TANet: Towards Fully Automatic Tooth Arrangement

Guodong Wei[1,2], Zhiming Cui[2], Yumeng Liu[2], Nenglun Chen[2], Runnan Chen[2], Guiqing Li[1], and Wenping Wang[2]

[1] South China University of Technology, Guangzhou, China
{csgdwei@mail, ligq}.scut.edu.cn
[2] The University of Hong Kong, Hong Kong
{zmcui, lym29, nolenc, rnchen2, wenping}@cs.hku.hk

Abstract. Determining optimal target tooth arrangements is a key step of treatment planning in digital orthodontics. Existing practice for specifying the target tooth arrangement involves tedious manual operations with the outcome quality depending heavily on the experience of individual specialists, leading to inefficiency and undesirable variations in treatment results. In this work, we proposed a learning-based method for fast and automatic tooth arrangement. To achieve this, we formulate the tooth arrangement task as a novel structured 6-DOF pose prediction problem and solve it by proposing a new neural network architecture to learn from a large set of clinical data that encode successful orthodontic treatment cases. Our method has been validated with extensive experiments and shows promising results both qualitatively and quantitatively.

Keywords: deep learning, orthodontics, tooth arrangement, 6D pose prediction, structure, graph neural network.

1 Introduction

Irregular tooth arrangements cause not only aesthetic issues but also compromised masticatory functions. Incorrect bite relationship, such as overjet or crowded teeth, may lead to disorders in chewing, which often induces other secondary diseases. With the growing concern of oral health, there is tremendous demand for orthodontic treatment. Although the number of people seeking orthodontic care is increasing rapidly, there is in general a severe lack of certified orthodontists to meet the demand. Currently, orthodontic treatment involves tedious manual operations and training professional orthodontists is a lengthy and costly process. Moreover, the quality of diagnosis and treatment depends in a large degree on the skills and experiences of individual orthodontists. Hence, it is imperative to develop a fully automated system for fast recommendation of optimal tooth arrangements to improve the efficiency and quality of orthodontic treatment planning.

Tooth arrangement is an essential step of orthodontic treatment. Given a set of ill-positioned teeth of a patient, tooth arrangement aims to predict an

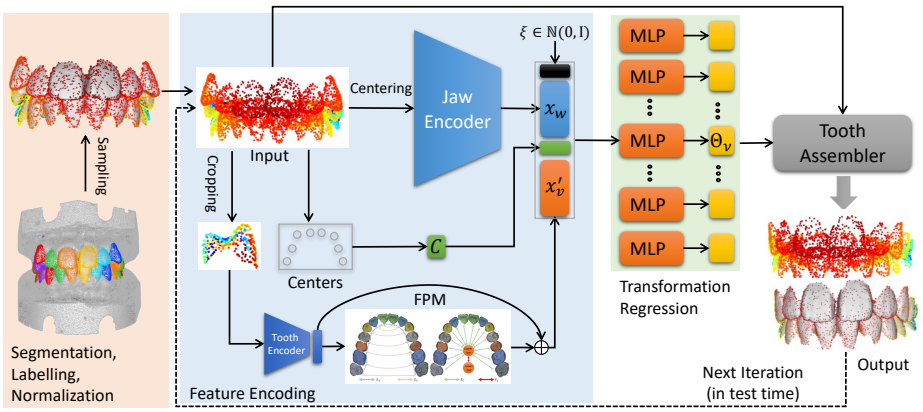


Fig. 1. The overall pipeline and the network architecture of our method. At the first stage, the input 3D dental model is automatically segmented to produce the label and point-cloud representation of each tooth. The point sets of each tooth is then normalized and sampled. The second stage involves a network consisting of four components: feature encoding, feature propagation, transformation (i.e. pose) regression, and tooth assembler modules. The final output is a rearranged dentition. For clarity, only one tooth-level encoder is illustrated here.

ideal tooth layout that serves as the target arrangement to achieve through orthodontic treatment. In order to produce a satisfactory arrangement, multiple factors need to be taken into consideration. This makes tooth arrangement a complex task with its outcome quality heavily dependent on professional skills and subjective judgement of orthodontists.

Existing computer-aided systems used in orthodontic treatment planning provide a user interface for visualizing and manually editing individual teeth. As a related work in prosthodontics, Dai [8] performs complete denture tooth arrangement according to a set of heuristic rules, with teeth selected from a pre-specified set. In contrast we intend to solve a different and more challenging problem of tooth arrangement with patient-specific dentition for orthodontic treatment. The work in [6] automatically establishes proper dental occlusion by treating the upper teeth and lower teeth as two rigid objects, while our tooth arrangement problem requires pose adjustment of each individual tooth in a dentition.

To automatically determine the ideal positions of teeth for each specific patient is extremely challenging. Even though clinical rules like "Andrew's six keys" [1] suggest the necessary conditions for proper tooth alignment, the actual layout of patient's teeth may prevent the accessibility to the theoretically ideal poses. Therefore, a mathematical model developed by rule-based method can hardly lead to a clinically feasible outcome. Apart from this, detecting landmarks or other human-defined features on dental models is a tedious process and may also introduce errors at the very beginning of pose prediction. Moreover, dental

models are texture-less and lack of sharp features, especially when we only have dental crowns, i.e. teeth outside gums. These characteristics make it hard to define orientation, position or other low-level features of a tooth precisely and consistently, while they are the prerequisites for a rule-based method.

We proposed a learning-based approach to predict an optimal treatment target from the initial irregular tooth positions of a patient before treatment, and thereby developed the first method for automatic tooth arrangement for orthodontic treatment. We formulate the tooth arrangement task as a structured 6D poses prediction problem, which has not been fully explored by the computer vision community. Our network aims to approximate the mapping from an input dental model representing the initial tooth arrangement to an ideal target poses via supervised learning. The network consists of four main components: a feature encoding module for information at the jaw-level and the tooth-level, a feature propagation module for information passing among teeth, a pose regression module for 6-DOF pose prediction and a differentiable tooth assembler module for rigid transformations. The loss function of the network is specially designed to capture intrinsic differences between different arrangements, enhance compact spatial relation and model the uncertainties in ground truth.

To summarize, the main contributions of this work are:

- We developed the first automatic tooth arrangement framework based on deep learning;
- We proposed the use of a graph-based feature propagation module to update features extracted by PointNet to provide crucial contextual information for successfully solving the structured poses prediction problem arising from the tooth arrangement task.
- We proposed a novel loss function that is able to provide effective supervision for aligning teeth by capturing intrinsic differences, spatial relations and uncertainties in the distribution of malaligned tooth layouts.

2 Related Work

6-DOF Pose Estimation Problem The pose estimation problem has been extensively studied in recent years. It aims to infer the three-dimensional pose, which has six degrees of freedom, of an object present in an RGB image, [3, 7, 45, 5, 33, 34, 18, 25], RGB-D image [39, 40, 35], or point cloud data [26, 44, 29, 30]. Existing methods can be roughly categorized into the object coordinate regression approach and the template matching approach. The methods based on coordinate regression estimates the object's surface corresponding to each object at the pixel level, with the assumption that the corresponding 3D model is known for training [36]. The methods based on template matching perform alignment between known 3D models and image observations using various techniques, such as Iterative Closest Point (ICP) [4]. All these previous works do not consider multiple objects and their relative relationships, while the tooth arrangement problem that we face needs to predict 6-DOF poses of all the teeth (i.e. multiple

objects) at the same time to form a regular layout, Most importantly, the 6-DOF pose estimation problem is concerned the relation between the pose of a known 3D shape and its image observation, In contrast, we aim to solve a more challenging problem of predicting the poses of regularly arranged teeth by learning from clinical data of orthodontic treatment.

Furniture Arrangement or Placement Problem There have recently been many studies on how to automatically generate an optimized indoor scene composed of various furniture objects [42, 10, 21, 14, 38, 37]. To simplify the problem, most of these methods use bounding boxes as proxies to roughly approximate the input objects, without taking into account the fine-grained geometric details of the objects. The work in [42] optimizes the configurations of given 3D models using learned priors. The core of their method is an energy function defined with a set of heuristic rules. The method in [31] addresses the problem of placing one 3D object with respect to others, assuming that all the objects are pre-aligned with the same orientation, thus only translation of the newly added component needs to be predicted. Since man-made furniture shapes usually have distinct sharp features, the orientations of these objects can easily be defined. As a comparison, we consider dental models which lack such distinct features, which makes it hard to precisely define orientations. Furthermore, most works on the furniture arrangement problem attempt to generate diverse arrangements for a given indoor scene, while the goal of the tooth arrangement problem is the best tooth arrangement for each specific patient.

3D Shape Generation Problem aims to generate realistic 3D shapes from user specifications or by inferring from images or partial models. Conditional generative methods [22] can generate realistic images based on the input condition. With the advance in geometric learning and 3D representation methods, various generative models have been proposed as powerful tools to process 3D shapes [32, 27, 28, 24]. The problem of conditional 3D shape generation and the problem of automatic tooth arrangement both aim to generate 3D shapes according to given conditions. Recent works of conditional 3D shape generation [11, 23, 17, 13] focus on generating realistic structured shapes that are able to adapt diverse shape variations. However they do not preserve the geometries of input objects, while this is a hard constraint in the tooth arrangement problem.

3 Method

3.1 Overview

An illustrated in Fig. 1, our proposed method contains two main stages. The first is a preprocessing stage that segments dental crowns from the whole model and then semantically labeling each individual tooth crown. The second stage uses a network with four main components to perform the following functions: a) a set of PointNet-based point feature encoders for jaw-level and tooth-level feature extraction; b) a graph-based feature propagation module that transfers information among teeth; c) the regressor for each tooth combines its corresponding tooth-level features, global features and a random conditional vector

as input, and outputs the 6D transformation relative to the input position of this tooth; d) an assembler to map the 3D rotations represented in the axis-angle representation into rotation matrices for transforming the points, and output the rearranged point cloud. The details are described in the subsequent sections.

3.2 Preprocessing

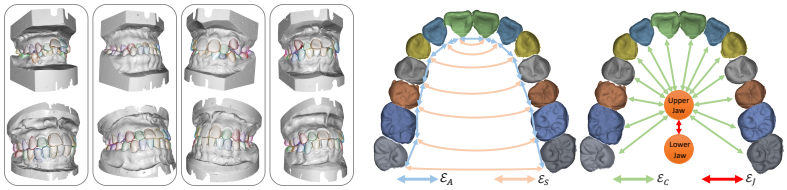
Segmentation and labeling are critical as preprocessing operations for our tooth arrangement algorithm. There exist many off-the-shelf methods [41, 20, 43] for accurate automatic semantic segmentation and labeling on 3D dental meshes. We use the method in [41]. The tooth labels are assigned according to *FDI two-digit notation* for permanent teeth. Note that we only keep the crowns for all the teeth for use in our tooth arrangement computing. A local coordinate system is then defined for the model consisting of these crowns to normalize the position and orientation by coarsely aligning it with the world coordinate system. The resulting tooth set is denoted $X = \{X_v \subseteq \mathbb{R}^3 | v \in \mathcal{V}\}$, where $\mathcal{V}$ is the set of tooth crown labels and X_v is the point cloud of the crown with label v.

3.3 Network

Tooth Centering. Since the input teeth are sparsely distributed in the space, this may increase the difficulty in capturing the features among teeth that are far away from each other. So we first translate all the teeth to the origin so that $\widetilde{X}_v = \{p' = p - c_v | p \in X_v\}$ and $\widetilde{X} = \{\widetilde{X}_v | v \in \mathcal{V}\}$, where $c_v \in C$ is the geometric center of tooth v. This measure is key to decoupling the center positions of teeth from other features so as to enable the encoder to focus on extracting geometric features, such as shape details, orientation and size, in a translation-independent manner.

Feature Encoder. The feature encoders in our network are based on PointNet [27]. Using symmetric functions, PointNet achieves permutation invariance of point sets and is able to efficiently extract local features for each point and global features for the whole point cloud. Here, we will use its global features thus extracted. The quality of tooth arrangement is determined by the position and orientation of every tooth with respect to the others, and the information of each tooth and that of the whole dentition are equally important. We therefore extract jaw-level features $\mathrm{x}_w = E_w(\widetilde{X}, C)$ and tooth-level features $\mathrm{x}_v = E_v(\widetilde{X}_v)$, where E_w represents the encoder for the whole tooth crown set and E_v the encoder for individual teeth.

Feature Propagation Module. Note that the jaw-level features are rather sparse and do not capture many geometric details, as shown in Fig. 8(b) and discussed in Section 4.5. The tooth-level features capture details, however, are encoded independently and so oblivious to the information from other teeth. So it is hard to achieve an accurate alignment of teeth by using these features. We introduce a graph-based feature propagation module (FPM) that allow geometric detail information to transfer among teeth via the connections of the graph.



(a) (b)

Fig. 2. (a) Representative examples of tooth arrangement in orthodontics. Each column contains dental models before(top) and after(bottom) the treatment; (b) The tooth graph used for feature propagation. We show the teeth connection in upper jaw here. The connection between jaws are also demonstrated.

Our feature propagation module G is based on the propagation model in [19]. First, we define a tooth graph as $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{H})$, where $\mathcal{N}$ is the set of nodes, each corresponding to a tooth v, $\mathcal{E}$ the set of undirected edges of the graph, and $\mathcal{H}$ the node embedding. In addition, two super nodes are created for the upper jaw and lower jaw. The node embedding of these two super nodes are set to zero vectors initially. The embedding h_v of any other node is initialized with its feature x_v. As illustrated in Fig. 2(b), $\mathcal{E}$ consists of four types of edges $\mathcal{E}_A, \mathcal{E}_S, \mathcal{E}_C, \mathcal{E}_J$, namely, $\mathcal{E} = \mathcal{E}_A \cup \mathcal{E}_S \cup \mathcal{E}_C \cup \mathcal{E}_J$, where $\mathcal{E}_A$ contains relationships between adjacent teeth in the same jaw, $\mathcal{E}_S$ connects left and right symmetric teeth in the same jaw, $\mathcal{E}_C$ consists of connections between each tooth node and its supper node of the corresponding jaw, and $\mathcal{E}_J$ is a set including single edge between two super nodes. Finally, local features x_v are updated in K iteration, each iteration with a fixed number of steps T, as follows:

$$\mathrm{m}_v^{k,t+1} = \sum_{w \in N(v)} \mathrm{A}_{e_{vw}}^k \mathrm{h}_w^{k,t}, \tag{1}$$

$$\mathrm{h}_v^{k+1,t+1} = \mathrm{GRU}(\mathrm{h}_v^{k,t}, \mathrm{m}_v^{k,t+1}), \tag{2}$$

where $N(v)$ denotes the set of neighboring nodes of v and $\mathrm{A}_{e_{vw}}^k$ is a learned matrix for each type of edge in the graph $\mathcal{E}$. Both K and T are set to 3 in our experiments.

To further improve the network performance, we add a residual connection with the original feature. The final updated tooth feature x'_v is obtained after a residual operation,

$$\mathrm{x}'_v = \mathrm{x}_v + \mathrm{h}_v^{K+1,T+1} \tag{3}$$

Pose Regressor. Considering that many other factors, such as the subjective judgment of clinical orthodontists or the age, gender and face appearance of patients, may also affect the layout of the optimal tooth arrangement, we generate a set of candidates instead of giving only one result. Inspired by the MoN loss [9], which is originally proposed to model the uncertainty in 3D recovery from a single image, we introduce a conditional weighting (CW) scheme.

This scheme is designed to allow the network to generate multiple plausible arrangements and still be able to recommend a most appropriate one. To make the CW scheme work, here in the pose regressor, we only need to append a random vector $\xi \in \mathbb{N}(0, \mathrm{I})$ to the input features in training, where $\mathbb{N}(0, \mathrm{I})$ is a zero-mean Gaussian distribution. We set ξ to a zero vector in testing. Another part of the CW scheme lies in the loss function (Section 3.4). All the features are combined and fed into the corresponding pose regressors to predict 6D transformation parameters,

$$\Theta_v = \Psi_v \left(C, \mathrm{x}_w, \mathrm{x}_v^{'}, \xi \right) \tag{4}$$

where Θ_v consists of $\mathrm{r}_v = (r_v^x, r_v^y, r_v^z)$ in axis-angle representation for rotation and $\mathrm{t}_v \in \mathbb{R}^3$ for translation.

Tooth Assembler. The predicted transformation parameters are then passed to the assembler Φ to transform, assemble and generate the final output. This module maps the axis-angle representation of rotation $\mathrm{r}_v \in \mathcal{R}^3$ back into a rotation matrix $R_v \in SO(3)$ through an exponential map, which is differentiable. The exponential map $\exp : so(3) \to SO(3)$ connects the Lie algebra with the Lie group by

$$\exp(\mathrm{r}_\times) = \mathrm{I}_{3\times 3} + \frac{sin\theta}{\theta}\mathrm{r}_\times + \frac{1 - cos\theta}{\theta^2}\mathrm{r}_\times^2, \tag{5}$$

where $\theta = \|\mathrm{r}\|_2$ is the rotation angle. Let $\mathrm{r} = (r^x, r^y, r^z)$ be a rotation vector in axis-angle representation, with the associated skew-symmetric matrix

$$r_\times = \begin{bmatrix} 0 & -r^z & r^y \\ r^z & 0 & -r^x \\ -r^y & r^x & 0 \end{bmatrix}. \tag{6}$$

Then, given r_v for a tooth v, the assembler first maps it to a rotation matirx R_v using Equation 5, and then applies the transformation to the input points to get the final output point cloud

$$X^* = \left\{ R_v p_v + c_v + t_v | v \in \mathcal{V}, p_v \in \widetilde{X}_v \right\}. \tag{7}$$

3.4 Loss Function

Geometric Reconstruction Loss. Based on the observation that teeth remain almost rigid during the treatment process and our network also keeps the shape of each tooth in input, we use iterative closest points method to align each pair of teeth in the prediction and ground truth (Fig. 4(c)). Then, for points in X_v^*, we find their correspondences $P_{\bar{X}}(X_v^*)$ by searching for the closest points in ground truth $\bar{X}_v$ based on the rigid alignment result. The function $P_{\bar{X}}(\cdot)$ represents this correspondence searching process. To eliminate the loss induced by global rigid transformation and reveal the intrinsic between two arrangements, we solve for

a global rigid transformation Π to align the prediction and ground truth (Fig. 4(d)) by minimizing the following energy,

$$\underset{\Pi}{\operatorname{argmin}} \sum_{v \in \mathcal{V}} \|[\mathrm{X}_v^*|1]^\top - \Pi[P_{\bar{X}}(X_v^*)|1]^\top\|_2^2, \tag{8}$$

where X_v^* is the coordinate matrix of X_v^*. We solve the above problem by orthogonal Procrustes analysis. Finally, the reconstruction loss is calculated as

$$L_{recon}(X^*, \bar{X}) = \sum_{v \in \mathcal{V}} \|[\mathrm{X}_v^*|1]^\top - \Pi[P_{\bar{X}}(X_v^*)|1]^\top\|_S, \tag{9}$$

where $\|\cdot\|_S$ represents the $Smooth_{L1}$ norm [12].

Geometric Spatial Relation Loss. To emphasize the fact that a good arrangement is mostly determined by the mutual spatial relation between all the teeth, we define the geometric spatial relation between two point sets $S_1, S_2 \subseteq \mathbb{R}^3$ as

$$V_{S_1,S_2} = \bigcup_{\substack{i \neq j \\ 1 \leq i,j \leq 2}} \left\{ x - y^* | y^* = \underset{y \in S_i}{\operatorname{argmin}} \|x - y\|_2^2, x \in S_j \right\}. \tag{10}$$

Based on the simple observation that the distance between two teeth should not be larger than a threshold σ if the dentition is aesthetically and functionally satisfactory, we calculate the clamped V_{S_1,S_2} by clamping all elements into $[-\sigma, +\sigma]$ and denote as V_{S_1,S_2}^c. We empirically set $\sigma = 5.0$ in all our experiments. Finally, the geometric spatial relation loss is calculated as,

$$L_{spatial}(X^*, \bar{X}) = \sum_{q \in \mathcal{N}} \sum_{e \in \mathcal{P}(q)} \|V_{X_q^*, X_e^*}^c - V_{P_{\bar{X}}(X_q^*), P_{\bar{X}}(X_e^*)}^c\|_S, \tag{11}$$

where $\mathcal{P}(q) = \mathrm{NBR}(q) \cup \mathrm{OPS}(q)$. The functions $\mathrm{NBR}(q)$ and $\mathrm{OPS}(q)$ return neighboring nodes and the opposite jaw, respectively. If node q is a super node for a jaw, then X_q^* is the set of points of all teeth belongs to that jaw.

Conditional Weighting Loss. As discussed in Section 1, we introduce a mechanism that allows the network to model uncertainty and generate a distributional output for one input given a conditional vector ξ. Our approach is inspired by the MoN loss [9] with the following variation. We enable the network to recommend a most likely arrangement by using a conditional weighting loss, which is defined as follows

$$\sum \min_{\substack{\xi_j \sim \mathbb{N}(\mathbf{0}, \mathbf{I}) \\ 1 \leq j \leq n}} \{ \frac{1}{e^{\|\xi_j\|}} \cdot Loss(X^*, \bar{X}) \}, \tag{12}$$

where $Loss = L_{recon} + L_{spatial}$ and n is set to 2 in our experiments.

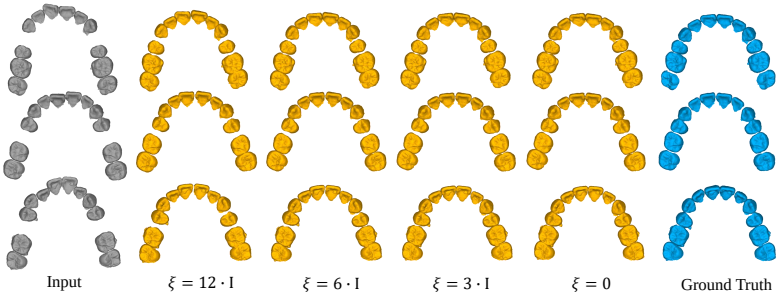


Fig. 3. For the same input model in test time, we give different values of ξ as conditions for the regressor, which result in predictions that have different distances with respect to the ground truth. It is observed that the prediction with $\xi = 0$ is often the most satisfactory arrangement.

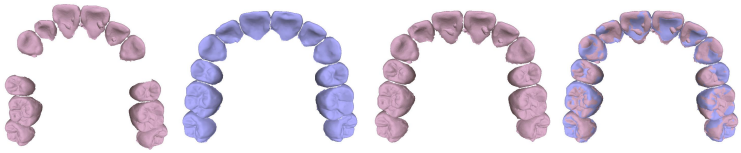


Fig. 4. (a) A pre-treatment model; (b) The corresponding post-treatment model (c) The aligned model with its tooth shapes from (a) and tooth poses from (b). It is used in the definition of our reconstruction loss (see Section 3.4); (d) Superposing (b) on (c) to visualize their differences.

3.5 Implementation and training details

Network Details. The dimensions of features encoded by the global and local PointNet encoders are 1024 and 512, respectively. The length of node embedding in FPM is set to 512. The random condition ξ is a 32-dimensional vector. The pose regressors consist of 3 linear layers with ReLU activator and dropout (0.3) in the first two layers. Only the Tanh activator is used in the last linear layer. The weights in the last layer of the regressors are initialized as zeros, as we assume that the teeth are more likely unmoved.

Training Details. Searching for corresponding point pairs is done before the training begins, since corresponding point pairs do not change due to rigid movement of teeth. Teeth that do not appear in both before and after treatment models are regarded as missing or extracted. We randomly sample 400 points on each tooth as the input. As for missing teeth, we set their positions with zeros. To augment the training data, all individual teeth of the input models, including pre-treatmemt and post-treatment models, are randomly rotated by an angle, within [-30,+30] in our experiments, in a random direction and translated by a distance vector from the zero-mean Gaussian distribution $\mathbb{N}(0, 1^2)$. The complete set of teeth is also augmented by a random global rotation. Note that these augmented models are only used to enlarge the set of the simulated pre-

Table 1. Ablation study. The mean errors of translation ΔT_{avg}, rotation $\Delta\theta_{avg}$, ADD and PA-ADD together with their AUC scores are reported. The coordinate unit is $millimeter(mm)$ except for $\Delta\theta_{avg}$, which is in $degree(^\circ)$.

	ΔT_{avg}/AUC	$\Delta\theta_{avg}$/AUC	ADD/AUC	PA-ADD/AUC
NetBL+Lrecon	1.09/73.47	9.26/57.13	1.200/70.825	1.038/74.719
NetGL+Lchamfer	1.06/73.47	7.08/65.78	1.133/71.795	0.992/76.082
NetGL+Lrecon	1.03/74.43	6.70/67.30	1.096/72.821	0.957/77.032
NetGL+FPM+Lrecon	0.99/75.46	**6.64**/67.67	1.057/73.864	0.893/78.195
NetGL+FPM+Lrecon+CW	0.98/75.61	6.71/67.32	1.051/73.953	**0.886**/**78.512**
NetCom+Lcom	**0.97**/**76.00**	**6.64**/**67.71**	**1.036**/**74.362**	0.893/78.456

treatment models. We assume that an augmented pre-treatment dental model $\mathcal{M}^*$ should be still mapped to the corresponding post-treatment model $\bar{\mathcal{M}}$, and an augmented treated model $\bar{\mathcal{M}}^*$ should also be mapped to its corresponding original post-treatment model before augmentation $\bar{\mathcal{M}}$.

Our network is implemented with PyTorch and trained on a server using one 1080-Ti GPU. We use Adam optimizer. The batch size is set to be 16 with a learning rate initially equal to $1.0e-4$ and dropping down by 0.5 when the validation loss stops improving.

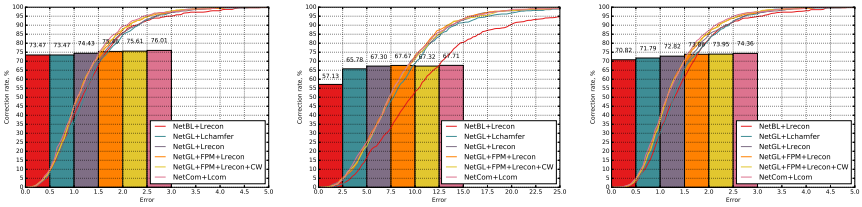
4 Experiments

4.1 Dataset

Our dataset consists of dental models of 300 patients, with males (47%) and females (53%) of age ranging from 6 to 18 years old. For each patient, there are two models scanned before and after treatment. All the three types of malocclusions (i.e., Class I, II, III) are observed in our dataset, according to Angle's classification [2]. Some examples are shown in Fig. 2(a). For network training, we randomly divide the 300 pairs of dental models of our dataset into three groups: 200 for training, 30 for validation, and 70 for testing.

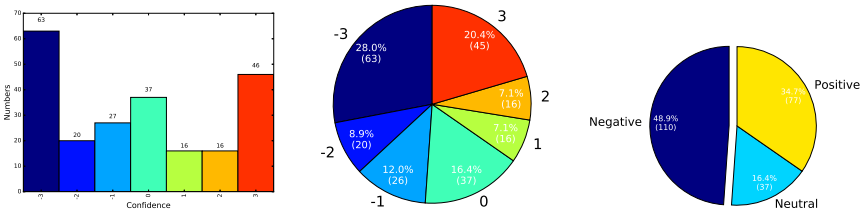
4.2 Evaluation Metric

We evaluate the precision of our network prediction using the ADD metric [15], which is the mean point-wise distance between the predicted and ground truth models. We also report PA-ADD which is the ADD calculated after the rigid alignment between predicted jaw and ground truth jaw using Procrustes Analysis. In addition, we define PCT@K metric as the percentage of tooth predicted by the network with the error smaller than a threshold K. The error can be the shape reconstruction error, rotation or translation estimation error, etc. Similar to AUC [39] for 6-DOF pose estimation, we define PCT-AUC as the area under the PCT curve, which is the integral of PCT with respect to K. The PCT-AUC for shape reconstruction, rotation and translation errors are denoted



(a) ΔT/AUC (b) $\Delta\theta$/AUC (c) ADD/AUC

Fig. 5. The quantitative evaluation on the effectiveness of different components.



(a) (b) (c)

Fig. 6. Statistics of the user study.

as ADD-AUC, $\Delta\theta$-AUC and ΔT-AUC, respectively. We set the maximum K of PCT-AUC to be $5mm$ for translation or reconstruction errors and 25° for rotation error. For ADD and PA-ADD metrics, the smaller values mean better precision. For the various AUC metrics, larger values indicate better precision.

4.3 Ablation Study

In this section, we will show the effectiveness of different components of our proposed network and the impact of different terms in our loss function.

The following three basic network architectures are used in the ablation study. They are the baseline network that has only the jaw-level global feature encoder in feature extracting stage(NetBL); the network with both jaw-level global feature encoder and tooth-level local feature encoders(NetGL); the complete network model proposed(NetCom), which contains all levels of feature encoders, the feature propagation module(FPM) and the conditional weighting mechanism(CW). Different losses are: our reconstruction loss L_{recon}(Lrecon); loss that replaces the $Smooth_{L1}$ with Chamfer distance(Lchamfer); the complete loss function we propose(Lcom). We have conducted six experiments with different combinations of the above networks with different loss functions. The results are reported in Table 1.

Global and Local Feature Integration. As shown in the 1st and 3rd rows in Table 1, introducing tooth-level local feature encoders to the network brings a significant improvement on the result, the PCT-AUC goes from 70.825

to 72.821. The improvement is almost completely caused by the growth in rotation estimation accuracy($\Delta\theta$-AUC is increased by more than 10 points). As will be discussed in Section 4.5 later, this is mainly caused by the local feature extractors that help capture more details of the teeth, so that the rotations can be determined more accurately.

Feature Propagation. The feature propagation module (FPM) is introduced to make arrangements more compact. The 3rd and 4th rows of Table 1 validate the effectiveness of our feature propagation module. The improvement is mainly attributed to the translation estimation accuracy, which is increased by around 1 point in ΔT-AUC.

Conditional Weighting. The conditional weighting mechanism is designed to generate a distribution of predictions, so as to relieve the network from the ambiguities of ground truth due to subjective judgments of different dentists or insufficient input information. The 4th and 5th rows in Table 1 show that this mechanism have larger improvement on PA-AUC than ADD-AUC, because the CW may also mitigate ambiguities introduced by global rotations.

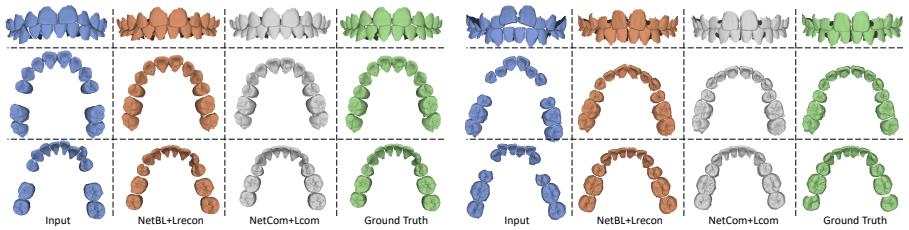
Reconstruction Loss. Based on the assumption that individual teeth have the same shape in each corresponding pair of pre-treatment model and the post-treatment model, we proposed to use MSE in reconstruction loss calculation. The 2nd and 3rd rows of Table 1 indicate that our loss is significantly better than the commonly used Chamfer Distance loss.

Spatial Relation Loss. The ablation study seems to suggest that the network learns better by emphasizing the reconstruction of the spatial relations between teeth. As can be seen in the last two rows of Table 1, the Add-AUC is increased by about 0.4 points. Although the improvement seems small in numerical value, we argue that this is significant in terms of shape variation because humans are visually sensitive to even slight misalignment of teeth.

Our complete model is able to achieve accurate tooth arrangement with around $0.97mm$ translation error, 6.64° rotation error and $0.89mm$ shape difference. A qualitative comparison between our complete method (NetCom+Lcom) and the baseline method (NetBL+Lrecon) is illustrated in Fig. 7. The results of our complete approach are significantly better than those of the baseline method. We show a more comprehensive comparison of these methods using different metrics in Fig. 5.

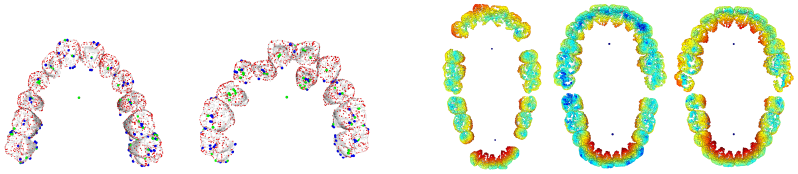
4.4 User Study

In order to evaluate the user perception of our results, we have conducted a user study. We randomly sampled 25 pairs of data from our test set, We recruited 9 students in dentistry and asked them to select the better one between the ground truth solutions and our predictions. The network predictions and ground truth were presented in random orders, with the original malaligned pre-treatment models also presented as reference. Besides, they were asked to score their confidences for each of their selections with numbers between 0 to 3. The confidence score 3 indicates the selected one was much better than the other one, while score 0 indicates that they cannot tell which one is better.



(a) (b)

Fig. 7. A qualitative comparison between our complete method (NetCom+Lcom) and the baseline method (NetBL+Lrecon). Here we show two examples (a-b). Each example includes 3 rows and 4 columns. From top to bottom, the 3 rows are the complete dentition, the upper jaw and the lower jaw of a patient, respectively.



(a) (b)

Fig. 8. (a) The critical points. Red: locally critical, Green: globally critical, Blue: both globally and locally critical; (b) The occlusion fields of the input, network output, and ground truth, respectively. Red: maximum distance; Green: minimum distance.

As shown in Fig. 6(c), in 51.1% of totally $25 \cdot 9 = 225$ ratings, our network predictions are better than or equal to the post-treatment arrangements designed by dentists. In order to take the participants' confidences into account, we sum up these ratings weighted by their confidence scores, with the signs of the ratings set to negative if they prefer the arrangement by dentists (Fig. 6(a, b)). Normalized by $255 \cdot 3$, the final score is a weighted average of the ratings in the range of in $[-1, 1]$, where 1 indicates that our predictions are better and 0 indicates that our predictions and the ground truth judged to be of equal quality. The final score thus computed for our user study is -0.1037.

4.5 Visualization

Critical Points. To provide a better understanding of what our network has learned, we visualize the critical points related to local (tooth-level) and global (jaw-level) features following the method in [27]. As shown in Fig. 8(a), the jaw-level feature extractor captures sparse features around the crown boundaries and the centers of teeth which can be helpful for the coarse arrangement of teeth, while the tooth-level local feature extractors capture denser features that describe the shape details of teeth much better and are beneficial for a more precise and compact arrangement.

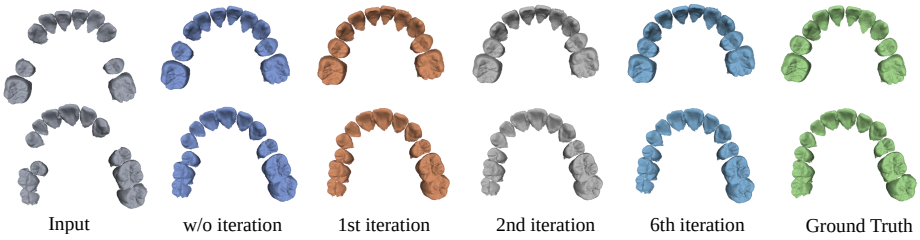


Fig. 9. By iteratively feeding the network output as input to the network, further improvement of arrangement is produced.

Occlusion Field. The occlusion relationship is an important aspect to evaluate the quality of our network prediction. We visualize the occlusion relationship by displaying the minimal distance of every point in one jaw with respect to the opposite jaw, called the *occlusion field*. As illustrated in Fig. 8(b), the occlusion relationship in our prediction is improved significantly compared to the arrangement before treatment.

Distributional Output. We give different vectors ξ as input conditions to the network in test time to generate multiple predictions for an input. Interestingly, it turns out that the predictions are getting closer to the ground truth as the input condition vectors are closer to zero vectors (see Fig. 3).

5 Discussion

Failure Cases. Our method will fail if the input dental models deviate severely from the distribution of training data. To alleviate this problem, during testing, we feed the unsatisfactory output predictions as input back into the network again. As shown in Fig. 9, the arrangement is iteratively refined in this way.

Physical Constraints. Enforcing physical constraints in neural networks is an outstanding problem. Although we have encoded the left-right symmetry prior in FPM and propose $L_{spatial}$ for enhancing compact spatial relation, our network outputs do not guarantee to be physical feasible. Hence, a postprocessing procedure is needed to resolve these problems, such as penetration. See the supplementary materals for more details.

6 Conclusion

We present the first learning-based approach for automatic tooth arrangement in orthodontic treatment planning. By modeling the task as a structured 6-DOF poses prediction problem, we propose a network architecture composed of PointNet encoders and a graph-based feature propagation module, that is able to effectively capture crucial features for a compact alignment. Our novel loss function captures intrinsic geometric difference and uncertainties in ground truth. Extensive experiments validated that our method is able to achieve tooth alignments in quality comparable to those designed by orthodontists.

References

1. Andrews, L.F.: The six keys to normal occlusion. Am J Orthod **62**(3), 296–309 (1972)
2. Angle, E.H.: Classification of malocclusion. Dent. Cosmos. **41**, 350–375 (1899)
3. Aubry, M., Maturana, D., Efros, A.A., Russell, B.C., Sivic, J.: Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3762–3769 (2014)
4. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. International Society for Optics and Photonics (1992)
5. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6d object pose estimation using 3d object coordinates. In: European conference on computer vision. pp. 536–551. Springer (2014)
6. Chang, Y.B., Xia, J.J., Gateno, J., Xiong, Z., Zhou, X., Wong, S.T.: An automatic and robust algorithm of reestablishment of digital dental occlusion. IEEE transactions on medical imaging **29**(9), 1652–1663 (2010)
7. Collet, A., Martinez, M., Srinivasa, S.S.: The moped framework: Object recognition and pose estimation for manipulation. The International Journal of Robotics Research **30**(10), 1284–1306 (2011)
8. Dai, N., Yu, X., Fan, Q., Yuan, F., Liu, L., Sun, Y.: Complete denture tooth arrangement technology driven by a reconfigurable rule. PloS one **13**(6), e0198252 (2018)
9. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017)
10. Fisher, M., Ritchie, D., Savva, M., Funkhouser, T., Hanrahan, P.: Example-based synthesis of 3d object arrangements. ACM Transactions on Graphics (TOG) **31**(6), 135 (2012)
11. Gao, L., Yang, J., Wu, T., Yuan, Y.J., Fu, H., Lai, Y.K., Zhang, H.: Sdm-net: Deep generative network for structured deformable mesh. ACM Transactions on Graphics (TOG) **38**(6), 243 (2019)
12. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
13. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: Atlasnet: A papier-m\^ ach\'e approach to learning 3d surface generation. arXiv preprint arXiv:1802.05384 (2018)
14. Guerrero, P., Jeschke, S., Wimmer, M., Wonka, P.: Learning shape placements by example. ACM Transactions on Graphics (TOG) **34**(4), 108 (2015)
15. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Asian conference on computer vision. pp. 548–562. Springer (2012)
16. Hwang, J.J., Azernikov, S., Efros, A.A., Yu, S.X.: Learning beyond human expertise with generative models for dental restorations. arXiv preprint arXiv:1804.00064 (2018)
17. Li, J., Xu, K., Chaudhuri, S., Yumer, E., Zhang, H., Guibas, L.: Grass: Generative recursive autoencoders for shape structures. ACM Transactions on Graphics (TOG) **36**(4), 52 (2017)

18. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: Deepim: Deep iterative matching for 6d pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 683–698 (2018)
19. Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R.: Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493 (2015)
20. Lian, C., Wang, L., Wu, T.H., Liu, M., Durán, F., Ko, C.C., Shen, D.: Meshsnet: Deep multi-scale mesh feature learning for end-to-end tooth labeling on 3d dental surfaces. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 837–845. Springer (2019)
21. Majerowicz, L., Shamir, A., Sheffer, A., Hoos, H.H.: Filling your shelves: Synthesizing diverse style-preserving artifact arrangements. IEEE transactions on visualization and computer graphics **20**(11), 1507–1518 (2013)
22. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
23. Mo, K., Guerrero, P., Yi, L., Su, H., Wonka, P., Mitra, N., Guibas, L.J.: Structurenet: hierarchical graph networks for 3d shape generation. arXiv preprint arXiv:1908.00575 (2019)
24. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. arXiv preprint arXiv:1901.05103 (2019)
25. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4561–4570 (2019)
26. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 918–927 (2018)
27. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017)
28. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems. pp. 5099–5108 (2017)
29. Song, S., Xiao, J.: Sliding shapes for 3d object detection in depth images. In: European conference on computer vision. pp. 634–651. Springer (2014)
30. Song, S., Xiao, J.: Deep sliding shapes for amodal 3d object detection in rgb-d images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 808–816 (2016)
31. Sung, M., Su, H., Kim, V.G., Chaudhuri, S., Guibas, L.: Complementme: Weakly-supervised component suggestions for 3d modeling. ACM Transactions on Graphics (TOG) **36**(6), 226 (2017)
32. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2088–2096 (2017)
33. Tekin, B., Sinha, S.N., Fua, P.: Real-time seamless single shot 6d object pose prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 292–301 (2018)
34. Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., Birchfield, S.: Deep object pose estimation for semantic robotic grasping of household objects. arXiv preprint arXiv:1809.10790 (2018)

35. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3343–3352 (2019)
36. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2642–2651 (2019)
37. Wang, K., Lin, Y.A., Weissmann, B., Savva, M., Chang, A.X., Ritchie, D.: Planit: planning and instantiating indoor scenes with relation graph and spatial prior networks. ACM Transactions on Graphics (TOG) **38**(4), 132 (2019)
38. Wang, K., Savva, M., Chang, A.X., Ritchie, D.: Deep convolutional priors for indoor scene synthesis. ACM Transactions on Graphics (TOG) **37**(4), 70 (2018)
39. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017)
40. Xu, D., Anguelov, D., Jain, A.: Pointfusion: Deep sensor fusion for 3d bounding box estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 244–253 (2018)
41. Xu, X., Liu, C., Zheng, Y.: 3d tooth segmentation and labeling using deep convolutional neural networks. IEEE transactions on visualization and computer graphics **25**(7), 2336–2348 (2018)
42. Yu, L.F., Yeung, S.K., Tang, C.K., Terzopoulos, D., Chan, T.F., Osher, S.: Make it home: automatic optimization of furniture arrangement. ACM Trans. Graph. **30**(4), 86 (2011)
43. Zanjani, F.G., Moin, D.A., Claessen, F., Cherici, T., Parinussa, S., Pourtaherian, A., Zinger, S., et al.: Mask-mcnet: Instance segmentation in 3d point cloud of intra-oral scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 128–136. Springer (2019)
44. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4490–4499 (2018)
45. Zhu, M., Derpanis, K.G., Yang, Y., Brahmbhatt, S., Zhang, M., Phillips, C., Lecce, M., Daniilidis, K.: Single image 3d object detection and pose estimation for grasping. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). pp. 3936–3943. IEEE (2014)