A Revised Approach to Orthodontic Treatment Monitoring From Oralscan Video

Yan Tian, Guotang Jian, Jialei Wang, Hong Chen, Lei Pan, Zhaocheng Xu, Jianyuan Li, and Ruili Wang

***Abstract*—Research on orthodontic treatment monitoring from oralscan video is a new direction in dental digitalization. We designed an approach to reconstruct, segment, and estimate the pose of individual teeth to measure orthodontic treatment. To handle the semantic gap in heterogeneous data on the condition that they are combined linearly, we present a multimedia interaction network (MIN) to combine heterogeneous information in point cloud segmentation by extending the graph attention mechanism. Moreover, a structure-aware quadruple loss is designed to explore the relation between multiple and diverse unmatched points in point cloud registration. The performance of our approach is evaluated on multiple tooth registration datasets, and extensive experiments show that our approach improves the accuracy by a margin of 1.4% in the inlier ratio on the Aoralscan3 dataset when it is compared with prevailing approaches.**

***Index Terms*—Digital dentistry, instance segmentation, point cloud registration, deep learning, computer vision.**

Manuscript received 11 December 2022; revised 25 July 2023, 3 September 2023, and 22 September 2023; accepted 23 September 2023. Date of publication 26 September 2023; date of current version 6 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 61972351 and 62111530300, in part by Zhejiang Province R&D Key Project under Grant 2022C03149, in part by Special Project for Basic Business Expenses of Zhejiang Provincial Colleges and Universities under Grant JRK22003, in part by the General Project Funds from the Health Department of Zhejiang Province under Grant 2021KY480, and in part by the Opening Foundation of State Key Laboratory of Virtual Reality Technology and System of Beihang University under Grant VRLAB2023B02. *(Corresponding authors: Yan Tian; Jianyuan Li.)*

Yan Tian and Guotang Jian are with the School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310018, China (e-mail: tianyan@zjgsu.edu.cn; 1183328120@qq.com).

Jialei Wang is with the Shining3D Tech Company, Ltd., Hangzhou 311200, China (e-mail: wangjialei@shining3d.com).

Hong Chen is with the Department of Stomatology, Zhejiang Hospital, Hangzhou 310013, China (e-mail: hssy81@126.com).

Lei Pan is with the Center for Plastic & Reconstructive Surgery, Department of Plastic & Reconstructive Surgery, Zhejiang Provincial People's Hospital, Hangzhou Medical College, Hangzhou 310014, China (e-mail: prs_dr_panlei@163.com).

Zhaocheng Xu and Ruili Wang are with the School of Mathematical and Computational Sciences, Massey University, Auckland 0632, New Zealand (e-mail: 285064360@qq.com; ruili.wang@massey.ac.nz).

Jianyuan Li is with the School of Computer and Computing Science, Zhejiang University City College, Hangzhou 310015, China (e-mail: lijy@zucc.edu.cn).

Digital Object Identifier 10.1109/JBHI.2023.3319361

I. Introduction

ORTHODONTIC treatment monitoring compares the current orthodontic degree with the schedule to make a decision regarding whether the orthodontic plan should be adaptively adjusted. The traditional method requires the patient to periodically visit a dental clinic, and the status of teeth is manually examined by a dentist. However, it requires an abundance of human resources, and the degree of orthodontic treatment cannot be quantitatively measured.

Recently, some deep learning-based methods [1], [2] have been proposed to automatically measure orthodontic treatment by reconstructing, segmenting, and estimating the pose of individual teeth, as illustrated in Fig. 1(a). Given a mesh that is reconstructed by using RGB images and depth images from an intraoral scanner, geometric data (3D coordinates of points) and visual data (color of points) are independently analyzed and then linearly fused to segment each tooth on the jaw for further tooth registration. However, geometry and visual data lie on different manifolds, and relations among points learned from one field do not match those in the other field; as a result, incompatible point relations from different fields confuse the segmentor. Moreover, in the registration stage, the number of unmatched points is far greater than the number of matched points. Nevertheless, cues on unmatched points are not fully exploited in the learning procedure.

Motivated by the recent development of graph attention [3], we argue that heterogeneous features can be represented and combined by designing multiple graphs to enhance the discriminant capacity. For the challenge of unmatched points in point cloud registration, using the relation between unmatched points to improve the model is also investigated.

In this article, a revised framework is proposed to measure orthodontic treatment, including 3D reconstruction, tooth instance segmentation, and point cloud-based tooth registration. The framework of the proposed approach is illustrated in Fig. 2. Given the oralscan RGB-D video, the 3D jaw model is first reconstructed by an on-the-shelf simultaneous localization and mapping (SLAM) method [4]. Then, tooth instance segmentation based on SoftGroup [5] is performed, in which a multimedia interaction network (MIN) is designed to use local cues within a single graph and across different graphs, which is illustrated in Fig. 1(b). After that, a quadruple loss exploring structure information is proposed in tooth registration, in which multiple and diverse

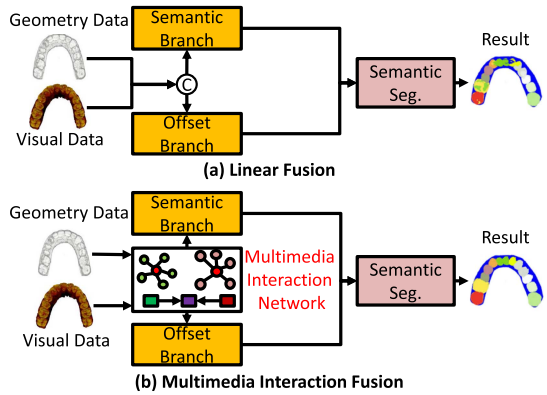


Fig. 1. Illustration of the challenge of heterogeneous data fusion encountered in tooth instance segmentation. (a) Geometry data and visual data are independently analyzed and then linearly fused. (b) Our approach uses the multimedia interaction network to learn the point relation on geometry and visual graphs for fusion.

unmatched points are simultaneously exploited to improve the robustness of tooth pose inference. Finally, the predicted tooth pose is compared to the target pose to check whether the orthodontic treatment is consistent with the schedule.

The contribution is highlighted by the following:

- We propose a deep learning-based framework to measure orthodontic treatment, which reconstructs, segments, and estimates the pose of individual teeth.
- We present a novel multimedia interaction network in point cloud segmentation that extends graph attention to effectively combine heterogeneous data by propagating information though different graphs.
- We design a new loss function in point cloud registration to explore the relation between the unmatched points by measuring diversity among negative samples.

Experiments on tooth segmentation datasets and tooth registration datasets verify that our approach improves accuracy by a margin of 1.4% in the inlier ratio when it is compared with state-of-the-art approaches.

II. Related Work

We briefly introduce the recent literature on orthodontic treatment monitoring, point cloud instance segmentation, and point cloud registration.

A. Orthodontic Treatment Monitoring

Orthodontic treatment monitoring is a new topic in dental digitalization, and only a few studies have focused on it [6]. Some work monitored the status of the periodontal ligament by using micro-Raman spectroscopy [7] or cone-beam computed tomography (CBCT); nevertheless, the specific source or expensive equipment limited the scope of use. Therefore, oralscan video that is captured from a smartphone is used to predict the maxillary and mandibular arches [8]. However, the need for a maxillary expander decreases the comfortability and simplicity of a patient. Recent works detected [9] or segmented [1], [2] individual teeth from the intraoral video and predicted the corresponding 6D pose by using a deep learning method. Nevertheless, these approaches infer tooth pose by comparing the posed rendered image and the observed image, neglecting the exploration of geometry knowledge in 3D space. For instance, the pose of individual teeth can be regressed by a multilayer perceptron (MLP) layer [10] or by detecting and matching tooth landmarks on a 3D model [11], [12].

B. Point Cloud-Based Instance Segmentation

Point cloud-based instance segmentation has become a research hotspot owing to the development of geometry deep learning and dataset construction [13], [14]. It can be generally classified into 2 categories: proposal-free and proposal-based methods. The comparison of different categories of methods is conducted in recent work [15].

Proposal-based methods convert the original instance segmentation task into instance localization and mask inference subtasks. The generative shape proposal network (GSPN) [16] uses simulation to generate proposals, while Mask-MCNet [17] and 3D-BoNet [18] directly infer bounding boxes from global features. However, these methods generate redundant proposals and are usually computationally expensive.

Proposal-free methods convert the instance segmentation task into a grouping procedure after center prediction. A milestone is pointGroup [19] and 3D-MPA [20], where the semantic segmentation and the point offset toward the center of an instance are simultaneously explored. Based on this framework, some works, such as SSTNet [21] and HAIS [22], adjust the proposal size by using the structure information. Specifically, DyCo3D [23] and instance kernels [24] design dynamic kernels for unordered and unstructured point cloud data to improve the localization and representation capability, while Kd-Network [25] extracts and aggregates features according to the subdivisions of the point clouds on Kd-trees. TSegNet [26] employs a cascade network structure to improve the predicted mask by combining prediction confidence. However, the segments lack objectness without explicitly recognizing boundaries. Recently, to handle the low overlap between a real instance and a predicted instance caused by hard one-hot semantic predictions, SoftGroup [5] employed soft semantic scores for grouping. Nevertheless, the effectiveness is restricted when the semantic gap occurs because of linear combination in heterogeneous data. TSGCNet [27] is a pioneering work using a graph-based method for data fusion; however, the original approach is designed for the 3D position and normal of point clouds, and the extension to other types of data is unexplored.

C. Point Cloud Registration

Point cloud registration aligns two point clouds and estimates the relative pose between them. Generally, there are four kinds of point cloud registration methods:

The classical solution detects and extracts feature descriptors for salient keypoints and then obtains putative correspondences via feature matching. After that, the relative pose between two point clouds is calculated by using random sampling consensus

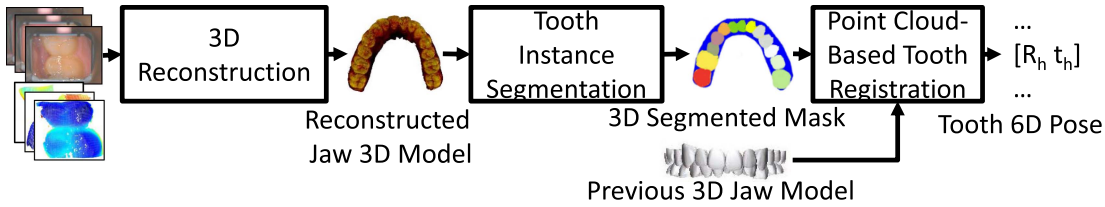


Fig. 2. Illustration of the proposed framework, including SLAM-based 3D reconstruction, tooth instance segmentation, and point cloud-based tooth registration. The input is RGB images and depth images from an intraoral scanner. R_h and t_h represent the rotation and transition of tooth h.

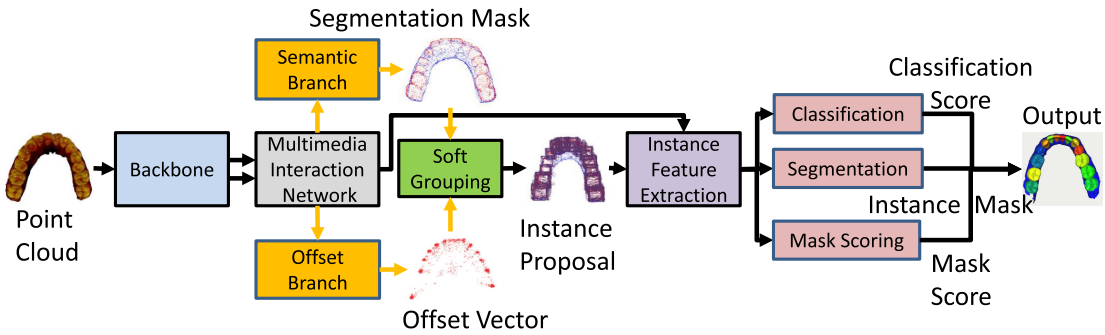


Fig. 3. Illustration of the proposed approach in tooth instance segmentation. A multimedia interaction network is proposed to alleviate the semantic gap. Some yellow arrows are employed to avoid misunderstandings caused by the intersection of arrows.

(RANSAC) and solving the Procrustes problem. Scan completion and pairwise matching are employed for low overlap situations [28]. Nevertheless, these methods depend highly on keypoint localization, and inaccurate keypoints deteriorate pose inference.

Recent methods based on deep CNN estimate dense correspondences with confidence scores between two point clouds and select the top-k confident correspondences for rigid transformation estimation [29], [30], [31]. These learning-based methods can be faster and more robust than classical methods. However, the number of unmatched points is far greater than the number of matched points, and the information on unmatched points is not fully exploited.

Some approaches regard point cloud registration as a regression task, and directly learn rigid transformation from the point cloud pair to position in relative pose space by using T-net in PointNet [32] or learn the nonrigid mapping by dividing objects into several rigid parts and inferring the relative pose in each part [33]. Nevertheless, these methods achieve weak performance owing to the neglect of geometric knowledge.

The last kind of method employs the trial-and-evaluation strategy, iteratively refining the pose of the source point cloud and matching the transformed source point cloud to the target point cloud [34], [35], [36]. Another pipeline employs deep reinforcement learning to select transform action according to the policy network and reward values, for instance, the learning-based iterative closest point (ICP) [37], ReAgent [38], and the cross-entropy method (CEM) [39]. Iteration methods have the advantage of generalization capacity. Nevertheless, multiple iterations are needed for optimum solution determination.

III. Our Approach

We designed a deep learning-based approach for orthodontic treatment monitoring, including SLAM and tooth instance segmentation and registration, and details are shown in Fig. 2.

A. Tooth Instance Segmentation

Our approach is based on SoftGroup [5], which follows the multitask framework [40] and introduces soft semantic scores for grouping. A multimedia feature interaction network is proposed to alleviate the semantic gap between visual and geometric features, which is illustrated in Fig. 3.

1) Review of SoftGroup: Given N points, including color and coordinate values, a 3D U-Net network based on submanifold sparse convolution (SSC) is employed as the backbone to extract point features. Then, point features go through two parallel branches to simultaneously predict semantic scores $\mathbf{S} \in \mathbb{R}^{N \times N_c}$ and offset vectors $\mathbf{O} \in \mathbb{R}^{N \times 3}$, which are optimized by cross-entropy loss and L1-norm regression loss, respectively, where $N_c = 2$ separates tooth and gum regions in the jaw. After that, offset vectors $\mathbf{O}$ are added to point coordinates to obtain corresponding instance centers. Point subsets belonging to each class are obtained by scanning semantic scores and comparing them with a fixed threshold τ. Instance proposals are generated by creating links between points with geometric distances smaller than a threshold b. Finally, an instance feature extraction network is employed to extract and enhance semantic features, and then, a multitask learning subnetwork is implemented, inferring mask scores, instance masks, and classification scores, which are optimized by the L2-norm regression, binary cross-entropy, and cross-entropy losses, respectively.

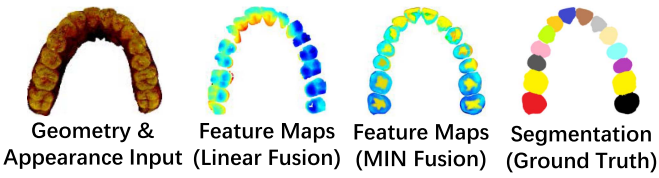


Fig. 4. Comparison of fused feature maps using different fusion methods. The input point cloud comprises the 3D coordinates and color of each point. Activation values of point features from low to high are represented by blue, green, yellow, and red. Linear fusion of geometry and appearance generates indiscriminant point features (weak in locating tooth center), while MIN fusion improves the discriminant capacity (good at locating tooth center).

The advantage of SoftGroup is that it employs soft semantic scores for grouping, which effectively handles the low overlap between the predicted instance and the real instance caused by the hard one-hot semantic predictions.

2) Multimedia Interaction Network: Geometric and visual data describe the 3D model from multiple aspects. They lie on different manifolds, and relations among points learned from one field do not match those in the other field. If these incompatible relations from different fields are directly fused, conflict information may puzzle the segmentor to make any correct decision. However, this principle is ignored in SoftGroup, which uses a linear combination to combine heterogeneous data with different patterns. An example is illustrated in Fig. 4. Activation values of point features from low to high are represented by blue, green, yellow, and red. Low activation values in the tooth center demonstrate that features obtained by linear fusion lead to an attenuation in discrimination.

Therefore, the MIN is proposed to improve context feature extraction in instance segmentation using a graph-based method, which is illustrated in Fig. 5. As visual and geometric data lie on different manifolds, two graphs are constructed to represent the point relation in visual and geometric fields, respectively. Local information is propagated across different graphs by using graph attention to interact with heterogeneous features by mapping them into a common space.

We assume that the color and 3D coordinates of points can be used to construct two graphs $G^1(\mathbf{V}^{1,l}, \mathbf{E}^{1,l})$ and $G^2(\mathbf{V}^{2,l}, \mathbf{E}^{2,l})$ in layer l via K-nearest neighbors, where matrices $\mathbf{V}^{1,l} = \{\mathbf{m}_1^{1,l}, \mathbf{m}_2^{1,l}, \ldots, \mathbf{m}_N^{1,l}\}$ and $\mathbf{E}^{1,l} \subseteq |\mathbf{V}^{1,l}| \times |\mathbf{V}^{1,l}|$ indicate sets of nodes and edges constructed by using only the point color information. The symbol $\mathbf{m}_i^{1,l}$ is node i in layer $l \in \{1, 2, \ldots, L_{\max}\}$; its corresponding feature is $\mathbf{f}_i^{1,l}$. The symbol $\mathbf{m}_{ij}^1$ is the edge of nodes i and j in layer l, and its corresponding feature is $\mathbf{f}_{ij}^{1,l}$. Vectors $\mathbf{V}^{2,l}$, $\mathbf{E}^{2,l}$, $\mathbf{m}_i^{2,l}$, $\mathbf{m}_{ij}^{2,l}$, $\mathbf{f}_i^{2,l}$, and $\mathbf{f}_{ij}^{2,l}$ indicate similar meanings but only use point coordination.

Local features in the same graph and across different graphs are used to update features via multilayer perceptron (MLP)

$$\hat{\mathbf{f}}_i^{1\to1,l} = MLP_u^{1\to1,l}([\mathbf{f}_i^{1,l}, \mathbf{f}_{ij}^{1,l}]), \tag{1}$$

$$\hat{\mathbf{f}}_i^{2\to1,l} = MLP_u^{2\to1,l}([\mathbf{f}_i^{1,l}, \mathbf{f}_{ij}^{2,l}]), \tag{2}$$

where [.] is the feature concatenation. After that, local features are aggregated by using graph attention. Let i and j represent

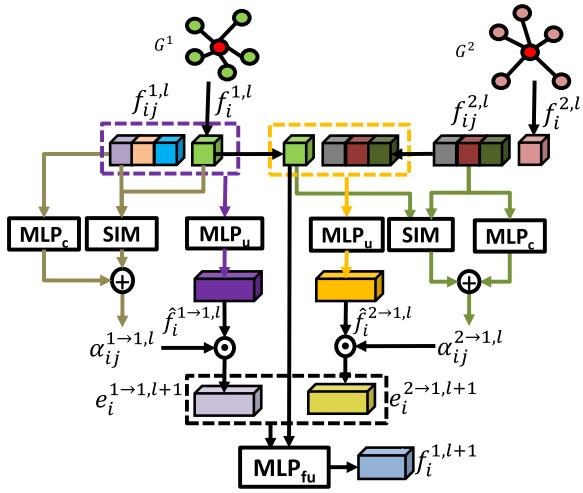


Fig. 5. Illustration of the heterogeneous feature interaction module. Note that only messages from graph G^1 and G^2 to graph G^1 are visualized for simplicity. The orange and purple arrows represent the node updates. The golden and green arrows are interactions within the same graph and across different graphs. The dot black box indicates the feature concatenation. 'SIM' means similarity measure and 'MLP' denotes multilayer perception.

node indices in graph 1 and graph 2, respectively. Then, the relation between these two points is expressed by node and edge features $sim(\mathbf{f}_i^{1,l}, \mathbf{f}_{ij}^{2,l}) = (\mathbf{W}^{1,l}\mathbf{f}_i^{1,l})^T(\mathbf{W}^{2,l}\mathbf{f}_{ij}^{2,l})$. The multi-dim attention weights are obtained by

$$\alpha_{ij}^{1\to1,l} = sim(\mathbf{f}_i^{1,l}, \mathbf{f}_{ij}^{1,l}) + MLP_c^{1,l}(\mathbf{f}_{ij}^{1,l}), \tag{3}$$

$$\alpha_{ij}^{2\to1,l} = sim(\mathbf{f}_i^{1,l}, \mathbf{f}_{ij}^{2,l}) + MLP_c^{2,l}(\mathbf{f}_{ij}^{2,l}), \tag{4}$$

where $MLP_c^{1,l}(.)$ and $MLP_c^{2,l}(.)$ score the importance of each dimension of features in graph 1 and graph 2, respectively. Then, context features $\mathbf{e}_i^{1\to1,l+1}$ in the same graph and $\mathbf{e}_i^{2\to1,l+1}$ across the graph are calculated by

$$\mathbf{e}_i^{1\to1,l+1} = \sum_{m_{ij}} \alpha_{ij}^{1\to1,l} \odot \hat{\mathbf{f}}_i^{1\to1,l}, \tag{5}$$

$$\mathbf{e}_i^{2\to1,l+1} = \sum_{m_{ij}} \alpha_{ij}^{2\to1,l} \odot \hat{\mathbf{f}}_i^{2\to1,l}, \tag{6}$$

where $\odot$ indicates the elementwise production. Finally, information across graphs is ensembled by a mapping $MLP_{fu}^{1,l+1}(.)$,

$$\mathbf{f}_i^{1,l+1} = MLP_{fu}^{1,l+1}([\mathbf{e}_i^{1\to1,l+1}, \mathbf{e}_i^{2\to1,l+1}], \mathbf{f}_i^{1,l}). \tag{7}$$

Compared with other interaction approaches, the proposed approach has several advantages: 1) The relation between visual and geometry is progressively explored to alleviate the semantic gap in heterogeneous data. 2) It is efficient to employ graph attention to aggregate the local features in and across graphs to improve the capability in geometry prediction. 3) The proposed relation method can be used as a general module for heterogeneous data fusion.

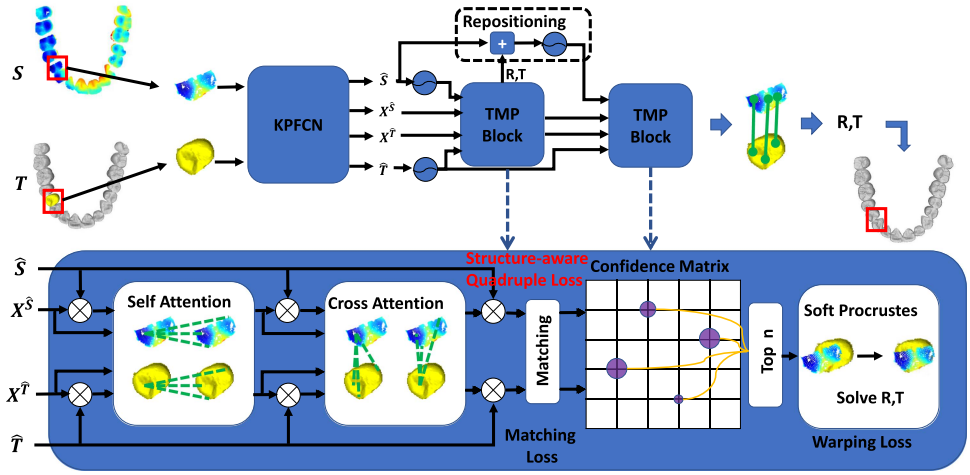


Fig. 6. Illustration of the point cloud registration module. Lepard is the baseline, and structure-aware quadruple loss is proposed. Given the source point cloud $\mathbf{S}$ and target point cloud $\mathbf{T}$, the KPFCN backbone samples the point cloud $\hat{\mathbf{S}}, \hat{\mathbf{T}}$ and extracts geometric features $\mathbf{x}^{\hat{\mathbf{S}}}, \mathbf{x}^{\hat{\mathbf{T}}}$. The position codes and geometry features are then processed by several transform-matching-Procrustes (TMP) blocks, each including a transformer layer with self and cross attentions, a differentiable matching layer, and a soft Procrustes layer to estimate the rigid fitting. The rigid fitting obtained from the previous TMP block is used to update the source point cloud positions and features in the next TMP block.

B. Point Cloud-Based Tooth Registration

Our framework is illustrated in Fig. 6, where Lepard [29] is employed as the baseline to handle repetitive geometry patterns. The structure-aware quadruple loss is proposed to explore additional hard negative examples for matching learning.

1) Review of Lepard: Given the source point cloud $\mathbf{S} \in \mathbb{R}^{n\times 3}$ and target point cloud $\mathbf{T} \in \mathbb{R}^{m\times 3}$, where n and m are the point numbers in each point cloud. The kernel point fully convolutional network (KPFCN) [41] backbone $\Phi(.)$ extracts point clouds $\hat{\mathbf{S}} = \Phi(\mathbf{S}) \in \mathbb{R}^{\hat{n}\times 3}$, $\hat{\mathbf{T}} = \Phi(\mathbf{T}) \in \mathbb{R}^{\hat{m}\times 3}$ and local geometry features $\mathbf{x}^{\hat{\mathbf{S}}} \in \mathbb{R}^{\hat{n}\times d}$, $\mathbf{x}^{\hat{\mathbf{T}}} \in \mathbb{R}^{\hat{m}\times d}$, where $\hat{n}$ and $\hat{m}$ are the number of points $\hat{\mathbf{S}}$ and $\hat{\mathbf{T}}$ and d is the dimension of features. Then, rotary positional encoding [42] is employed to enhance local features with position embedding. After that, several transform-matching-Procrustes (TMP) blocks are employed to infer the rigid transformation, each including a transformer layer, a differentiable matching layer, and a soft Procrustes layer. The transformer layer uses self-attention [43], [44] to explore the global context features and cross-attention to further interact with source and target point clouds. The matching layer obtains the confidence matrix $\mathbf{C}$ between the source and target point clouds. In the soft Procrustes layer, correspondence points with top scores in confidence matrix $\mathbf{C}$ are selected and employed to infer the rigid transformation by using the perspective-n-point (PnP) method. The total loss in Lepard optimizes the warping loss for point cloud warping and matching loss over the confidence matrix $\mathbf{C}$.

2) Structure-Aware Quadruple Loss: In point cloud registration, the number of positive samples (matching pairs between different point clouds) is much lower than the number of negative samples (nonmatching pairs between different point clouds). However, the imbalance between negative and positive samples is ignored in Lepard. Moreover, it is unnecessary to use all points to take part in the learning procedure to prevent outliers and long training times.

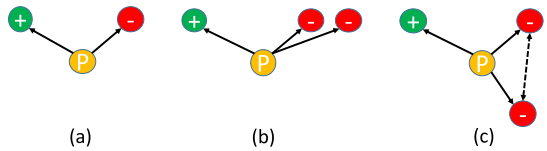


Fig. 7. Illustrations of different contrastive loss functions. (a) Cycle loss. (b) Quadruple loss. (c) Structure-aware quadruple loss. '+' represents the positive point, '−' represents the negative points, and 'P' is the anchor point. The dotted arrow represents the structural relation between negative points.

Motivated by structure knowledge in negative examples, we propose extending the cycle loss to structure-aware quadruple loss, exploring two or more hard negative examples for matching learning, which is illustrated in Fig. 7. Additional knowledge from hard negative samples constrains the matching function to be smooth on the manifold and partially avoids the chance of overfitting with limited resource consumption.

We assume that confidence $\mathbf{C}(i, j)$ measures the matching degree between point i with features $\mathbf{x}_i^{\hat{\mathbf{S}}}$ in the source point cloud and point j with features $\mathbf{x}_j^{\hat{\mathbf{T}}}$ in the target point cloud. Hard sample mining can be adopted to distinguish the most difficult negative examples for training. Then, the cycle loss is calculated as follows:

$$L_f(i,j) = -\mathbb{E}\left[\log \frac{\mathbf{C}(i,p_i)}{\mathbf{C}(i,p_i) + \max_j \mathbf{C}(i,n_i^j)}\right], \quad (8)$$

where p_i and n_i^j indicate points indices in the target point cloud marching (positive) and not matching (negative) point i in the source point cloud.

To fully exploit the knowledge in negative examples, knowledge from multiple hard negative examples is supplemented to

constrain the matching relation. Note that features of different hard samples may be neighbors in the confidence space. Therefore, we design a mechanism to efficiently increase the diversity in negative samples. Assume that $n_i^1 = \max_j \mathbf{C}(i, n_i^j)$ is the hardest sample selected in the proposal set. The second negative sample is determined by the constraint that the relation to point i in the source point cloud is strong, but the relation to the hardest sample n_i^1 is weak:

$$n_i^2 = arg\max_j \{\mathbf{C}(i, n_i^j) - \mathbf{C}(n_i^1, n_i^j)\}. \tag{9}$$

As a result, the feature loss in TMP block bt becomes a function of one positive sample and two negative samples

$$L_f^{bt}(i,j) = -\mathbb{E}\left[\log \frac{\mathbf{C}(i,p_i)}{\mathbf{C}(i,p_i) + \mathbf{C}(i,n_i^1) + \mathbf{C}(i,n_i^2)}\right]. \tag{10}$$

The total loss is a linear combination of the matching loss, warping loss, and structure-aware quadruple loss as follows:

$$L_{total} = \sum_{bt}(\lambda_m L_m^{bt} + \lambda_w L_w^{bt} + L_f^{bt}), \tag{11}$$

where λ_m and λ_w balance the effects of matching loss and warping loss, respectively.

Our structure-aware feature loss has some advantages: 1) The structural knowledge in the point space (especially negative points) is exploited to guide the optimization process; 2) Diverse negative samples revise the confidence matrix from multiple aspects and potentially improve the convergence; 3) Multiple hard samples can be selected to increase the flexibility of the approach.

IV. Experiments and Results

We evaluate the effectiveness of our approach and compare it with other approaches on publicly available datasets.

A. Materials and Methods

1) Dataset: We evaluate the proposed instance segmentation approach on the Shining3D tooth segmentation dataset [45] and the Aoralscan3 tooth segmentation dataset [46]. We evaluate the proposed tooth point cloud registration approach on the Shining3D tooth pose dataset [47] and the Aoralscan3 tooth registration dataset [48]. Jaw models are generated from hospital patients by oral scanning. The ground truth of the relative pose of each tooth is generated by adding random jittering to the tooth models. The training, validation, and testing sets in the Shining3D tooth pose dataset contain 1,689, 150, and 150 samples, respectively. The constructed Aoralscan3 tooth registration dataset includes 1,667 samples for training, 156 samples for validation, and 176 samples for testing.

2) Evaluation Criteria: In tooth instance segmentation, the mean average precision (mAP) at intersection-over-union(IoU) thresholds of 25% and 50% are used as the evaluation criteria. In tooth point cloud registration, the inlier ratio (IR), registration recall (RR), and mean absolute error (MAE) are used as the evaluation criteria. IR measures the fraction of correct correspondences (threshold 0.1 mm) among the putative matches [49], RR measures the fraction of correctly registered point cloud pairs [49], and $\mathrm{MAE_R}$/$\mathrm{MAE_T}$ measure the deviation between the predicted tooth rotation (3D)/transition (3D) and the corresponding ground truth. The units of $\mathrm{MAE_R}$ and $\mathrm{MAE_T}$ are degrees and millimeters, respectively.

TABLE I
Experimental Results With Different Backbones on the Shining3D Tooth Segmentation Dataset

Method	Shining3D Tooth Segmentation Dataset	
	mAP@50	Time (ms)
PointNet [32]	44.6	906
PointNet++ [51]	49.8	982
3D U-Net [52]	**59.5**	874
SSC+3D U-Net [53]	59.4	**233**

The values in bold represent the best results among different approaches.

3) Implementation Details: We conduct experiments on a workstation with 2 Intel i7-4790 3.6 GHz CPUs, 64 GB of memory, and 4 NVIDIA RTX 3090Ti GPUs. Different approaches are implemented based on PyTorch [50] to verify and compare the performance.

In tooth instance segmentation, the number of points on the jaw model is downsampled to 100 k by using random sampling. Data augmentation, such as vertical/horizontal flipping, translation, rotation, and scaling, is employed to enlarge the dataset. The SSC-based 3D U-Net is used as the backbone, including 4 layers, each downsampling the width and height of feature sizes and doubling the channel dimension. The number of interaction layers is selected as $L = 4$ according to the performance on the evaluation set. The score threshold τ is set to 0.4, and the grouping bandwidth b is set to 0.2 mm. The weights are updated by Adam, with a weight decay of 3×10^{-3} and a momentum of 0.9. The learning rate is adjusted to 4.0×10^{-3} for the first 50 k iterations and scheduled by cosine annealing for the following 70 k iterations. Each minibatch contains 4 samples.

In tooth point cloud registration, the number of input points is downsampled to 4 k by using the RS. Data augmentation methods such as rotation and translation are employed to enlarge the dataset. The network is trained with Adam for 100 epochs and the weight decay is 10^{-4}. The batch size is 4. The learning rate is adjusted to 2.0×10^{-2} and decays exponentially by 5×10^{-3} in each epoch. We set 0.1 mm as the threshold of the matching radius.

B. Ablation Study

Extensive experiments are conducted to verify the effect of several contributions in the proposed approach on the Shining3D tooth segmentation or tooth pose dataset.

1) Tooth Instance Segmentation: Backbone: We select the backbone according to its effectiveness, and the comparison is reported in Table I. Note that values of mAP are multiplied by 100. All approaches share the same setting except for the backbone. Although PointNet and PointNet++ have been widely used in point cloud analysis, operating on unordered point sets limits the ability to capture fine-grained details. SSC-based 3D U-Net is chosen as the backbone when both efficiency and effectiveness are taken into consideration.

Parameters: Parameters are determined by grid searching, which is illustrated in Fig. 8. Parameter values are listed on the x-axis, and the performance expressed by mAP at an IoU

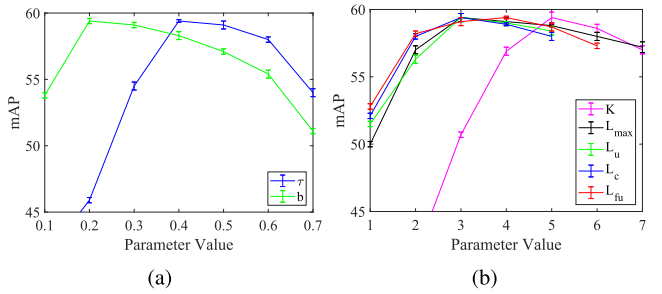


Fig. 8. Parameter selection for the Shining3D tooth segmentation dataset. Quantitative analysis of (a) score threshold τ and grouping bandwidth b, (b) number of nearest neighbors K, number of the network $L_{\max}$, layer number L_u, L_c, and L_{fu}.

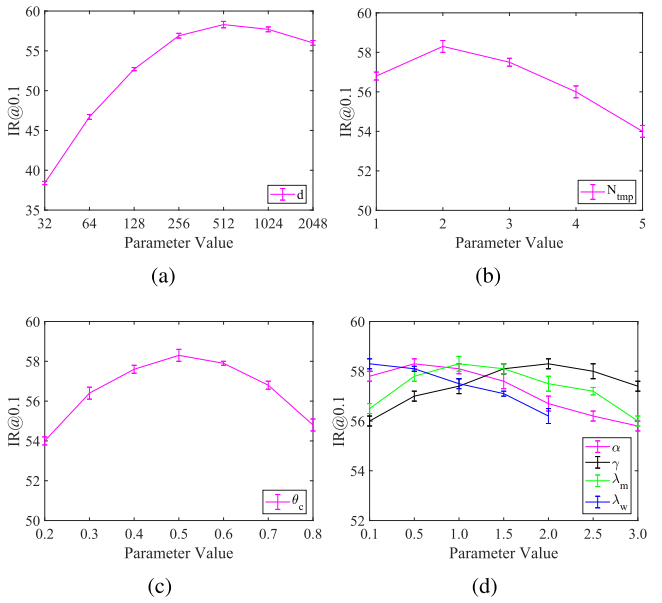


Fig. 9. Parameter selection for the Shining3D tooth pose dataset. Quantitative analysis of (a) the dimension d of feature maps, (b) the TMP block number N_{tmp}, (c) threshold θ_c, and (d) weights α, γ, λ_m, and λ_w in the loss function.

threshold of 50% is listed on the y-axis. The score threshold $\tau = 0.4$, grouping bandwidth $b = 0.2$ mm, number of nearest neighbors $K = 5$, number of networks $L_{\max} = 3$, number of layers in MLP_u ($L_u = 3$), number of layers in MLP_c ($L_c = 3$), and number of layers in MLP_{fu} ($L_{fu} = 4$) are selected in the instance segmentation module.

2) Point Cloud-Based Tooth Registration: Parameters: The dimension number $d = 512$, TMP block number $N_{tmp} = 2$, and confidence threshold $\theta_c = 0.5$ are determined by grid searching. The performance comparison of different parameters is illustrated in Fig. 9(a)–(c). Low dimensions have limited representation capability, but unnecessarily high dimensions easily induce overfitting.

Loss Function: Weights α and γ in the matching loss and weights λ_m and λ_w to combine corresponding loss functions are evaluated to select optimum values to control the effect of various factors. The results of the evaluation experiments are illustrated in Fig. 9(d). Only weights outperforming other settings are used in the remaining experiments.

TABLE II
EFFECTIVENESS COMPARISON AMONG IMPORTANT MODULES ON THE AORALSCAN3 TOOTH REGISTRATION DATASET AND SHINING3D TOOTH POSE DATASET. THE SYMBOLS 'MIN' AND 'SQL' INDICATE THE MULTIMEDIA INTERACTION NETWORK AND STRUCTURE-AWARE QUADRUPLE LOSS, RESPECTIVELY

Configurations			Aoralscan3		Shining3D	
Baseline	+MIN	+SQL	IR@0.1	RR@0.1	IR@0.1	RR@0.1
✓			55.7	97.2	55.4	96.8
✓	✓		57.6	99.0	57.2	98.7
✓		✓	57.5	98.8	57.0	98.3
✓	✓	✓	**58.6**	**100.0**	**58.3**	**99.4**

The values in bold represent the best results among different approaches.

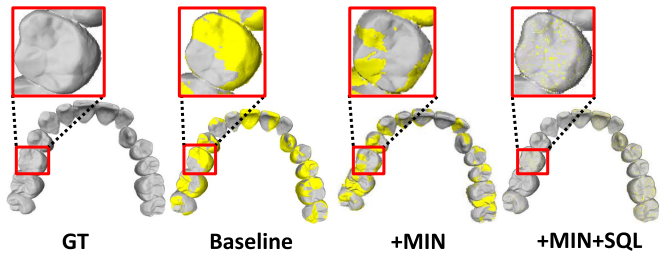


Fig. 10. Evaluations of different contributions on the Shining3D tooth pose dataset. The symbols 'GT', 'MIN', and 'SQL' represent the ground truth, multimedia interaction network, and structure-aware quadruple loss, respectively. Red boxes highlight differences in results obtained by the new module. Yellow regions represent unmatched regions between the warped source and target point cloud.

Effectiveness: We evaluate the effectiveness of the multimedia interaction network and structure-aware quadruple loss in point cloud-based tooth registration. Comparisons of IR and RR at the threshold of 0.1 mm are reported in Table II. Note that values of IR and RR are multiplied by 100, and upper values are better. The baseline uses SoftGroup [5] for instance segmentation and Lepard [29] for tooth registration. When the multimedia interaction network is employed in SoftGroup, accurate segmented masks help to improve the robustness in point cloud registration; that is, the IR increases by 1.8% on the Shining3D tooth pose dataset. Further IR improvement (1.1%) is achieved when the structure-aware quadruple loss is employed in Lepard. The outputs of different contributions are shown in Fig 10. Yellow regions represent unmatching regions between the warped source and target point cloud (fewer yellow regions are better). Multimedia interaction considers multiple cues to infer proper tooth shapes, improving pose inference by avoiding the local optimum. Structural knowledge in confidence space enhances the capability to match correspondence points by simultaneously exploring associations between positive and negative points and relations among negative points.

Tooth Group: We compare the accuracy of point cloud-based tooth registration on the Shining3D tooth pose dataset by dividing teeth into different groups, for example, incisors, canines, premolars, and molars. The experimental results are reported in Table III. Incisors obtain the best accuracy as more frames are captured in the 3D reconstruction stage to jointly optimize the shape of the tooth. Molars are weak in accuracy partially because the oralscan device cannot capture part of the molars, and incomplete shapes deteriorate the results.

TABLE III
EFFECTIVENESS COMPARISON AMONG IMPORTANT MODULES ON THE SHINING3D TOOTH POSE DATASET

Measure	Tooth Group				
	Incisors	Canines	Premolars	Molars	Mean
IR@0.1	**58.5**	58.4	58.2	58.0	58.3
RR@0.1	**99.6**	99.5	99.3	99.1	99.4

The values in bold represent the best results among different approaches.

TABLE IV
EXPERIMENTAL RESULTS ON THE TEST SET OF THE AORALSCAN3 TOOTH SEGMENTATION DATASET AND SHINING3D TOOTH SEGMENTATION DATASET

Method	Aoralscan3		Shining3D	
	mAP@50	mAP@25	mAP@50	mAP@25
3D-MPA [20]*	50.4	69.6	49.5	68.7
PointGroup [19]	50.7	70.1	50.2	69.2
OccuSeg [54]*	51.2	70.7	50.7	69.6
DyCo3D [23]	52.3	71.2	51.3	69.8
PES [55]	53.2	71.6	52.2	70.1
HAIS [22]	54.4	72.0	52.9	70.5
SSTNet [21]	55.2	72.2	53.7	71.0
LOA [45]*	56.5	73.8	55.0	72.4
DKNet [24]	57.7	74.6	56.3	73.2
SoftGroup [5]	58.6	75.9	57.8	74.7
Ours	**60.5**	**77.1**	**59.4**	**76.2**

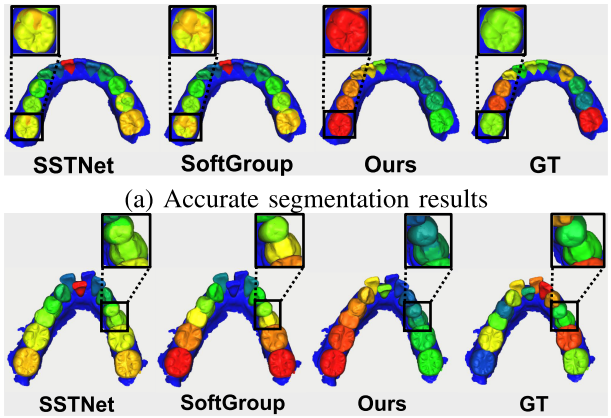
The values in bold represent the best results among different approaches.

C. Evaluation of Tooth Instance Segmentation

We report the comparison obtained on the Shining3D tooth segmentation dataset and Aoralscan3 tooth segmentation dataset in Table IV. Codes are obtained from original papers if codes are released by authors; otherwise, we reimplemented them. '*' denotes the corresponding approach is reimplemented. Multitask learning (semantic segmentation and offset prediction in parallel branches) is a popular framework in point cloud instance segmentation. Our approach improves the mAP@50 by 1.9% on the Aoralscan3 dataset and 1.6% on the Shining3D dataset when it is compared with SoftGroup [5], which verifies that multimedia interaction improves the semantic gap in heterogeneous data.

Some results obtained using SSTNet, SoftGroup, and our approach are illustrated in Fig. 11. Examples of accurate results are illustrated in Fig. 11(a). Points of the second molar are error prone in SSTNet and SoftGroup because of conflicts between visual and geometric data. More reasonable outputs are obtained in our approach in part because discrepancies in heterogeneous data are explicitly represented and propagated to each other on graphs, employing an attention mechanism to measure discrepancy.

Examples of failure cases in tooth instance segmentation are illustrated in Fig. 11(b). Although the semantic gap is alleviated in our approach, the segmentation quality is partially affected by the degree of teeth crowding. For instance, SSTNet obtains poor result in the 1st premolar and our approach obtains inaccurate result in the 2nd premolar, owing to the overlap between neighboring teeth, which distorts the geometric features of normal teeth. A potential solution is to use the geometric information of the tooth root as guidance to discover the overlap between teeth and rescue the segmentation results.



(a) Accurate segmentation results

(b) Inaccurate segmentation results

Fig. 11. Illustration of the results of SSTNet, SoftGroup, and our approach obtained on the Shining3D tooth segmentation dataset. (a) Accurate segmentation results. (b) Inaccurate segmentation results. Different teeth are rendered by different colors. Black boxes highlight discrepancies produced by different approaches.

D. Evaluation of Point Cloud-Based Tooth Registration

We report the results obtained on the Shining3D tooth pose dataset and Aoralscan3 tooth registration dataset in Table V. Recent data driven methods, including methods based on iteration [35], sparse correspondence [56], [57] and dense correspondence [29], [30], [31], obtain an obvious improvement over traditional geometric methods such as ICP. For instance, GeoTransformer obtains an IR of approximately 57.2% at the threshold of 0.1 mm in both the Aoralscan3 tooth registration dataset and the Shining3D tooth pose dataset. Dense correspondence shows additional benefits when compared to sparse correspondence in matching thanks to the robustness of occlusion handling. Our approach improves the IR by 1.4% on the Aoralscan3 dataset and 1.2% on the Shining3D dataset when it is compared with GeoTransformer [31], demonstrating that diversity negative sample mining helps to improve the confidence in the matching stage by exploring structure knowledge in confidence space.

Some examples of comparisons between our approach and other point cloud registration methods are illustrated in Fig. 12. Columns from left to right indicate the source point cloud, the target point cloud, and the estimated results of REGTR, Lepard, and our approach. Fig. 12(a) shows that our approach improves effectiveness because of the use of the structure information of negative points in confidence space.

From the inaccurate cases in Fig. 12(b), it can be found that 1) Although it is effective in terms of the structure-aware loss function, the proposed approach is sensitive to the space between teeth, especially gaps between molars and premolars. 2) The proposed approach adopts cases in which teeth are arranged with low gaps between each other.

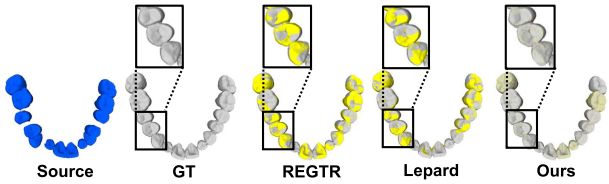
E. Discussion

Orthodontic treatment monitoring is a novel and important topic in digital dentistry. The technical details of this topic are limited. Therefore, we design a framework to reconstruct,

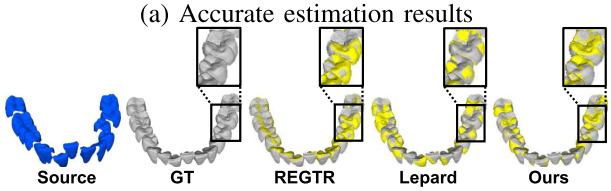
TABLE V
EXPERIMENTAL RESULTS ON THE TEST SET OF THE AORALSCAN3 TOOTH REGISTRATION DATASET AND THE SHINING3D TOOTH POSE DATASET

Method	Aoralscan3				Shining3D			
	IR@0.1	RR@0.1	MAE_R	MAE_T	IR@0.1	RR@0.1	MAE_R	MAE_T
ICP	54.33	94.81	1.75	0.20	53.92	94.31	1.77	0.21
YOHO [56]	55.62	96.35	1.57	0.18	55.12	96.01	1.64	0.19
SGP [35]	55.91	96.65	1.44	0.17	55.63	96.14	1.56	0.18
REGTR [30]	56.25	97.04	1.36	0.16	55.81	96.64	1.47	0.17
Pointdsc [57]	56.43	97.50	1.26	0.14	56.01	97.12	1.33	0.15
Lepard [29]	56.91	98.44	1.07	0.12	56.62	98.22	1.10	0.14
GeoTransformer [31]	57.22	98.51	0.92	0.11	57.13	98.31	0.94	0.13
Ours	**58.62**	**100.00**	**0.77**	**0.09**	**58.33**	**99.43**	**0.78**	**0.10**

The values in bold represent the best results among different approaches.



(a) Accurate estimation results



(b) Inaccurate estimation results

Fig. 12. Illustration of estimated results obtained on the Shining3D tooth pose dataset. (a) Accurate estimation results. (b) Inaccurate estimation results. Columns from left to right indicate the source point cloud, the target point cloud, and the estimated results of REGTR, Lepard, and our approach. Yellow regions represent unmatching regions between the warped source and target point cloud.

segment, and register the tooth model by using only the oralscan video. After a series of evaluation experiments, the effectiveness of each module is demonstrated.

By using our treatment monitoring approach, the current orthodontic degree can be compared with the schedule to make a decision regarding whether the orthodontic plan should be adaptively adjusted. The patient no longer needs to periodically visit a dental clinic to measure the status of teeth by a dentist, which is especially useful in the epidemic stage.

In tooth instance segmentation, both visual and geometric information play an important role in pointwise classification. The relation between vision and geometry is progressively explored to alleviate the semantic gap in heterogeneous data; Moreover, employing graph attention to aggregate the local features in and across graphs is efficient in improving the capability of geometry prediction. However, segmentation quality is partially affected by the degree of teeth crowding. The overlap between neighboring teeth distorts the geometric features of normal teeth. The geometric information of the tooth root is a potential solution to discover the overlap between teeth and improve segmentation quality.

In point cloud-based tooth registration, the structural knowledge in the point space (especially negative points) is exploited to guide the matching process. Diverse negative samples revise the confidence matrix from multiple aspects and potentially improve convergence. However, the proposed approach is sensitive to the space between teeth, especially gaps between molars and premolars. Moreover, the proposed approach adopts cases in which teeth are arranged with low gaps between teeth.

Future work includes combining CBCT data to explore the tooth root information to rectify the segmentation and registration results in cases of overlap between teeth. We will also collect new data to extend our dataset with the aim of enhancing the diversity of samples.

V. CONCLUSION

We propose a novel approach for orthodontic treatment monitoring from oralscan video. We also propose a method that combines heterogeneous features by exploiting graph attention to alleviate the semantic gap. In addition, we propose a new loss function that uses diversity and negative sample mining to improve matching accuracy. In reconstructed tooth registration datasets, our approach obtains an approximately 1.2–1.4% improvement over state-of-the-art methods.

REFERENCES

[1] S. Caruso, S. Caruso, M. Pellegrino, R. Skafi, A. Nota, and S. Tecco, "A knowledge-based algorithm for automatic monitoring of orthodontic treatment: The dental monitoring system. two cases," *Sensors*, vol. 21, no. 5, pp. 1856–1865, 2021.

[2] I. Hansa, V. Katyal, S. J. Semaan, R. Coyne, and N. R. Vaid, "Artificial intelligence driven remote monitoring of orthodontic patients: Clinical applicability and rationale," *Seminars Orthodontics*, vol. 27, no. 2, pp. 138–156, 2021.

[3] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks3?," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 586–594.

[4] Z. Teed and J. Deng, "DROID-SLAM: Deep visual slam for monocular, stereo, and RGB-D cameras," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 16558–16569.

[5] T. Vu, K. Kim, T. M. Luu, X. T. Nguyen, and C. D. Yoo, "Softgroup for 3D instance segmentation on point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1586–1594.

[6] Y. Tian et al., "RGB oralscan video-based orthodontic treatment monitoring," *Sci. China Inf. Sci.*, vol. 66, no. 12, pp. 1–16, 2023.

[7] L. Perillo et al., "Monitoring biochemical and structural changes in human periodontal ligaments during orthodontic treatment by means of micro-raman spectroscopy," *Sensors*, vol. 20, no. 2, 2020, Art. no. 497.

[8] H. B. Moylan, C. K. Carrico, S. J. Lindauer, and E. Tüfekçi, "Accuracy of a smartphone-based orthodontic treatment–monitoring application: A pilot study," *Angle Orthod.*, vol. 89, no. 5, pp. 727–733, 2019.

[9] S. Talaat et al., "The validity of an artificial intelligence application for assessment of orthodontic treatment need from clinical images," *Seminars Orthodontics*, vol. 27, no. 2, pp. 164–171, 2021.

[10] G. Wei et al., "Tanet: Towards fully automatic tooth arrangement," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 481–497.

[11] G. Wei et al., "Dense representative tooth landmark/axis detection network on 3D model," *Comput. Aided Geometric Des.*, vol. 94, 2022, Art. no. 102077.

[12] T.-H. Wu et al., "Two-stage mesh deep learning for automated tooth segmentation and landmark localization on 3D intraoral scans," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3158–3166, Nov. 2022.

[13] A. X. Chang et al., "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.

[14] L. Yi et al., "A scalable active framework for region annotation in 3D shape collections," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–12, 2016.

[15] A. Jana, H. M. Subhash, and D. Metaxas, "Automatic tooth segmentation from 3D dental model using deep learning: A quantitative analysis of what can be learnt from a single 3D dental model," in *Proc. SPIE*, vol. 12567, pp. 42–51, 2023.

[16] L. Yi, W. Zhao, H. Wang, M. Sung, and L. J. Guibas, "GSPN: Generative shape proposal network for 3D instance segmentation in point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3947–3956.

[17] F. G. Zanjani et al., "Mask-MCNet: Tooth instance segmentation in 3D point clouds of intra-oral scans," *Neurocomputing*, vol. 453, pp. 286–298, 2021.

[18] B. Yang et al., "Learning object bounding boxes for 3D instance segmentation on point clouds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1132–1141.

[19] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "Pointgroup: Dual-set point grouping for 3D instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4867–4876.

[20] F. Engelmann, M. Bokeloh, A. Fathi, B. Leibe, and M. Nießner, "3D-MPA: Multi-proposal aggregation for 3D semantic instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9031–9040.

[21] Z. Liang, Z. Li, S. Xu, M. Tan, and K. Jia, "Instance segmentation in 3D scenes using semantic superpoint tree networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2783–2792.

[22] S. Chen, J. Fang, Q. Zhang, W. Liu, and X. Wang, "Hierarchical aggregation for 3D instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15467–15476.

[23] T. He, C. Shen, and A. van den Hengel, "DyCo3D: Robust instance segmentation of 3D point clouds through dynamic convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 354–363.

[24] Y. Wu, M. Shi, S. Du, H. Lu, Z. Cao, and W. Zhong, "3D instances as 1D kernels," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1132–1141.

[25] R. Klokov and V. Lempitsky, "Escape from cells: Deep KD-networks for the recognition of 3D point cloud models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 863–872.

[26] Z. Cui et al., "TSegNet: An efficient and accurate tooth segmentation network on 3D dental model," *Med. Image Anal.*, vol. 69, 2021, Art. no. 101949.

[27] L. Zhang et al., "TSGCNet: Discriminative geometric feature learning with two-stream graph convolutional network for 3D dental model segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6699–6708.

[28] Z. Yang, J. Z. Pan, L. Luo, X. Zhou, K. Grauman, and Q. Huang, "Extreme relative pose estimation for RGB-D scans via scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4531–4540.

[29] Y. Li and T. Harada, "Lepard: Learning partial point cloud matching in rigid and deformable scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5554–5564.

[30] Z. J. Yew and G. H. Lee, "REGTR: End-to-end point cloud correspondences with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6677–6686.

[31] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, and K. Xu, "Geometric transformer for fast and robust point cloud registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11143–11152.

[32] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[33] W. Feng, J. Zhang, H. Cai, H. Xu, J. Hou, and H. Bao, "Recurrent multi-view alignment network for unsupervised surface registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10297–10307.

[34] S. Agostinho, A. Ošep, A. Del Bue, and L. Leal-Taixé, "(Just) a spoonful of refinements helps the registration error go down," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6108–6117.

[35] H. Yang, W. Dong, L. Carlone, and V. Koltun, "Self-supervised geometric perception," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14350–14361.

[36] Z. J. Yew and G. H. Lee, "RPM-Net: Robust point matching using learned features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11824–11833.

[37] H. Izadinia and S. M. Seitz, "Scene recomposition by learning-based ICP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 930–939.

[38] D. Bauer, T. Patten, and M. Vincze, "ReAgent: Point cloud registration using imitation and reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14586–14594.

[39] H. Jiang, Y. Shen, J. Xie, J. Li, J. Qian, and J. Yang, "Sampling network guided cross-entropy method for unsupervised point cloud registration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6128–6137.

[40] Y. Tian, G. Cheng, J. Gelernter, S. Yu, C. Song, and B. Yang, "Joint temporal context exploitation and active learning for video segmentation," *Pattern Recognit.*, vol. 100, 2020, Art. no. 107158.

[41] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6411–6420.

[42] J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," 2021, *arXiv:2104.09864*.

[43] Y. Tian, J. Gelernter, X. Wang, J. Li, and Y. Yu, "Traffic sign detection using a multi-scale recurrent attention network," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4466–4475, Dec. 2019.

[44] Y. Tian, Y. Zhang, D. Zhou, G. Cheng, W.-G. Chen, and R. Wang, "Triple attention network for video segmentation," *Neurocomputing*, vol. 417, pp. 202–211, 2020.

[45] Y. Tian et al., "3D tooth instance segmentation learning objectness and affinity in point cloud," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 18, no. 4, pp. 1–16, 2022.

[46] D. Liu, Y. Tian, Y. Zhang, J. Gelernter, and X. Wang, "Heterogeneous data fusion and loss function design for tooth point cloud segmentation," *Neural Comput. Appl.*, vol. 34, no. 20, pp. 1–10, 2022.

[47] G. Jian, "Shining3d tooth pose dataset," 2023, doi: 10.21227/jbj7-tv82.

[48] G. Jian, "Aoralscan3 tooth registration dataset," 2023, doi: 10.21227/aa5w-kz03.

[49] F. Yang, L. Guo, Z. Chen, and W. Tao, "One-inlier is first: Towards efficient position encoding for point cloud registration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 6982–6995.

[50] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[51] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet : Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108.

[52] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. 19th Med. Image Comput. Comput.- Assist. Interv.*, 2016, pp. 424–432.

[53] C. Liu and Y. Furukawa, "Masc: Multi-scale affinity with sparse convolution for 3D instance segmentation," 2019, *arXiv:1902.04478*.

[54] L. Han, T. Zheng, L. Xu, and L. Fang, "Occuseg: Occupancy-aware 3D instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2940–2949.

[55] B. Zhang and P. Wonka, "Point cloud instance segmentation using probabilistic embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8883–8892.

[56] H. Wang, Y. Liu, Z. Dong, and W. Wang, "You only hypothesize once: Point cloud registration with rotation-equivariant descriptors," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 732–741.

[57] X. Bai et al., "PointDSC: Robust point cloud registration using deep spatial consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15859–15869.