone. By doing so, we can encourage the feature extractor backbone to learn robust global tooth representations while allowing the two heads to learn disentangled representations and focus on the estimation of rotation or translation independently. In particular, based on the learned representation $\boldsymbol{z}=\boldsymbol{\Theta}(\boldsymbol{P})$, one estimation head in our TP-Net will output a 3D vector $\boldsymbol{y}_t=\boldsymbol{\Phi}_t(\boldsymbol{z})\in\mathbb{R}^3$ for the translation estimation, and the other head will produce a 6D vector $\boldsymbol{y}_r=\boldsymbol{\Phi}_r(\boldsymbol{z})\in\mathbb{R}^6$ for rotation estimation. The translation estimation follows the usual 6-DoF pose estimation convention by predicting the 3D translation along the three Cartesian axes. The estimation of rotation we adopted here is a little different, as described below.

In TP-Net, we estimated a 6-DoF estimation $\boldsymbol{y}_r$ for the 3D rotation. Compared to methods directly regressing the quaternion or Euler angles with discontinuous representations, our method maintains the continuity of the SO(3) representation, which is easier to approximate with deep neural networks (Zhou et al., 2019). Moreover, the estimation $\boldsymbol{y}_r$ is equivalent to a 3×3 rotation matrix while requiring no post-processing for orthogonalization. Specifically, given the output $\boldsymbol{y}_r=[\boldsymbol{y}_1, \boldsymbol{y}_2]\in\mathbb{R}^6$, where $\boldsymbol{y}_1, \boldsymbol{y}_2\in\mathbb{R}^3$, we can construct the 3-DoF rotation matrix $\hat{\boldsymbol{R}}=[\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3]^{\mathrm{T}}\in\mathbb{R}^{3\times 3}$ as follows: $\boldsymbol{r}_1=\mathrm{Norm}(\boldsymbol{y}_1)$, $\boldsymbol{r}_2=\mathrm{Norm}(\boldsymbol{y}_2-(\boldsymbol{r}_1\circ\boldsymbol{y}_2)\boldsymbol{r}_1)$, $\boldsymbol{r}_3=\boldsymbol{r}_1\times\boldsymbol{r}_2$, where Norm(·) denotes the unit normalization for vectors, and ∘ and × denote the inner and cross products of vectors, respectively. Finally, our estimated rigid transformation can be represented as $\hat{\boldsymbol{T}}=[\hat{\boldsymbol{R}}, \boldsymbol{y}_t]$.

We designed a novel loss function for tooth pose estimation with TP-Net, a weighted sum of three parts—a translation loss, a rotation loss, and a geodesic loss. Given a tooth $\boldsymbol{P}$, the translation loss measures the displacement error between the predicted translation $\hat{\boldsymbol{y}}_t$ and the true translation $\boldsymbol{y}_t$, which is defined as $L_t=\|\hat{\boldsymbol{y}}_t-\boldsymbol{y}_t\|^2$, where $\|\cdot\|$ denotes the Frobenius norm. Similarly, the rotation loss measures the difference between the predicted rotation $\hat{\boldsymbol{R}}$ and the true rotation $\boldsymbol{R}$, which is mathematically defined by a squared error function, i.e., $L_R=\|\hat{\boldsymbol{R}}-\boldsymbol{R}\|^2$. In addition, we proposed to use another geodesic loss function to measure the minimum angular difference between two rotations to

better characterize the difference between two rotations and to help the network learn better representations. The geodesic loss is defined as $L_g = \arccos(0.5(\mathrm{tr}(\boldsymbol{G})-1))$, where $\boldsymbol{G}=\hat{\boldsymbol{R}}^{\mathrm{T}}\boldsymbol{R}\in\mathbb{R}^{3\times 3}$ and $\mathrm{tr}(\boldsymbol{G})$ denotes the trace of matrix $\boldsymbol{G}$. The final loss function for neural network training is a weighted sum of these three losses, i.e., $L=\lambda_1 L_R + \lambda_2 L_t + \lambda_3 L_g$, where λ_1, λ_2, and λ_3 are hyperparameters determined by a grid search in the experiments. We performed a comprehensive hyperparameter search for these three hyperparameters and finally set $\lambda_1=10$, $\lambda_2=100$, $\lambda_3=10$.

2.3 Experimental setup and evaluation

We conducted a comprehensive ablation study on a small dataset, consisting of 1500 patients (80% for training and 20% for testing) randomly selected from the whole dataset, to evaluate the effectiveness of each loss function. To the best of our knowledge, we are the first to conduct research on tooth pose estimation with a large-scale dataset and using deep learning methods, so we cannot find any baseline. Thus, for the experiments, we chose a popular deep learning method, PointNet, for point cloud processing, as our baseline. The overall performance is reported with the large dataset.

We evaluated the translation and rotation results for tooth pose estimation separately. The translation error was measured by the average L_1 and L_2 norms between the estimated and actual translation vectors. As for the rotation, we reported the average geodesic distance L_g, as well as the average Euler angle error along the Cartesian X-, Y-, and Z-axis, i.e., X_e, Y_e, Z_e as defined in the supplementary materials. We further demonstrated the clinical and industrial utility by comparing the inference time of our method with that of human experts. More statistical analyses and visualizations are also reported.

3 Results

3.1 Tooth pose estimation performance

The performance of our model for tooth pose estimation is reported in Table 1. We use M_s and M_l to denote TP-Net trained with the small and large datasets, respectively. Overall, we can see that our methods significantly outperform the PointNet baseline in terms of all evaluation metrics. As for our method, M_l performs much better than M_s, achieving errors L_g, X_e, Y_e, Z_e that are roughly half of those of M_s, with Euler angle errors of only 4.7° to 6.0° and L1/L2 translation errors of 0.633/0.446 mm. According to the feedback from a committee of experienced dentists, an angle error of 5° to 7° is generally acceptable for real-world orthodontics treatment, as even human experts might generate two annotations with 5° of inconsistency for the same tooth. The ablation study over the three loss functions further shows that learning rotation estimations with a single L_g or L_r loss can achieve a comparable performance, while the model trained on our weighted sum loss can further reduce rotation errors by 0.7° to 1.8°, demonstrating the effectiveness of our loss function.

3.2 Statistical analysis

We further evaluated the effectiveness of our method by computing the mean and maximum rotation angle errors and reported the percentile at both the

Table 1 Pose estimation results for TP-Net*

Model	Loss	G_e	X_e (°)	Y_e (°)	Z_e (°)	L_1 (mm)	L_2 (mm)
PointNet	r+t+g	14.469	11.326	13.068	9.202	6.215	4.259
PointNet++	r+t+g	14.786	10.114	11.703	8.167	1.254	0.842
PointCNN	r+t+g	14.957	12.865	13.024	8.839	0.952	0.644
MinkCNN	r+t+g	12.368	10.631	11.524	8.271	1.017	0.718
M_s	r+t+g	11.607	9.628	10.061	7.355	1.313	0.884
M_s	r+t	13.229	10.721	11.807	8.184	0.963	0.650
M_s	g+t	13.236	10.722	11.812	8.182	1.060	0.715
M_l	r+t+g	6.863	5.265	5.979	4.780	0.633	0.446

* Tested on 300/2598 patients. r: rotation loss; t: translation loss; g: geodesic loss. s: small dataset with 1200 patients for training; l: large dataset with 10 393 patients for training. G_e: geodesic error; X_e, Y_e, Z_e: angle errors along the X, Y, and Z axes for rotation, respectively. L_1, L_2: L1 and L2 loss for translation, respectively

tooth level and the patient level (Table 2). At the tooth level, 84.85% of the teeth had a maximum angle error of lower than 10°, and 92.47% of the teeth had a mean angle error of lower than 10°. Such a high estimation accuracy can save a lot of manual annotation time, as most of the teeth no longer require manual annotation, which is demonstrated by the integration of our method into real-world orthodontics treatment planning software.

We also computed a mean angle error for each patient defined as $\widehat{\theta}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{3}\left(X_{\mathrm{e}}^{i}+Y_{\mathrm{e}}^{i}+Z_{\mathrm{e}}^{i}\right)$, where n is the total number of the patient's teeth. We computed the percentile for $\widehat{\theta}$, with results reported in Table 2. The results showed that, in 99.92% of the patients, the mean angle error was less than 13°, which again demonstrated the effectiveness of our method. We further investigated cases with a mean angle error larger than 13°. The large $\widehat{\theta}$ was attributed mainly to teeth with substantial angle errors, e.g., a tooth with an angle error of more than 90°. An extensive manual examination revealed that a large part of these substantial angle errors was due to annotation errors from human experts; e.g., some human experts might wrongly annotate the buccolingual axis along the opposite direction, as shown in Fig. 2b. A detailed discussion will be presented in Section 3.4.

We further conducted a comprehensive statistical analysis on the best-performing model M_1 to characterize its performance. The results are reported in Table 3, with more details in Table S2 in the supplementary materials. We can see that the third molars in each quadrant had the most significant angle errors in most measurements. Their rotation estimations led to angle errors of about 7°–11°, while most of the other teeth usually exhibited errors of about 3°–6°. The primary reason for the relatively inferior performance is that the number of third molars is relatively small in our dataset, i.e., two orders of magnitude smaller than the numbers of other teeth, as shown in Table S1. Hence, the data-hungry deep neural network might not be well-trained to precisely estimate the pose of third molars.

Additionally, we can observe that the orientations along the buccolingual direction for teeth 35 and 45 exhibited significant errors, i.e., more than 9°. This is mainly because the two premolars in each quadrant from some patients are very similar; hence, it is quite difficult to tell the difference between their buccolingual directions. Meanwhile, many human annotation errors occurred in teeth 35 and 45, which again demonstrates that it is challenging to perform pose estimation of these two teeth. Another interesting observation is that the median of the angle errors for almost all teeth was smaller than the mean angle error, as shown in Table S3 in the supplementary materials. We conducted a detailed analysis, showing that there always existed several teeth whose angle errors were enormous, i.e., more than 90°, leading to more significant mean angle errors than median angle errors.

Table 2 Percentiles of the estimated angle errors from M_1^*

Angle error	$p_{\max}$	p_{mean}	Angle error	$\dot{p}$
≤ 5°	41.56%	56.25%	≤ 4°	5.39%
≤10°	84.85%	92.47%	≤ 5°	42.19%
≤ 15°	95.95%	98.29%	≤ 6°	78.29%
≤ 20°	98.52%	99.32%	≤ 7°	93.03%
≤ 25°	99.22%	99.66%	≤ 8°	97.69%
≤ 40°	99.80%	99.90%	≤ 10°	99.54%
≤ 90°	99.94%	99.96%	≤ 13°	99.92%
> 90°	0.06%	0.04%	≤ 30°	100%

* Tested on 2598 patients/77 870 teeth. $p_{\max}, p_{\mathrm{mean}}$: percentiles of the maximum and mean angle errors $\theta_{\max}$ and θ_{mean} for all teeth, respectively. $\dot{p}$: percentile of the mean angle error $\dot{\theta}$ for all patients. For each tooth, $\theta_{\max}=\max(X_{\mathrm{e}}, Y_{\mathrm{e}}, Z_{\mathrm{e}})$, $\theta_{\mathrm{mean}}=\frac{1}{3}(X_{\mathrm{e}}+Y_{\mathrm{e}}+Z_{\mathrm{e}})$. For each patient, $\widehat{\theta}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{3}(X_{\mathrm{e}}^{i}+Y_{\mathrm{e}}^{i}+Z_{\mathrm{e}}^{i})$, where n is the number of teeth for the patient

3.3 Clinical and industrial utility

We evaluated the clinical and industrial utility of our method by comparing the inference time of our method with the annotation time of human experts. Based on our data collection statistics, it takes about 15 min for a human expert to annotate all the teeth of a patient. In stark contrast, our TP-Net needed only 0.08 s for each tooth with an NVIDIA 3090 GPU, i.e., less than 3 s if we annotate all the teeth sequentially without parallel computing frameworks. We can see

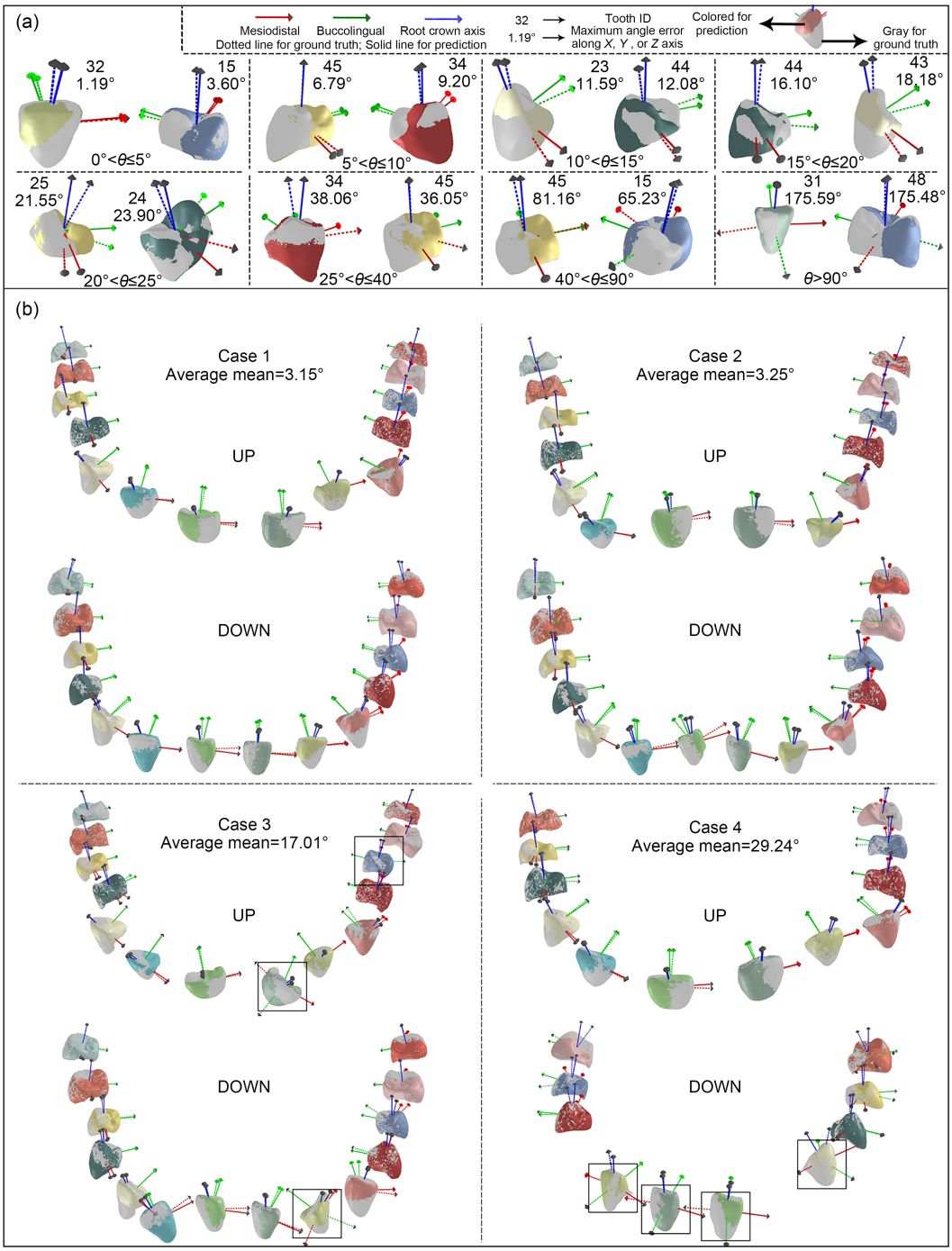

Fig. 2 Visualizations of estimated orientation and ground truth: (a) visualization of multiple teeth with different scales of Euler angle errors; (b) visualization of four patients with small or large mean angle errors (References to color refer to the online version of this figure)

that our method is two orders of magnitude faster than human experts. Though it might generate some mistakes, the great efficiency would make it invaluable for dentists and clinicians by saving a huge amount of time during treatment planning.

3.4 Visualization

We visualized the tooth pose estimation results for multiple single teeth as well as for four patients to demonstrate the superiority of our method, and presented its limitations (Fig. 2). In Fig. 2a, we visualized

Table 3 Mean and standard deviation of the predicted angle error from M_1^*

Tooth	X_e (°)	Y_e (°)	Z_e (°)	Tooth	X_e (°)	Y_e (°)	Z_e (°)
11	3.479 (4.242)	4.504 (4.655)	4.107 (2.723)	31	3.374 (5.540)	4.990 (5.641)	4.699 (2.488)
12	5.218 (3.958)	5.769 (4.077)	4.436 (2.667)	32	3.600 (5.381)	4.079 (5.547)	3.796 (2.448)
13	5.995 (4.994)	6.293 (5.017)	4.678 (2.861)	33	6.412 (4.849)	6.092 (4.666)	4.576 (3.027)
14	5.173 (5.391)	6.102 (5.667)	5.182 (2.902)	34	7.014 (5.711)	8.357(5.871)	6.287 (3.869)
15	5.778 (8.018)	6.445 (8.049)	4.782 (2.899)	35	8.213 (9.143)	**9.501 (9.241)**	6.569 (4.084)
16	3.751 (2.589)	4.096 (2.730)	3.375 (2.127)	36	4.576 (2.970)	5.955 (3.301)	4.537 (2.896)
17	5.738 (3.666)	5.425 (3.597)	5.069 (3.141)	37	5.831 (3.987)	6.679 (4.250)	5.543 (3.592)
18	8.993 (5.879)	8.497 (5.384)	7.389 (4.574)	38	**10.870 (8.715)**	**10.640 (7.608)**	**9.793 (7.251)**
21	3.249 (2.446)	4.323 (3.229)	4.249 (2.905)	41	3.632 (4.261)	4.239 (4.466)	3.558 (2.444)
22	4.667 (4.202)	5.319 (4.522)	4.322 (2.867)	42	3.609 (2.688)	4.848 (3.185)	4.418 (2.736)
23	6.093 (6.306)	6.473 (6.235)	5.180 (2.980)	43	5.720 (4.897)	5.644 (4.838)	4.277 (2.836)
24	5.252 (5.203)	5.739 (5.275)	4.546 (2.721)	44	6.537 (4.834)	7.649 (5.026)	5.682 (3.465)
25	6.619 (10.750)	6.822 (10.780)	4.596 (2.657)	45	7.567 (7.493)	**9.574 (7.468)**	6.532 (3.925)
26	4.102 (2.902)	4.298 (2.963)	3.215 (2.017)	46	4.173 (3.070)	5.153 (3.603)	4.165 (2.864)
27	5.916 (3.728)	5.875 (3.663)	5.177 (3.226)	47	5.264 (3.688)	6.433 (4.071)	5.744 (3.615)
28	**9.291 (5.347)**	**9.468 (5.334)**	7.877 (4.472)	48	**9.246 (8.158)**	**9.089 (6.812)**	8.179 (6.154)

* Trained with weighted sum loss and tested on 2598 patients, and the standard deviation is included in the brackets. X_e, Y_e, and Z_e denote angle errors along the X, Y, and Z axes, respectively. Bold numbers indicate the worst angle errors or errors larger than 9°

16 teeth with various scales of estimated angle errors. As we can see, the teeth with angle errors less than 15° did not exhibit significant differences between the estimates and the ground truth from human experts. For teeth with errors of 15°−40°, the cause is the prediction error of our model, which reveals the limitations of our method. An interesting observation is for teeth with errors near 90° or 180°, due mainly to the confusion of different Cartesian axes, e.g., predicting the X axis as the Y axis, or due to the opposite axis directions. In fact, these errors might also come from human experts during annotation, as shown in Fig. 2b and discussed below.

We showed four cases of result visualizations at the patient level. For cases 1 and 2 with minor mean angle errors, we noticed that the predicted orientations matched the ground truth with indistinguishable differences. For cases 3 and 4 with larger mean angle errors, most teeth had minimum angle errors, while only a few teeth exhibited relatively large angle errors. The inferior performance could be partially attributed to the estimation error but may come from the annotation errors of human beings. Although we have an extensive dataset, a few annotation errors might still exist despite following a strict data annotation pipeline. These annotation errors with opposite axis directions are anatomically valid as well, as highlighted with black boxes in Fig. 2b. Such mistakes would not impede the clinical or industrial applicability of our method. More visualizations are attached in Figs. S1 and S2 in the supplementary materials.

4 Discussion

Traditional methods play an important role in the estimation of tooth pose. Mok et al. (2002) introduced a technique using 2D view matching to ascertain tooth orientation, while Ulrich et al. (2022) presented a novel approach for achieving complete 6-DoF pose estimation of a circular fiducial marker. However, such methods usually lack large-scale data for validation.

This paper introduces the first approach using deep learning methods for accurate and automatic 6-DoF tooth pose estimation, which has already been

integrated into clinical software for orthodontics in China. The key novelty of our TP-Net is that it identifies the tooth pose with an end-to-end system that integrates a feature extractor backbone, a two-branch estimation head, and a novel loss function, leading to impressive performance in comprehensive experiments on a newly collected large dataset. To the best of our knowledge, we are the first to present deep learning models for tooth pose estimation and to verify their effectiveness with a large-scale dataset.

Although our deep learning model performed well with respect to estimation accuracy and efficiency, there are some limitations. First, the performance of our model on rotation estimation needs to be improved. For example, the orientation estimation of the third molars could be significantly boosted. Besides acquiring more data samples, we need to develop novel and effective tooth pose estimation methods for such a low-resource regime. Meanwhile, the performance of teeth other than third molars needs to be improved; i.e., with our current estimation angle errors of 3°–6°, an inclusive rotation angle error of less than 3° is expected. However, note that even two annotations of the same tooth from the human experts might have a deviation of 3°. Hence, we need to set a reasonable goal for rotation estimation with clinical applicability.

Currently, our method is used mainly for treatment planning in orthodontics, e.g., for precise simulation of the optimal force magnitude and directions. To investigate the clinical applicability of our method, additional experiments are required. Many real-world clinical applications require accurate tooth pose estimation, such as implants, yet the clinical applicability of our method in various scenarios needs to be explored.

Lastly, we use only the tooth crowns collected by 3D IOS, while orientation estimation for teeth with both crown and root plays a pivotal role in many dental applications, such as designing orthodontics treatment plans or evaluating treatment effectiveness. Extensive research needs to be conducted for more accurate and comprehensive tooth pose estimation, motivated by real-world clinical applications.

5 Conclusions

We proposed an accurate and automatic 6-DoF tooth pose estimation system. The system shows good feasibility with a large dataset and has been integrated into a real-world clinical orthodontic system to assist dental diagnosis in orthodontics. Our work reveals that deep learning and artificial intelligence possess great potential for developing more intelligent and efficient digital dental solutions.

Contributors

Wanghui DING, Mengfei YU, and Zuozhu LIU designed and initiated the project. Zuozhu LIU, Jianhua LI, and Wanghui DING designed the model architecture. Kaiwei SUN and Hangzheng LIN conducted validation experiments. Kaiwei SUN, Jianhua LI, and Hangzheng LIN created the dataset. Hangzheng LIN, Kaiwei SUN, and Yang FENG developed the software. Wanghui DING and Kaiwei SUN drafted the paper. Wanghui DING and Zuozhu LIU revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Compliance with ethics guidelines

This article does not contain any studies with human or animal subjects performed by any of the authors.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

References

Cai M, Reid I, 2020. Reconstruct locally, localize globally: a model free method for object pose estimation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3150-3160. https://doi.org/10.1109/CVPR42600.2020.00322

Chen QM, Wang YH, Shuai J, 2023. Current status and future prospects of stomatology research. *J Zhejiang Univ-Sci B (Biomed & Biotechnol)*, 24(10):853-867. https://doi.org/10.1631/jzus.B2200702

Gu CH, Ren XF, 2010. Discriminative mixture-of-templates for viewpoint classification. 11th European Conf on Computer Vision, p.408-421. https://doi.org/10.1007/978-3-642-15555-0_30

Herrmann W, 1967. On the completion of Fédération Dentaire Internationale Specifications. *Zahn Mitteil*, 57(23):1147-1149 (in German).

Hinterstoisser S, Cagniart C, Ilic S, et al., 2012. Gradient response maps for real-time detection of textureless objects.

IEEE Trans Patt Anal Mach Intell, 34(5):876-888. https://doi.org/10.1109/TPAMI.2011.206

Kendall A, Grimes M, Cipolla R, 2015. PoseNet: a convolutional network for real-time 6-DOF camera relocalization. Proc IEEE Int Conf on Computer Vision, p.2938-2946. https://doi.org/10.1109/ICCV.2015.336

Li ZG, Wang G, Ji XY, 2019. CDPN: coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation. Proc IEEE/CVF Int Conf on Computer Vision, p.7677-7686. https://doi.org/10.1109/ICCV. 2019.00777

Liebelt J, Schmid C, Schertler K, 2008. Viewpoint-independent object class detection using 3D feature maps. IEEE Conf on Computer Vision and Pattern Recognition, p.1-8. https://doi.org/10.1109/CVPR.2008.4587614

Mok V, Ong SH, Foong KWC, et al., 2002. Pose estimation of teeth through crown-shape matching. Proc SPIE 4684, Medical Imaging 2002: Image Processing, p.955-964. https://doi.org/10.1117/12.467048

Newell A, Yang KY, Deng J, 2016. Stacked hourglass networks for human pose estimation. 14th European Conf on Computer Vision, p.483-499. https://doi.org/10.1007/978-3-319-46484-8_29

Oberweger M, Rad M, Lepetit V, 2018. Making deep heatmaps robust to partial occlusions for 3D object pose estimation. Proc 15th European Conf on Computer Vision, p.125-141. https://doi.org/10.1007/978-3-030-01267-0_8

Park K, Patten T, Vincze M, 2019. Pix2Pose: pixel-wise coordinate regression of objects for 6D pose estimation. Proc IEEE/CVF Int Conf on Computer Vision, p.7667-7676. https://doi.org/10.1109/ICCV.2019.00776

Peng SD, Liu Y, Huang QX, et al., 2019. PVNet: pixel-wise voting network for 6DoF pose estimation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4556-4565. https://doi.org/10.1109/CVPR.2019.00469

Qi CR, Litany O, He KM, et al., 2019. Deep Hough voting for 3D object detection in point clouds. Proc IEEE/CVF Int Conf on Computer Vision, p.9276-9285. https://doi.org/10.1109/ICCV.2019.00937

Su H, Qi CR, Li YY, et al., 2015. Render for CNN: viewpoint estimation in images using CNNs trained with rendered 3D model views. Proc IEEE Int Conf on Computer Vision, p.2686-2694. https://doi.org/10.1109/ICCV.2015.308

Sun M, Bradski G, Xu BX, et al., 2010. Depth-encoded Hough voting for joint object detection and shape recovery. 11th European Conf on Computer Vision, p.658-671. https://doi.org/10.1007/978-3-642-15555-0_48

Ulrich J, Alsayed A, Arvin F, et al., 2022. Towards fast fiducial marker with full 6 DOF pose estimation. Proc 37th ACM/SIGAPP Symp on Applied Computing, p.723-730. https://doi.org/10.1145/3477314.3507043

Wang C, Xu DF, Zhu YK, et al., 2019. DenseFusion: 6D object pose estimation by iterative dense fusion. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3338-3347. https://doi.org/10.1109/CVPR.2019.00346

Wang H, Sridhar S, Huang JW, et al., 2019. Normalized object coordinate space for category-level 6D object pose and size estimation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2637-2646. https://doi.org/10.1109/CVPR.2019.00275

Wang Y, Sun YB, Liu ZW, et al., 2019. Dynamic graph CNN for learning on point clouds. *ACM Trans Graph*, 38(5): 146. https://doi.org/10.1145/3326362

Wei GD, Cui MZ, Liu YM, et al., 2020. TANet: towards fully automatic tooth arrangement. 16th European Conf on Computer Vision, p.481-497. https://doi.org/10.1007/978-3-030-58555-6_29

Xiang Y, Schmidt T, Narayanan V, et al., 2018. PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes. Proc 14th Robotics: Science and Systems.

Zhou Y, Tuzel O, 2018. VoxelNet: end-to-end learning for point cloud based 3D object detection. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.4490-4499. https://doi.org/10.1109/CVPR.2018.00472

Zhou Y, Barnes C, Lu JW, et al., 2019. On the continuity of rotation representations in neural networks. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5738-5746. https://doi.org/10.1109/CVPR.2019.00589

Zhu JJ, Yang YX, Wong HM, 2023. Development and accuracy of artificial intelligence-generated prediction of facial changes in orthodontic treatment: a scoping review. *J Zhejiang Univ-Sci B (Biomed & Biotechnol)*, 24(11):974-984. https://doi.org/10.1631/jzus.B2300244

List of supplementary materials

Table S1 Tooth axis data statistics

Table S2 Median, minimum, and maximum values of the predicted angle error from M_1

Table S3 Average occlusal angle for teeth pairs

Fig. S1 Visualization of estimated orientation and ground truth for Example 1

Fig. S2 Visualization of estimated orientation and ground truth for Example 2