Visual Intelligence

RESEARCH Open Access

A full-set tooth segmentation model based on improved PointNet++

Li Yuan[1†], Xinyi Liu[1†], Jiannan Yu[2†] and Yanfeng Li[2*]

Abstract

Segmentation of a complete set of teeth from three-dimensional (3D) intra-oral scanner images is a crucial step in tooth identification procedures. In large-scale disasters with many victims, teeth are often the preferred and reliable source for victim identification due to their hard and non-deformable characteristics. In this paper we present a study on the automatic segmentation of a complete set of teeth from intra-oral scanner images. We propose a tooth segmentation method based on an improved PointNet++ architecture. To address the problem of inadequate segmentation capability of the teeth-gingival boundary of PointNet++, we introduce a single-point preliminary feature extraction (SPFE) module to better preserve the subtle details that may be overlooked by the original PointNet++ model. In addition, a weighted-sum local feature aggregation (WSLFA) mechanism is proposed to replace the max pooling in PointNet++ to better perform feature aggregation. The experimental results on 52 testing datasets using the network trained on 160 annotated 3D intra-oral scanner images demonstrate that our improved PointNet++ method achieves a segmentation accuracy of 97.68%, and performs well under different dental conditions.

Keywords: 3D intra-oral scanning (IOS), Tooth segmentation, 3D point cloud, PointNet++

1 Introduction

Three-dimensional (3D) intra-oral scanning (IOS) is a small-sized optical scanning technology that allows clinicians to use digital intro-oral scanners to obtain relevant information about teeth, mucosa, and the associated soft and hard tissues, generating a 3D model of the oral cavity. It is commonly used to assist in oral examinations, teeth alignment, restoration, and treatment. Compared to cone beam computerized tomography (CBCT), 3D IOS has many advantages, such as no radiation exposure and easy acquisition. With the development of medical technology and peoples' increasing attention to their oral health, the IOS technology is being widely used by orthodontists for significantly improving treatment efficiency in modern dentistry.

The tooth part segmented from the intra-oral scanner images serves as a personalized structure that can be used for personal identification. Tooth identification is of great significance in the identification of victims of natural disasters or crimes because teeth, as one of the hardest tissues in the human body, are not easily deformed, highly individualized and can be well-preserved after severe disasters or violent crimes. The percentage of identified victims using tooth identification methods in some large-scale disasters ranges from 60.63% to 100% [1]. As soft tissues such as gums are prone to deformation and decay, tooth identification requires tooth segmentation technology to accurately segment the entire set of teeth from the intra-oral scanner images. Therefore, we study the accurate segmentation of the entire set of teeth from the intra-oral scanner images for tooth identification purposes.

Automated segmentation of teeth from intra-oral scanner images is a challenging task due to the complex boundary between teeth and gingiva, as well as the significant variations in tooth shapes and appearances among differ-

*Correspondence: m.god@yeah.net
[2] Department of Stomatology, the Fourth Medical Center, Chinese PLA General Hospital, Beijing, China
Full list of author information is available at the end of the article †Equal contributors

Springer

ent subjects, such as missing or misaligned teeth. Early tooth segmentation methods often relied on hand-crafted features, including curvature-based methods [2], skeleton-based methods [3], and harmonic field-based methods [4]. However, these methods lack robustness and are difficult to adapt to the diverse tooth arrangements of different individuals, often requiring human interaction to complete the segmentation. With the development of 3D deep learning techniques, many deep learning-based tooth segmentation methods have been proposed. One approach is to transform the unorganized 3D intra-oral scanner images (point cloud or mesh data) into two-dimensional (2D) images [5] or octree grids [6], and then use 2D or 3D convolutional neural network (CNN) for segmentation. However, these methods generate additional computational load and cause some information loss due to the conversion of data. Another approach is to directly apply deep learning networks to point cloud or mesh models for segmentation. PointNet [7] and PointNet++ [8] are representative methods for point cloud segmentation that use multi-layer perceptron (MLP) and max pooling for feature extraction. To extract features at different scales, PointNet++ also employs a multi-scale local feature extraction strategy. However, PointNet++ uses a strategy where the point cloud model is divided into overlapping local regions, and the most distinctive features within each region are extracted using max pooling. This approach may not accurately capture important features at the gingival boundary of each individual tooth, leading to a coarse segmentation result. Lian et al. [9] designed MeshSegNet for tooth segmentation based on the mesh structure of 3D intra-oral scanner images. This model uses a graph neural network (GNN) to process the mesh structure, which operates on the graph representation of the mesh. However, MeshSegNet does not have the encoder-decoder structure of PointNet++, which means that the resolution of the input mesh model is not compressed throughout the network and leads to a higher number of parameters compared with PointNet++. Simplification of the input mesh model is usually necessary for this network to be used. Some studies attempt to simplify the structure of such networks as CNN and Transformer. Li et al. [10] used dynamic networks to reduce computational redundancy by automatically adjusting their architectures for different inputs, and they made further improvements to dynamic networks by pre-defining dense weight slices of varying importance in a dynamic super-net using nested residual learning.

In this paper, we propose an improved network structure based on PointNet++ for the full-set tooth segmentation of 3D intra-oral scanner images. To address the problem of inadequate segmentation capability of the teeth-gingival boundary of PointNet++, a single-point preliminary feature extraction (SPFE) module is added to better preserve the subtle details that may be overlooked by the original PointNet++ model. In addition, inspired by Li et al. [10] using dynamic weights to adjust network architectures, we use dynamic weights to aggregate features and propose a weighted-sum local feature aggregation (WSLFA) mechanism to replace the max pooling in PointNet++, thus enabling better feature aggregation. The proposed method can achieve an accuracy of 97.68% for tooth segmentation.

2 Related works

2.1 Point cloud deep learning

A point cloud is a collection of points in space used to represent a 3D shape. Due to the unordered and non-structural nature of point clouds, it is difficult to directly apply standard CNNs in the task of tooth segmentation. The PointNet series utilizes symmetry operations to handle the disorder and non-structure of point clouds for classification and segmentation tasks. Specifically, PointNet [7] made groundbreaking work by using MLP to extract features from each point and aggregating features using max pooling. Since MLP and max pooling are both symmetric operations, they help to handle the permutation invariance of point clouds. PointNet++ [8] divides the point cloud into hierarchical groups and uses the same MLP and max pooling as PointNet does to extract features at different levels. Features learned from multiple scales and layers are combined to obtain better robustness. Other methods attempt to apply convolution on point clouds, such as PointCNN [11], which uses MLP to learn a transformation matrix, normalizes the point cloud with this matrix, and then extracts features using CNN. In addition, some graph-based methods treat each point in a point cloud as a vertex in a graph and establish edges between these vertices to create a graph structure. For example, edge conditioned convolution (ECC) [12] performs convolution-like operations on graph-structured data in spatial domains. DGCNN [13] constructs directed graphs in both the original point cloud and feature space, and dynamically updates features after each layer in the network. EdgeConv, which was proposed in DGCNN, captures local geometric structures and is dynamically implemented in each layer of the network. Furthermore, to improve performance and reduce model size, LDGCNN [14] removes the transformation network learned from different layers in DGCNN and links hierarchical features learned from different layers in DGCNN to improve performance and reduce model size.

2.2 3D intra-oral scanner images segmentation

The traditional method for segmenting 3D intra-oral scanner images usually involves pre-defining geometric standards to separate teeth from the intra-oral scanner images. For example, Zou et al. [4] used a harmonic field defined on a triangular mesh to iteratively annotate teeth on the tooth surface model. Kumar et al. [2] adopted curvature to segment teeth. Wu et al. [3] defined the morphologic skeleton of the scanned teeth grid and used region growing operations to segment teeth from the intra-oral scanner images iteratively. Although these methods are intuitive, they typically depend on expert prior knowledge and require tedious manual operations, leading to sensitivity to changes in surface appearance. To fully automate tooth segmentation and improve segmentation robustness, an increasing number of deep learning methods are being applied to precise segmentation of teeth from 3D intra-oral scanner images. Xu et al. [5] developed a two-stage tooth segmentation model that includes teeth-gingival segmentation and inter-teeth segmentation. The method first extracts 600-dimensional geometric features (coordinates, curvature, principal component analysis (PCA), etc.) for each facet of intra-oral scanner images and packs them into a 20 × 30 image, and then performs segmentation using the two-stage CNN network. However, this method ignores the disorder and different packing orders of the hand-designed geometric features, which affects the segmentation results. Tian et al. [6] applied sparse octree methods to voxelize unordered 3D meshes and then used 3D CNN for tooth segmentation, but voxelization can cause loss of model information. Lian et al. [9] designed MeshSegNet, which uses the characteristics of mesh models to combine PointNet with graphs and a multi-scale graph-constrained learning module for simulating CNN multi-scale feature extraction. Li et al. [15] established a multi-scale bilateral enhancement network and adopted a bilateral enhancement module for multi-scale feature extraction. However, these two methods produce a large number of model parameters. As highly accurate intra-oral scanner images may have a large number of mesh grids, simplification of the input mesh model is usually necessary. Other scholars have used instance segmentation methods to segment individual teeth to avoid the problem of uncertain semantic numbers caused by different numbers of teeth. For example, Zanjani et al. [16] presented Mask-MCNet, which for the first time applies instance segmentation to 3D intra-oral scanner images. The network first predicts 3D bounding boxes of teeth, and then performs segmentation of the points that belong to each individual tooth instance. Tian et al. [17] introduced a point cloud-based 3D tooth instance segmentation method and used an instance-aware module based on attention mechanisms to extract local and global features to better distinguish different tooth instances. Cui et al. [18] proposed TSegNet, which represents tooth segmentation as two sub-problems: tooth centroid prediction and individual tooth segmentation in order to segment 3D tooth models quickly and accurately. These works segment single tooth from intra-oral scanner images, and the segmented models are primarily used for orthodontics, dental diagnosis, etc. Multi-modal learning such as utilizing visual content from videos in unsupervised machine translation [19] has also been applied to tooth segmentation. Jang et al. [20] used both 2D and 3D images for tooth segmentation and developed a hierarchical multi-step model that first generates regions of interest from 2D images and then performs segmentation on 3D models.

In order to facilitate the use of teeth for identification in forensic medicine, our paper aims to segment the full set of teeth from the intra-oral scanner images to retain the holistic identification features. Therefore, we do not conduct experiments on the segmentation of a single tooth, but instead conduct experiments on the segmentation of the full set of teeth.

3 Full-set tooth segmentation model based on improved PointNet++

The network structure of the proposed model is illustrated in Fig. 1. Similar to PointNet++, our network has an encoder-decoder structure. In the encoder part, the input intra-oral scanner images are gradually down-sampled and local features are extracted, with the block responsible for down-sampling and local feature extraction called the set abstraction (SA) layer. In the decoder part, up-sampling is performed to restore the original model resolution, with the block responsible for up-sampling and feature back-propagation called the feature propagation (FP) layer. In this paper, we propose the following two improvements based on PointNet++:

1) A single-point preliminary feature extraction (SPFE) module is added to address the problem of directly extracting local region features and ignoring subtle details in PointNet++, allowing detailed information to be better preserved.

2) A weighted-sum local feature aggregation (WSLFA) mechanism is proposed to better balance the fusion of various useful information in the local region, and to retain important features of teeth-gingival boundaries that are more useful for segmentation, which receive better preservation under the proposed aggregation mechanism.

Let N be the number of points in the input intra-oral scanner images. Before down-sampling, a SPFE module is applied to extract $N \times 64$ dimensional features. The encoder part includes three SA layers. As demonstrated in Fig. 2, the SA layer first constructs local regions, which include down-sampling N_i center points on the basis of the previous layer and constructing a spherical neighborhood with a radius of r for each center point. Then, for each point in the local region, the network learns a feature vector and a weight. These feature vectors and weights are then used to obtain a weighted sum of the features for all points in the region, resulting in a global feature that represents the entire spherical region. After one SA layer, the N_i feature vectors obtained are sent to the next SA layer for further down-sampling and feature extraction, including three SA layers in total. The decoder part includes three

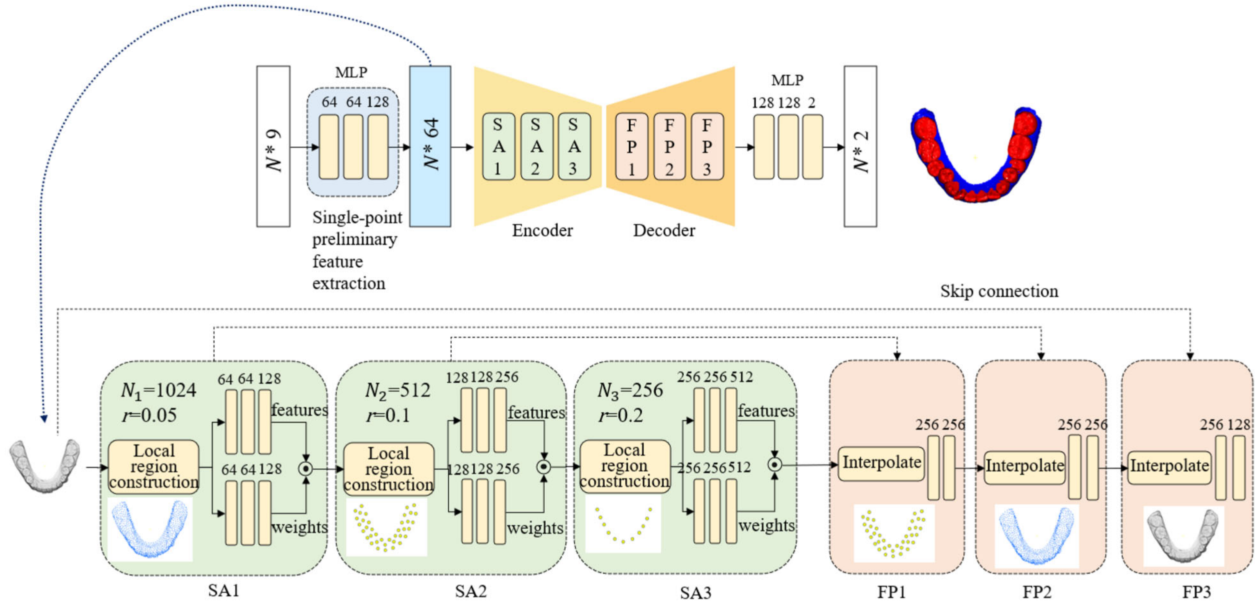


Figure 1 Network structure of the improved PointNet++. SA represents set abstraction, FP represents feature propagation, and MLP denotes multiple layer perceptron

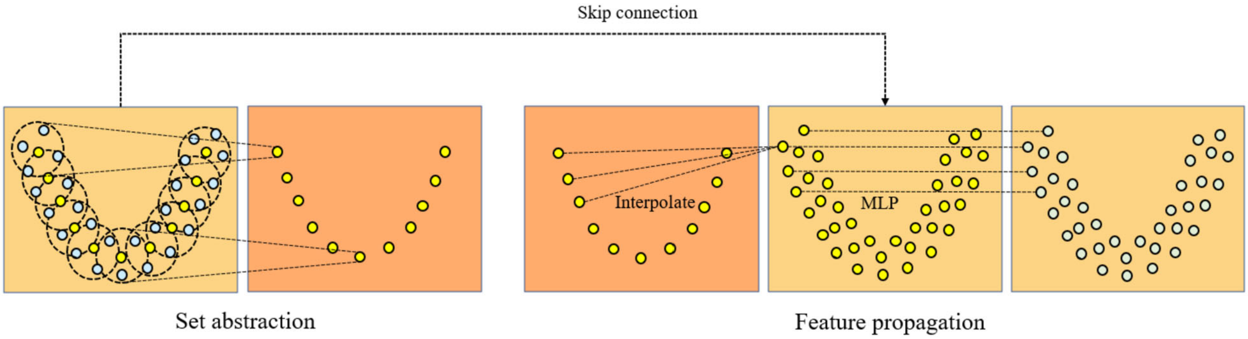


Figure 2 Structure of the set abstraction (SA) layer and the feature propagation (FP) layer. MLP denotes multiple layer perceptron

FP layers to gradually reconstruct the original number of points of the input model. As presented in Fig. 2, the FP layer first interpolates to restore the point number of the previous SA layer, and then makes a skip connection with this SA layer. Next, MLP is applied to learn a new feature vector that is sent to the next FP layer. Finally, MLP compresses the feature dimension to two categories (teeth and gingiva), outputs the probability of each category, and predicts the category label for each point after restoring the original number of points.

3.1 Network input and single-point preliminary feature extraction

The input of the network is the point cloud data of the intra-oral scanner images, which can be represented as an $N \times 9$ matrix. Here, N represents the number of points in the point cloud model, and each point is represented by a 9-dimensional vector, including 3D coordinates, 3D normal vectors, and 3D zero-mean coordinates. In PointNet++, the input point cloud model directly enters the SA layer for down-sampling and extracting features representing local regions. This approach ignores the detailed information of the point cloud, which reduces the accuracy of PointNet++ in tooth segmentation. Therefore, we add a SPFE module before the SA layer. In this module, the 9-dimensional vector of each point is sent to the MLP for preliminary feature extraction, and a new 64-dimensional feature vector is obtained, which then enters the SA layer for local feature extraction. Due to the extraction of 64 dimensional features from each point, more detailed information can be mined. Every point in our method contributes to the segmentation while only the points that are sampled and in the designated local areas are used in PointNet++; thus our method solves the problem of inaccurate segmentation in detailed area.

To visualize the results of SPFE, the 64-dimensional features extracted are reduced to one-dimensional features by PCA. Figure 3(a) and Fig. 3(b) show two feature maps of different intra-oral scanner images after dimension reduction. It is noted that the feature maps indicate that the SPFE module retains more detailed information of the teeth-gingival boundary, dental groove and dental gap, thus improving the network's ability to extract features from point clouds. Moreover, the SPFE module makes preliminary distinctions between different regions of teeth (Fig. 3, the front teeth are red and the upper jaw is blue), which is beneficial for subsequent segmentation.

Figure 4(a) demonstrates the features extracted by the SA1 layer after using the SPFE module, while Fig. 4(b) shows the features extracted by the SA1 layer directly using the original 9-dimensional vector. Similarly, these 128-dimensional features are visualized after PCA dimensional reduction to one dimension. It is observed that the features of teeth and gingiva after using the SPFE module are more distinguishable compared with directly using the original 9-dimensional vector, so the SPFE module can enhance the effectiveness of subsequent SA1 layer feature extraction.

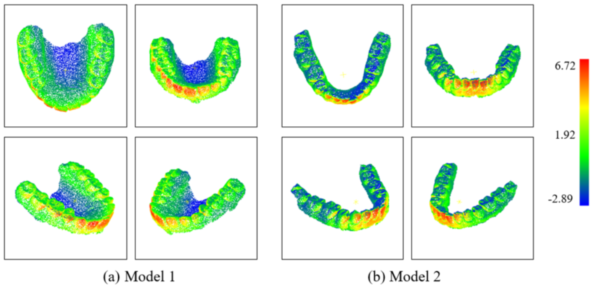


Figure 3 Visualization of feature maps of the SPFE module

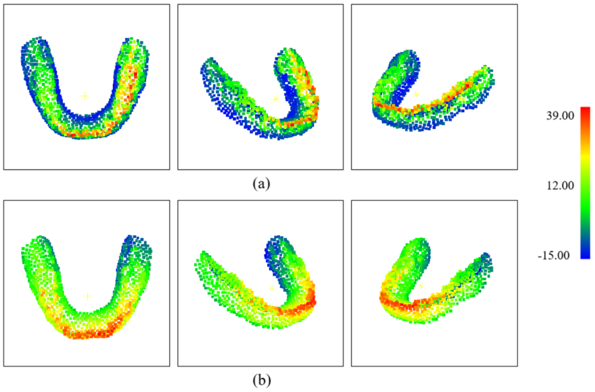


Figure 4 Comparison of feature maps of the SA1 layer: (a) Using SPFE; (b) Directly using the original 9-dimensional vector

3.2 Local region construction

In the SA layer, local region construction is first performed to divide the point cloud into overlapping local regions as presented in Fig. 5, preparing for subsequent feature extraction. The input point cloud coordinates are downsampled to obtain the center point of each local region; then, a sphere with a certain radius is constructed around these points. The number of sampled points N_i and the sphere radius r in each SA layer are adjustable parameters. The sampling algorithm used is farthest point sampling (FPS) [6], ensuring that the sampling points are uniformly distributed. The sampled center points and their spherical neighborhoods constitute a local region, and representative features are extracted for each region.

3.3 Local feature extraction and weighted-sum local feature aggregation

The local feature extraction and aggregation module is illustrated in Fig. 6. Let N be the number of input points. After local region construction, the set of center points $P_{\text{center}} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots \boldsymbol{p}_{N'}\}$ contains a total of N' sampled center point coordinates, where each center point $\boldsymbol{p}_i$ has a neighborhood point set $P^i_{\text{local}} = \{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}, \ldots \boldsymbol{p}_{ik}\}$ consisting of k neighboring point coordinates. Each point $\boldsymbol{p}_{ij}$ has a feature vector $\boldsymbol{f}_{ij}$ extracted from the previous layer, where the coordinate dimension is d and the feature dimension is C. The input size of this module is $N' \times k \times (d + C)$. Local feature extraction first extracts C'-dimensional new

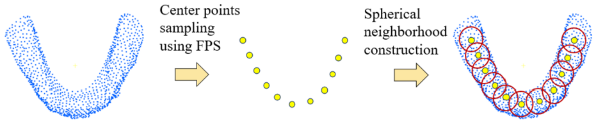


Figure 5 Local region construction. FPS represents farthest point sampling

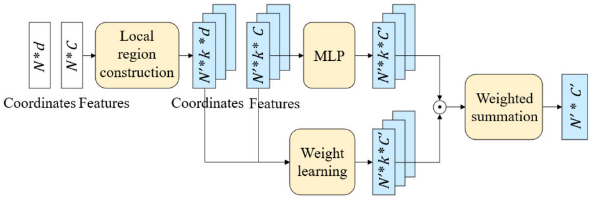


Figure 6 Local feature extraction and aggregation using the weighted-sum local feature aggregation (WSLFA) mechanism. MLP denotes multiple layer perceptron

features $\boldsymbol{f}'_{ij}$ for each point $\boldsymbol{p}_{ij}$ in the set P^i_{local}:

$$\boldsymbol{f}'_{ij} = \mathbf{MLP}\big((\boldsymbol{p}_{ij} - \boldsymbol{p}_i) \oplus \boldsymbol{f}_{ij}\big), \tag{1}$$

where $\boldsymbol{p}_{ij} - \boldsymbol{p}_i$ represents the neighborhood point coordinates minus the corresponding center point coordinates, that is, in each spherical neighborhood, the coordinates of the points are standardized relative to the center point.

After extracting features $\boldsymbol{f}'_{ij}$, it is necessary to aggregate these point features into global features representing the local regions. The PointNet series adopts the method of max pooling; however, max pooling can only capture the most distinctive feature in the region and cannot retain more internal details of the region. Therefore, we propose an adaptive method to learn the weight of each feature in a sub-network, and then perform weighted summation, called weighted-sum local feature aggregation named WSLFA, thereby preserving the internal details of the region. Our method can adaptively adjust weights during the training process, effectively weighting the features of each part based on their contribution to segmentation. By comparison, PointNet++ does not distinguish the contribution of each part's features, thus causing the effective features to be ignored and resulting in poor segmentation results.

Our method first uses the coordinates of each point $\boldsymbol{p}_{ij}$ in the local region and its learned new feature $\boldsymbol{f}'_{ij}$ to learn a weight vector $\boldsymbol{\alpha}_{ij}$ of the same dimension as $\boldsymbol{f}'_{ij}$:

$$\boldsymbol{\alpha}_{ij} = \mathbf{MLP}\big((\boldsymbol{p}_{ij} - \boldsymbol{p}_i) \oplus (\boldsymbol{f}'_{ij} - \boldsymbol{f}'^{\text{mean}}_i)\big), \tag{2}$$

where, $\boldsymbol{p}_{ij} - \boldsymbol{p}_i$ represents the coordinate difference between the neighbor points and the corresponding center point. $\boldsymbol{f}'^{\text{mean}}_i$ is the mean of all features $\boldsymbol{f}'_{ij}$ in the region, that is:

$$\boldsymbol{f}'^{\text{mean}}_i = \frac{\sum_{j=1}^{k} \boldsymbol{f}'_{ij}}{k}. \tag{3}$$

The global feature $\boldsymbol{f}'_i$ of the region is obtained by weighted summation of the point features $\boldsymbol{f}'_{ij}$ and weight vectors $\boldsymbol{\alpha}_{ij}$, expressed as:

$$\boldsymbol{f}'_i = \sum_{j=1}^{k} \boldsymbol{\alpha}_{ij} \odot \boldsymbol{f}'_{ij}, \tag{4}$$

where $\boldsymbol{f}'_i$ is the Hadamard product of the weight vectors and the point features.

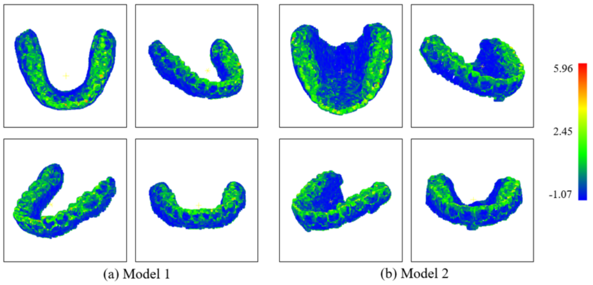


Figure 7 Weight vectors learned in the SA1 layer using the weighted-sum local feature aggregation (WSLFA) mechanism

The weight vectors learned by the aforementioned process in the SA1 layer is illustrated in Fig. 7. The 128-dimensional weight vectors are shown after PCA dimensionality reduction to one dimension, and Fig. 7(a) and Fig. 7(b) show two different intra-oral scanner images. Weights at the teeth-gingival boundary are larger (green), highlighting the features of the finer details. The larger weights of dental crown also enhance the contrast between teeth and gingival features. This helps to better retain useful feature information.

After one layer of local feature extraction and aggregation, a global feature is extracted for each local region, resulting in N' feature vectors representing different regions. These features and N' center point coordinates $P_{\text{center}} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots \boldsymbol{p}_{N'}\}$ form a new point cloud with N' points of the size $N' \times (d + C')$, which will be used for the next layer of local feature extraction and aggregation. Our network includes a total of three layers of local feature extraction and aggregation.

3.4 Feature backpropagation

In the feature backpropagation stage, the locally aggregated features are gradually restored to the original size of the point cloud for segmentation prediction. This includes three steps: interpolation, skip connection, and MLP. The first step is to restore the output point number of the $(l-1)$-th SA layer from the l-th SA layer through interpolation. Let the original point set be $\boldsymbol{P}_l$, and the restored point set be $\boldsymbol{P}_{l-1}$. Each point in $\boldsymbol{P}_l$ contains a 3D coordinate $\boldsymbol{p}^l_i$ and a feature vector $\boldsymbol{f}^l_i$, and the restored coordinate $\boldsymbol{p}^{l-1}_i$ is the same as the coordinate of the $(l-1)$-th SA layer. The restored feature $\boldsymbol{f}^{l-1}_i$ can be represented as the weighted average of the features of its three nearest original points:

$$\boldsymbol{f}^{l-1}_i = \sum_{j=1}^{3} \frac{1/\|\boldsymbol{p}^{l-1}_i - \boldsymbol{p}^l_j\|}{\sum_{j=1}^{3}(1/\|\boldsymbol{p}^{l-1}_i - \boldsymbol{p}^l_j\|)} \boldsymbol{f}^l_j. \tag{5}$$

Feature $\boldsymbol{f}^{l-1}_i$ obtained after interpolation is concatenated with the feature obtained from the $(l-1)$-th SA layer through skip connection, and the concatenated features obtained are then compressed by MLP to reduce the feature dimension. The above mentioned operations are repeated until the original number of points N is restored. Finally, the $N \times 2$ segmentation prediction score matrix is output by MLP, which predicts the probabilities of teeth and gingiva categories, and the maximum predicted probability is selected as the final segmentation category for each point.

3.5 Loss function

The loss function used in our model is the negative log-likelihood (NLL) loss function. When the model outputs the probability distribution of two classes (teeth and gingiva) for each point, the probability distribution is used to measure the difference between the predicted results and true labels. Specifically, it measures the error by taking the negative log of the probability of the true class label for each point, and averaging the negative log errors within a batch. NLL loss for one point can be expressed as:

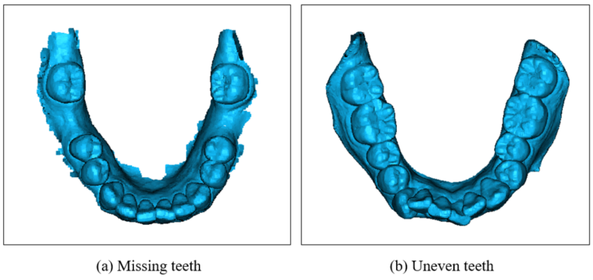
$$NLL = \frac{1}{batch\_size} \sum_{i=1}^{batch\_size} -a \log(p_1) - (1-a) \log(p_2), \tag{6}$$

where p_1 and p_2 represent the probabilities of the point being teeth or gingiva, respectively. a takes the value of 0 or 1, where $a = 1$ means the true label of the point is teeth and $a = 0$ means the true label is gingiva. When the predicted values and the true labels are not consistent, the corresponding probability will be small, resulting in a larger negative log probability for that class, thus increasing the NLL loss value. Therefore, by minimizing the NLL loss value, the model can predict the labels of the input samples more accurately.

4 Experimental results and analysis

4.1 Datasets

Our experimental data consist of 212 3D intra-oral scanner images that are manually labeled as either teeth or gingiva. A total of 160 of these examples are used for training, while the remaining 52 are used for testing. Specifically, we select 13 intra-oral scanner images with poor dental conditions (missing teeth, uneven dentition, etc.) as shown in Fig. 8 and use a total of 52 testing datasets to discuss the generalization and robustness of our method. Each model is sampled to contain 32,768 points, with each point containing 3D positional information (x, y, z) and a corresponding 3D normal vector. Additionally, the training data are augmented with the following operations: (1) random rotations and (2) random translations in coordinates. Each training example undergoes these two operations before participating in network training.


(a) Missing teeth

(b) Uneven teeth

Figure 8 Intra-oral scanner images with poor dental condition

4.2 Implementation details

The network is implemented using PyTorch, with a GPU version of Tesla V100 and an Ubuntu operating system. The Adam optimizer is used during training, with the NLL loss function and an initial learning rate of 0.001. The learning rate is reduced by a factor of 0.7 every 20 epochs, with a minimum learning rate of 0.00001. The batch size is set to 4 during training, and the network is trained for a total of 100 epochs.

4.3 Experimental results

This section includes two experiments. Section 4.3.1 tests the effectiveness of our method on datasets with different dental conditions and compares it with other methods. Section 4.3.2 tests the effectiveness of our method under different sampling points compared with PointNet++.

4.3.1 Comparison of experimental results with other methods

Table 1 presents the experimental results of our method and other methods (PointNet, PointNet++ and PointCNN) on the whole dataset, while Table 2 shows the experimental results on the dataset with only poor tooth conditions. The training loss curve is demonstrated in Fig. 9. The number of points sampled for each layer in the model is 1024/512/256, and the radius of the spherical neighborhood for each layer is 0.05, 0.1, and 0.2 (normalized), respectively. Compared with PointNet++, our method achieves significantly higher segmentation accuracy and mean intersection over union (mIoU), and the loss curve decreases more rapidly. Our method performs equally well on datasets with different dental conditions, demonstrating its robustness. The visualization of the segmentation results with different dental conditions is presented in Fig. 10, and our method achieves more accurate segmentation of the teeth-gingival boundary with smoother boundary curves regardless of the condition of the teeth. Due to the effects of SPFE and WSLFA, our method significantly eliminates jagged edges and mis-segmentation compared with the original PointNet++.

Although our method has achieved good segmentation results on most of the teeth-gingival boundaries, it struggles to predict the wisdom teeth. This is because the degree of wisdom tooth germination varies among individuals, and the intra-oral scanner images are more blurry and may be incomplete in the area of wisdom teeth and

Table 1 Experimental results of PointNet, PointCNN and PointNet++ on the whole dataset

Model	Accuracy	mIoU	IoU		Precision		Recall		*F*1	
			Teeth	Gingiva	Teeth	Gingiva	Teeth	Gingiva	Teeth	Gingiva
PointNet	0.9274	0.8644	0.863	0.866	0.936	0.919	0.918	0.937	0.927	0.928
PointCNN	0.9390	0.8856	0.882	0.889	0.966	0.915	0.910	0.968	0.937	0.941
PointNet++	0.9585	0.9201	0.919	0.921	0.969	0.949	0.947	0.970	0.958	0.959
Our method	0.9768	0.9547	0.954	0.955	0.984	0.970	0.969	0.985	0.976	0.977

Table 2 Experimental results of PointNet, PointCNN and PointNet++ on the dataset with poor dental condition

Model	Accuracy	mIoU	IoU		Precision		Recall		*F*1	
			Teeth	Gingiva	Teeth	Gingiva	Teeth	Gingiva	Teeth	Gingiva
PointNet	0.9259	0.8644	0.859	0.865	0.946	0.908	0.904	0.948	0.925	0.928
PointCNN	0.9258	0.8615	0.855	0.868	0.971	0.888	0.878	0.974	0.922	0.929
PointNet++	0.9549	0.9139	0.913	0.915	0.967	0.943	0.942	0.968	0.954	0.955
Our method	0.9735	0.9586	0.948	0.949	0.983	0.964	0.964	0.983	0.973	0.973

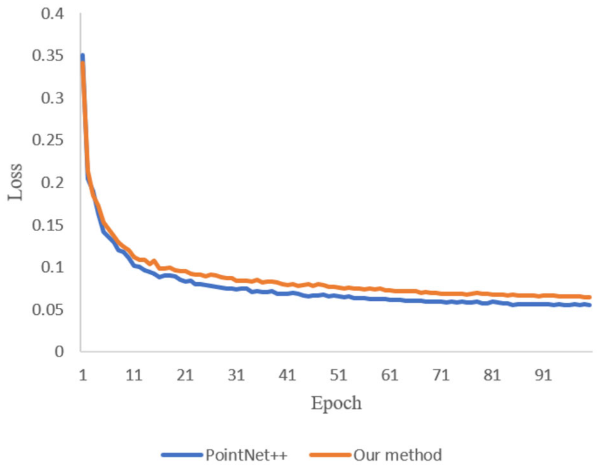


Figure 9 Loss curves of our method and PointNet++

nearby gingiva. Therefore, our method sometimes mistakenly identifies a portion of the gingiva as wisdom teeth as shown in Fig. 11. To avoid this issue, branch networks can be used to first predict the center points of each tooth, and then perform semantic segmentation, which is the next direction of our work. In addition, the intra-oral scanner images only include the crown portion of the teeth and cannot reveal the root portion beneath the gingiva, and using CBCT and intra-oral scanner images for multimodal learning can compensate for this deficiency.

4.3.2 *Segmentation performance comparison under different sampling points*

Due to limitations in computational load, some devices find it difficult to use large sampling points for segmentation, which results in a decrease in segmentation accuracy. To test the robustness of our method under different sampling points, we change the points' number and compare our method with PointNet++. The experimental results show that our method is less sensitive to changes in the number of sampled points compared with PointNet++. Figure 12 illustrates the comparison of mIoU for tooth segmentation using PointNet++ and our method under different sampled point numbers. The number of sampled points in the first, second, and third layers are 4096/1024/512, 1024/512/256, 512/256/128, and 256/128/64, respectively. When the number of sampled points decreases, PointNet++ shows a faster decline in mIoU, while our method is less sensitive to this variation, because SPFE and WSLFA can make better use of detailed information. When the number of sampled points decreases from 4096/1024/512 to 256/128/64, the mIoU of our method decreases by less than 2%. Figure 13 demonstrates the segmentation results of an intra-oral scanner image under different sampling points using our method and PointNet++. It can be observed that our method outperforms PointNet++ and has better robustness to the decrease in sampling points, especially on the teeth-gingival boundary.

4.4 Ablation study

We conduct ablation experiments on the SPFE and the WSLFA mechanism, as displayed in Table 3. The number of sampling points used by the model is 1024/512/256 for each layer, and the radius of each layer's spherical neighborhood is 0.05, 0.1, and 0.2 (normalized). Models 1, 2, 3, and 4 represent the model with neither SPFE nor WSLFA, the model with only WSLFA, the model with only SPFE, and the model with both SPFE and WSLFA, respectively. Models 1 and 3 use max pooling instead of WSLFA. Comparing models 1 and 2, 3 and 4, it can be found that the addition of WSLFA improved the segmentation effect of Models 1 and 3, with the mIoU increasing by approximately 1%, because WSLFA makes better use of the information inside each local region. Comparing Models 1 and

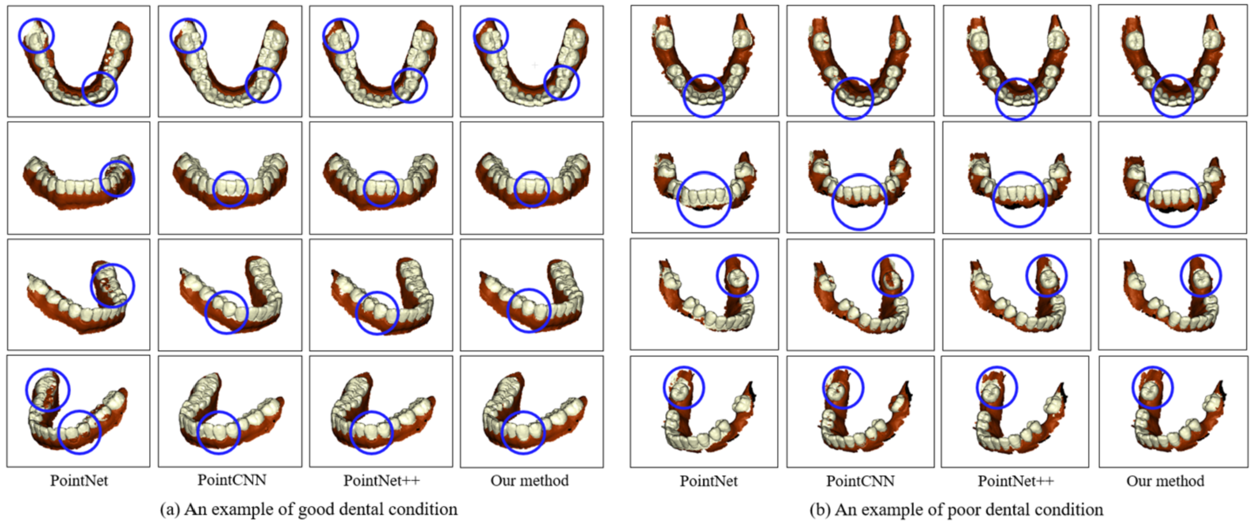

PointNet
PointCNN
PointNet++
Our method
(a) An example of good dental condition
PointNet
PointCNN
PointNet++
Our method
(b) An example of poor dental condition

Figure 10 Visualization of the segmentation results

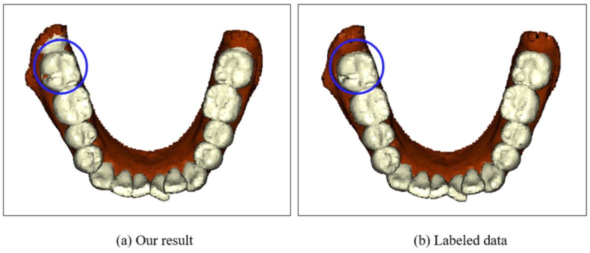


Figure 11 Segmentation results of an intra-oral scanner image that mistakenly identifies a portion of the gingiva as wisdom teeth

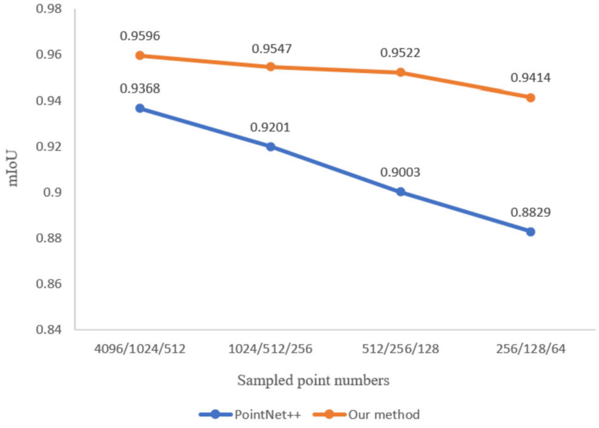


Figure 12 Comparison of mIoU for tooth segmentation using PointNet++ and our method under different sampled point numbers

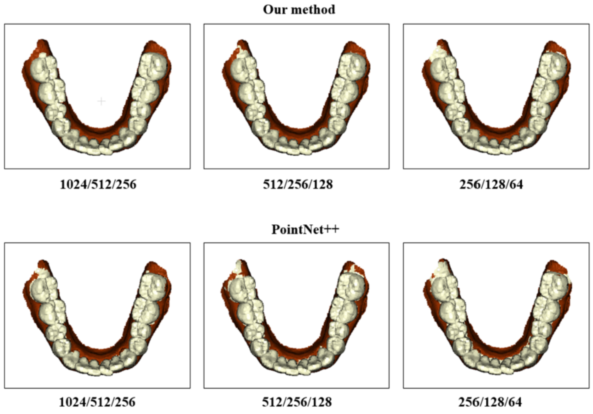


Figure 13 Segmentation results of an intra-oral scanner image under different sampling points using our method and PointNet++

3, 2 and 4, it can be seen that the addition of SPFE has improved both Models 1 and 2, with the mIoU increasing by more than 2%, because SPFE captures more detailed feature of the intra-oral scanner images. Using both SPFE and WSLFA simultaneously increases the mIoU by approximately 3.5%.

5 Conclusion

In this paper, an improved PointNet++ based method is proposed for full-set tooth segmentation of 3D intra-oral scanner images. The method first extracts preliminary features from individual points, retaining detailed features as much as possible. Then, multi-scale local regions are constructed, and a weighted-sum local feature aggregation mechanism is proposed to better integrate various useful information in local regions. These two methods effectively solve the problem of imprecise tooth segmenta-

Table 3 Ablation experiments

Model	SPFE	WSLFA	Accuracy	mIoU
1	-	-	0.9585	0.9201
2	-	✓	0.9672	0.9321
3	✓	-	0.9726	0.9453
4	✓	✓	0.9768	0.9547

tion, and achieve good segmentation results through clinical data experiments. For future research, adaptive adjustment of feature aggregation radius will be considered to better adapt to the complex teeth-gingival boundaries and further improve the accuracy of the method, and branch networks can be used to improve the accuracy of wisdom tooth segmentation. In addition, post-processing methods such as conditional random fields can be added to refine the boundary curve and improve its smoothness. Based on the above tooth segmentation work, we will conduct research on identity recognition using the segmented tooth parts, with the aim of utilizing the tooth model to recognize identity.

Funding
This work was supported by the 2022 Beijing Natural Science Foundation - Haidian Original Innovation Joint Fund (No. L222110).

Abbreviations
CBCT, cone beam computerized tomography; CNN, convolutional neural networks; DGCNN, dynamic graph CNN; ECC, edge conditioned convolution; FP, feature propagation; FPS, farthest point sampling; GNN, graph neural network; IOS, intra-oral scanning; LDGCNN, linked dynamic graph CNN; mIoU, mean intersection over union; MLP, multi-layer perceptron; PCA, principal components analysis; SA, set abstraction; SPFE, single-point preliminary feature extraction; WSLFA, weighted-sum local feature aggregation.

Availability of data and materials
The data that support the findings of this study are available from the Department of Stomatology, the Fourth Medical Center, Chinese PLA General Hospital but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Department of Stomatology, the Fourth Medical Center, Chinese PLA General Hospital.

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
LY and XL performed the data analyses and wrote the manuscript. YL and JY collected data. All authors read and approved the final manuscript.

Author details
[1]University of Science and Technology Beijing, Beijing, China. [2]Department of Stomatology, the Fourth Medical Center, Chinese PLA General Hospital, Beijing, China.

Received: 31 March 2023 Revised: 26 September 2023
Accepted: 26 September 2023 Published online: 10 October 2023

References

1. Hinchliffe, J. (2011). Forensic odontology, part 2. Major disasters. *British Dental Journal*, *210*, 269–274.
2. Kumar, Y., Janardan, R., Larson, B., & Moon, J. (2011). Improved segmentation of teeth in dental models. *Computer Aided Design*, *43*(2), 211–224.
3. Wu, K., Chen, L., Li, J., & Zhou, Y. (2014). Tooth segmentation on dental meshes using morphologic skeleton. *Computers & Graphics*, *38*(1), 199–211.
4. Zou, B. J., Liu, S. J., Liao, S. H., Ding, X., & Liang, Y. (2015). Interactive tooth partition of dental mesh base on tooth-target harmonic field. *Computers in Biology and Medicine*, *56*, 132–144.
5. Xu, X., Liu, C., & Zheng, Y. (2019). 3D tooth segmentation and labeling using deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics*, *25*(7), 2336–2348.
6. Tian, S., Dai, N., Zhang, B., Yuan, F., Yu, Q., & Cheng, X. (2019). Automatic classification and segmentation of teeth on 3D dental model using hierarchical deep learning networks. *IEEE Access*, *7*, 84817–84828.
7. Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 77–85). Piscataway: IEEE.
8. Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). PointNet++: deep hierarchical feature learning on point sets in a metric space. In I. Guyon, U. Von Luxburg, & S. Bengio, et al. (Eds.), *Advances in neural information processing systems* (Vol. *30*, pp. 5099–5108). Red Hook: Curran Associates.
9. Lian, C. F., Wang, L., Wu, T. H., Liu, M., Durán, F., Ko, C.-C., et al. (2019). MeshSNet: deep multi-scale mesh feature learning for end-to-end tooth labeling on 3D dental surfaces. In *Proceedings of the 22nd international conference on medical image computing and computer-assisted intervention* (pp. 837–845). Cham: Springer.
10. Li, C., Wang, G., Wang, B., Liang, X., Li, Z., & Chang, X. (2023). DS-Net++: dynamic weight slicing for efficient inference in CNNs and vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4430–4446.
11. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., & Chen, B. (2018). PointCNN: convolution on X-transformed points. In S. Bengio, H. Wallach, H. Larochelle, et al. (Eds.), *Advances in neural information processing systems* (Vol. *31*, pp. 820–830). Red Hook: Curran Associates.
12. Simonovsky, M., & Komodakis, N. (2017). Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference of computer vision and pattern recognition* (pp. 29–38). Piscataway: IEEE.
13. Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, *38*(5), 1–12.
14. Zhang, K., Hao, M., Wang, J., de Silva, C. W., & Fu, C. (2019). Linked dynamic graph CNN: learning on point cloud via linking hierarchical features. arXiv preprint. arXiv:1904.10014.
15. Li, Z., Liu, T., Wang, J., Zhang, C., & Jia, X. (2022). Multi-scale bidirectional enhancement network for 3D dental model segmentation. In *Proceedings of the 19th IEEE international symposium on biomedical imaging* (pp. 1–5). Piscataway: IEEE.
16. Zanjani, F.G., Pourtaherian, A., Zinger, S., Moin, D.A., Claessen, F., Cherici, T., et al. (2021). Mask-MCNet: tooth instance segmentation in 3D point clouds of intra-oral scans. *Neurocomputing*, *453*, 286–298.
17. Tian, Y., Zhang, Y., Chen, W.-G., Liu, D., Wang, H., Xu, H., et al. (2022). 3D tooth instance segmentation learning objectness and affinity in point cloud. *ACM Transactions on Multimedia Computing Communications and Applications*, *18*(4), 1–16.
18. Cui, Z., Li, C., Chen, N., Wei, G., Chen, R., Zhou, Y., et al. (2020). TSegNet: an efficient and accurate tooth segmentation network on 3D dental model. *Medical Image Analysis*, *69*, 101949.
19. Li, M., Huang, P.-Y., Chang, X., Hu, J., Yang, Y., & Hauptmann, A. (2023). Video pivoting unsupervised multi-modal machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3918–3932.

20. Jang, T. J., Kim, K. C., Cho, H. C., & Seo, J. K. (2022). A fully automated method for 3D individual tooth identification and segmentation in dental CBCT. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6562–6568.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.