SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2024, Vol. 67, Iss. 1, 112107:1–112107:16
https://doi.org/10.1007/s11432-023-3847-x

RGB oralscan video-based orthodontic treatment monitoring

Yan TIAN[1,2*], Hanshi FU[1,2], Hao WANG[1,2], Yuqi LIU[3], Zhaocheng XU[4], Hong CHEN[5], Jianyuan LI[6*] & Ruili WANG[1*]

[1]*School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou* 310018*, China;*
[2]*Shining3D Tech Co., Ltd., Hangzhou* 311258*, China;*
[3]*School of Information and Technology, Monash University, Melbourne* 3800*, Australia;*
[4]*School of Mathematical and Computational Sciences, Massey University, Auckland* 0632*, New Zealand;*
[5]*Department of Stomatology, Zhejiang Provincial People's Hospital, Hangzhou Medical College, Hangzhou* 310014*, China;*
[6]*School of Computer and Computing Science, Zhejiang University City College, Hangzhou* 310015*, China*

Received 17 January 2023/Revised 24 March 2023/Accepted 8 August 2023/Published online 27 December 2023

Abstract Orthodontic treatment monitoring involves using current images and previous 3D models to estimate the relative position of individual teeth before and after orthodontic treatment. This process differs from image-based object 6D pose estimation due to the gingiva deformation and varying pose offsets for each tooth during treatment. Motivated by the fact that the poses of molars remain relatively fixed in implicit orthodontics, we design an approach that employs multiview pose evaluation and bidirectional temporal propagation for jaw pose estimation and then employs an iteration-based method for tooth alignment. To handle changes in tooth appearance or location with weak texture across frames, we also introduce an instance propagation module that leverages positional and semantic information to explore instance relations in the temporal domain. We evaluated the performance of our approach using both the Shining3D tooth pose dataset and the Aoralscan3 tooth registration dataset. Our experimental results demonstrate remarkable accuracy improvements compared with existing methods.

Keywords digital dentistry, object 6D pose estimation, deep learning, computer vision

1 Introduction

Orthodontic treatment plays a crucial role in dental care as malocclusion not only increases the likelihood of caries but also causes psychological discomfort, poses a health risk, and reduces the quality of life [1]. Despite its benefits, patients must regularly visit the dental clinic to monitor the progress of orthodontic treatment and ensure that it aligns with the schedule arranged by dentists. This process can be cumbersome and complex, particularly during times of epidemic outbreaks.

The development of artificial intelligence has led to the emergence of remote monitoring of orthodontic patients, which enables patients to capture and scan their dental situation using simple and portable RGB/RGB-D equipment. This approach has gained attention from medical and academic communities due to its ability to reduce time and societal costs, as well as lessening inconvenience to patients [2]. To facilitate remote monitoring, some data-driven methods based on deep learning have been proposed to automatically segment and estimate the pose of individual objects [3–14]. However, orthodontic treatment monitoring differs from image-based object 6D pose estimation in several ways: (1) In orthodontic treatment monitoring, the current observed images are compared with the previous 3D model that was reconstructed in the last period of treatment, while in object 6D pose estimation, the 3D model and images at the same moment are compared. (2) Orthodontic treatment monitoring focuses on the relative pose of individual teeth before and after treatment, while object 6D pose estimation infers the absolute camera pose across different frames. (3) Changes in the appearance or location of each tooth with weak texture across frames make the segmentation of each tooth challenging in orthodontic treatment monitoring. An illustrative comparison is provided in Figure 1.

* Corresponding author (email: tianyan@zjgsu.edu.cn, Lijy@zucc.edu.cn, Prof.ruili.wang@gmail.com)

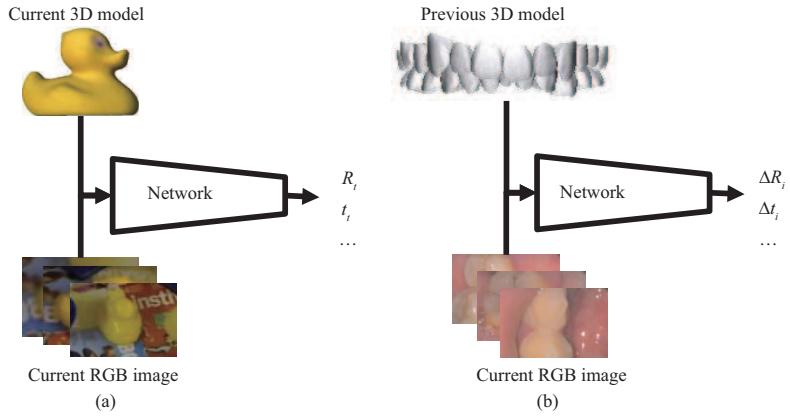

Figure 1 (Color online) Illustration of differences between (a) object 6D pose estimation and (b) orthodontic treatment monitoring. $\boldsymbol{R}_t$ and $\boldsymbol{t}_t$ indicate the absolute camera pose (rotation and transition) in frame t. $\Delta\boldsymbol{R}_i$ and $\Delta\boldsymbol{t}_i$ indicate the relative poses (rotation and transition) of tooth i before and after orthodontic treatment.

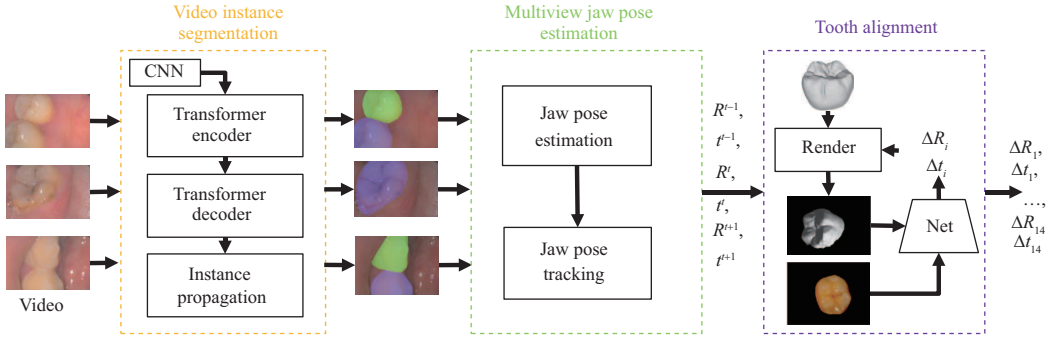

Figure 2 (Color online) Illustration of the proposed framework. Given an oral scanned video, the region of each tooth is divided by video instance segmentation. Multiview jaw pose estimation uses multiview pose evaluation and bidirectional pose tracking to infer rotation $\boldsymbol{R}_t$ and transition $\boldsymbol{t}_t$ in frame t. Tooth alignment is performed in an iteration manner to estimate rotation offset $\triangle\boldsymbol{R}_i$ and transition offset $\triangle\boldsymbol{t}_i$ of tooth i.

Motivated by the observation that some teeth, such as the third molars, have relatively fixed poses in implicit orthodontics. To take advantage of this, we designed an approach that uses multiview pose evaluation and bidirectional temporal propagation as intermediate modules to match a previous 3D model with current RGB images at the object level. In addition, for tooth segmentation, we believe that the temporal relation of the instance can be modeled by relevant factors in a transformer architecture, since the predicted box and class score contain compact semantic and location knowledge.

In this paper, we propose a novel approach for accurately measuring the degree of orthodontics during a period of treatment. The framework, as shown in Figure 2, utilizes an oral scanned video and a transformer architecture method to segment the region of each tooth. To construct the instance relation in the temporal domain, we propose an instance propagation module that utilizes box position, class scores, and instance queries. To ensure temporal consistency, we design a temporal consistency loss that learns embeddings with high similarity across frames for the same instance. Next, we employ a data-driven method to estimate the jaw poses of frames with fixed teeth, followed by improving the estimates with multiview geometry knowledge. Jaw poses of frames with orthodontic teeth are inferred by pose tracking with bidirectional constraints. Finally, an iteration-based method is used to estimate the relative pose of each orthodontic tooth.

The contribution of this paper is exemplified as follows:

- An approach of orthodontic treatment monitoring is proposed that uses the previous 3D tooth model to estimate pose offsets of orthodontic teeth, rather than relying on the current 3D tooth model.
- An instance propagation module is introduced in video instance segmentation, which models the temporal association of an instance by three factors, leading to a significant reduction in computational complexity and memory consumption.
- A temporal consistency loss is designed in video instance segmentation to strengthen the relations between positive samples, while reducing the importance of relations between negative samples.

• Poses of the jaw with rigid-transformed orthodontic teeth and nonrigid-transformed gingiva in a video sequence are estimated, in which orthodontic knowledge is utilized as an intermediate factor to guide the registration process.

The experimental results obtained from the Shining3D tooth pose dataset and the Aoralscan3 tooth registration dataset demonstrate that the proposed approach achieves outstanding performance.

The rest of this paper is organized as follows. Section 2 reviews previous studies on video instance segmentation and video-based object pose estimation. Section 3 introduces the proposed approach. Section 4 discusses the experimental results. Concluding remarks are described in Section 5.

2 Related work

In this section, we briefly review the recent literature on video instance segmentation and video-based object 6D pose estimation.

2.1 Orthodontic treatment monitoring

Some studies have utilized micro-Raman spectroscopy [15] or cone-beam computed tomography (CBCT) to monitor the periodontal ligament status. However, the limited sources of these methods restrict their scope of use. To address this limitation, Ref. [16] leveraged oralscan videos captured by smartphones to predict the maxillary and mandibular arches. However, the use of a maxillary expander in patients can lead to discomfort and complexity. More recent studies have focused on detecting [17] or segmenting [2,18] individual teeth from intraoral videos and using deep learning methods to predict the corresponding pose parameters. Tooth identity is crucial in orthodontic treatment monitoring, and to improve the recognition results provided by the detection or segmentation modules, a separate classification network [19] is used to accurately identify the teeth. Similar classification methods have also been introduced in the index of orthodontic treatment needs (IOTN) [20] assessment, which determines whether an individual is eligible for further orthodontic treatment. However, the individual modules used in these approaches have not been fully investigated.

2.2 Video instance segmentation

Video instance segmentation (VIS) is a task that involves simultaneously classifying, segmenting, and tracking object instances of interest in a video sequence. There are two main categories of VIS methods: frame-based methods and clip-based methods.

Frame-based methods, also known as online methods, segment each frame independently and then associate segmented masks of each instance across frames using a postprocessing step [21–23]. However, the association process can be sensitive to motion blurs and occlusions, which are common in videos. Additionally, separating image-level instance segmentation and association of instances across consecutive frames can increase the risk of local optimization.

Clip-based methods or offline methods extract a 3D spatiotemporal volume from a video clip and directly segment the 3D mask for each instance. Recently, transformer-based approaches, such as VisTR [24], have been used to generate a sequence of masks for each instance in an end-to-end manner. The sequential transformer (SeqFormer) [25] uses temporal box queries to learn a powerful representation of instance queries. The video mask transfiner (VMT) [26], on the other hand, employs a 3D incoherence detector and temporal refinement transformer to rectify false instance masks obtained by SeqFormer. The use of clip-based methods presents a challenge in image and video segmentation research because it disconnects the two fields. To address this issue, Cheng et al. [27] proposed an extension of Mask2Former [28] that leverages 3D spatiotemporal features and instance tracking over time. However, the computational complexity and memory storage requirements pose a significant hindrance to the practical application of this approach. To address the challenges associated with frame-to-frame communication, interframe communication transformers (IFC) [29] have been developed to enrich and correlate features in each frame through memory tokens. Despite these advancements, the temporal association of an instance remains complex and memory-intensive.

2.3 Multiview object pose estimation

Current methods with regard to multiview object pose estimation can be divided into four categories: (1) tracking or recurrent prediction of object poses in multiple frames; (2) independent inference of object pose in each frame, followed by joint optimization through frame association; (3) direct and joint optimization of object poses in multiple frames; and (4) 3D volume reconstruction.

Pose tracking. Kalman filters [6] or particle filters [12] are commonly used to associate corresponding objects in neighboring frames and update the object pose in the next frame. However, recent approaches have explored alternative methods to model temporal relations at the feature level using long short-term memory (LSTM) [5, 10], convolution LSTM [3, 4], or graph neural networks (GNNs) [10], owing to their ability to nonlinearly fuse historic information. To address occlusion issues, some methods explicitly predict the full object mask [7] or attention map [8, 30] before estimating the object pose based on prior information. For textureless objects, a contour part model [13, 14, 31] is employed for matching between rendered and observed images. To reduce the time-consuming process of reprojection, Ref. [32] employed geometric contours and local regions for pose prediction and refinement, while minimizing the reprojection process to a single time. However, the accuracy of inference is limited due to the neglect of bidirectional motion modeling.

Bundle adjustment. After independently predicting the object poses in each frame, global consistency is then used to jointly optimize them. The traditional approach [33] selects the pose with the highest voting score among all the candidates. In some methods, the initial predicted pose is jointly optimized through an association of neighboring frames. The relative transformations between camera viewpoints can be estimated by matching objects in different images [9], and the 6D poses of objects in different frames can be improved through a global refinement procedure based on either an object-level [9] or point-level [11] bundle adjustment.

Joint prediction of multiview poses. RotationNet [34] is an unsupervised method that generates viewpoint-specific category likelihoods for predefined discrete viewpoints and selects the object pose with the maximum integrated object category likelihood. However, the effectiveness of this approach is limited by the lack of supervision signals. To address this limitation, alternative approaches have been proposed. The object detection, association, and mapping (ODAM) method [35] utilizes a maximum a posteriori (MAP) framework to globally optimize the object pose, while Vid2CAD [36] optimizes the sum of geometric losses over total frames. Nonetheless, these methods do not explore the relationship between neighboring objects to evaluate the prediction of each object.

3D volume reconstruction. Multiview images are used to explicitly [37] or implicitly [38] reconstruct a 3D model, and object poses in new images are predicted by comparing rendered and observed images [38]. To enhance the discriminative capacity in feature maps, graph attention [37] or ray-traced transformers [39] are employed to explore context information in 2D/3D representations, and to decrease computational complexity, 2D-3D matching may only be performed on keypoints [37, 40]. While 3D volume-based methods are useful, they are generally resource-intensive and not suitable for mobile applications such as orthodontic treatment monitoring.

3 Our approach

To estimate the degree of orthodontic treatment needed, we propose a three-stage approach that utilizes an oralscan RGB video. First, a multiview pose estimation method is used to infer the jaw pose in each frame. Subsequently, an iteration-based framework is employed to compare the rendered and observed tooth images and update the tooth pose based on the comparison result, as depicted in Figure 2.

The performance gap in video instance segmentation can be attributed to the changes in the appearance or location of each instance across frames in a video sequence. Furthermore, the memory consumption of spatial-temporal attention needs to be constrained to adapt to the device capability.

Drawing inspiration from the dynamic anchor box DETR (DAB-DETR) [41], we introduce a prior propagation module that uses sparse reference boxes and content queries to represent object queries. This approach not only reduces computational complexity and storage consumption in cross-attention but also provides semantic knowledge of the instance through class scores. Instead of relying on complex data-association methods, such as spatiotemporal attention, our approach constructs the instance relation in neighboring frames using reference boxes, class scores, and instance queries. For further details, refer

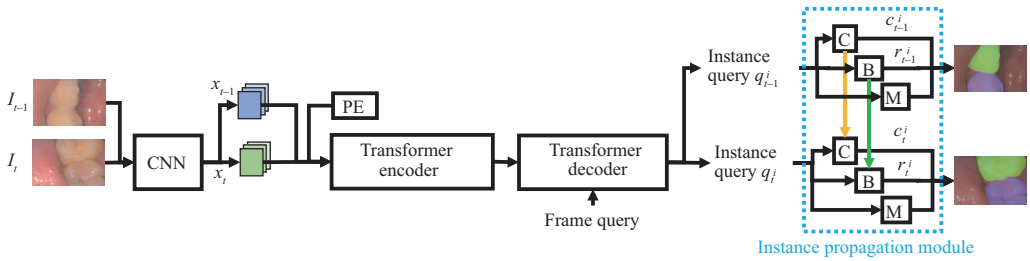

Figure 3 (Color online) Illustration of the video instance segmentation module. Given an oral scanned video containing frames $\{\boldsymbol{I}_t\}$, where t is the frame index, a backbone is employed to extract feature maps $\{\boldsymbol{X}_t\}$. Then, the transformer encoder refines the feature maps, and the transformer decoder generates instance queries $\{\boldsymbol{q}_t^i\}$, where i is the instance index. Finally, the instance propagation module uses instance queries to segment and classify each instance. Symbol 'PE' represents the positional embedding. Symbols 'C', 'B', and 'M' represent the classification, localization, and segmentation heads, respectively. Arrows in yellow and green are important factors in our approach.

to Figure 3.

3.1 Video instance segmentation

We assume that the input video consists of a sequence of image frames $\{\boldsymbol{I}_t\} \in \mathbb{R}^{H\times W\times 3}$, where $t \in \{1, 2, \ldots, T\}$ is the frame index, and H and W are the height and width of the image, respectively. The feature maps $\{\boldsymbol{x}_t\}$ are extracted by a backbone network, such as ResNet-50 [42]. We then utilize a transformer encoder with 5 blocks to refine the feature maps $\{\boldsymbol{x}_t\}$, with positional encoding added. The transformer decoder is fed frame queries, which include positional and content queries, to probe instances. In the instance propagation module, we assume that the instance query i at frame t is $\boldsymbol{q}_t^i$. This representation is propagated to frame $t+1$ by initializing query weights of $\boldsymbol{q}^i t+1$. Moreover, we consistently propagate prior locations and scales by learning to determine the offset of reference boxes in each frame. We obtain reference boxes $\boldsymbol{r}_t^i$ with index i in frame t by

$$\boldsymbol{r}_t^i = \sigma(\boldsymbol{W}_r(\boldsymbol{q}_t^i) \times \boldsymbol{r}_{t-1}^i), \tag{1}$$

where $\sigma(\cdot)$ indicates a sigmoid function, and $\boldsymbol{W}_r$ is a weight matrix to be learned. In the first frame of the video sequence ($t = 1$), we obtain reference boxes $\boldsymbol{r}_t^i$ through a mapping function of instance query $\boldsymbol{q}_t^i$.

However, the probability distribution of an instance changes due to the inconsistent appearance in subsequent frames of the video sequence. The class probability $\boldsymbol{c}_t^i$ of instance i at time t is

$$\boldsymbol{c}_t^i = \sigma(\boldsymbol{W}_c(\boldsymbol{q}_t^i))\mathrm{Softmax}(\boldsymbol{W}_t[\boldsymbol{c}_f^i]_{f=t-d}^{t-1}), \tag{2}$$

where $\boldsymbol{W}_c$ and $\boldsymbol{W}_t$ are weight matrices to be learned, $[\cdot]$ represents a concatenation operation, d is the frame number stored in the memory, and f is the index of the stored frame.

Loss function. In DAB-DETR, the Hungarian algorithm is used to match the instance query with a real instance. In video instance segmentation based on DAB-DETR, the classification loss L_{cls}, box loss L_{box}, and mask loss L_{mask} are optimized as follows:

$$L_{\mathrm{cls}} = \sum_i \sum_t \tilde{\boldsymbol{c}}_t^i \log(\boldsymbol{c}_t^i), \tag{3}$$

$$L_{\mathrm{box}} = \sum_i \sum_t |\boldsymbol{r}_t^i - \tilde{\boldsymbol{r}}_t^i|, \tag{4}$$

$$L_{\mathrm{mask}} = \sum_p \sum_t \tilde{\boldsymbol{m}}_t^p \log(\boldsymbol{m}_t^p), \tag{5}$$

where $\boldsymbol{c}_t^i$ and $\boldsymbol{r}_t^i$ are the predicted class vector and box position of instance i in frame t, $\boldsymbol{m}_t^p$ is the predicted class vector of pixel p in frame t, and $\tilde{\boldsymbol{c}}_t^i$, $\tilde{\boldsymbol{r}}_t^i$, and $\tilde{\boldsymbol{m}}_t^p$ represent the corresponding ground-truth class, box position, and pixel class, respectively. However, the importance of temporal consistency in instances cannot be ignored, as the similarity of the same instance across different frames is expected to be greater than the similarity between different instances within the same frame.

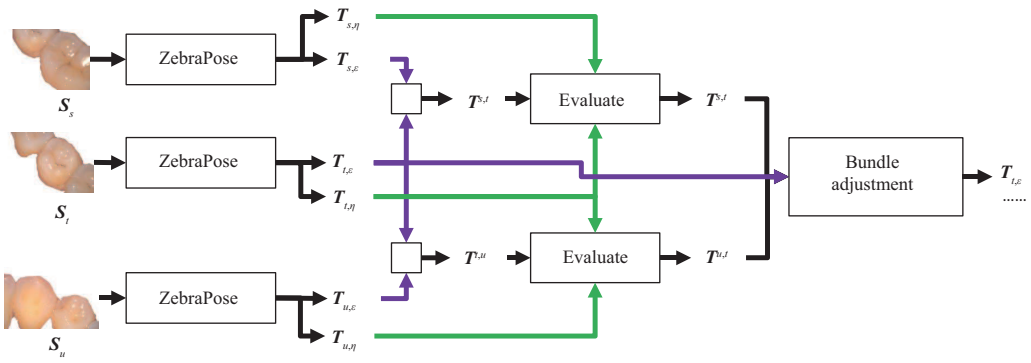

Figure 4 (Color online) Illustration of multiview jaw pose estimation on fixed teeth. Given foreground frames $\boldsymbol{S}_s$, $\boldsymbol{S}_t$, and $\boldsymbol{S}_u$, the 6D poses $\boldsymbol{T}_{t,\eta}$ and $\boldsymbol{T}_{t,\epsilon}$ of orthodontic tooth η and fixed tooth ϵ with respect to the camera are predicted by ZebraPose. The camera motion $\boldsymbol{T}^{s,t}$ between frames s and t is employed for multiview pose evaluation. Finally, bundle adjustment is employed for pose refinement. Arrows in purple and green represent operations on fixed teeth and orthodontic teeth, respectively.

Therefore, we introduce a temporal consistency loss, denoted by L_{tc}. Specifically, we define a positive set comprising the same instance appearing in different frames and a negative set consisting of different instances appearing in different frames. For example, i at time t, the temporal consistency loss L_{tc} is calculated as follows:

$$L_{\mathrm{tc}} = \sum_{i} \sum_{f=t-d}^{t-1} \log \frac{\exp\left(\boldsymbol{q}_t^i \cdot \boldsymbol{q}_f^i / \tau\right)}{\sum_{j=0}^{k} \exp\left(\boldsymbol{q}_t^i \cdot \boldsymbol{q}_f^j / \tau\right)}, \tag{6}$$

where d is the frame number stored in the memory, f is the index of the stored frame, k is the number of instances with greater class scores, and τ is an attenuation coefficient.

The total loss is a linear combination of the classification loss L_{cls}, the box loss L_{box}, the mask loss L_{mask}, and the temporal consistency loss L_{tc} as follows:

$$L_{\mathrm{total}} = \lambda_{\mathrm{c}} L_{\mathrm{cls}} + \lambda_{\mathrm{b}} L_{\mathrm{box}} + \lambda_{\mathrm{m}} L_{\mathrm{mask}} + L_{\mathrm{tc}}, \tag{7}$$

where λ_{c}, λ_{b}, and λ_{m} are coefficients to balance different constraints.

Our approach offers several advantages. (1) Representation propagation enables the exploration of semantic instance knowledge, while reference boxes provide detailed position and scale information. (2) The position and scale information is updated by calculating the offset in each frame. (3) The memory consumption is constrained to adapt to device capability. (4) Temporal consistency is employed to learn embeddings that maintain high similarity across frames.

3.2 Multiview jaw pose estimation

Accurately aligning the previous 3D model with the current observed image to measure the discrepancy in tooth pose requires a precise jaw pose as the alignment base. Unfortunately, current object 6D pose estimation methods are not suitable for predicting jaw pose due to the deformation that occurs during orthodontic treatment. This deformation causes a natural mismatch between the 3D model and the observed image, rendering the existing methods ineffective.

Motivated by the observation that certain teeth, such as molars, remain fixed in specific stages of orthodontic treatment, we design an approach to register the previous 3D jaw model with current observed images through multiview pose evaluation. Additionally, we designed a bidirectional pose tracking approach to propagate the anchor pose to temporal neighboring frames.

3.2.1 *Multiview 6D pose estimation on fixed teeth*

We present a method for multiview 6D pose estimation on fixed teeth, as illustrated in Figure 4. Assuming that the input video consists of foreground frames $\{\boldsymbol{S}_t\}_{t=1}^{T}$ extracted by video instance segmentation, where t indicates the frame index, we determine the corresponding tooth between frames after segmentation. The 3D jaw model $\boldsymbol{M}$ generated in the last orthodontic period contains the gingiva part and multiple teeth with tooth index $l = 1, 2, \ldots, 14$. We use the fixed and known intrinsic parameter matrix $\boldsymbol{\pi}$ of a camera to predict the 6D pose of fixed teeth associated with each frame.

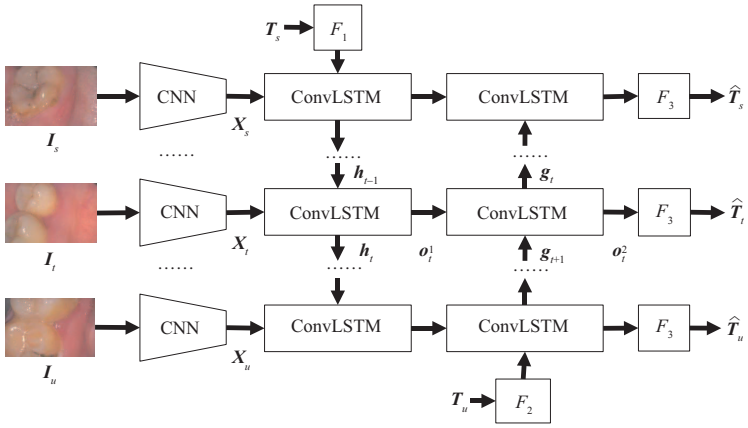

Figure 5 (Color online) Illustration of the pose tracking module. Frames $\boldsymbol{I}_s$, $\boldsymbol{I}_t$, and $\boldsymbol{I}_u$ go through the backbone to extract features $\boldsymbol{x}_s$, $\boldsymbol{x}_t$, and $\boldsymbol{x}_u$. $\boldsymbol{T}_s$ and $\boldsymbol{T}_u$ represent 6D poses obtained by using a fixed tooth. Poses $\boldsymbol{T}_t$ in the remaining frames are inferred by using bidirectional convLSTM.

The initial 6D pose $\boldsymbol{T}_{t,\epsilon} \in SE(3)$ of a fixed tooth ϵ such as the third molar is a 4×4 homogeneous matrix, including a 3D rotation matrix $\boldsymbol{R}_{t,\epsilon}$ and a 3D translation vector $\boldsymbol{t}_{t,\epsilon}$ of the fixed tooth ϵ in frame t, and $\boldsymbol{T}_{t,\epsilon}$ is initially estimated by ZebraPose [43], which encodes the object surface by incorporating a hierarchical binary grouping, matches predicted codes with the object surface, and estimates the 6DoF pose by using a PnP solver.

After instance segmentation, we assume that both frames s and t contain the same fixed tooth ϵ, and then a relative camera pose hypothesis is obtained by $\boldsymbol{T}^{s,t} = \boldsymbol{T}_{s,\epsilon}\boldsymbol{T}_{t,\epsilon}^{-1}$. To evaluate the relative pose hypothesis, we assume that n is the index of the neighboring tooth of the fixed tooth ϵ that both teeth appear in the same frame, and then the distance $D(\cdot,\cdot)$ between $\boldsymbol{T}_{s,n}$ and $\boldsymbol{T}^{s,t}\boldsymbol{T}_{t,n}$ is measured

$$D(\boldsymbol{T}_{s,n}, \boldsymbol{T}^{s,t}\boldsymbol{T}_{t,n}) = \frac{1}{|\mathcal{Y}_n|} \sum_{\boldsymbol{y}\in\mathcal{Y}_n} ||\boldsymbol{T}_{s,n}\boldsymbol{y} - \boldsymbol{T}^{s,t}\boldsymbol{T}_{t,n}\boldsymbol{y}||_2, \tag{8}$$

where tooth n is associated with a set of 3D points $\boldsymbol{y} \in \mathcal{Y}_n$. The distance $D(\cdot,\cdot)$ is compared with a given threshold C to evaluate whether the relative camera pose is reasonable.

Finally, unique and consistent poses $\boldsymbol{T}_{t,\epsilon}$ of some fixed teeth ϵ are recovered by a bundle adjustment as follows:

$$L = \frac{1}{|\mathcal{X}_\epsilon|} \sum_s \sum_{\boldsymbol{x}\in\mathcal{X}_\epsilon} ||\boldsymbol{S}_{s,\epsilon} - \pi(\boldsymbol{T}^{s,t}\boldsymbol{T}_{t,\epsilon}\boldsymbol{y})||_2, \tag{9}$$

where $\boldsymbol{S}_{s,\epsilon}$ is a foreground image of tooth ϵ in frame s extracted by instance segmentation.

3.2.2 *Bidirectional jaw 6D pose tracking*

The poses of the jaw have been established in frames that include teeth with fixed poses, but in other frames, the jaw poses are still unknown. To address this issue, pose tracking is performed, starting from the frame with the fixed tooth, to obtain the initial pose. Further details of the pose tracking procedure are depicted in Figure 5. Instead of the transformer architecture, we adopt ConvLSTM for pose tracking because it has a built-in memory mechanism that enables it to capture long-term dependencies in the input data, making it suitable for modeling intricate, long-term relationships between input and output data.

We assume that the input video contains image frames $\{\boldsymbol{I}_s, \ldots, \boldsymbol{I}_t, \ldots, \boldsymbol{I}_u\} \in \mathbb{R}^{H\times W\times 3}$, where s, t, and u are frame indices. The backbone ResNet-50 is employed to extract visual features $\{\boldsymbol{x}_s, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_u\}$. We also assume that jaw poses $\boldsymbol{T}_s$ and $\boldsymbol{T}_u$ in frames s and u are obtained using fixed teeth without loss of generality. The hidden vector $\boldsymbol{h}_t$ and output vector $\boldsymbol{o}_t^1$ in the first convLSTM are estimated by incorporating the current visual feature $\boldsymbol{x}_t$ and previous hidden vector $\boldsymbol{h}_{t-1}$; the hidden vector $\boldsymbol{g}_t$ and output vector $\boldsymbol{o}_t^2$ in the second convLSTM are inferred by using the output vector $\boldsymbol{o}_t^1$ and hidden vector $\boldsymbol{g}_{t+1}$ in the next frame.

$$\{\boldsymbol{h}_t, \boldsymbol{o}_t^1\} = \text{ConvLSTM}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}), \tag{10}$$

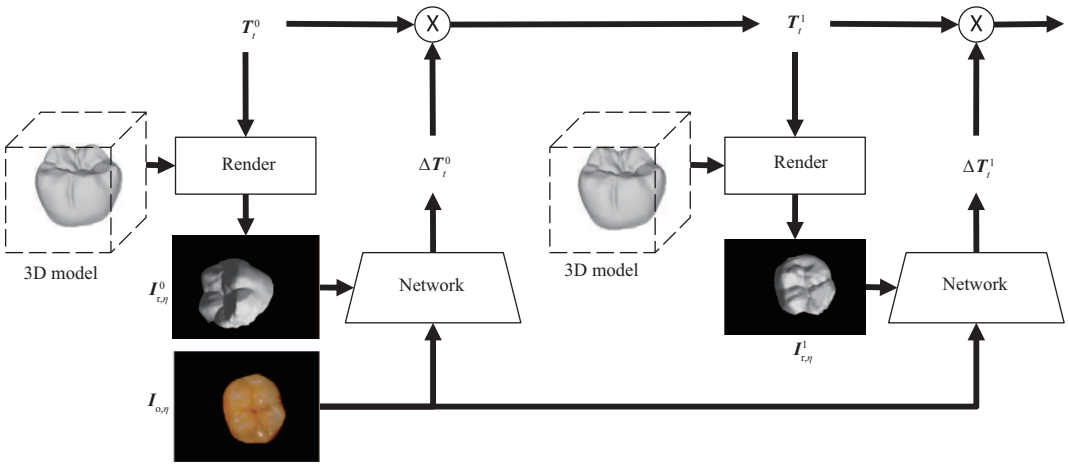

Figure 6 (Color online) Illustration of the iteration-based tooth alignment module. In iteration ν, the rendered image $\boldsymbol{I}^{\nu}_{\mathrm{r},\eta}$ and observed image $\boldsymbol{I}_{\mathrm{o},\eta}$ of orthodontic tooth η are input to the mapping network to infer a relative transformation $\Delta\boldsymbol{T}^{\nu}_{\eta}$.

$$\{\boldsymbol{g}_t, \boldsymbol{o}^2_t\} = \mathrm{ConvLSTM}(\boldsymbol{o}^1_t, \boldsymbol{g}_{t+1}), \tag{11}$$

$$\hat{\boldsymbol{T}}_t = F_3(\boldsymbol{o}^2_t), \tag{12}$$

where the pose $\boldsymbol{T}_t$ in frame t is obtained by a linear mapping layer $F_3(\cdot)$. The initial hidden vector $\boldsymbol{h}_{s-1}$ in the first frame and $\boldsymbol{g}_{u+1}$ in the last frame are obtained by using linear mappings $F_1(\cdot)$ and $F_2(\cdot)$, and 6D poses $\boldsymbol{T}_s$ and $\boldsymbol{T}_u$ obtained from fixed teeth, respectively,

$$\boldsymbol{h}_{s-1} = F_1(\boldsymbol{T}_s), \tag{13}$$

$$\boldsymbol{g}_{u+1} = F_2(\boldsymbol{T}_u). \tag{14}$$

In the training stage, the L1-norm loss in the 6D pose space is optimized. We assume that $\boldsymbol{T}_t$ is the ground truth of the 6D pose in frame t; then,

$$L_u = \sum_t ||\hat{\boldsymbol{T}}_t - \boldsymbol{T}_t||_1. \tag{15}$$

To enhance the smoothness of the predicted poses across frames, we added a regularization term as follows:

$$L_p = \sum_t ||(\hat{\boldsymbol{T}}_t - \hat{\boldsymbol{T}}_{t-1}) - (\boldsymbol{T}_t - \boldsymbol{T}_{t-1})||_1. \tag{16}$$

The total loss is a linear combination of these two terms as follows:

$$L_{\mathrm{total}} = L_u + \lambda_p L_p, \tag{17}$$

where λ_p is a combination weight to control the effect of the smooth term.

Our approach offers several advantages: (1) The 3D model and observed images from different periods can be aligned using 6D pose estimation in a specific frame and bidirectional information propagation. (2) Multiview cues are used to enhance the robustness of the estimation model. (3) Our approach is simple and straightforward to implement.

3.3 Iteration-based tooth alignment

After obtaining the initial pose $\boldsymbol{T}^0_{\eta,t}$ of the orthodontic tooth η in frame t, we can further improve the alignment of the orthodontic tooth η by updating its pose progressively and comparing the projection of the 3D tooth model with the observed foreground image. The process is demonstrated in Figure 6.

Given a jaw 3D model $\mathcal{Z}$ that is constructed in a previous orthodontic period, a 3D instance segmentation method [44] is employed to divide the 3D region of each tooth and classify its corresponding tooth category; i.e., the orthodontic tooth model $\mathcal{Z}_\eta$ with index η is obtained. In iteration ν, the rendered tooth image $\boldsymbol{I}^{\nu}_{\mathrm{r},\eta}$ is obtained by using the orthodontic tooth model $\mathcal{Z}_\eta$ and a 6D pose $\boldsymbol{T}^{\nu-1}_{\eta}$ that is estimated

from the previous iteration $\nu-1$. The observed tooth image $\boldsymbol{I}_{\mathrm{o},\eta}$ is obtained by extracting the region of tooth η via video instance segmentation.

The rendered tooth image $\boldsymbol{I}^{\nu}_{\mathrm{r},\eta}$ and observed tooth image $\boldsymbol{I}_{\mathrm{o},\eta}$ are input into the mapping network based on the deepim [45], and the relative transformation $\triangle\boldsymbol{T}^{\nu}_{\eta}$ is inferred.

The pixel matching loss punishes the 2D discrepancy between the rendered image and the observed image, which is calculated as follows:

$$L_{\mathrm{pm}}=\sum_{i}|\pi\triangle\boldsymbol{T}^{\nu}_{\eta}\boldsymbol{T}^{\nu}_{\eta}\boldsymbol{z}^{i}_{\eta}-\boldsymbol{I}^{i}_{\mathrm{o},\eta}|, \tag{18}$$

where $\boldsymbol{z}^{i}_{\eta}$ is a 3D point i on the orthodontic tooth model $\mathcal{Z}_{\eta}$. The pixel matching loss finds the relative transformation $\Delta\boldsymbol{T}^{\nu}_{\eta}$ that the rendered tooth image $\boldsymbol{I}^{\nu}_{\mathrm{r},\eta}$ and the observed tooth image $\boldsymbol{I}_{\mathrm{o},\eta}$ are matched.

4 Results

In this section, we compare the performance of each proposed module with that of relevant approaches.

4.1 Hardware and software environment

We use a workstation with 4 NVIDIA RTX 3090 GPUs. Our approach is implemented based on the PyTorch [46] platform.

4.2 Dataset

The proposed instance segmentation approach is verified on two datasets, namely the Shining3D tooth segmentation dataset [47] and the Aoralscan3 tooth segmentation dataset [48]. The Shining3D dataset consists of 1866, 272, and 272 videos for the training, validation, and testing sets, respectively, while the Aoralscan3 dataset includes 1573, 244, and 244 videos for the corresponding sets. Both datasets have an image size of 640 × 480 pixels. LabelMe software is employed to accurately mark the boundary and classify the region of each tooth in the datasets.

The proposed jaw and tooth pose estimation approach is evaluated on the Shining3D tooth pose dataset and the Aoralscan3 tooth registration dataset. The jaw models used in our approach were generated via oral scanning of hospital patients. For each tooth, ground truth relative pose information was generated by introducing random jittering to the tooth models. The Shining3D tooth pose dataset was split into 1689 training samples, 150 validation samples, and 150 testing samples. Meanwhile, the Aoralscan3 tooth registration dataset was constructed using 1667 samples for training, 156 samples for validation, and 176 samples for testing.

4.3 Evaluation criteria

In video instance segmentation, we employ average precision (AP) metrics under intersection over union (IoU) thresholds of 50% and 75% as the evaluation criteria to compare its performance to that of state-of-the-art approaches.

On the other hand, for jaw and tooth pose estimation, we utilize the ADD-S and ADD(-S) metrics, as well as their corresponding area under the curve (AUC), as the evaluation criteria.

4.4 Implementation details

In video instance segmentation, data augmentation methods are employed to expand the dataset to approximately 160000 images, including vertical/horizontal flipping, translation variance, and scaling. ResNet-50 serves as the backbone, and the transformer encoder and decoder consist of $L=5$ blocks. To update the weights, AdamW is used with a momentum of 0.9 and a weight decay of 2.0×10^{-3}. Each minibatch contains 4 samples, and the learning rate starts at 4.0×10^{-3} and is then decreased to 2.0×10^{-3} after 70000 iterations.

To enhance the dataset for jaw and tooth pose estimation, various data augmentation methods, such as rotation and translation, are commonly employed. In our experiment, we use 20 consecutive images to construct a sequence, even though our tracking module can accept dynamic lengths of inputs. This approach not only helps to constrain accumulative errors in the tracking module but also provides an

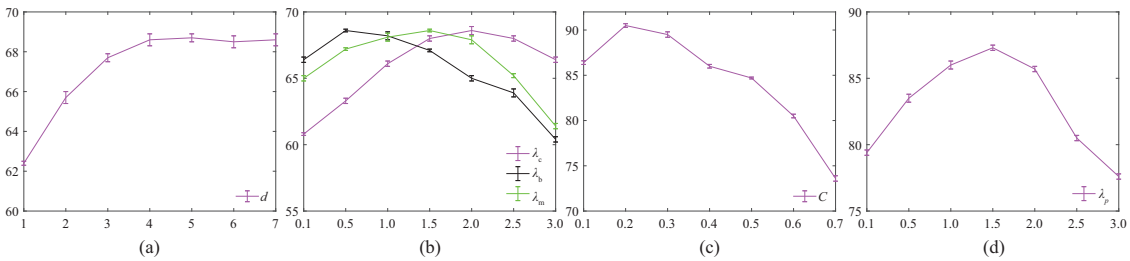

Figure 7 (Color online) Parameter selection on the Shining3D tooth segmentation dataset and Shining3D tooth pose estimation dataset. Quantitative analysis of (a) the frame number d stored in memory, (b) the class weight λ_{c}, box weight λ_{b}, and mask weight λ_{m}, (c) the distance threshold C, and (d) the weight λ_p in smooth terms.

Table 1 Evaluation of different numbers of fixed teeth on the Shining3D tooth pose estimation dataset and Aoralscan3 tooth registration dataset[a)]

	Shining3D		Aoralscan3	
Number of fixed teeth	AUC of ADD-S	AUC of ADD(-S)	AUC of ADD-S	AUC of ADD(-S)
1	84.1	78.4	84.6	89.7
2	86.9	81.5	87.2	92.0
3	88.7	83.8	89.0	94.1
4	**90.5**	**85.6**	**91.2**	**96.2**

a) The values in bold represent the best results among different number of fixed teeth.

instantaneous response to the patient. During the training process, we use the AdamW optimizer for 100 epochs with a weight decay of 1.0×10^{-4} and a batch size of 4. The learning rate is adjusted to 2.0×10^{-2} and decays exponentially by 4.0×10^{-3} in each epoch. We set the matching radius threshold to 0.1 mm.

4.5 Ablation study

We perform ablation studies to demonstrate the effectiveness of several key components in our approach. All experiments in this subsection are conducted on the Shining3D tooth segmentation or Shining3D tooth pose dataset.

Parameters. The grid search approach is employed to select various parameters. For video instance segmentation, the optimal number of frames d stored in memory is determined to be $d=4$, as shown in the experimental results in Figure 7(a). The results suggest that adding more frames does not improve performance but instead consumes unnecessary resources. Moreover, the weights for different components, including class weight λ_{c}, box weight λ_{b}, and mask weight λ_{m}, are selected as $\lambda_{\mathrm{c}}=2.0$, $\lambda_{\mathrm{b}}=0.5$, and $\lambda_{\mathrm{m}}=1.5$, respectively, as evaluated in the experiments shown in Figure 7(b). Regarding jaw poses estimation, the distance threshold is set to $C=0.2$ mm based on the results in Figure 7(c). Finally, the selection of weights $\lambda_p=1.5$ in smooth terms is depicted in Figure 7(d).

Number of fixed teeth. The accuracy of jaw pose estimation is influenced by the number of fixed teeth. Table 1 reports the experimental results for 1–4 fixed teeth. While having more fixed teeth would improve effectiveness, it is not feasible in real-world situations.

Effectiveness. We evaluated the effectiveness of multiple components in video instance segmentation, jaw pose estimation, and jaw pose tracking. A comparison of their effectiveness can be found in Table 2.

By employing the instance propagation module and temporal consistency loss in video instance segmentation, we observed an increase in ADD-S of 1.6 and 1.7 in the two datasets, respectively. Further improvements in ADD-S can be achieved by implementing neighboring tooth evaluation in multiview jaw pose estimation and bidirectional propagation in pose tracking, resulting in margins of approximately 1.6 and 2.1 in the Shining3D tooth pose estimation dataset, respectively. The effectiveness of these different modules is illustrated in Figure 8.

4.6 Evaluation of video instance segmentation

We compared various approaches for video instance segmentation using the Shining3D tooth segmentation dataset and the Aoralscan3 tooth segmentation dataset, and the results are presented in Table 3. Overall, transformer-based methods exhibit strong discrimination capacity and are competitive with other approaches.

Table 2 Effectiveness comparison among important modules on the Shining3D tooth pose estimation dataset and Aoralscan3 tooth registration dataset[a)]

Configurations				Shining3D		Aoralscan3	
+IPM	+TCL	+NTE	+BP	AUC of ADD-S	AUC of ADD(-S)	AUC of ADD-S	AUC of ADD(-S)
–	–	–	–	83.8	74.0	76.9	73.8
✓	–	–	–	84.9	75.2	77.7	74.5
–	✓	–	–	85.1	75.8	77.9	75.3
✓	✓	–	–	85.4	76.2	78.6	75.9
–	–	✓	–	85.4	75.4	78.6	75.8
–	–	–	✓	85.9	75.7	79.2	76.1
–	–	✓	✓	86.7	76.6	80.1	76.9
✓	✓	✓	✓	**87.7**	**77.8**	**81.2**	**78.2**

a) Symbols 'IPM' and 'TCL' represent the instance propagation module and temporal consistency loss in video instance segmentation. Symbols 'NTE' and 'BP' represent the neighboring tooth evaluation in multiview jaw pose estimation and bidirectional propagation in pose tracking. The values in bold represent the best results among different configurations.

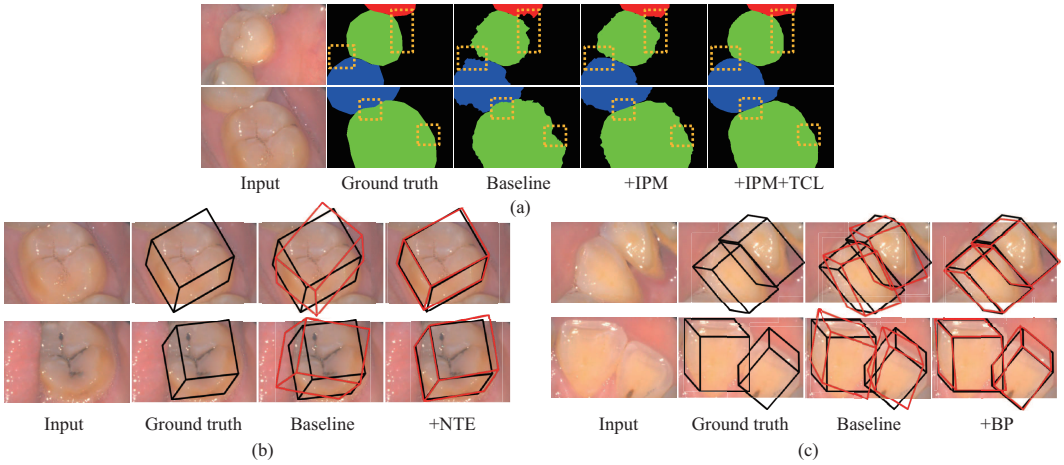

Figure 8 (Color online) Experimental results of different modules on the Shining3D tooth dataset. The symbols 'IPM', 'TCL', 'NTE', and 'BP' represent the instance propagation module, temporal consistency loss, neighboring tooth evaluation, and bidirectional propagation, respectively. Orange dotted boxes highlight differences in results. Red straight boxes represent predicted 6D poses. (a) Video instance segmentation; (b) multiview jaw 6D pose estimation; (c) jaw 6D pose tracking.

Table 3 Effectiveness comparison of video instance segmentation on the Shining3D tooth segmentation dataset and the Aoralscan3 tooth segmentation dataset[a)]

Approach	Shining3D			Aoralscan3		
	AP↑	AP50↑	AP75↑	AP↑	AP50↑	AP75↑
CSipMask [49]	36.4	55.5	39.9	37.2	57.3	40.1
PCAN [50]	37.3	58.4	39.7	39.6	59.7	40.3
CrossVIS [51]	38.2	61.2	38.9	40.3	60.1	40.8
VisTR [24]	38.3	58.8	40.9	40.6	59.7	42.9
IFC [29]	39.5	60.3	42.5	41.3	60.8	44.4
DeVIS [52]	40.3	61.1	43.4	42.4	61.2	45.4
InstanceFormer [53]	40.4	60.4	42.0	42.6	61.9	45.9
SeqFormer [25]	41.4	59.5	46.5	43.5	63.0	47.1
Mask2Former [27]	42.3	60.4	46.9	44.4	63.4	47.8
MinVIS [54]	42.8	60.9	47.4	44.9	64.2	48.8
VITA [55]	43.5	62.2	47.9	45.5	66.0	49.8
IDOL [56]	44.0	66.1	48.2	46.4	68.2	50.3
Ours	**45.6**	**68.6**	**49.6**	**47.9**	**70.3**	**51.4**

a) '↑' means upper is better. The values in bold represent the best results among different approaches.

However, existing approaches, such as CSipMask [49], lack the ability to effectively explore temporal relations or rely on complex data-association methods that require large memory consumption, as seen in VITA [55] and IDOL [56]. In contrast, our approach utilizes an instance propagation module to

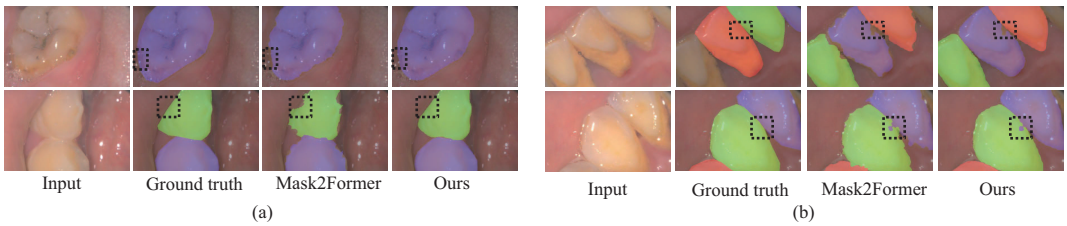

Figure 9 (Color online) Illustration of the instance segmentation results on the Shining3D tooth segmentation dataset. The 1st (left)–4th (right) columns contain input images, ground truth, results predicted by Mask2Former, and results predicted by our approach, respectively. (a) Accurate prediction results. (b) Inaccurate prediction results. Black dotted boxes highlight discrepancies produced by different approaches.

Table 4 Effectiveness comparison of fixed teeth pose estimation on the Shining3D tooth pose dataset and the Aoralscan3 tooth registration dataset[a)]

Approach	Shining3D		Aoralscan3	
	AUC of ADD-S	AUC of ADD(-S)	AUC of ADD-S	AUC of ADD(-S)
RotationNet [34]	85.2	80.1	86.0	81.0
DifRender [11]	87.0	81.9	88.1	82.8
KeyPose [57]	88.5	83.3	89.3	84.1
ODAM [35]	88.7	83.6	89.7	84.7
CosyPose [9]	89.0	83.8	90.2	84.6
Vid2CAD [36]*	89.3	84.0	90.2	85.1
Ours	**90.5**	**85.6**	**91.2**	**86.2**

a) '*' means the approach we reimplemented. The values in bold represent the best results among different approaches.

Table 5 Effectiveness comparison of jaw pose tracking on the Shining3D tooth pose dataset and the Aoralscan3 tooth registration dataset[a)]

Approach	Shining3D		Aoralscan3	
	AUC of ADD-S	AUC of ADD(-S)	AUC of ADD-S	AUC of ADD(-S)
MotionNet [3]*	81.1	78.0	81.6	78.6
AttentionTracking [8]	82.8	79.1	83.1	79.6
3DOT [6]	83.2	79.6	83.6	80.2
MaskTracking [7]*	84.2	80.4	84.6	80.8
PoseRBPF [12]	84.7	80.8	85.0	81.6
PTP [10]*	84.9	80.9	85.3	81.2
CPM [13]*	85.3	81.4	85.5	81.8
GeometricContour [32]*	85.6	81.8	86.0	82.2
SRT3D [14]	85.9	82.4	86.3	82.6
Ours	**87.3**	**84.1**	**87.4**	**84.0**

a) '*' means the approach we reimplemented. The values in bold represent the best results among different approaches.

enhance temporal relations and achieves superior performance on both datasets, as evidenced by multiple evaluation metrics.

A visual comparison of the segmentation mask produced by our approach, Mask2Former, and the corresponding ground truth is provided in Figure 9. Specifically, Figure 9(a) demonstrates accurate prediction results, showcasing the effectiveness of our approach.

4.7 Evaluation of jaw pose estimation & pose tracking

We compared several multiview 6D pose estimation methods using the Shining3D tooth pose dataset and the Aoralscan3 tooth registration dataset, and the corresponding results are presented in Table 4. Our approach, which incorporates the evaluation of neighboring teeth, shows a significant improvement in the AUC of the ADD-S, with a margin of 1.0%–1.2%.

We compared various pose tracking methods on the Shining3D tooth pose dataset and the Aoralscan3 tooth registration dataset, and the corresponding results are presented in Table 5. Contour-based methods outperform other approaches since the texture of the tooth is not strong enough to be utilized solely for alignment purposes. By imposing bidirectional constraints on camera motion, our proposed approach achieves an improvement in the AUC of ADD-S by 1.1%–1.4%.

Table 6 Effectiveness comparison of different approaches of orthodontic treatment monitoring on the Shining3D tooth pose dataset and the Aoralscan3 tooth registration dataset[a)]

Approach	Shining3D		Aoralscan3			
	AUC of ADD-S	AUC of ADD(-S)	AUC of ADD-S	AUC of ADD(-S)	GFLOPs	Seconds
IOTN [20]*	79.9	70.0	73.4	70.4	**994**	**1.5**
YOLO [17]*	81.9	72.2	75.1	72.2	1516	2.1
KBA [18]*	82.5	72.7	75.9	72.8	1955	2.4
AIRM [2]*	82.5	72.9	76.2	73.0	2042	2.5
DeepID [19]*	83.9	74.1	77.8	74.6	2650	3.1
Ours	**87.7**	**77.8**	**81.2**	**78.2**	3280	5.5

a) '*' means the approach we reimplemented. The values in bold represent the best results among different approaches.

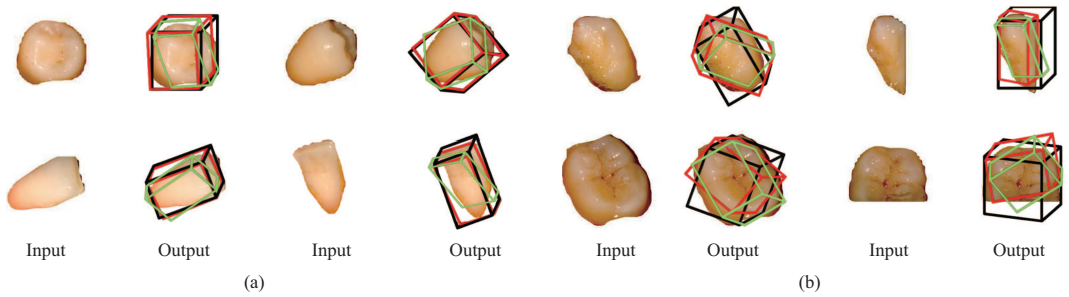

Figure 10 (Color online) Illustration of tooth alignment results on the Shining3D tooth pose dataset. (a) Accurate prediction results. (b) Inaccurate prediction results. Green, red, and black boxes represent predicted 6D poses obtained by AIRM, our approach, and the corresponding ground truth, respectively.

4.8 Evaluation of tooth alignment

To align teeth, we utilize the deepim method [45] to predict relative transformation by comparing the tooth masks of the rendered foreground image and the observed foreground image. We assess the effectiveness of various approaches for monitoring orthodontic treatment using the Shining3D tooth pose dataset and the Aoralscan3 tooth registration dataset, with the corresponding results presented in Table 6. In KBA [18] and AIRM [2], tooth instance segmentation is leveraged to provide pixel-level details for registration, outperforming object-level bounding boxes provided by the YOLO detector [17]. The experimental results indicate that our approach achieves remarkable performance, with AUCs of ADD-S of 87.7 and 81.2 on the Shining3D tooth pose dataset and the Aoralscan3 tooth registration dataset, respectively, while the AUCs of ADD(-S) in both datasets are 77.8 and 78.2, respectively.

Regarding efficiency, we evaluated the computational complexity of different approaches on the Aoralscan3 tooth registration dataset using Giga floating point operations (GFLOPs) and seconds in the inference stage. Our experimental results demonstrate that our approach achieves remarkable improvements at the cost of an increment in computational complexity.

To further demonstrate the effectiveness of our approach, Figure 10 provides some examples of the alignment of AIRM, our approach, and the corresponding ground truth. Figure 10(a) shows that our approach achieves robustness in various classes and poses of teeth, even in challenging scenarios, due to multiple components in our approach, such as temporal consistency exploration in video instance segmentation and neighboring tooth evaluation in multiview jaw pose estimation.

4.9 Discussion

Orthodontic treatment monitoring is a critical topic in both geometry analysis and medical image analysis. The goal is to measure the degree of orthodontic treatment for each tooth using the 3D jaw model constructed from the previous period and observed images in the current period. However, this task is more complex than the object 6D pose estimation, as each tooth undergoes a unique rigid deformation, while the gingiva part experiences nonrigid deformation during each period of orthodontic treatment. To address these challenges, we propose a framework that involves segmenting and registering each orthodontic tooth by aligning the jaw model in each frame as an intermediate stage.

In the first stage of our approach, we utilize both semantic knowledge and detailed position information in representation learning for video instance propagation. To ensure that the method is compatible

with different device capabilities, we constrain memory consumption. Furthermore, we employ temporal consistency to learn embeddings that maintain high similarity between instances across frames. However, as shown in Figure 9(b), there are cases where the boundary may be defective due to the weak texture of the tooth, which decreases the discrimination between neighboring teeth. To address this issue, the shape prior is a potential solution that can guide boundary extraction, for example, through explicit edge detection [58] or implicit edge modeling [59].

During the intermediate stage of our approach, we align the 3D jaw model using fixed teeth. While multiview 6D pose estimation only employs object-level information, evaluating poses in neighboring teeth is an effective way to detect and correct outliers. We can then use the supervision information of the fixed teeth to predict the jaw in the remaining frames. Bidirectional propagation is also a powerful strategy that helps to constrain the deviation in pose space.

In the final stage, the rendered tooth image is compared with the observed foreground image to iteratively predict the pose offset. However, the accuracy of this approach is highly dependent on the quality of instance segmentation in both 3D and 2D space. As shown in Figure 10(b), inaccurate segmentation masks can lead to pose errors since teeth have weak texture and shape cues are relied upon for alignment. It should be noted that rotation errors are generally greater than translation errors due to the significant influence of tooth appearance on rotation.

In the future, research will be conducted to study the shape prior to improving robustness in segmentation and registration. Additionally, new multitask learning methods [60–62] will be explored to predict translation and rotation, as these two tasks have distinct characteristics and are influenced by different factors. Furthermore, the proposed work will be expanded to measure the level of deformation in nonrigid objects, thereby extending its potential applications.

5 Conclusion

In this paper, we present an approach to measure the degree of orthodontic treatment for each individual tooth. We accomplish this by predicting the jaw pose in each frame using a combination of fixed teeth assumptions and bidirectional pose tracking constraints. Furthermore, our proposed approach has potential applications beyond orthodontics, as it can be extended to measure the degree of deformation of nonrigid objects.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61972351, 62111530300), Zhejiang Province R&D Key Project (other categories) (Grant No. 2022C03149), Special Project for Basic Business Expenses of Zhejiang Provincial Colleges and Universities (Grant No. JRK22003), and Opening Foundation of State Key Laboratory of Virtual Reality Technology and System of Beihang University (Grant No. VRLAB2023B02). The authors would like to thank AJE (www.aje.com) for its linguistic assistance during the preparation of this manuscript.

References

1 Li P, Kong D, Tang T, et al. Orthodontic treatment planning based on artificial neural networks. Sci Rep, 2019, 9: 2037
2 Hansa I, Katyal V, Semaan S J, et al. Artificial intelligence driven remote monitoring of orthodontic patients: clinical applicability and rationale. In: Proceedings of Seminars in Orthodontics, 2021. 138–156
3 Leeb F, Byravan A, Fox D. Motion-Nets: 6D tracking of unknown objects in unseen environments using RGB. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2019. 474–484
4 Xue F, Wang X, Li S, et al. Beyond tracking: selecting memory and refining poses for deep visual odometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 8575–8583
5 Hu H N, Cai Q Z, Wang D, et al. Joint monocular 3D vehicle detection and tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 5390–5399
6 Weng X, Wang J, Held D, et al. 3D multi-object tracking: a baseline and new evaluation metrics. In: Proceedings of International Conference on Intelligent Robots and Systems, 2020. 10359–10366
7 Zhong L, Zhang Y, Zhao H, et al. Seeing through the occluders: robust monocular 6-DoF object pose tracking via model-guided video object segmentation. IEEE Robot Autom Lett, 2020, 5: 5159–5166
8 Marougkas I, Koutras P, Kardaris N, et al. How to track your dragon: a multi-attentional framework for real-time RGB-D 6-DOF object pose tracking. In: Proceedings of the European Conference on Computer Vision, 2020. 682–699
9 Labbé Y, Carpentier J, Aubry M, et al. CosyPose: consistent multi-view multi-object 6D pose estimation. In: Proceedings of the European Conference on Computer Vision, 2020. 574–591
10 Weng X, Yuan Y, Kitani K. PTP: parallelized tracking and prediction with graph neural networks and diversity sampling. IEEE Robot Autom Lett, 2021, 6: 4640–4647
11 Shugurov I, Pavlov I, Zakharov S, et al. Multi-view object pose refinement with differentiable renderer. IEEE Robot Autom Lett, 2021, 6: 2579–2586
12 Deng X, Mousavian A, Xiang Y, et al. PoseRBPF: a rao-blackwellized particle filter for 6-D object pose tracking. IEEE Trans Robot, 2021, 37: 1328–1342
13 Sun X, Zhou J, Zhang W, et al. Robust monocular pose tracking of less-distinct objects based on contour-part model. IEEE Trans Circuits Syst Video Technol, 2021, 31: 4409–4421

14 Stoiber M, Pfanne M, Strobl K H, et al. SRT3D: a sparse region-based 3D object tracking approach for the real world. Int J Comput Vis, 2022, 130: 1008–1030
15 Perillo L, d'Apuzzo F, Illario M, et al. Monitoring biochemical and structural changes in human periodontal ligaments during orthodontic treatment by means of micro-Raman spectroscopy. Sensors, 2020, 20: 497
16 Moylan H B, Carrico C K, Lindauer S J, et al. Accuracy of a smartphone-based orthodontic treatment-monitoring application: a pilot study. Angle Orthod, 2019, 89: 727–733
17 Talaat S, Kaboudan A, Talaat W, et al. The validity of an artificial intelligence application for assessment of orthodontic treatment need from clinical images. In: Proceedings of Seminars in Orthodontics, 2021. 164–171
18 Caruso S, Caruso S, Pellegrino M, et al. A knowledge-based algorithm for automatic monitoring of orthodontic treatment: the dental monitoring system. Two cases. Sensors, 2021, 21: 1856
19 Li S, Guo Z, Lin J, et al. Artificial intelligence for classifying and archiving orthodontic images. Biomed Res Int, 2022, 2022: 1473977
20 Murata S, Ishigaki K, Lee C, et al. Towards a smart dental healthcare: an automated assessment of orthodontic treatment need. In: Proceedings of HealthInfo, 2017. 35–39
21 Tian Y, Gelernter J, Wang X, et al. Traffic sign detection using a multi-scale recurrent attention network. IEEE Trans Intell Transp Syst, 2019, 20: 4466–4475
22 Tian Y, Wang X, Wu J, et al. Multi-scale hierarchical residual network for dense captioning. J Artif Intell Res, 2019, 64: 181–196
23 Liu D, Tian Y, Zhang Y, et al. Heterogeneous data fusion and loss function design for tooth point cloud segmentation. Neural Comput Applic, 2022, 34: 17371–17380
24 Wang Y, Xu Z, Wang X, et al. End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 8741–8750
25 Wu J, Jiang Y, Bai S, et al. SeqFormer: sequential transformer for video instance segmentation. In: Proceedings of the European Conference on Computer Vision, 2022. 553–569
26 Ke L, Ding H, Danelljan M, et al. Video mask transfiner for high-quality video instance segmentation. In: Proceedings of the European Conference on Computer Vision, 2022. 474–491
27 Cheng B, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 1290–1299
28 Cheng B, Choudhuri A, Misra I, et al. Mask2Former for video instance segmentation. 2021. ArXiv:2112.10764
29 Hwang S, Heo M, Oh S W, et al. Video instance segmentation using inter-frame communication transformers. In: Proceedings of the Advances in Neural Information Processing Systems, 2021. 13352–13363
30 Tian Y, Hu W, Jiang H, et al. Densely connected attentional pyramid residual network for human pose estimation. Neurocomputing, 2019, 347: 13–23
31 Wang B, Zhong F, Qin X. Robust edge-based 3D object tracking with direction-based pose validation. Multimed Tools Appl, 2019, 78: 12307–12331
32 Li J, Song X, Zhong F, et al. Fast 3D texture-less object tracking with geometric contour and local region. Comput Graphics, 2021, 97: 225–235
33 Li C, Bai J, Hager G D. A unified framework for multi-view multi-class object pose estimation. In: Proceedings of the European Conference on Computer Vision, 2018. 254–269
34 Kanezaki A, Matsushita Y, Nishida Y. RotationNet for joint object categorization and unsupervised pose estimation from multi-view images. IEEE Trans Pattern Anal Mach Intell, 2019, 43: 269–283
35 Li K, DeTone D, Chen Y F S, et al. ODAM: object detection, association, and mapping using posed RGB video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 5998–6008
36 Maninis K K, Popov S, Niesner M, et al. Vid2CAD: CAD model alignment using multi-view constraints from videos. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 1320–1327
37 Sun J, Wang Z, Zhang S, et al. OnePose: one-shot object pose estimation without CAD models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 6825–6834
38 Park K, Mousavian A, Xiang Y, et al. LatentFusion: end-to-end differentiable reconstruction and rendering for unseen object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 10710–10719
39 Tyszkiewicz M J, Maninis K K, Popov S, et al. RayTran: 3D pose estimation and shape reconstruction of multiple objects from videos with ray-traced transformers. 2022. ArXiv:2203.13296
40 Kaskman R, Shugurov I, Zakharov S, et al. 6 DOF pose estimation of textureless objects from multiple RGB frames. In: Proceedings of the European Conference on Computer Vision, 2020. 612–630
41 Liu S, Li F, Zhang H, et al. DAB-DETR: dynamic anchor boxes are better queries for DETR. In: Proceedings of International Conference on Learning Representations, 2022. 998–1008
42 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
43 Su Y, Saleh M, Fetzer T, et al. ZebraPose: coarse to fine surface encoding for 6DoF object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 6738–6748
44 Tian Y, Zhang Y, Chen W G, et al. 3D tooth instance segmentation learning objectness and affinity in point cloud. ACM Trans Multimedia Comput Commun Appl, 2022, 18: 1–16
45 Li Y, Wang G, Ji X, et al. DeepIM: deep iterative matching for 6D pose estimation. In: Proceedings of the European Conference on Computer Vision, 2018. 683–698
46 Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the Advances in Neural Information Processing Systems, 2019
47 Tian Y, Cheng G, Gelernter J, et al. Joint temporal context exploitation and active learning for video segmentation. Pattern Recognition, 2020, 100: 107158
48 Tian Y, Zhang Y, Zhou D, et al. Triple attention network for video segmentation. Neurocomputing, 2020, 417: 202–211
49 Qi J, Gao Y, Hu Y, et al. Occluded video instance segmentation: a benchmark. Int J Comput Vis, 2022, 130: 2022–2039
50 Ke L, Li X, Danelljan M, et al. Prototypical cross-attention networks for multiple object tracking and segmentation. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 1192–1203
51 Yang S, Fang Y, Wang X, et al. Crossover learning for fast online video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 8043–8052

52 Caelles A, Meinhardt T, Brasó G, et al. DeVIS: making deformable transformers work for video instance segmentation. 2022. ArXiv:2207.11103
53 Koner R, Hannan T, Shit S, et al. InstanceFormer: an online video instance segmentation framework. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2023
54 Huang D A, Yu Z, Anandkumar A. MinVIS: a minimal video instance segmentation framework without video-based training. In: Proceedings of Conference on Neural Information Processing Systems, 2022. 1766–1774
55 Heo M, Hwang S, Oh S W, et al. VITA: video instance segmentation via object token association. In: Proceedings of Conference on Neural Information Processing Systems, 2022. 766–774
56 Wu J, Liu Q, Jiang Y, et al. In defense of online models for video instance segmentation. In: Proceedings of the European Conference on Computer Vision, 2022. 588–605
57 Liu X, Jonschkowski R, Angelova A, et al. KeyPose: multi-view 3D labeling and keypoint estimation for transparent objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 11602–11610
58 Tian Y, Wang H, Wang X. Object localization via evaluation multi-task learning. Neurocomputing, 2017, 253: 34–41
59 Wang P, Tian Y, Liu N, et al. A tooth surface design method combining semantic guidance, confidence, and structural coherence. IET Comput Vision, 2022, 16: 727–735
60 Tian Y, Gelernter J, Wang X, et al. Lane marking detection via deep convolutional neural network. Neurocomputing, 2018, 280: 46–55
61 Liu D, Tian Y, Xu Z, et al. Handling occlusion in prohibited item detection from X-ray images. Neural Comput Applic, 2022, 34: 20285–20298
62 Wang B, Tian Y, Wang J, et al. Detect occluded items in X-ray baggage inspection. Comput Graphics, 2023, 115: 148–157