Self-Supervised Learning with Masked Autoencoders for Teeth Segmentation from Intra-oral 3D Scans

Amani Almalki   Longin Jan Latecki
Department of Computer and Information Sciences, Temple University, Philadelphia, USA
{amani.almalki,latecki}@temple.edu

Abstract

In modern dentistry, teeth localization, segmentation, and labeling from intra-oral 3D scans are crucial for improving dental diagnostics, treatment planning, and population-based studies on oral health. However, creating automated algorithms for teeth analysis is a challenging task due to the limited availability of accessible data for training, particularly from the point of view of deep learning. This study extends the self-supervised learning framework of the mesh masked autoencoder (MeshMAE) transformer. While the MeshMAE loss measures the quality of reconstructed masked mesh triangles, the loss of the proposed DentalMAE evaluates the predicted deep embeddings of masked mesh triangles. This yields a better generalization ability on a very limited number of 3D dental scans, as documented by our results on teeth segmentation of intra-oral scans. Our results show that masking-based unsupervised learning methods may, for the first time, provide convincing transfer learning improvements on 3D intra-oral scans, increasing the overall accuracy over both MeshMAE and prior self-supervised pre-training.

1. Introduction

Computer-aided design (CAD) tools have gained significant popularity in modern dentistry, especially in orthodontic or prosthetic CAD systems, for accurate treatment planning. Advanced intra-oral scanners (IOS) are widely used to obtain precise digital surface models of dentition. The IOSs produce 3D surface reconstructions of the teeth either in the form of a point cloud or in a mesh format, or both. These models are invaluable in simulating teeth extraction, movement, deletion, and rearrangement, enabling dentists to predict treatment outcomes with greater ease. Consequently, digital teeth models have the potential to alleviate dentists' time-consuming and tedious tasks.

Tooth segmentation from intra-oral scans is a key step in computer-aided dentistry. It can help in recognizing and classifying different dental/oral conditions like gingivitis, caries, and white lesions. While tooth segmentation and labeling is a first step in digital dentistry, it is difficult due to the inherent similarities between teeth shapes and the ambiguity surrounding their positions on jaws. Furthermore, variations in teeth position and shape across different individuals present additional challenges in this process. Other challenges involved in tooth mesh segmentation, such as - crowded teeth, misaligned teeth, and missing teeth. The size of teeth can also vary widely across meshes. The second and third molars may evade capturing due to their being in the deep intra-oral regions. Or the second/third molar might not be fully formed. Different teeth and gum conditions, like recession, enamel loss, etc, can also alter the appearance of the teeth significantly.

Furthermore, the manual process of segmenting and labeling teeth is a time-consuming task that can potentially miss important data. This has led to a growing interest in leveraging computer vision and computer science to automate these processes. Multiple automatic tooth mesh segmentation algorithms have been proposed [37,44,50]. They include convolutional neural networks (CNNs) for teeth segmentation from 3D intra-oral scans [14–16, 40, 49, 52]. Recently, the use of CNNs in the analysis of medical images has experienced significant growth due to advancements in computational hardware, algorithms, and expansion in the amount of data [19]. However, CNNs are constrained in their overall capability due to the inherent inductive biases they possess [7].

Recent advancements in self-supervised learning have demonstrated the effectiveness of masked image modeling (MIM) [3, 10, 39] as a pre-training strategy for the Vision Transformer (ViT) [7] and the hierarchical Vision Transformer using shifted windows (Swin) [1, 2, 20]. MIM involves the masking and subsequent reconstruction of image patches, allowing the network to infer the masked regions by leveraging contextual information. We believe that the ability to aggregate contextual information is crucial in the context of 3D dental scan analysis. Among various MIM frameworks, the Masked Autoencoder (MAE) [10] stands out as a simple yet effective approach. MAE employs an encoder-decoder architecture, with a ViT encoder that receives only visible tokens and a lightweight decoder that reconstructs the masked patches using the encoder's patch-wise output and trainable mask tokens.

This paper introduces a novel approach to teeth segmentation in 3D dental scans called Dental Masked Autoencoder (DentalMAE) based self pre-training, which works for 3D dental meshes analysis. We apply DentalMAE pre-training on the same dataset, referred to as the train set, which is used for the downstream task. We term this approach self pre-training, which is particularly advantageous in scenarios where acquiring suitable pre-training data is challenging. Additionally, self pre-training eliminates the domain discrepancy between the pre-training and fine-tuning stages by unifying the training data. Our experiments focus on teeth segmentation in 3D intra-oral scans [4].

Specifically, We extend the self-supervised learning framework of the mesh masked autoencoder (MeshMAE) transformer [17]. While the MeshMAE loss measures the quality of reconstructed masked mesh triangles, the loss of the proposed DentalMAE evaluates the predicted deep embeddings of masked mesh triangles. After pre-training, the decoder is discarded, and the encoder is applied to the downstream task, i.e., teeth segmentation. We compare three ViT Transformer initializations, including our proposed DentalMAE, MeshMAE [17], and a mesh transformer without any self-pre-training. The experimental results demonstrate that DentalMAE self-pre-training significantly enhances dental scan segmentation performance compared to the baselines. Our main contributions are threefold:

- We utilize self-supervised learning with masked autoencoders to alleviate the problem of small data for 3D intra-oral scans.
- We replace the MeshMAE reconstruction of masked mesh patches with the reconstruction of mesh patch embeddings. Hence our loss is simply the L_2 distance between the predicted and computed embeddings over the masked patches, which is much simpler than the loss used by MeshMAE.
- Our proposed method leads to a significant performance improvement. DentalMAE outperforms all state-of-the-art methods on the tooth mesh segmentation task.

2. Related work

Most of the existing research in this field can be categorized into two groups: approaches based on handcrafted features and approaches based on learning.

2.1. Handcrafted features-based approaches

Previous methods primarily focused on extracting manually designed geometric features to segment 3D dental scans. These methods can be classified into three types: surface curvature-based methods, contour line-based methods, and harmonic field-based methods. Surface curvature is particularly useful for describing tooth surfaces and identifying tooth/gum boundaries in IOS. Zhao *et al*. [50] proposed a semi-automatic teeth segmentation method based on curvature thresholding, followed by gum separation and identification of 3D teeth boundary curves. Another approach by Yuan *et al*. [45] used minimum surface curvature calculation to extract individual teeth regions and separate them. Wu *et al*. [37] presented a morphological skeleton-based method for teeth segmentation in IOS, utilizing area growing operations. Similarly, Kronfeld *et al*. [12] introduced a system that detects tooth-gingiva boundaries using active contour models. Contour line-based methods involve manual selection of tooth boundary landmarks, followed by contour line generation based on geodesic information, as demonstrated in studies such as Sinthanayothin *et al*. and Yaqi *et al*. [31, 42]. Harmonic field methods require less user interaction, as they allow a limited number of surface points to be selected prior to the segmentation process, as seen in studies by Zou *et al*. [54] and Liao *et al*. [18].

However, these approaches have limitations in achieving robust and fully automated segmentation of dental 3D scans. Setting the optimal threshold for surface curvature-based methods is challenging, and they are sensitive to noise. Incorrect threshold selection can significantly impact segmentation accuracy, leading to over- or under-segmentation. Moreover, the manual threshold selection makes these methods unsuitable for fully automatic segmentation. Contour line-based methods are time-consuming, difficult to use, and rely heavily on human interaction. Harmonic field techniques involve complex and computationally intensive preprocessing steps.

2.2. Learning-based approaches

Recent advancements in deep learning techniques have shifted the focus of teeth segmentation from handcrafted features to learned features. It is now widely recognized that data-driven feature extraction, using techniques like convolutional neural networks (CNNs), outperforms handcrafted features in various computer vision tasks, including object detection [30] and image classification [35]. The same applies to 3D teeth segmentation and labeling. Learning-based approaches can be divided into two main categories based on the input data: 2D image segmentation and 3D mesh segmentation.

For 2D image segmentation, CNNs have been extensively used to extract relevant features. Cui *et al*. [6] introduced a two-stage deep supervised neural network architecture for tooth segmentation and identification in Cone-Beam Computed Tomography (CBCT) images. They employed an autoencoder CNN to extract edge maps from CBCT slices, which were then fed into a Mask R-CNN network for tooth segmentation and recognition. Similarly, Miki *et al*. [23] fine-tuned a pre-trained AlexNet network on CBCT dental slices for automatic teeth classification. Rao *et al*. [29] proposed a symmetric fully convolutional residual neural network for tooth segmentation in CBCT images. They incorporated dense conditional random field techniques and a deep bottleneck architecture for teeth boundary smoothing and segmentation enhancement, respectively. Zhang *et al*. [48] isomorphically mapped 3D dental scans into a 2D harmonic parameter space and used a CNN based on the U-Net architecture for tooth image segmentation.

Learning-based methods applied directly to 3D dental meshes have also been explored. Sun *et al*. [32] used a graph CNN-based architecture called FeaStNet for automated tooth segmentation and labeling from 3D dental scans. They extended this architecture to propose an end-to-end graph convolutional network-based model that achieved tooth segmentation and dense correspondence in 3D dental scans. Xu *et al*. [41] introduced a multi-stage framework based on a deep CNN architecture for 3D dental mesh segmentation. They employed two independent CNNs for teeth-gingiva and inter-teeth labeling. Zanjani *et al*. [47] proposed an end-to-end deep learning system based on the PointNet network architecture for semantic segmentation of individual teeth and gingiva from point clouds. They also used a secondary neural network as a discriminator in an adversarial learning setting to refine teeth labeling. Lian *et al*. [16] modified the PointNet architecture by incorporating graph-constrained learning modules to extract multi-scale local contextual features for teeth segmentation and labeling in 3D intra-oral scans. Tian *et al*. [33] introduced a preprocessing step that encoded input 3D scans using sparse voxel octree partitioning. They then employed three-level hierarchical CNNs for the segmentation process and another two-level hierarchical CNNs for teeth recognition. Other studies, such as Cui *et al*. [5] and Zanjani *et al*. [46], proposed pipeline-based architectures combining multiple CNNs for teeth localization, segmentation, and labeling. Ma *et al*. [21] suggested a deep neural network architecture for pre-detected teeth classification based on adjacency similarity and relative position feature vectors, explicitly modeling spatial relationships between adjacent teeth.

Zhao *et al*. [53] proposed an end-to-end network utilizing graph attentional convolution layers and a global structure branch for fine-grained local geometric feature extraction and global feature learning from raw mesh data. These features were fused to perform segmentation and labeling tasks. In another study, Zhao *et al*. [51] introduced a two-stream graph convolutional network (TSGCN). The first stream captured coarse structures of teeth from 3D coordinate information, while the second stream extracted distinctive structural details from normal vectors. To address the reliance on expensive point-wise annotations in current learning-based methods, Qiu *et al*. [27] presented the Dental Arch (DArch) method for 3D tooth segmentation using weak low-cost annotated data. The DArch consists of two stages: tooth centroid detection and segmentation. It generates the dental arch using Bezier curve regression and refines it using a graph-based convolutional network (GCN).

To the best of our knowledge, there have been no studies in the literature that specifically employ transformer models, such as the Vision Transformer (ViT) [7], for 3D dental scan analysis. Additionally, the application of self-supervised learning techniques to ViT on intra-oral scans is also unprecedented.

Transformer models, originally introduced in natural language processing tasks [34], have shown remarkable success in various computer vision domains, including image classification, object detection, and image segmentation. The ViT architecture, in particular, has gained attention for its ability to effectively process 2D images by leveraging self-attention mechanisms.

However, the application of transformer models to 3D dental scans and the use of self-supervised learning techniques on intra-oral scans have not been explored in the existing literature. This indicates a research gap and an opportunity to investigate the potential benefits and challenges of utilizing ViT and self-supervised learning in the context of 3D dental scan analysis.

By applying self-supervised learning to ViT on intra-oral scans, it becomes possible to mitigate the limited number of available intra-oral scans. This can help overcome the limitations of traditional supervised learning approaches, which rely heavily on large data for training. Self-supervised learning enables the model to learn from the inherent structure and properties present in the data, leading to improved generalization and potentially reducing the need for extensive manual labeling.

The application of transformer models and self-supervised learning techniques to 3D dental scans, specifically intra-oral scans, has the potential to advance the field by providing new insights and improved performance in tasks such as segmentation, labeling, and analysis of dental structures. Further research in this direction could pave the way for more accurate and efficient automated dental scan analysis, benefiting various clinical applications and oral healthcare practices.

3. Methods

In this paper, we use the Mesh Transformer framework for tooth mesh segmentation, which extends the Vision Transformer to mesh analysis. We propose a novel self-supervised learning pre-training strategy, which is based on mesh masked autoencoding. Fig. 1 illustrates the DentalMAE framework. DentalMAE divides the input mesh into non-overlap patches, these patches are embedded us-

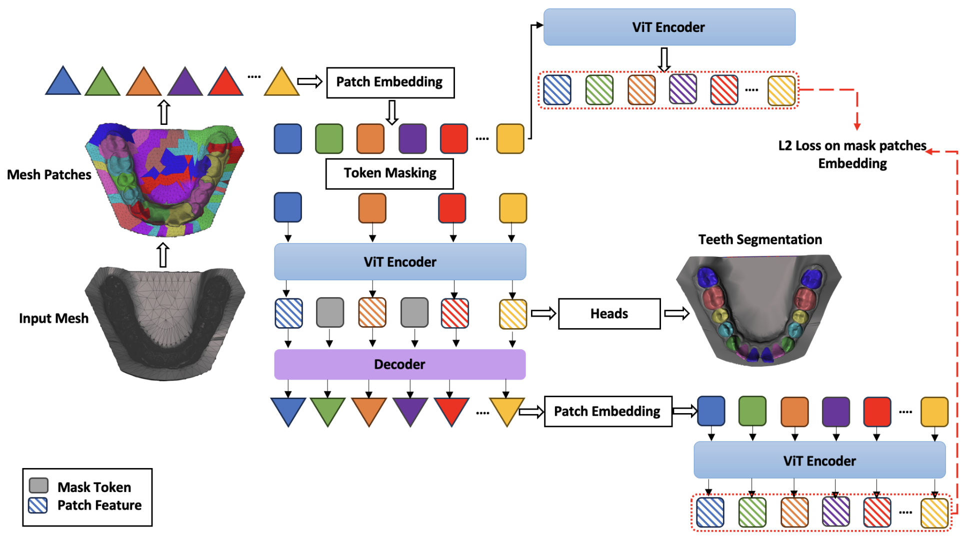

Figure 1. **The teeth segmentation pipeline for DentalMAE self-pre-training.** Initially, the input mesh is divided into non-overlap patches. These patches are then embedded using an MLP. During the pre-training phase, the patch embeddings are randomly masked, and only the visible embeddings are utilized by the transformer. Subsequently, the masked embeddings are combined with the encoded embeddings and sent to the decoder. The objective of the decoder is to reconstruct the vertices and face features of the masked patches, followed by the prediction of the patch embeddings of the masked patches. The L_2 loss is used to compare the masked patch embeddings. After the completion of pre-training, the decoder is discarded, and the encoder is employed for segmentation.

ing an MLP, and certain random patches are replaced with mask tokens. Only the visible patches are utilized by the ViT encoder. Subsequently, the mask tokens are combined with the encoded embeddings and are input to the decoder. The primary objective of the decoder is to reconstruct the vertices and face features of the masked patches, followed by the prediction of the patch embeddings of the masked patches. We do the two-stage process of reconstructing vertices and face features followed by computing embeddings because it performs better than directly predicting the embeddings as shown in the supplementary materials. Compared to MeshMAE [17], its loss measures the quality of reconstructed masked mesh triangles, while the loss of the proposed DentalMAE evaluates the predicted deep embeddings of masked mesh triangles. Following the pre-training phase, the decoder is discarded, and the encoder is employed for the specific task of tooth segmentation.

3.1. Mesh Transformer

Mesh Patch Split. The faces of a 3D mesh establish connections between vertices, allowing us to utilize geometric information from each face to represent their features. Similar to SubdivNet [11], we define a 10-dimensional vector for each face f_i comprising the face area (1-dim), three interior angles of the triangle (3-dim), face normal (3-dim), and three inner products between the face normal and three vertex normals (3-dim).

Transformers, with their self-attention-based architectures, simplify the process of designing feature aggregation operations for 3D meshes. However, applying self-attention to all faces incurs a prohibitively high computational cost due to quadratic complexity. To overcome this, the faces are grouped into non-overlapping patches before applying transformers. Unlike regular image data that can be divided into grid-like patches, mesh data is irregular, and faces are typically unordered.

To address this challenge, we utilize a ”re-meshing” step to regularize and hierarchically structure the original mesh. We employ the MAPS algorithm [13] to simplify the mesh into a coarser base mesh with a varying number of faces N faces within a specific range ($96 \leq N \leq 256$ in our experiments). Although less accurate in shape representation, the resulting base mesh serves as a foundation. To refine it, we further subdivide all faces in the base mesh t times in a 1-to-4 manner, resulting in a more detailed mesh called $t-$mesh. By grouping the faces of the $t-$mesh corresponding to the same face in the base mesh, we create non-overlapping patches. In our implementation, we perform three subdivisions, yielding patches consisting of 64 faces each. The process is illustrated in Fig. 2.

Transformer Backbone. The transformer serves as the backbone network for the Mesh Transformer. It consists of multi-headed self-attention layers and feedforward network (FFN) blocks. To represent each patch, we concatenate the feature vectors of the constituent faces belonging to that patch. The order of concatenation is determined by the

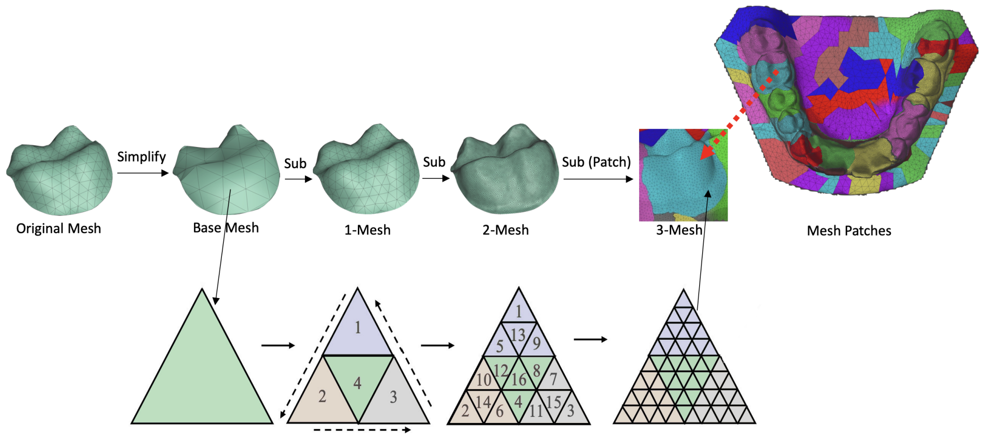

Figure 2. **The remeshing operation** involves several steps. Initially, the input mesh undergoes a simplification process. Subsequently, a mapping is established between the original mesh and the base mesh. The base mesh is then subdivided three times, and the newly generated vertices are projected back onto the input mesh.

re-meshing process, which guarantees a consistent and predictable face order. Consequently, an MLP is employed to project the feature vector of each patch into a representation denoted as $\{e_i\}_{i=1}^{g}$, where g denotes the number of patches. These representations serve as inputs to the transformer.

In addition to shape information captured by the input features, transformer-based methods often rely on positional embeddings to provide spatial information. Since mesh data contains 3D spatial coordinates for each face, we leverage the center 3D coordinates of the faces to compute the positional embeddings. To accomplish this, we calculate the center point coordinates $\{c_i\}_{i=1}^{g}$ for each patch and apply an MLP to obtain the positional embedding $\{p_i\}_{i=1}^{g}$ associated with each patch.

Formally, the input embeddings $X = \{x_i\}_{i=1}^{g}$ are defined as the combination of the patch embeddings $E = \{e_i\}_{i=1}^{g}$ and positional embeddings $P = \{p_i\}_{i=1}^{g}$. This results in an overall input sequence denoted as $H^0 = x_1, x_2, ..., x_g$. The encoder network consists of L layers of transformer blocks, and the output of the last layer $H^L = h_1^L, ..., h_g^L$ represents the encoded representations of the input patches.

3.2. Mesh Pre-training Task

In this section, we provide a detailed description of the mesh pre-training task, which employs a masked modeling strategy based on the Mesh Transformer architecture. The task aims to predict deep embeddings of masked mesh triangles from embeddings of visible mesh triangles. We outline the components of the pre-training task, including the encoder and decoder networks, masked sequence generation, and prediction.

Encoder and Decoder. The encoder and decoder networks used in the pre-training task are composed of several transformer blocks. The Mesh Transformer serves as the encoder, consisting of 12 layers, while a lightweight decoder with 6 layers is employed. During pre-training, a predefined masking ratio is applied to randomly mask a subset of patches in the input mesh. The visible patches are fed into the encoder, and a shared mask embedding is used to replace the masked embeddings in the input before feeding them into the decoder. The positional embeddings are added to both the masked and visible patches to provide location information. It is important to note that the decoder is only used during pre-training for mesh reconstruction tasks, while the encoder is utilized in downstream tasks.

Masked Sequence Generation. Mesh embeddings, represented by E, have corresponding indices denoted as I. Following the MAE approach, we randomly mask a subset of patches by sampling indices I_m from I with a ratio r. Masked embeddings are represented as E_m, while unmasked embeddings are denoted as E_{um}. We replace the masked embeddings E_m with a shared learnable mask embedding E_{mask} without altering their positional embeddings. Finally, the corrupted mesh embeddings E_c are formed by combining E_{um} with the sum of E_{mask} and positional embeddings p_i for each index i in I_m. These corrupted embeddings are then inputted into the encoder for further processing.

Prediction. MeshMAE [17] recovers the shape of the masked patches as the reconstruction target. It predicts 3D relative coordinates of vertices to match the ground truth positions, where the reconstruction loss is calculated using the Chamfer distance [8] between the predicted relative coordinates and the ground truth relative coordinates. It also predicts the face-wise features using a linear layer behind the decoder. It uses face-wise mean squared error (MSE) loss to evaluate the reconstruction effect of the features.

The overall optimization objective of MeshMAE combines the Chamfer distance loss $\mathcal{L}_{CD}$ and the MSE loss $\mathcal{L}_{MSE}$ to $\mathcal{L} = \mathcal{L}_{MSE} + \lambda \cdot \mathcal{L}_{CD}$, where λ is the loss weight. In contrast, our loss is simpler in that it does not require any meta parameter λ. We simply compute the L_2 loss between the original and predicted embeddings of the mask triangle patches.

4. Experiments

4.1. Dataset

We use the public dataset 3D Teeth Seg Challenge 2022 [4]. There are a total of 1800 3D intra-oral scans collected for 900 patients covering their upper and lower jaws separately. They are separated into training (1200 scans, 16004 teeth) and test data (600 scans, 7995 teeth). The task is tooth segmentation from the 3D dental model. Throughout the paper, we use the color coding shown in Fig. 3 to visualize the teeth labels. There are 8 different semantic parts, indicating the central incisor (T7), lateral incisor (T6), canine/cuspid (T5), 1st premolar/bicuspid (T4), 2nd premolar/bicuspid (T3), 1st molar (T2), 2nd molar (T1), and background/gingiva (BG).

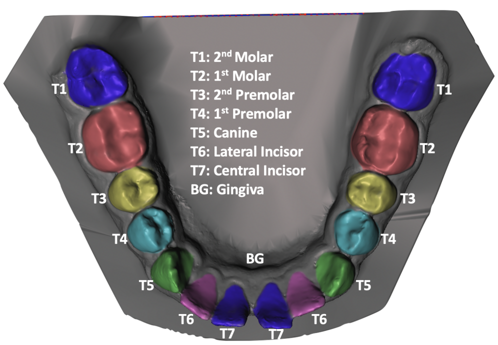

Figure 3. Tooth segmentation and the corresponding color coding.

4.2. Evaluation metric

We use Dice Score(DSC), Overall Accuracy (OA), sensitivity (SEN), and Positive Predictive Value (PPV) to evaluate the performance of our model.

4.3. Implementation details

Data Pre-processing. The dataset is processed by the re-meshing operation, and the face labels are obtained from the mapping between the re-meshed data and the raw meshes using the nearest face strategy.

Data Augmentation. We employ three data augmentation techniques: 1) random rotation, 2) random translation, and 3) random rescaling. By applying these techniques, we generate 40 augmented versions for each data point, resulting in the creation of 40 additional samples for every jaw scan.

Training Details. For pre-training, We utilize ViT-Base [7] as the encoder network with very slight modification, e.g., the number of input features' channels. And following [10], we set a lightweight decoder, which has 6 layers. We employ an AdamW optimizer, using an initial learning rate of 1e-4 with a cosine learning schedule. The weight decay is set as 0.05, and the batch size is set as 32. We set the same encoder network of pre-training in the downstream task. For our segmentation task, we utilize two segmentation heads to provide a two-level feature aggregation. Specifically, we concatenate the output of the encoder with the feature embedding of each face to provide a fine-grained embedding. We set the batch size as 32 and employed an AdamW optimizer with an initial learning rate of 1e-4. The learning rate is decayed by a factor of 0.1 at 80 and 160 epochs.

5. Results and analysis

5.1. Quantitative results

Table 1 presents the quantitative results of tooth segmentation using various methods, and it clearly shows that DentalMAE outperforms other state-of-the-art methods.

Comparing the Dice Scores of ViT with the other methods, it is evident that ViT achieves higher scores on almost all tooth labels (T1-T7) and the background (BG). ViT achieves Dice Scores ranging from 0.885 to 0.985, indicating its effectiveness in accurately segmenting tooth structures. This demonstrates the capability of the Vision Transformer to capture relevant features and contextual information, leading to improved segmentation results.

The results of ViT+MeshMAE outperform the standard ViT, indicating further improvements. The combination of ViT and MeshMAE enhances the segmentation accuracy and ensures more precise delineation of tooth boundaries.

Our method, DentalMAE, surpasses not only the other methods but also the standalone ViT and its enhanced version MeshMAE. It is evident that our method consistently achieves the highest Dice Scores across all tooth labels (T1-T7) and the background (BG). The Dice Scores range from 0.921 to 0.995, highlighting the effectiveness of incorporating the loss on mask patches embedding for tooth structure reconstruction.

All ViT variants outperform traditional methods like PointNet [25], PointNet++ [26], DGCNN [36], and MeshSegNet [16], as well as advanced methods such as MeshSegNet+GCO [16], TSGCNet [49], GAC [52], BAAFNet [28], pointMLP [22], PCT [9], MBESegNet [14], and CurveNet [38]. It also performs better than state-of-the-art self-supervised learning methods, Point-MAE [24] and Point-BERT [43]. This indicates the superiority of our proposed methods in accurately segmenting tooth structures and surpassing the performance of existing state-of-the-art approaches.

Table 2 presents additional quantitative results for tooth segmentation, evaluating various methods based on Overall Accuracy (OA), Dice Score (DSC), Sensitivity (SEN), and

Method	BG	T1	T2	T3	T4	T5	T6	T7
PointNet [25]	0.947	0.793	0.920	0.895	0.925	0.903	0.909	0.933
PointNet++ [26]	0.924	0.780	0.903	0.876	0.883	0.837	0.782	0.837
DGCNN [36]	0.968	0.847	0.944	0.936	0.945	0.941	0.939	0.947
MeshSegNet [16]	0.922	0.712	0.799	0.775	0.860	0.831	0.684	0.794
MeshSegNet+GCO [16]	0.957	0.850	0.904	0.902	0.926	0.879	0.778	0.906
TSGCNet [49]	0.962	0.642	0.915	0.916	0.945	0.937	0.916	0.926
GAC [52]	0.909	0.643	0.819	0.759	0.828	0.846	0.823	0.845
BAAFNet [28]	0.511	0.465	0.677	0.639	0.673	0.655	0.586	0.682
pointMLP [22]	0.975	0.865	0.959	0.950	0.969	0.959	0.945	0.953
PCT [9]	0.789	0.307	0.524	0.459	0.330	0.375	0.459	0.588
MBESegNet [14]	0.818	0.420	0.708	0.695	0.739	0.661	0.556	0.535
CurveNet [38]	0.964	0.783	0.923	0.917	0.939	0.922	0.918	0.939
Point-MAE [24]	0.971	0.802	0.956	0.924	0.949	0.943	0.942	0.948
Point-BERT [43]	0.976	0.835	0.962	0.939	0.952	0.951	0.951	0.957
ViT	0.985	0.885	0.971	0.966	0.959	0.969	0.959	0.968
ViT+MeshMAE	<u>0.990</u>	<u>0.908</u>	<u>0.982</u>	<u>0.976</u>	<u>0.978</u>	<u>0.985</u>	<u>0.961</u>	<u>0.983</u>
Ours	**0.995**	**0.921**	**0.989**	**0.988**	**0.986**	**0.992**	**0.974**	**0.990**

Table 1. The tooth segmentation results from different methods in terms of the label-wise Dice Score.

Positive Predictive Value (PPV). The results further confirm the superior performance of our proposed method, DentalMAE, compared to other state-of-the-art techniques.

Method	OA	DSC	SEN	PPV
PointNet [25]	0.926	0.903	0.913	0.912
PointNet++ [26]	0.892	0.853	0.864	0.865
DGCNN [36]	0.933	0.915	0.923	0.923
MeshSegNet [16]	0.901	0.873	0.888	0.879
MeshSegNet+GCO [16]	0.931	0.918	0.929	0.911
TSGCNet [49]	0.936	0.895	0.924	0.902
GAC [52]	0.855	0.809	0.818	0.844
BAAFNet [28]	0.601	0.611	0.755	0.594
pointMLP [22]	0.943	0.927	0.936	0.931
PCT [9]	0.629	0.479	0.509	0.586
MBESegNet [14]	0.716	0.642	0.710	0.644
CurveNet [38]	0.939	0.912	0.922	0.923
Point-MAE [24]	0.945	0.927	0.942	0.936
Point-BERT [43]	0.949	0.935	0.948	0.944
ViT	0.955	0.945	0.950	0.957
ViT+MeshMAE	<u>0.971</u>	<u>0.954</u>	<u>0.966</u>	<u>0.983</u>
Ours	**0.983**	**0.970**	**0.977**	**0.989**

Table 2. The tooth segmentation results from different methods in terms of the Overall Accuracy, the Dice Score, the Sensitivity, and the Positive Predictive Value.

Our method, DentalMAE, achieves an OA value of 0.983. This score indicates the overall accuracy of the tooth segmentation results obtained by our method. It is evident that DentalMAE outperforms all other SOTA methods.

The Dice Score measures the similarity between the predicted and ground truth tooth segmentations. In terms of DSC, our method, DentalMAE, achieves a score of 0.970. These scores demonstrate the accuracy and overlap of the segmented tooth structures compared to the ground truth. Notably, our method consistently outperforms all other methods, including the top-performing MeshMAE method.

SEN and PPV evaluate the ability of the segmentation methods to correctly identify tooth structures (SEN) and the precision of the predicted tooth segmentations (PPV). Our method exhibits high SEN and PPV scores, with a SEN value of 0.977, and a PPV value of 0.989. These results indicate the robustness and accuracy of our method in identifying tooth structures while minimizing false positives and false negatives.

Parameter Setting and Masking Strategies. The experiments conducted in Table 3 explore the effects of different masking strategies and ratios on teeth segmentation. In contrast to the high mask ratios commonly used in 3D natural models [17], the segmentation task for teeth exhibits distinct preferences regarding the mask ratio. Notably, we consistently observe performance improvements as the mask ratio decreases from 50% to 20%. This finding suggests that reducing the mask ratio is beneficial for training the model, potentially because relevant features in 3D intra-oral models tend to be smaller in scale.

Additionally, the random masking strategy outperforms the block and grid strategies, emphasizing its effectiveness in generating masks during the training process. These findings contribute to our understanding of optimal parameter settings for teeth segmentation and inform the development of more accurate and efficient segmentation models in this domain.

5.2. Qualitative results

Figure 4 presents qualitative examples that showcase the enhanced performance achieved through pre-training the ViT mesh transformer with DentalMAE for teeth segmentation. The observed improvements in segmentation align

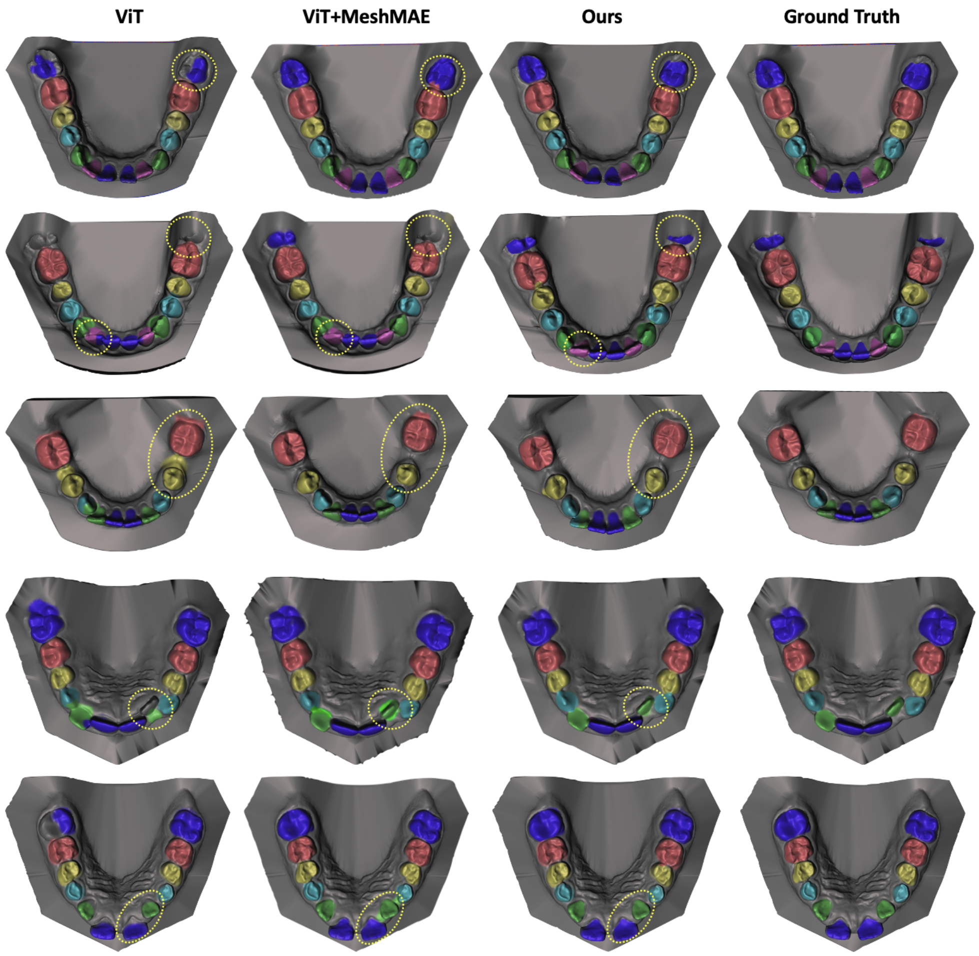

Figure 4. Comparison of teeth segmentation of DentalMAE and baselines. The first three rows show samples of the lower jaw, while the last two rows show the upper jaw.

Mask ratio	strategy	OA	DSC
50%	random	0.947	0.936
50%	block	0.931	0.930
50%	grid	0.943	0.932
40%	random	0.955	0.939
30%	random	0.959	0.941
20%	random	**0.971**	**0.954**
10%	random	0.958	0.943

Table 3. The influence of Mask Ratios/strategies on teeth segmentation of our DentalMAE.

with the quantitative findings discussed in Section 5.1.

6. Conclusions

We have demonstrated that DentalMAE pre-training improves SOTA segmentation performance on 3D dental scan analysis. Importantly, DentalMAE self-pre-training outperforms existing methods on a small dataset, something that has not previously been explored. Our results also suggest that parameters, including mask ratio and strategy, should be tailored when applying masked autoencoders pretraining to the 3D dental scan domain. Together, these observations suggest that DentalMAE can further improve the already impressive performance of mesh ViTs in intra-oral scan analysis. In future work, we will test the efficacy of DentalMAE pretraining in prognosis and outcome prediction tasks.

7. Acknowledgments

We would like to express our deepest thanks to Dr. Abdulrahman Almalki, a dental expert from the Department of Prosthetic Dental Science at Prince Sattam Bin Abdulaziz University, for his valuable discussions related to dentistry.

References

[1] Amani Almalki and Longin Jan Latecki. Enhanced masked image modeling for analysis of dental panoramic radiographs. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023. 1

[2] Amani Almalki and Longin Jan Latecki. Self-supervised learning with masked image modeling for teeth numbering, detection of dental restorations, and instance segmentation in dental panoramic radiographs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5594–5603, 2023. 1

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[4] Achraf Ben-Hamadou, Oussama Smaoui, Houda Chaabouni-Chouayakh, Ahmed Rekik, Sergi Pujades, Edmond Boyer, Julien Strippoli, Aurélien Thollot, Hugo Setbon, Cyril Trosset, et al. Teeth3ds: a benchmark for teeth segmentation and labeling from intra-oral 3d scans. *arXiv preprint arXiv:2210.06094*, 2022. 2, 6

[5] Zhiming Cui, Changjian Li, Nenglun Chen, Guodong Wei, Runnan Chen, Yuanfeng Zhou, and Wenping Wang. Tsegnet: an efficient and accurate tooth segmentation network on 3d dental model. *Medical Image Analysis*, 69:101949, 2020. 3

[6] Zhiming Cui, Changjian Li, and Wenping Wang. Toothnet: Automatic tooth instance segmentation and identification from cone beam ct images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377, 2019. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1, 3, 6

[8] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 5

[9] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021. 6, 7

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1, 6

[11] Shi-Min Hu, Zheng-Ning Liu, Meng-Hao Guo, Jun-Xiong Cai, Jiahui Huang, Tai-Jiang Mu, and Ralph R Martin. Subdivision-based mesh convolution networks. *ACM Transactions on Graphics (TOG)*, 2021. 4

[12] Thomas Kronfeld, David Brunner, and Guido Brunnett. Snake-based segmentation of teeth from virtual dental casts. *Computer-Aided Design and Applications*, 7(2):221–233, 2010. 2

[13] Aaron WF Lee, Wim Sweldens, Peter Schröder, Lawrence Cowsar, and David Dobkin. Maps: Multiresolution adaptive parameterization of surfaces. In *ACM SIGGRAPH*, pages 95–104, 1998. 4

[14] Zigang Li, Tingting Liu, Jun Wang, Changdong Zhang, and Xiuyi Jia. Multi-scale bidirectional enhancement network for 3d dental model segmentation. In *IEEE 19th Int. Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022. 1, 6, 7

[15] Chunfeng Lian, Li Wang, Tai-Hsien Wu, Mingxia Liu, Francisca Durán, Ching-Chang Ko, and Dinggang Shen. Meshsnet: Deep multi-scale mesh feature learning for end-to-end tooth labeling on 3d dental surfaces. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 837–845. Springer, 2019. 1

[16] Chunfeng Lian, Li Wang, Tai-Hsien Wu, Fan Wang, Pew-Thian Yap, Ching-Chang Ko, and Dinggang Shen. Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3d intraoral scanners. *IEEE transactions on medical imaging*, 39(7):2440–2450, 2020. 1, 3, 6, 7

[17] Yaqian Liang, Shanshan Zhao, Baosheng Yu, Jing Zhang, and Fazhi He. Meshmae: Masked autoencoders for 3d mesh data analysis. In *ECCV*, 2022. 2, 4, 5, 7

[18] Sheng-hui Liao, Shi-jian Liu, Bei-ji Zou, Xi Ding, Ye Liang, and Jun-hui Huang. Automatic tooth segmentation of dental mesh based on harmonic fields. *BioMed research international*, 2015, 2015. 2

[19] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[21] Qian Ma, Guangshun Wei, Yuanfeng Zhou, Xiao Pan, Shiqing Xin, and Wenping Wang. Srf-net: Spatial relationship feature network for tooth point cloud classification. In *Computer Graphics Forum*, volume 39, pages 267–277. Wiley Online Library, 2020. 3

[22] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022. 6, 7

[23] Yuma Miki, Chisako Muramatsu, Tatsuro Hayashi, Xiangrong Zhou, Takeshi Hara, Akitoshi Katsumata, and Hiroshi Fujita. Classification of teeth in cone-beam ct using deep convolutional neural network. *Computers in biology and medicine*, 80:24–29, 2017. 3

[24] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, 2022. 6, 7

[25] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 6, 7

[26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 6, 7

[27] Liangdong Qiu, Chongjie Ye, Pei Chen, Yunbi Liu, Xiaoguang Han, and Shuguang Cui. Darch: Dental arch prior-assisted 3d tooth instance segmentation with weak annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20752–20761, 2022. 3

[28] Shi Qiu, Saeed Anwar, and Nick Barnes. Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1757–1767, 2021. 6, 7

[29] Yunbo Rao, Yilin Wang, Fanman Meng, Jiansu Pu, Jihong Sun, and Qifei Wang. A symmetric fully convolutional residual network with dcrf for accurate tooth segmentation. *IEEE Access*, 8:92028–92038, 2020. 3

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 2

[31] Chanjira Sinthanayothin and Wichit Tharanont. Orthodontics treatment simulation by teeth segmentation and setup. In *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTICON'08)*, volume 1, pages 81–84. IEEE, 2008. 2

[32] D. Sun, Y. Pei, G. Song, Y. Guo, G. Ma, T. Xu, and H. Zha. Tooth segmentation and labeling from digital dental casts. In *IEEE International Symposium on Biomedical Imaging (ISBI'20)*, April, Iowa City, IA, USA 2020. 3

[33] Sukun Tian, Ning Dai, Bei Zhang, Fulai Yuan, Qing Yu, and Xiaosheng Cheng. Automatic classification and segmentation of teeth on 3d dental model using hierarchical deep learning networks. *IEEE Access*, 7:84817–84828, 2019. 3

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[35] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2016. 2

[36] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 6, 7

[37] Kan Wu, Li Chen, Jing Li, and Yanheng Zhou. Tooth segmentation on dental meshes using morphologic skeleton. *Computers & Graphics*, 38:199–211, 2014. 1, 2

[38] Tiange Xiang, Chaoyi Zhang, Yang Song, Jianhui Yu, and Weidong Cai. Walk in the cloud: Learning curves for point clouds shape analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 915–924, 2021. 6, 7

[39] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[40] Xiaojie Xu, Chang Liu, and Youyi Zheng. 3d tooth segmentation and labeling using deep convolutional neural networks. *IEEE transactions on visualization and computer graphics*, 25(7):2336–2348, 2018. 1

[41] X. Xu, C. Liu, and Y. Zheng. 3d tooth segmentation and labeling using deep convolutional neural networks. *IEEE transactions on visualization and computer graphics*, 25:2336–2348, 2018. 3

[42] Ma Yaqi and Li Zhongke. Computer aided orthodontics treatment by virtual segmentation and adjustment. In *2010 International Conference on Image Analysis and Signal Processing*, pages 336–339. IEEE, 2010. 2

[43] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, 2022. 6, 7

[44] Tianran Yuan, Wenhe Liao, Ning Dai, Xiaosheng Cheng, and Qing Yu. Single-tooth modeling for 3d dental model. *International journal of biomedical imaging*, 2010, 2010. 1

[45] Tianran Yuan, Wenhe Liao, Ning Dai, Xiaosheng Cheng, and Qing Yu. Single-tooth modeling for 3d dental model. *International journal of biomedical imaging*, 2010, 01 2010. 2

[46] F. G. Zanjani, D. A. Moin, F. Claessen, T. Cherici, S. Parinussa, A. Pourtaherian, and S. Zinger. Mask-mcnet: Instance segmentation in 3d point cloud of intra-oral scans. pages 128–136, 2019. 3

[47] Farhad Ghazvinian Zanjani, David Anssari Moin, Bas Verheij, Frank Claessen, Teo Cherici, Tao Tan, et al. Deep learning approach to semantic segmentation in 3d point cloud intra-oral scans of teeth. In *International Conference on Medical Imaging with Deep Learning*, pages 557–571. PMLR, 2019. 3

[48] Jianda Zhang, Chunpeng Li, Qiang Song, Lin Gao, and Yu-Kun Lai. Automatic 3D Tooth Segmentation using Convolutional Neural Networks in Harmonic Parameter Space. *Elsevier Graphical Models*, 39:101071, 2020. 3

[49] Lingming Zhang, Yue Zhao, Deyu Meng, Zhiming Cui, Chenqiang Gao, Xinbo Gao, Chunfeng Lian, and Dinggang Shen. Tsgcnet: Discriminative geometric feature learning with two-stream graph convolutional network for 3d dental model segmentation. In *CVPR*, 2021. 1, 6, 7

[50] Mingxi Zhao, Lizhuang Ma, Wuzheng Tan, and Dongdong Nie. Interactive tooth segmentation of dental models. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 654–657. IEEE, 2006. 1, 2

[51] Yue Zhao, Lingming Zhang, Yang Liu, Deyu Meng, Zhiming Cui, Chenqiang Gao, Xinbo Gao, Chunfeng Lian, and

Dinggang Shen. Two-stream graph convolutional network for intra-oral scanner image segmentation. *IEEE Transactions on Medical Imaging*, 41, 2022. 3

[52] Yue Zhao, Lingming Zhang, Chongshi Yang, Yingyun Tan, Yang Liu, Pengcheng Li, Tianhao Huang, and Chenqiang Gao. 3D dental model segmentation with graph attentional convolution network. *Pattern Rec. Letters*, 152, 2021. 1, 6, 7

[53] Yue Zhao, Lingming Zhang, Chongshi Yang, Yingyun Tan, Yang Liu, Pengcheng Li, Tianhao Huang, and Chenqiang Gao. 3d dental model segmentation with graph attentional convolution network. *Pattern Rec. Letters*, 152, 2021. 3

[54] Bei-ji Zou, Shi-jian Liu, Sheng-hui Liao, Xi Ding, and Ye Liang. Interactive tooth partition of dental mesh base on tooth-target harmonic field. *Computers in biology and medicine*, 56:132–144, 2015. 2