DArch: Dental Arch Prior-assisted 3D Tooth Instance Segmentation with Weak Annotations

Liangdong Qiu[1,3]* Chongjie Ye[1]* Pei Chen[1] Yunbi Liu[1,3] Xiaoguang Han[1,2]† Shuguang Cui[1,2,3]
[1]SSE, CUHK-Shenzhen [2]FNii, CUHK-Shenzhen [3]Shenzhen Research Institute of Big Data
{liangdongqiu, chongjieye, peichen}@link.cuhk.edu.cn, ybliu1994@gmail.com,
{hanxiaoguang, shuguangcui}@cuhk.edu.cn

Abstract

Automatic tooth instance segmentation on 3D dental models is a fundamental task for computer-aided orthodontic treatments. Existing learning-based methods rely heavily on expensive point-wise annotations. To alleviate this problem, we are the first to explore a low-cost annotation way for 3D tooth instance segmentation, i.e.*, labeling all tooth centroids and only a few teeth for each dental model. Regarding the challenge when only weak annotation is provided, we present a dental arch prior-assisted 3D tooth segmentation method, namely DArch. Our DArch consists of two stages, including tooth centroid detection and tooth instance segmentation. Accurately detecting the tooth centroids can help locate the individual tooth, thus benefiting the segmentation. Thus, our DArch proposes to leverage the dental arch prior to assist the detection. Specifically, we firstly propose a coarse-to-fine method to estimate the dental arch, in which the dental arch is initially generated by Bezier curve regression, and then a graph-based convolutional network (GCN) is trained to refine it. With the estimated dental arch, we then propose a novel Arch-aware Point Sampling (APS) method to assist the tooth centroid proposal generation. Meantime, a segmentor is independently trained using a patch-based training strategy, aiming to segment a tooth instance from a 3D patch centered at the tooth centroid. Experimental results on* $4,773$ *dental models have shown our DArch can accurately segment each tooth of a dental model, and its performance is superior to the state-of-the-art methods.*

1. Introduction

Dental models, obtained by direct intraoral scanning of the dentition, are commonly used in computer-aided dentistry. Computer-aided dentistry requires dental models as input to assist dentists to analyze and evaluate patient-specific dental health and dental arrangement for the following treatments. Automatic tooth instance segmentation on dental models is an essential prerequisite step for computer-aided orthodontic treatments.

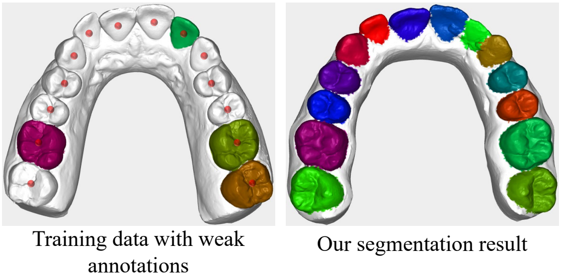

Figure 1. An example of our segmentation result when using training data with weak annotations for training models. Left: dental model with weak annotations, *i.e.*, labelling all tooth centroids and only a few teeth. Right: our segmentation result.

Although recent learning-based methods have achieved impressive performance on 3D tooth instance segmentation [28, 29, 39], they rely heavily on a large number of data with dense manual annotations, such as labeling all points of every individual tooth from a dental model. Since annotating such training data is particularly time-consuming, it is hard to collect a large enough dataset to cover complex dental models in real-world, thus largely limiting the generalization of those learning-based segmentation methods [30, 36, 38]. One of the main challenges in automatic 3D tooth instance segmentation is locating each tooth object on a variety of dental models, some of which have missing, crowding, or misaligned teeth. Cui et al. [3] found that in the tooth detection stage, the tooth centroid is a more reliable signal than the bounding box that is used to crop the detected tooth objects in the traditional approaches. Tooth detection thus can be converted to tooth centroids detection. Motivated by their work, we propose a feasible and low-cost annotation way as shown in the right of the Fig. 1, that is, **specifying 3D centroids for all tooth instances**

[*] L. Qiu and C. Ye contribute equally.
[†] Corresponding author

and labeling dense instance mask for only a few teeth for each dental model, to alleviate the demand for expensive point-wise annotations. In this paper, we present a novel *detect-and-segment* framework, including detecting tooth centroids and segmenting every teeth instance assigned to the corresponding teeth centroid. We mainly focus on the detection stage based on the intuition that the more accurate the detection, the better the segmentation. Furthermore, we adopt a patch-based training strategy to decrease the discriminating difficulty for our segmentor, which aims to segment a tooth instance from a 3D patch centered at the tooth centroid, rather than segment all tooth instances from a whole point cloud data of a dental model. In such a way, it makes a great demand on the tooth centroid prediction in our method. Previous detection methods for 3D point clouds [3, 13, 18] generally use the furthest point sampling (FPS) method to uniformly select the sampling points for generating proposals. For tooth centroid detection, the sampled points by FPS method generally contain irrelevant points, such as one located on the tooth crown and gingiva, which may lead to inaccurate proposals for tooth centroids.

To accurately and completely predict each tooth centroid of a dental model, we propose an arch-aware point sampling (APS) module for tooth centroid detection by introducing dental arch prior to assist the detection procedure. This is based on the observations that a dental arch naturally depicts one's overall dentition, and all tooth centroids will fall on it. To estimate the dental arch of each dental model, we first formulate the dental arch by representing it as a curve passing through teeth centroids and then adopt a lightweight 1-D convolutional network to refine the dental arch [10, 17]. Different from FPS method that performs uniform sampling from the whole tooth votes, we sample points along the estimated dental arch to filter out a majority of irrelevant points. Experiments have shown that our proposed APS strategy can largely improve the detection accuracy for tooth centroids compared to FPS and benefit the following segmentation.

To the best of our knowledge, this is the first attempt for 3D tooth instance segmentation on dental models with weak annotations. The main contributions of our work can be summarized as follows:

- We are the first to explore a low-cost annotation way for 3D tooth instance segmentation and propose a novel framework named DArch to handle this challenging task with weak annotations. We hope this attempt will inspire more learning-based methods in the weakly-annotated scenario.

- We propose a coarse-to-fine method to estimate the dental arch. Specifically, the dental arch is initially approximated by Bézier curve regression, and then a graph-based convolutional network (GCN) is used for further refinement.

- We introduce a dental arch-aware point sampling (APS) module for tooth centroid detection by introducing dental arch prior to assist the proposal generation.

- Extensive experiments have shown that our proposed DArch can vastly improve the performance of tooth centroid detection compared to other methods using other sampling strategies. As for the segmentation performance, our DArch is superior to the state-of-the-art methods in both weakly- and fully-annotated scenarios.

2. Related Work

2.1. 3D Understanding in Natural Scene

3D understanding in natural scenes usually involves object detection [9, 15, 27, 37], instance segmentation [4, 5], shape understanding [31, 33], part segmentation [7, 32] and so on, which is a fundamental problem in computer vision. In recent years, some deep learning-based methods have been proposed on different representations, such as volumetric data [12, 21, 41, 41], point cloud [25, 26, 35] and other representations [16, 24]. A point cloud is among one of the most popular ways to represent the 3D shape or object. PointNet [20] is an early representative attempt to design a novel deep network suitable for unordered point sets in 3D. PointNet++ [22] and PointCNN [8] extended PointNet by recursively applying it in a hierarchical fashion, so as to learn deep point set features efficiently and robustly. These two works inspired a lot of follow-up works. For example, VoteNet [18] propose to detect 3D objects by endowing point cloud deep networks (*i.e., PointNet++*) with a voting mechanism similar to the classical Hough voting. By voting, VoteNet essentially generates new points that lie close to objects centers, which can be grouped and aggregated to generate box proposals. Regarding the strong ability of feature representation of PointNet++ and the voting mechanism to generate objects centers in VoteNet, we adopt VoteNet as the basic architecture of our tooth centroid detection network and PointNet++ as the backbone network to exact the deep point features of the fine-grained tooth objects. To generate proposals from the votes in the proposal step, VoteNet used the furthest point sampling (FPS) to uniformly sample K vote clusters. Such a sampling strategy may select irrelevant vote clusters for tooth centroid detection, such as one located on the tooth crown and gingiva. To avoid this problem, instead of using FPS, we propose an arch-aware point sampling (APS) strategy to assist in generating proposals of tooth centroid by leveraging the dental arch prior.

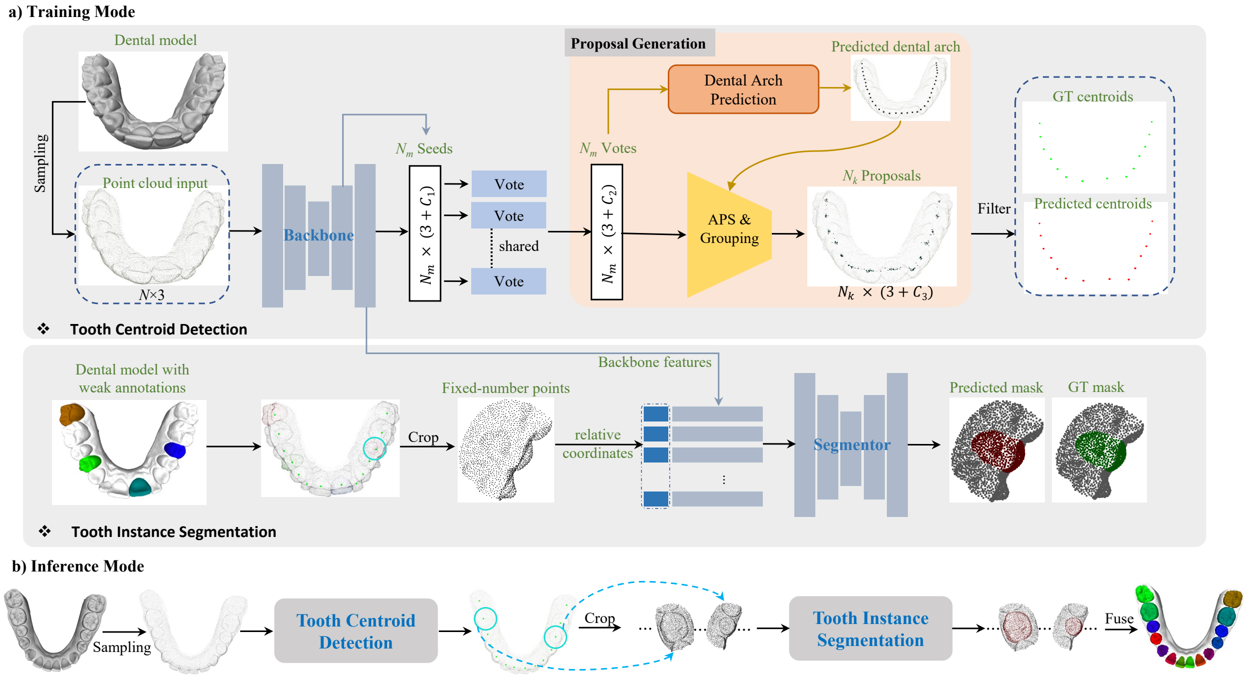

Figure 2. **Illustration of our DArch in both training and inference mode.** Our DArch consists of two parts of tooth centroid detection and tooth instance segmentation. In the inference mode, our DArch can segment all tooth instances by fusing the patch-based results. APS: Arch-aware point sampling.

2.2. 3D Tooth Understanding

Recently, deep learning-based methods have been popularly used to handle the task of tooth instance segmentation [30, 38, 39]. For example, Mask MCNet [38] proposed a framework that combines the Monte Carlo Convolutional Network (MCCNet) with Mask R-CNN to simultaneously locate each tooth object by predicting its bounding box and segment all the tooth points inside the box. Graph convolutional neural network-based frameworks (GCN) [28,29,40] have been proposed to learn more discriminative geometric features for 3D dental model segmentation. TSegNet [3] found that the tooth centroid is a more reliable signal than the bounding box in the tooth detection stage, and based on this observation proposed a novel pipeline that formulates the dental model segmentation as two sub-problems: robust tooth centroids prediction and accurate individual tooth segmentation on point cloud data. However, existing learning-based methods heavily depended on expensive dense point-wise annotations, that is, labeling all teeth of each dental model in the training data, to supervise the training process. Such a full annotation way brings a considerable burden for human labeling and increases the difficulty of collecting a large number of data, thus limiting these methods to real-world applications. In this paper, we are the first to study a 3D tooth instance segmentation problem with limited annotations. Motivated by those methods above, our proposed DArch includes a tooth centroid detection model to identify each tooth object and a tooth instance segmentation model to segment every tooth instance. To accurately detect each tooth centroid, we propose to estimate the dental arch and leverage the estimated dental arch to assist the proposal generation of tooth centroids.

3. Method

3.1. Overview

In this work, we propose a novel detect-and-segment framework, dubbed DArch, to tackle the challenging task of 3D tooth instance segmentation with weak annotations. Our DArch aims to segment all tooth instances given a point cloud input of a single dental model. As shown in Fig. 2, our DArch consists of two parts, including tooth centroid detection and tooth instance segmentation. In particular, to accurately predict all tooth centroids, we introduce a dental arch prediction module to estimate the dental arch and propose an arch-aware point sampling (APS) strategy to generate the centroid proposals. Our segmentation network adopts a patch-based training strategy, and in the inference phase the trained segmentor can predict all the tooth instances from a dental model by fusing all patch-based segmentation results. We will elaborate our detection and segmentation networks as follows.

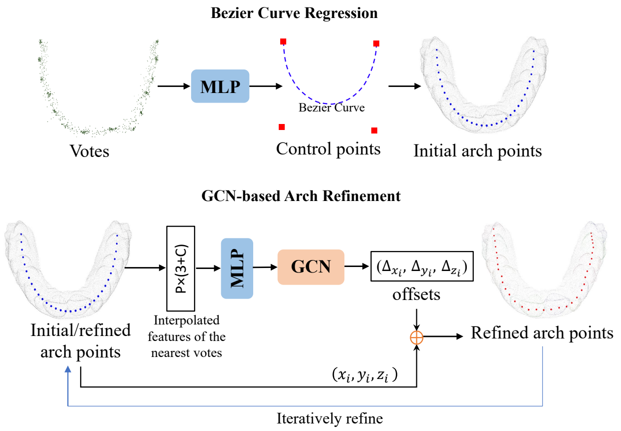

Figure 3. The overview of our proposed dental arch prediction method. Our method consists of two steps, including Bézier regression and GCN refinement. An initial curve is sampled from generated Bézier curve. Then, the points $\hat{\mathbf{X}}$ are refined by offsets iteratively.

3.2. Tooth Centroid Detection

Our tooth centroid detection network consists of a detection backbone and a branch of arch generation. We adopt the VoteNet [18] as our detection backbone regarding its solid architecture and voting mechanism. As we know, proposal generation is one of the most vital parts for 3D object center detection. Considering the fine-grained structures of teeth, we propose an APS method to replace the FPS sampling method used in VoteNet for generating tooth centroid proposals. In the following, we first review the work of VoteNet briefly and then propose our arch generation method. Finally, we present our APS method for proposal generation.

3.2.1 Review of VoteNet

The original VoteNet is proposed by Qi et al. [18]. It is tailored for 3D point cloud detection based on PointNet++ [22]. Given a set of input 3D points $\{p_i\}_{i=1}^{N}$, the backbone network of PointNet++ selects seed points and generate enriched C-dimensional feature vector. The point coordinates are embedded into the C-dimensional feature vector to represent the seeds $\{s_i\}_{i=1}^{M}$, where s_i is a $(3+C)$-dimentional feature vector. Then the seed points are fed into a shared multi-layer perception (MLP) to compute votes $\{v_i\}_{i=1}^{M}$. The generated voting v_i will be aggregated around object center.

After generating votes v_i, FPS is used to sample a subset of the votes to get $\{v_i\}_{i=1}^{K}$. Then by finding all near votes within a certain Euclidean distance, votes are generated into K clusters, followed by a three-layer MLP to generate the proposals. Finally, NMS is applied to filter out the overlapped proposals and generate the final prediction. The sampling method is very significant for generating reasonable proposals. For the specific task of tooth centroid detection, FPS may sample irrelevant points from the whole tooth votes, such as one located on the tooth crown and gingiva, due to its uniform and sparse sampling mechanism, resulting in inaccurate proposals. To address this issue, we propose to predict the dental arch that passes through all tooth centroids and then propose an APS method based on the predicted dental arch to replace with FPS for generating accurate proposals.

3.2.2 Dental Arch Prediction

The dental arch can describe the teeth arrangement of a dental model. To automatically predict the dental arch for each dental model, we propose a coarse-to-fine dental arch prediction method. As shown in Fig 3, our proposed dental arch prediction method first roughly predicts the dental arch by regressing a cubic Bézier curve and then adopts a GCN-based network to refine the arch. In the following, we present our dental arch prediction method in detail.

Bézier Curve Regression Recently, it has been shown that the human dental arch form is accurately represented mathematically by the beta function [14]. Motivated by [14], we select a simple function, cubic Bézier curve, from the beta function set to initially approximate dental arch. The specific cubic Bézier curve can be decided by four control points. The ground truth of control points are obtained by minimizing the distance between the synthesized Bézier curve and the teeth centroids. As shown in the top of Fig. 3, we use an MLP to predict 4 control points $\{x_i^{ctr}\}_{i=1}^{4}$. The loss is defined as

$$L_{ctr} = \frac{1}{4}\sum_{i=1}^{4} \ell_1\left(\hat{\mathbf{x}}_i^{ctr} - \mathbf{x}_i^{ctr}\right) \tag{1}$$

where $\mathbf{x}_i^{ctr}$ and $\hat{\mathbf{x}}_i^{ctr}$ are the i-th points corresponding to the target and predicted control points, respectively. By regressing the 4 control points, we can obtain the final synthesized Bézier curve to characterize the dental arch initially.

GCN-based Arch Refinement We generate the target dental arch by connecting all the line segments that pass sequentially through the teeth centroids and then sampling uniform points from the connected line segments. The target and predicted dental arch are denoted as $\{x_i^{gt}\}_{i=1}^{N}$ and $\{\hat{x}_i\}_{i=1}^{N}$, respectively, where N is the number of points comprising the dental arch curve and is set to 32. As shown in the bottom of Fig 3, we first initialize the arch curve with uniformly-sampled points along the synthesized Bézier curve above. The nearest three votes corresponding to each initial arch point are selected, and their features are interpolated to represent the corresponding arch point features. The interpolated features are aggregated by MLP and then fed into our GCN for generating the offsets. We add the coordinates of the initial arch points and the learned offsets to generate new arch points. The learning process for generating offsets is iteratively repeated 3 times to refine the initial dental arch prediction, generating the fine prediction of the dental arch. The loss function for arch points prediction can be formulated as follows:

$$L_{arch} = \frac{1}{N}\sum_{i=1}^{N}\ell_1\left(\hat{\mathbf{x}}_i - \mathbf{x}_i^{gt}\right) \tag{2}$$

3.2.3 Arch-aware Point Sampling (APS)

With the estimated dental arch, we design an APS method to expressly select the points around the tooth crown to tackle the issue above. This is based on our observation that all teeth are sequentially arranged on a dental arch, so are their centroids. As shown in Fig. 2, the APS and grouping module makes use of the predicted dental arch and generates final N_k teeth proposals. Specifically, we utilize the Hungarian [6] method to sample subsampled points in N_m votes. Hungarian method considers the distances among assigned points and samples points more uniformly, compared with KNN-like methods that directly sample the K-nearest points to the dental arch. The cost matrix $\mathbb{C}$ for Hungarian method consists of two parts:

$$\mathbb{C} = \alpha\mathbb{D}_{\text{arch}} + \beta\mathbb{D}_{\text{votes}} \tag{3}$$

The first matrix $\mathbb{D}_{\text{arch}}$ is the Euclidean distance between votes and dental arch points. The second part $\mathbb{D}_{\text{votes}}$ is the Euclidean distance of votes displacement. α and β are used to balance the importance of these two distance measurements for sampling. We experimentally set α and β to be 1 and 5, respectively. The effect of different sampling methods on both detection and segmentation are compared, and the results are attached in the *Supplementary*.

3.2.4 Loss Function

When training the networks, only annotations of teeth centroids are utilized. We use the Huber ℓ_1 loss [23] L_{offset} to supervise the offsets prediction to obtain points F from original subsampled points to their nearest annotated centroids. Next, we use Cross-Entropy loss L_{conf} to supervise the proposal confidence. We assume the ground truth confidence of proposals which distances to their closest teeth centroids less than 0.3 to be 1 and assign the corresponding teeth centroids to the proposals such as VoteNet [18]. In the end, base on the assigned teeth centroids, we compute the losses L_{centers} and L_{boxs} for learning centroids offset and regressing teeth objects box regression [19]. Specifically, loss for the teeth detection is as follows:

$$L_{\text{det}} = L_{\text{offset}} + L_{\text{conf}} + \gamma L_{\text{centers}} \tag{4}$$

where we empirically set γ to be 0.1.

3.3. Tooth Instance Segmentation

Our segmentor is build upon PointNet++ [22]. We adopt a patch-based training strategy to train the segmentor and the common cross-entropy loss function to optimize the training process. Given a centroid point, we crop the closest $M = 2048$ points to the centroid point from the original point cloud P. As shown in Fig. 2, the input of segmentor are the backbone features and the relative coordinates to the given centroid of the cropped 3D patch, and the output is the probability mask indicating the possibility of the points from the 3D patch being tooth point. The training data for training our segmentor is all 3D patches generated by cropping the closest M points to those tooth centroids of labeled tooth instances. For example, if three tooth instances are labelled in a dental model, we will generate three 3D patches by cropping the closest M points to the three tooth centroids of labelled tooth instances. The patch-based training strategy can augment the training samples and fully utilize the annotation information. In the inference stage, the well-trained segmentor can segment all tooth instances of the entire dental model by fusing the segmentation results on all patches that are generated based on each detected centroid.

3.4. Network Training

For training the tooth centroid detection network, we sample $N = 16,000$ points uniformly from each dental model, using their 3D coordinates as the unique feature input. We first train the detection backbone in the first 210 epochs and other network settings, such as the optimizer and the learning rate, follows [18]. Then we train the arch prediction branch for 100 epochs with the fixed detection backbone. With the estimated dental arch, we perform APS to generate accurate proposals and fine-tune the network of proposal generation, as the yellow trapezoid denoted in Fig. 2. Non-maximum suppression (NMS) is applied to these proposals to generate the final centroid prediction. For training the tooth instance segmentation network, tooth centroids of those annotated teeth masks in the training dental models are used to generate 3D patches by cropping the closest $M = 2,048$ points to them from the corresponding point clouds. A patch-based training strategy is used for our segmentor. Our segmentator is build upon PointNet++ [22] and follow the similar settings of [22] in the training phase. All trainings are conducted under a single RTX 3090 Nvidia GPU. Please refer to *Supplementary* for detailed information.

Method	Tooth centroid detection			Tooth instance segmentation			
	Acc.	Recall	C. Dist.	Full		Weak	
				IoU	Dice	IoU	Dice
VoteNet [18]	88.82	85.68	0.036	-	-	-	-
MLCVNet [34]	90.86	85.68	**0.033**	-	-	-	-
Group-free 3D [11]	91.14	**92.70**	0.035	-	-	-	-
TSegNet [3]	99.41	84.94	0.037	94.83	96.91	93.39	95.83
VoteNet & PointNet++ [22]	84.32	**85.40**	0.040	93.92	96.29	93.38	95.97
DArch (Ours)	**99.68**	85.39	0.037	**95.93**	**97.70**	**95.42**	**97.38**

Table 1. Tooth centroid detection and tooth instance segmentation results compared with state-of-the-art methods in weakly- and fully annotated scenarios. ”-” denotes the unavailable segmentation scores for those detection methods.

4. Experiments

4.1. Dataset and Annotation

We collected $4,773$ 3D dental models from $3,231$ patients before orthodontics. We randomly select $3,973$ models as the training models and the rest 800 models as the test models. All training dental models contain a total of $54,658$ in teeth instances. All the dental models are fully annotated, in which all tooth instances of each dental model are manually labeled by professional dentists. In our work, we propose a low-cost annotation way, that is, labeling all tooth centroids and only a few teeth for each dental model. To calculate the time spent for full annotation and our proposed weak annotation, one of the authors manually annotates 10 dental models with different annotation ways under the guidance of professional dentists. Although teeth centroids used in our experiments are calculated by the fully annotated teeth masks, we propose a new way to annotate teeth centroids by multi-view images, which is less time-consuming. We first render the dental model to three images of different views. Then, performed in a strict sequence, we select the center point of each tooth on these images respectively to calculate the coordinates of the to-be-annotated tooth centroid. For annotating a tooth to generate the mask, we use the popular and programmable 3D mesh editing software, Meshlab [2], as our annotation tool. We use the Z-painting tool provided by Meshlab by painting vertexes on each tooth instance. Fig 4 shows an example of full annotation and our proposed weak annotation and indicates the averaged annotation time on one dental model for both types of annotation. As shown in Fig 4, the weak annotation way used in our work can save time to a large extent compared to the fully annotated approach used in other learning-based methods.

4.2. Experimental Setup

Competing methods. We compare our approach with the state-of-the-art method on tooth centroid detection and tooth instance segmentation. As for the detection, our DArch is compared with the popular 3D detection methods(*i.e.*, VoteNet [18], MLCVNet [34] and Group-free 3D [11]). VoteNet is a general 3D detection method for point clouds. MLCVNet extends the VoteNet by leveraging multi-level context modules, *i.e.*, patch-to-patch, object-to-object and global scene. Group-free 3D further adopts a transformer-based proposal generation networks. As for the segmentation, we compare our DArch with the state-of-the-art 3D tooth instance segmentation method (*i.e.*, TSegNet [3]) and the combination of popular VoteNet and PointNet++. TSegNet is the start-of-the-art learning-based method for 3D tooth instance segmentation.

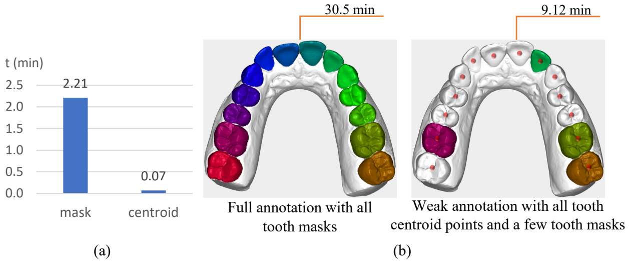

Figure 4. Illustration of time consumption for different annotation ways. (a) Comparison of time spent on labeling each tooth mask and centroid; (b) Comparison of time spent on labeling one dental model with full and weak annotations.

Metrics. We use the widely-used metrics-Accuracy (ACC) and Recall for evaluating the detection performance, as well as IoU and Dice metrics are used to evaluate the segmentation performance. Besides, we adopt an extra metric-Chamfer Distance [1] to measure the distance between the predicted centroids and the ground truth centroids. Given two point clouds $P_1 \subseteq R^3, P_2 \subseteq R^3$, The Chamfer Distance can be defined as

$$d_{CH}\left(P_1, P_2\right) = \sum_{x \in S_1} \min_{y \in P_2} \|x-y\|_2^2 + \sum_{y \in P_2} \min_{x \in P_1} \|x-y\|_2^2 \tag{5}$$

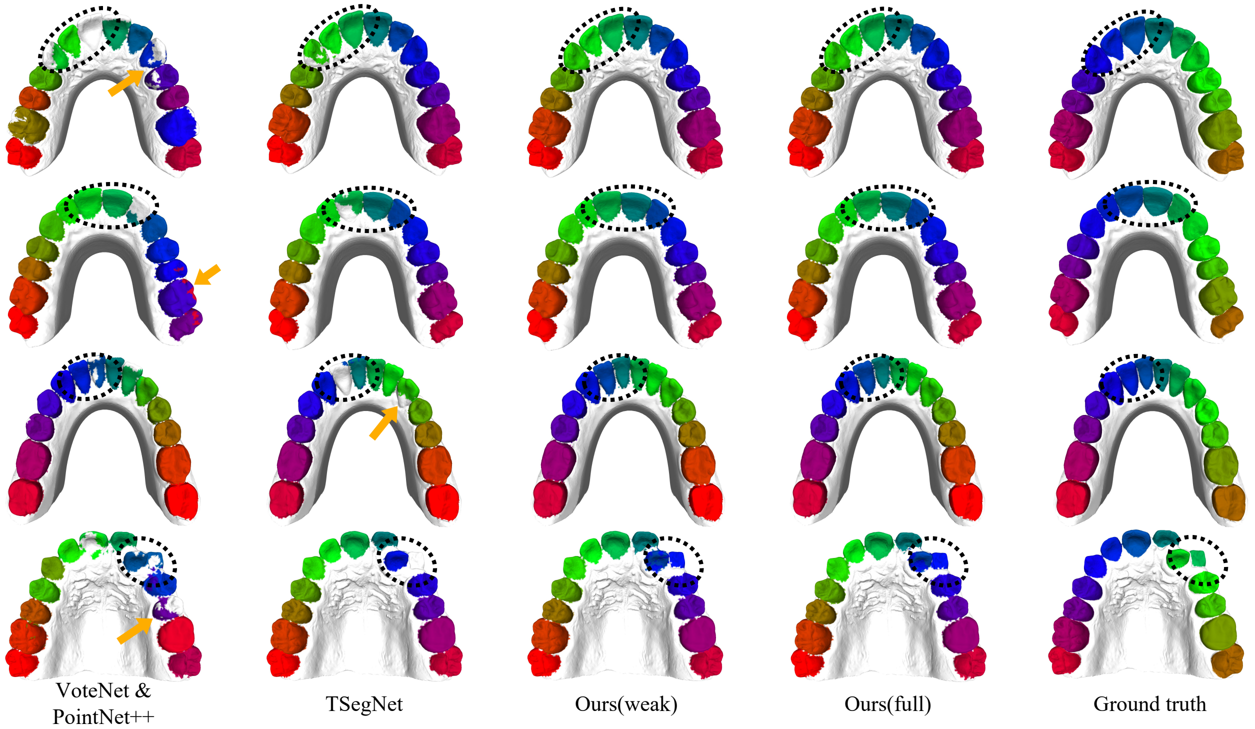

Figure 5. The visual comparison of dental model segmentation results produced by different methods, as well as the corresponding ground truth. From left to right are the results of other methods (1st-2nd columns) with full annotations, our results with weak annotations, our results with full annotations and the ground truth.

Method	Number	Tooth Centroid Detection			Tooth Instance Segmentation	
		Acc.	Recall	C. Dist.	IoU	Dice
FPS	20	84.32	85.40	0.040	93.38	95.97
	30	85.4	**85.66**	0.038	95.57	97.49
APS (Ours)	20	99.68	85.39	**0.037**	95.42	97.38
	30	**99.74**	85.37	**0.037**	**95.67**	**97.53**

Table 2. The detection and segmentation results when using different thresholding centroid number and sampling methods. 'Number' means the number of the detected tooth centroids in the detection stage.

4.3. Comparison with Competing Methods

Experimental setup. In this section, we compare our method with different competing methods. Note that all segmentation models of our DArch and another two competing methods, TSegNet and VoteNet PointNet++, adopt patch-based training strategy and fuse all patch-based segmentation results to produce the segmentation result of an entire dental model. The 3D patches that are used as the input of all segmentation models are generated by cropping the closest $2,048$ points to those detected tooth centroids. As we mentioned in Section 3.2.1, the tooth centroids detected by VoteNet and our DArch are generated by NMS filtering. By thresholding, VoteNet and our DArch can generate different numbers of the predicted tooth centroids. The number of detected tooth centroids can affect the detection and segmentation results. With a small increase in the number of the detected tooth centroids, the detection recall may increase, and the segmentation performance also may improve at the expense of decreased efficiency since the segmentation results from more patches are fused. Our experimental statistics yield an average number of the detected tooth centroids for TSegNet model of about 28.6. For fair comparison and taking into account model efficiency, we filter the proposals of VoteNet and our DArch and generate 20 tooth centroids for both methods.

Results. The overall detection and segmentation results are presented in Table 1, and we compared these competing methods in both weakly- (*i.e.*, only labeling 20% teeth instances from all tooth instances in the training dental models) and fully annotated scenarios. Since VoteNet [18], MLCVNet [34] and Group-free 3D [11] can only be used for detection, their segmentation metrics are default. From the table, we can see that our DArch achieves the best segmentation performance in both weakly- and fully-annotated scenarios. Compared to the state-of-the-art 3D tooth instance segmentation method, TSegNet, the proposed DArch improves the segmentation performance by 1.1% and 0.79% on the IoU and Dice, respectively, with full annotations, and by 2.03% and 1.55% on the IoU and Dice, respectively, with weak annotations. In the weakly-annotated scenario, our DArch improves more. The reason may be that our method can generate more accurate detection results. Owing to accurate detection results, our segmentation models perform well even in the weakly-annotated scenario. This also suggests that locating tooth objects is important for the segmentation, and our proposed weak annotation is feasible. The visual results of our method and other methods are shown in Fig 5. From this figure, we can find that even only weak annotations are available for our DArch, it can also produce visually better results than other methods with full annotations, especially in areas of small teeth.

4.4. Ablation Studies

4.4.1 Sampling

Thresholding Centroid Number and different sampling methods in the detection stage will affect the detection and segmentation performance. In this section, we investigate the effect of different thresholding centroid numbers (*i.e.*, 20 and 30) and sampling methods (*i.e.*, FPS and APS) on the tooth centroid detection and tooth instance segmentation. The results are reported in Table 2. From this table, we can observe that our proposed APS method achieves the best results in terms of the detection and segmentation results, especially the ACCs are much higher than that of other sampling methods in different centroid numbers. Besides, when the thresholding centroid number is low (*i.e.*, 20), our APS remains reflecting a relatively consistent detection and segmentation performance with the higher centroid number of 30, while FPS decreases more. This also suggests that by leveraging dental arch prior, our APS can detect more accurate centroid points than the conventional FPS method.

4.4.2 Dental Arch Prediction

In our work, we propose a coarse-to-fine method for predicting dental arches. We first synthesize a cubic Bézier curve using an MLP network to initially characterize the dental arch and then use a lightweight network to refine the initially estimated arch. To validate the effectiveness of the coarse-to-fine strategy, we predict the dental arch using different methods, such as direct prediction using an MLP, coarse Bézier curve regression and our proposed coarse-to-fine strategy. The results are reported in the Table 3. The results in this table indicate the effectiveness of our coarse-to-fine strategy on arch prediction. The analysis of hyperparameters is attached in the *Supplementary*.

Method	Acc.	Recall	MSE. (1e-4)
Direct*	93.13	85.12	7.50
Coarse	93.44	85.27	6.22
Coarse + Fine	**99.89**	**84.17**	**4.36**

Table 3. Ablation study of Arch prediction. Direct* denotes directly predicting the arch points using an MLP network. Coarse denotes predicting the arch points only by Bézier curve regression. Fine indicates further refining the coarse prediction.

5. Conclusion

In this work, we propose a novel tooth instance segmentation framework-*DArch*. Our DArch consists of two parts of tooth centroid detection and tooth instance segmentation. This method provides a novel dental arch estimation method and introduces an arch-aware point sampling (APS) module based on the estimated dental arch for tooth centroid detection. Owing to the impressive detection performance obtained by the detection stage, our DArch has achieved superior performance to the competing segmentation methods in both weakly- and fully-annotated scenarios. Our segmentor is trained in a fully-supervised manner and does not take full advantage of the weakly-annotated centroid information and our proposed dental arch prior. In the future, we will design a smarter segmentor by fully leveraging this information.

Broader Impact. The segmentor of our DArch is trained in a fully-supervised manner. The training data is limited when only a small amount of teeth are manually labeled, which will limit the generalization ability of the trained segmentor. The model may generate inaccurate segmentation results on unseen dental models from the real-world, thus adversely affecting the computer-aided orthodontic treatments.

Acknowledgements: The work was supported in part by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK ST Cooperation Zone, National Key RD Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, and by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104. This work was also supported by NSFC-62172348, 61902334, Shenzhen General Project (JCYJ20190814112007258) and High-Performance Computing Portal under the administration of CUHK(Shenzhen) Information Technology Services Office.

References

[1] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER, 1977. 6

[2] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, Guido Ranzuglia, et al. Meshlab: an open-source mesh processing tool. In *Eurographics Italian chapter conference*, volume 2008, pages 129–136. Salerno, Italy, 2008. 6

[3] 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, 2021. 1, 2, 3, 6

[4] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020. 2

[5] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4867–4876, 2020. 2

[6] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[7] Truc Le, Giang Bui, and Ye Duan. A multi-view recurrent neural network for 3d mesh segmentation. *Computers & Graphics*, 66:103–112, 2017. 2

[8] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31:820–830, 2018. 2

[9] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 641–656, 2018. 2

[10] Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with curve-gcn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5257–5266, 2019. 2

[11] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. *arXiv preprint arXiv:2104.00678*, 2021. 6, 8

[12] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928. IEEE, 2015. 2

[13] Ehsan Nezhadarya, Yang Liu, and Bingbing Liu. Boxnet: A deep learning method for 2d bounding box estimation from bird's-eye view point cloud. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1557–1564. IEEE, 2019. 2

[14] Hassan Noroozi, Tahereh Hosseinzadeh Nik, and Reza Saeeda. The dental arch form revisited. *The Angle Orthodontist*, 71(5):386–389, 2001. 4

[15] Guan Pang and Ulrich Neumann. 3d point cloud object detection with multi-view convolutional neural network. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 585–590. IEEE, 2016. 2

[16] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[17] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2020. 2

[18] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2, 4, 5, 6, 8

[19] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018. 5

[20] 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. 2

[21] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 2

[22] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 2, 4, 5, 6

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 5

[24] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3577–3586, 2017. 2

[25] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020. 2

[26] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 2

[27] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, pages 808–816, 2016. 2

[28] Diya Sun, Yuru Pei, Peixin Li, Guangying Song, Yuke Guo, Hongbin Zha, and Tianmin Xu. Automatic tooth segmentation and dense correspondence of 3d dental model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 703–712. Springer, 2020. 1, 3

[29] Diya Sun, Yuru Pei, Guangying Song, Yuke Guo, Gengyu Ma, Tianmin Xu, and Hongbin Zha. Tooth segmentation and labeling from digital dental casts. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 669–673. IEEE, 2020. 1, 3

[30] 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. 1, 3

[31] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017. 2

[32] Zongji Wang and Feng Lu. Voxsegnet: Volumetric cnns for semantic part segmentation of 3d shapes. *IEEE transactions on visualization and computer graphics*, 26(9):2919–2930, 2019. 2

[33] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2

[34] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. Mlcvnet: Multi-level context votenet for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10447–10456, 2020. 6, 8

[35] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3182, 2021. 2

[36] 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

[37] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Realtime 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018. 2

[38] Farhad Ghazvinian Zanjani, David Anssari Moin, Frank Claessen, Teo Cherici, Sarah Parinussa, Arash Pourtaherian, Svitlana Zinger, et al. Mask-mcnet: Instance segmentation in 3d point cloud of intra-oral scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 128–136. Springer, 2019. 1, 3

[39] Jianda Zhang, Chunpeng Li, Qiang Song, Lin Gao, and Yu-Kun Lai. Automatic 3d tooth segmentation using convolutional neural networks in harmonic parameter space. *Graphical Models*, 109:101071, 2020. 1, 3

[40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6699–6708, 2021. 3

[41] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 2