

# ToothNet: Automatic Tooth Instance Segmentation and Identification from Cone Beam CT Images

Zhiming Cui Changjian Li Wenping Wang

The University of Hong Kong

{zmcui, cjli, wenping}@cs.hku.hk

## Abstract

This paper proposes a method that uses deep convolutional neural networks to achieve automatic and accurate tooth instance segmentation and identification from CBCT (cone beam CT) images for digital dentistry. The core of our method is a two-stage network. In the first stage, an edge map is extracted from the input CBCT image to enhance image contrast along shape boundaries. Then this edge map and the input images are passed to the second stage. In the second stage, we build our network upon the 3D region proposal network (RPN) with a novel learned-similarity matrix to help efficiently remove redundant proposals, speed up training and save GPU memory. To resolve the ambiguity in the identification task, we encode teeth spatial relationships as an additional feature input in the identification task, which helps to remarkably improve the identification accuracy. Our evaluation, comparison and comprehensive ablation studies demonstrate that our method produces accurate instance segmentation and identification results automatically and outperforms the state-of-the-art approaches. To the best of our knowledge, our method is the first to use neural networks to achieve automatic tooth segmentation and identification from CBCT images.

## 1. Introduction

Digital dentistry has been developing rapidly in the past decade. The key to digital dentistry is the acquisition and segmentation of complete 3D teeth models; for example, they are needed for specifying the target setup and movements of individual teeth for orthodontic diagnosis and treatment planning. However, acquiring complete 3D input teeth models is a challenging task. Currently, there are two mainstream technologies for acquiring 3D teeth models: (1) Intraoral or desktop scanning; and (2) cone beam computed tomography (CBCT) [25]. Intraoral or desktop scanning is a convenient way to obtain surface geometry of tooth crowns but it cannot provide any information of tooth

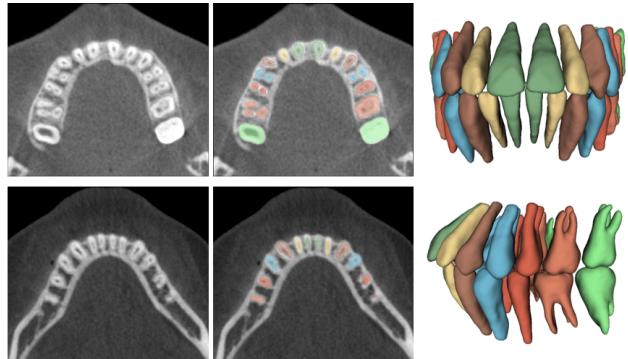


Figure 1. An example of tooth segmentation and tooth identification. The first column shows a CBCT scan in the axis view, the second column shows its segmentation results, and the last column shows the 3D segmentation results with different colors for different teeth respectively.

roots, which is needed for accurate diagnosis and treatment in many cases. In contrast, CBCT provides more comprehensive 3D volumetric information of all oral tissues, including teeth. Because of its high spatial resolution, CBCT is suitable for 3D image reconstruction and is widely used for oral surgery and digital orthodontics. In this paper, we focus on 3D tooth instance segmentation and identification from CBCT image data, which is a critical task for applications in digital orthodontics, as shown in Fig. 1.

Segmenting teeth from CBCT images is a difficult problem for the following reasons. (1) When CBCT is acquired in the nature occlusion condition (i.e., the lower teeth and upper teeth are in touch in the normal bite condition), it is hard to separate a lower tooth from its opposing upper teeth along their occlusal surface because of the lack of changes in gray values [18, 22]; (2) Similarly, it is hard to separate a tooth from its surrounding alveolar bone due to their highly similar densities; and (3) Adjacent teeth with similar shape appearance are likely to confuse the effort of identifying different tooth instances (See for example the two maxillary central incisors in Fig. 1). Hence, successful tooth segmentation can hardly be achieved by relying only on the intensity variation of CT images, as shown by many previ-

ous attempts tooth segmentation methods.

To address the above issues, some previous works exploit either the level-set method [11, 18, 13, 22] or the template-based fitting method [2] for tooth segmentation. The former methods are restricted by their need for a feasible initialization that requires tedious user annotations, and they produce unsatisfactory results when teeth are in natural occlusion condition. The later methods lack the necessary robustness in practice when there are large shape variations for different patients. Recently, many deep learning methods for medical image analysis [41, 42, 40], though have not been applied to tooth segmentation, have demonstrated promising performance over traditional methods in various tasks.

All these previous works have motivated us to solve the problem of tooth segmentation from CBCT by using a data-driven method which learns the shape and data priors simultaneously. Specifically, we present a novel learning-based method for automatic tooth instance segmentation and identification. That is, we aim to segment all the teeth from the surrounding tissues, separate the teeth from each other, and identify each tooth by assigning to it a correct label.

The core of our method is a two-stage deep supervised neural network. In the first stage, to enhance the boundary of blurring and low contrast signals, we train an edge extraction subnetwork. In the second stage, we devise a 3D region proposal network [33] with a novel learned similarity matrix which efficiently removes the duplicated proposals, speeds up and stabilizes the training process, and significantly cuts down the GPU memory usage. The input CBCT images combined with the extracted edge map are then sent to the 3D RPN followed by four individual branches for segmentation, classification, 3D bounding box regression, and identification tasks. To resolve the identification ambiguity, we take into consideration tooth spatial position by adding a spatial relation component to encode the position features to improve identification accuracy. To the best of knowledge, our method is the first to apply deep neural networks to automatic tooth instance segmentation and identification from CBCT images.

We train our neural networks on a proprietary data set of CBCT images collected by the radiologists in our team and validate our method with extensive experiments and comparisons with the state-of-the-art methods, as well as comprehensive ablation studies. The experiments and comparisons demonstrate that our method produces superior results and significantly outperforms other existing methods.

## 2. Related work

**Object Detection and Segmentation.** Driven by the effectiveness of deep learning, many approaches in object detection [33, 15, 28, 17] and instance segmentation [27, 7, 6, 32, 31, 21] have achieved promising results. In partic-

ular, R-CNN [[16]] introduces an object proposal scheme and establishes a baseline for 2D object detection. Faster R-CNN [33] advances the stream by proposing a Region Propose Network (RPN). Mask R-CNN [17] extends Faster R-CNN by adding an additional branch that outputs the object mask for instance segmentation. Following the set of representative 2D R-CNN based works, 3D CNNs have been proposed to detect objects and estimate poses relying on 3D bounding box detection [34, 36, 37, 8, 5, 14] on 3D voxelized data. Girdhar et al. [14] extend Mask R-CNN to the 3D domain by creating 3D RPN for key point tracking. Inspired by the success of region-based methods on object detection and segmentation, we exploit 3D Mask R-CNN as the base network.

**CNNs for Medical Image Segmentation.** CNNs-based methods for medical image analysis have demonstrated excellent performance on many challenging tasks, including classification [35], detection [38] and segmentation [9, 40]. Note that medical images usually appear in a volumetric form, e.g., 3D CT scans and MR images, many works employ 2D CNNs taking input the adjacent 2D slices [4] from the 3D volume. Though coping with 2D data will not consume too much GPU resources, the 3D spatial information lying in volumetric data is not fully exploited. To directly apply convolutional layers on 3D data, more 3D CNN-based algorithms [9, 3, 23] are proposed. However, the existing methods target only on semantic level segmentation or classification rather than instance level which is essential in orthodontic diagnosis.

**Tooth Segmentation from CBCT Images.** Accurate tooth segmentation from CBCT is a fundamental step for individual 3D tooth model reconstruction, which can assist doctors in orthodontic diagnosis and treatment planning [11, 43]. Many traditional algorithms are proposed for tooth segmentation, reflecting the importance of this application. Driven by the intensity distribution in CBCT images, previous approaches resort to region growing [1, 24] and level sets boosted variants [39, 22, 12]. By further considering the prior knowledge of tooth, statistical shape models [30, 2] become the most powerful and efficient choice. However, these methods always suffer from many artifacts or failures even with excellent manual initialization.

## 3. Methods

The core of our method is a two-stage deep neural network. In the first stage, we extract the edge map from CBCT images by a deep supervised network. In the second stage, we concat the learned edge map features with the original image features and send them to the 3D RPN. Then we propose one learned similarity matrix to filter the plenty of redundant proposals inner the 3D RPN module, and one spatial relationship component to further resolve the ambiguity

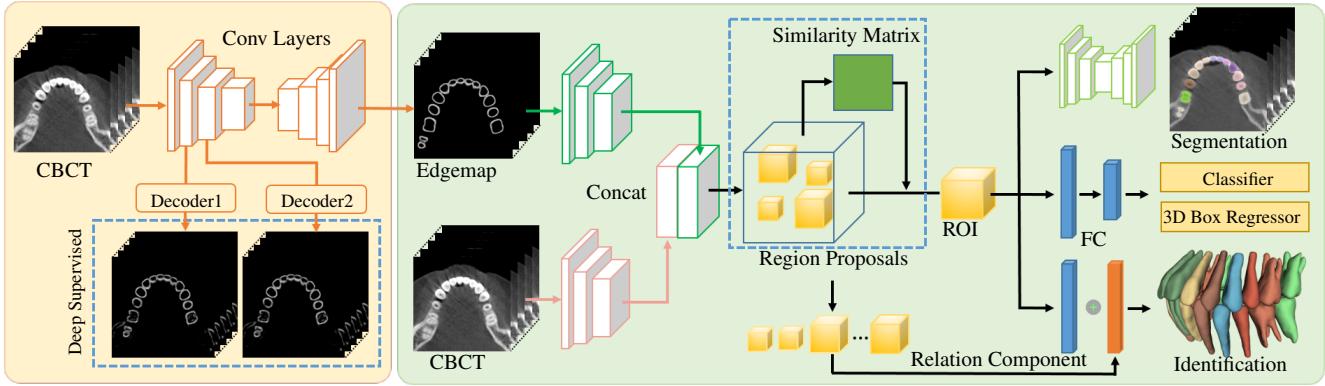


Figure 2. Our two-stage network architecture for tooth instance segmentation and identification. Given CBCT images, we first pass them to the edge map extraction network in stage one, where the deep supervised scheme is used. Then the detected edge map and the original CBCT images are sent to the region proposal network with a novel similarity matrix. Four individual branches are followed for tooth segmentation, classification, 3D box regressor and identification. In the identification branch, we further add the spatial relation component to help resolve the ambiguity.

of tooth identification. Fig. 2 shows an overview of the whole pipeline.

### 3.1. Base Network

Inspired by the excellent performance of R-CNN based networks for general object segmentation and classification, we extend the pipeline of Mask R-CNN [17] to a 3D version as our base network.

In the backbone feature extraction module, we first apply five 3D convolutional layers to CBCT images. Then the encoded features are fed into the 3D RPN module where we use the same structure as in [14] except the number of anchors. Since the teeth size is relatively similar with little variation, we tune the number of anchor to 1 at each sliding position. In addition, we add one more branch for identification as shown in Fig. 2.

Finally, the loss with multiple tasks for the base network is defined as:

$$\mathcal{L}_b = \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{seg} + \mathcal{L}_{id}, \quad (1)$$

where the classification loss  $\mathcal{L}_{cls}$ , the 3D bounding-box regression loss  $\mathcal{L}_{box}$  and the segmentation loss  $\mathcal{L}_{seg}$  are identical as defined in [17]. And  $\mathcal{L}_{id}$  is a log-softmax function for tooth identification.

### 3.2. Our Network

#### 3.2.1 Edge Map Prediction and Representation

The blurring signals in CBCT images make it hard to find the clear tooth boundary. Besides, the low contrast value of touching teeth prevents an accurate segmentation. To solve these problems, we propose to extract an edge map from CBCT images to enhance clear boundary information.

Given a CBCT volume data, which is annotated with a multi-label ground truth segmentation  $Y$  where  $y_i = k$  indicates the  $i$ th tooth has label  $k$ . Then a binary edge map

$E_B$  of the same size can be produced by setting the voxels on the tooth boundary to be 1 according to  $Y$ , others are set to 0. Finally, a Gaussian filter  $G$  with standard deviation  $\sigma = 0.1$  is applied to the binary edge map  $E_B$  to generate ground truth edge map  $E$ .

To obtain fast convergence and more accurate prediction, we employ deep supervised learning [26, 9] to train the edge map detection network and enforce the learning of ground truth edge map from three different feature levels as shown in Fig. 2. Specifically, the network consists of one encoder with nine convolutional layers and three branches of decoders linking with lower-level, middle-level and high-level features from the encoder. Then the loss function using mean squared error (MSE) is defined as:

$$\mathcal{L}_{EM} = \sum_{i=0,1,2} \|E'_i - E\|_2^2, \quad (2)$$

where  $E$  denotes the ground truth edge map and  $E'_i$  is the predicted edge map by different levels of features.

Having the edge map, we first apply three individual conv layers (not shared) to edge map which comes from the deepest decoder, and the original CBCT images. Then we concatenate them together with another five conv layers to be the input of the region proposal module.

#### 3.2.2 Similarity Matrix

In the base network, 3D RPN module generates a set of region proposals and removes the duplicate ones using non-maximum suppression (NMS) before sending to the 3D ROIAlign module. The challenge here is two-fold: 1) the huge consumption of GPU memory on 3D volumetric data prevents us setting big ROI number (32 (3D) vs. 512 (2D)); 2) the NMS method depends on regressed bounding box position to remove redundant proposals, which is somewhat inaccurate. To overcome these challenges, we propose a

nature similarity matrix component that exploits shape features directly to remove duplicate proposals efficiently.

In contrast to the NMS method utilizing simple relations of regressed candidate bounding boxes and scores, we train a similarity matrix  $S$  employing features of different proposals. To train the similarity matrix, we first obtain the top- $k$  ( $k=256$  is used) ranked proposals generated by 3D RPN, denoted as  $P = \{P_0, P_1, \dots, P_k\}$ .  $S$  has the dimension  $k \times k$ , and the element  $S_{ij}$  represents the possibility of proposals  $P_i$  and  $P_j$  containing the same tooth. In training stage, for any pair of proposals  $P_i$  and  $P_j$  in  $P$ , we first extract their corresponding features  $F_{P_i}$  and  $F_{P_j}$  in the backbone convolutional layer, and then we concatenate them together and send them to the fully-connected layers to output a binary classification probability, which is supervised by the ground truth similarity matrix introduced in the following.

The preparation of the ground truth similarity matrix  $S_G$  is divided into two steps. Suppose we have  $m$  ground truth bounding box in current patch, denoted as  $B = B_0, B_1, \dots, B_m$ , given the candidate proposal  $P_i \in P$ , we first calculate the Intersection-over-Union score between the bounding box of  $P_i$  and each bounding box in  $B$ . If the highest IoU score  $P_{iou}^i$  is derived between  $P_i$  and  $B_c$ ,  $P_i$  gets the object index  $c$  representing that  $P_i$  contains same tooth as in  $B_c$ . Then in step two, we fill the value of  $S_{ij}^G$  following the three rules: 1)  $S_{ij}^G = 1$  if the pair of proposals  $\{P_i, P_j\}$  has the same object index, and both of their IoU scores  $P_{iou}^i$  and  $P_{iou}^j$  are higher than  $\eta$ ; 2)  $S_{ij}^G = 0$  if the pair of proposals  $\{P_i, P_j\}$  has different object indices, and both of their IoU scores  $P_{iou}^i$  and  $P_{iou}^j$  are higher than  $\eta$ ; 3)  $S_{ij}^G = -1$  if one of the IoU scores  $P_{iou}^i$  or  $P_{iou}^j$  of the pair of proposals  $\{P_i, P_j\}$  is not higher than  $\eta$ , where  $\eta = 0.2$  in all of our experiments. With ground truth matrix, the network can learn the similarity matrix  $S$  via the loss functions defined as:

$$\mathcal{L}_{SM} = \sum_{(i,j) \in \varepsilon} S_{ij}^G \log S_{ij} + (1 - S_{ij}^G) \log(1 - S_{ij}), \quad (3)$$

where  $(i, j) \in \varepsilon$  indicates the set of elements  $(i, j)$  that satisfy  $G_{ij} \neq -1$ .

In testing stage, the learned similarity matrix  $S$  is treated as a look-up table. That is for any pair of proposals  $\{P_i, P_j\}$ , if the element  $S_{ij} > 0.5$ , we discard the duplicate proposal with a lower classification score. Eventually, the redundant proposals are removed efficiently and the selected proposals are sent to the following steps for tooth detection, segmentation, and identification.

### 3.2.3 Tooth Identification

To identify every tooth with a distinct label, we obey the ISO standard tooth numbering system as shown in Fig. 3,

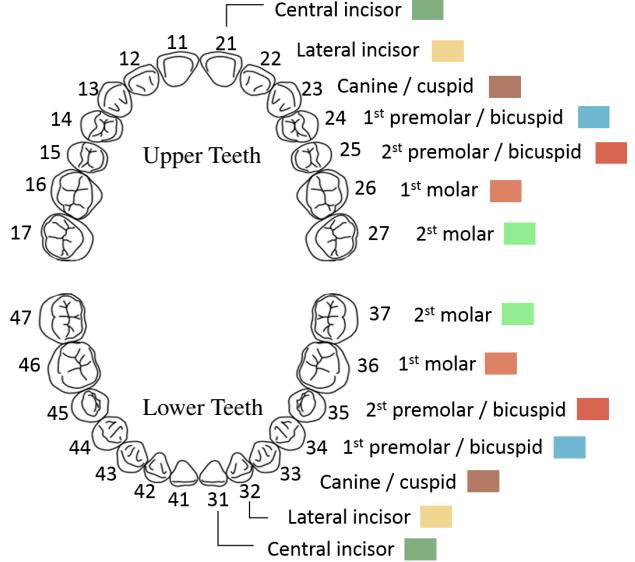


Figure 3. Tooth numbering system and the corresponding color coding.

where the mouth is split into four quadrants: upper right, upper left, lower right, and lower left respectively. Each quadrant has seven teeth with different types. The wisdom teeth are excluded from this study because of limited samples. Throughout the paper, we use the color coding shown in Fig. 3 to visualize the teeth labels. However, we observe that the general classifier would be confused if two neighboring teeth have similar shapes, e.g., molars and central incisors, without considering the spatial relationship.

To tackle this problem, we propose to encode the neighboring teeth spatial boxes and shape features as additional features for the identification task. Specifically, given a candidate proposal  $P_i$  ( $P_i \in \{P_1, P_2, \dots, P_n\}$ ,  $n$  equals to the number of ROI) after ROIAlign module, we first obtain the compacted shape feature. Then taking the neighboring spatial relations into consideration, we build the relation feature as a weighted sum of shape features from all other proposals. The relation weight indicates the impact from other proposals and can be calculated by the geometric box features following the idea [20]. Having the spatial relationship encoded, the identification branch takes the shape feature and relation feature as input, which is supervised by the ground truth label using a soft-max function.

**Final Loss Function.** In the end, with all these proposed novel components, our network is trained using the overall loss function combining the loss of base network and similarity matrix loss, defined as:

$$\mathcal{L} = \mathcal{L}_b + \lambda \mathcal{L}_{SM}, \quad (4)$$

where  $\lambda = 0.5$  for all experiments.

### 3.3. Dataset and Network Training

To train our network, we collect a CBCT dataset from some patients before or after orthodontics. The dataset contains 20 3D CT scans with a resolution varied from 0.25 mm to 0.35 mm (12 for training and 8 for testing). We then normalize the intensity of the CBCT image to the range of  $[0, 1]$ . To generate the training data, we randomly crop 150 patches of size  $128 \times 128 \times 128$  around the alveolar bone ridge in the CT scan and finally acquire about 1800 patches as training data. The ground truth of the dataset is annotated with a tooth-level bounding box, mask, and label. In the test phase, the overlapped sliding window method is applied to crop sub-volumes with a stride  $32 \times 32 \times 32$ . Then for two overlapped teeth, we use the one with a maximum value of  $P_{cls} \times P_{id}$  to be the final tooth prediction if the  $IoU$  of their teeth segmentation results is higher than 0.2, where  $P_{cls}$  and  $P_{id}$  indicate the tooth classification and identification probabilities respectively.

The network is trained in a two-step process. We first train the edge map extraction sub-network for 10 epochs in step one and fix it in step two, where we train the segmentation and identification sub-networks for 10 epochs as well. All the networks are implemented in PyTorch and trained on the server with an Nvidia GeForce 1080Ti GPU, using Adam solver with a fixed learning rate 0.001. Generally, the total training time is about 30 hours (6 hours for stage one and 24 hours for stage two respectively).

## 4. Results and Discussion

To evaluate our algorithm, we feed tooth CBCT images in our testing dataset to the two-stage network, and the complete 3D teeth model are reconstructed using 3D Slicer [10] given the labels from network outputs. Some representative results are shown in Fig. 1 and 7. Note that different colors indicate different tooth types as defined in Sec. 3.2.3. Furthermore, we conduct ablation studies (Sec. 4.1) and comparison with the state-of-the-art methods (Sec. 4.2) quantitatively and qualitatively.

**Error Metric.** We report three error metrics in this paper, i.e., the accuracy for tooth segmentation, detection and identification respectively. To evaluate tooth segmentation accuracy, we employ the widely used Dice similarity coefficient ( $DSC$ ) metric and the formulation is:

$$DSC = \frac{2 \times |Y \cap Z|}{|Y| + |Z|}, \quad (5)$$

where  $Y$  and  $Z$  refer to the voxelized prediction results and ground truth masks. Furthermore, we define the accuracy of detection and identification as follows: suppose  $G$  is the set of all teeth in ground truth data, and  $D$  is the set of teeth detected by our network, and within  $D$  we have  $L$  right teeth

Metric Network	DSC	DA	FA
bNet	89.73%	96.39%	90.54%
bENet	<b>91.98%</b>	<b>97.75%</b>	<b>92.79%</b>

Table 1. Accuracy comparison of bNet and bENet.

Nb <sub>ROI</sub>	Method	Metric		
		DSC	DA	FA
32	NMS	91.98%	97.75%	92.79%
	SM	92.10%	98.20%	93.24%
16	NMS	91.08%	95.49%	90.54%
	SM	92.07%	98.20%	93.24%
12	NMS	86.76%	83.33%	77.93%
	SM	91.77%	96.85%	90.99%
8	NMS	77.07%	68.92%	65.32%
	SM	89.86%	88.29%	82.91%

Table 2. Performance comparison between the NMS and our SM under different ROI numbers.

labels. The detection accuracy ( $DA$ ) and identification accuracy ( $FA$ ) are calculated as:

$$DA = \frac{|D|}{|D \cup G|} \quad \text{and} \quad FA = \frac{L}{|D \cup G|}. \quad (6)$$

All the experiments are performed on a machine with Intel(R) Xeon(R) E5-2628 1.90GHz CPU and 256GB RAM.

### 4.1. Ablation Study

To validate the effects of our two-stage network components, we have done additional experiments by augmenting the base network (Sec. 3.1) with our proposed novel components. All alternative networks are trained on the same dataset, and we report the accuracy on our test dataset for comparison.

**Edge Map.** To validate the effect of the edge map input, we augment the base network (bNet) with edge map detection stage, and the detected edge map is combined with original CBCT images as the input for the following tasks. Here we use bENet as the notation of this variation. We then compare the results from both networks as shown in Tab. 1 and Fig. 4. Statistically, we acquire higher accuracy on all our three subtasks and gain a remarkable 2.25% increasing in terms of segmentation accuracy, though the bNet has obtained promising results. And visually we select three typical cases, where the edge map has a great advantage. With the edge map, the accurate boundary on body part (the first row in Fig. 4), crown part (the second row in Fig. 4) with touching teeth and even root part (the third row in Fig. 4) with low contrast between tooth and alveolar bone can be found benefiting for the accurate teeth reconstructed.

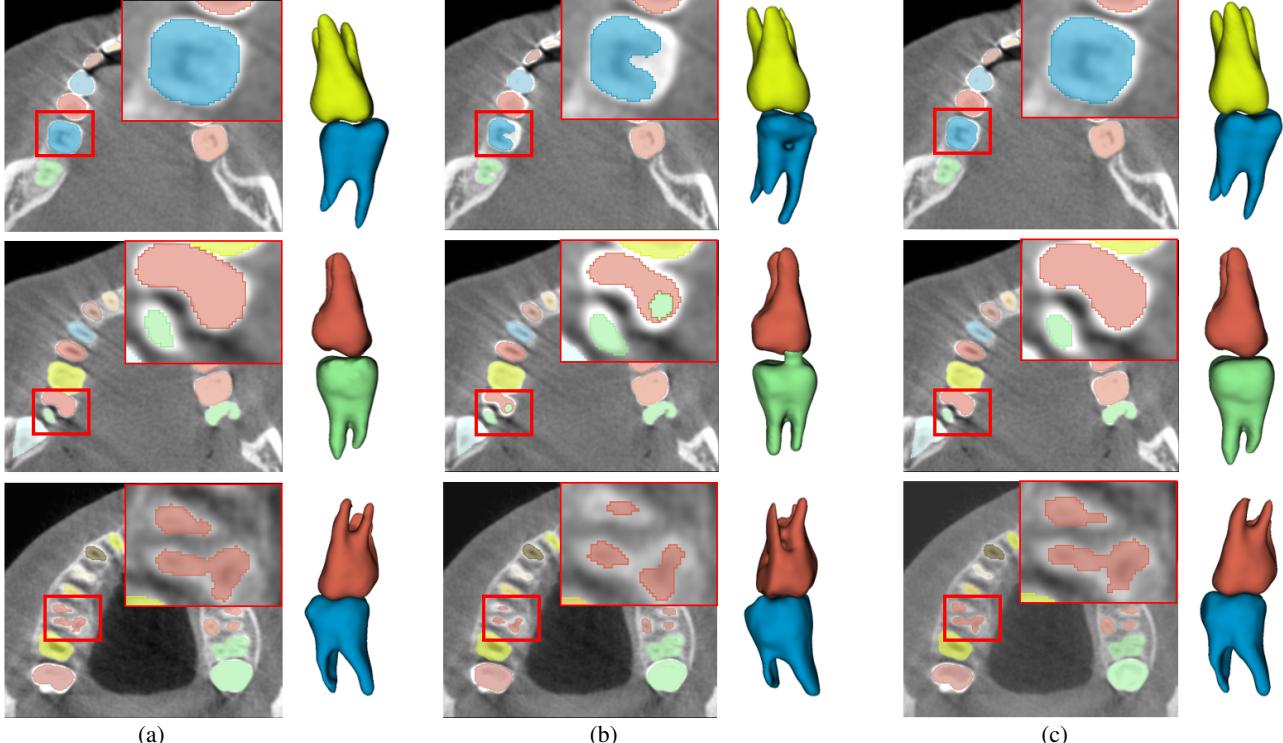


Figure 4. The visual comparison between networks w/wo edge map extraction subnetwork. We show the axial-aligned CT image with some details zooming in on some areas and the corresponding 3D reconstruction. (a) Results from ground truth data, (b) results from bNet, and (c) results from bENet. The comparison is performed row by row.

Nb <sub>ROI</sub>	32	16	12	8
Memory	9.3GB	6.3GB	5.4GB	4.6GB
T <sub>NMS</sub>	34.0h	28.0h	24.5h	22.5h
T <sub>SM</sub>	25.0h	24.0h	23.5h	22.0h

Table 3. The statistics of GPU memory usage and training time under different ROI numbers for both the NMS and our SM.

Metric Network \	DSC	DA	FA
bESNet	92.07%	98.20%	93.24%
fullNet	<b>92.37%</b>	<b>99.55%</b>	<b>96.85%</b>

Table 4. Accuracy comparison of networks w/wo the spatial relation component.

**Similarity Matrix (SM).** The huge amount of proposals in 3D RPN prevent us from setting bigger ROI in practice training, where bigger ROI number means more GPU memory usage but higher ability to include more object candidates. Thus we design the control experiments by applying our similarity matrix to replace the traditional NMS in bENet. We test both networks with various ROI numbers and the statistical results are shown in Tab. 2. We estimate the training time and GPU memory usage roughly and the statistics are reported in Tab. 3. Using the same ROI num-

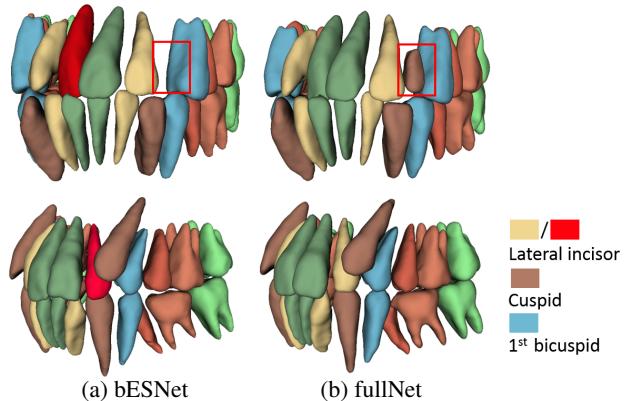


Figure 5. The qualitative comparison of tooth identification w/wo the spatial relation component (SR). Different tooth types are represented by different colors as defined in Sec. 3.2.3 and red color represents the wrong label.

ber, SM generates superior accuracy over NMS on all three accuracy metrics. And even using  $ROI = 12$ , SM produces comparable results with NMS using  $ROI = 32$  (91.77% vs. 91.98%). But SM uses as less as 44.7% training hours and 72.2% GPU memory (23.5 vs. 34.0 hours, 5.4 vs. 9.3 GB respectively), which we argue that our similarity matrix significantly speeds up the training process and saves GPU memory under specific quality control. One interesting observation is that when we set  $ROI = 8$ , the accuracy

of NMS decreases drastically, while we still can produce 89.86% segmentation accuracy which is comparable with bNet ( $ROI = 32$ , NMS is used). The reason here is that when setting small ROI number, NMS receives less instance objects while our SM encourages more instance objects efficiently using object features.

Using SM and setting  $ROI = 16$ , we already remove most of the redundant proposals, thus we only get a slightly increasing in terms of segmentation accuracy with  $ROI = 32$ , as shown in Tab. 2. In order to leverage the advantages of SM for efficient network training, we use  $ROI = 16$  in the following spatial relation component ablation test and our final full network.

Note that in NMS method, the IoU threshold  $N_t$  will affect the performance. We empirically set  $N_t = 0.2$  based on our substantial experiments to encourage better results.

**Spatial Relation Component.** To validate the effectiveness of the spatial relation component in resolving the identification ambiguity, we further augment bESNet (bENet with similarity matrix) with spatial relation component which is our final two-stage full network (fullNet) and compare the accuracy performance on our three subtasks (see statistics in Tab. 4). With spatial relation network, the performance of tooth identification and detection tasks earn about 3.61% and 1.35% growth with almost all teeth detected and labeled correctly.

We also present the visual comparison in Fig 5. Using spatial relation component, the two similar central incisors (the first row in Fig. 5) are correctly identified. Besides, the cuspid tooth grows in a wrong direction (the second row in Fig. 5), such that the lateral incisor and cuspid are spatially too close to the 1<sup>st</sup> bicuspids tooth. Without taking the spatial relation into consideration, the identification of lateral incisor tooth will be affected by the label of 1<sup>st</sup> bicuspids tooth which has bigger volume and is easy to recognize. Instead, with the spatial relation included, the label for the lateral incisor tooth is correctly predicted. Furthermore, spatial relation component has a positive effect on the segmentation task since it detects the tiny tooth as highlighted in the red box in Fig. 5, which attributes to the positive correlation between three subtasks.

## 4.2. Comparison

We compare our method with the state-of-the-art learning and non-learning methods.

**Learning-based methods.** Recently, Miki et al. [29] propose to use deep learning for tooth type labeling. They manually crop each tooth from one 2D slice of CBCT images and feed the cropped 2D image to the network for tooth type classification. In contrast, we perform instance segmentation and identification together in 3D domain, then not only the instance labels are found but also the accurate teeth shapes are built.

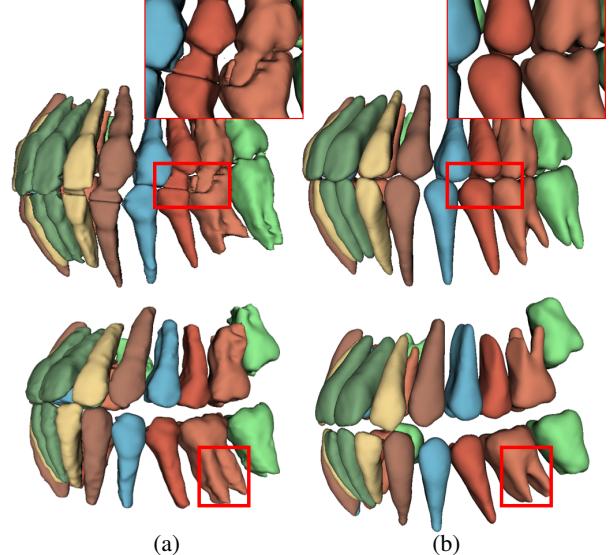


Figure 6. Comparison with the state-of-the-art method. (a) Segmentation results from [11]. (b) Our segmentation results. The first row shows results in a close bite position, while the second row shows the results in an open bite position.

**Non-learning based methods.** Many more non-learning methods target tooth shape segmentation [11, 2, 30], type classification, or both together [19]. Although [30] achieves a high dice score, it requires extra template teeth meshes and tedious user annotations, while [2] has a small average surface distance error, but it is not able to segment molar teeth. Thus, we compare the more recent state-of-the-art method [11] on tooth segmentation task, where they employ the level-set based method with manual initialization. There is a visual comparison in Fig. 6 showing that 1) they cannot find the correct tooth shape boundary in the close bite condition (the first row); 2) they can segment every tooth, but with noisy boundary in the open bite condition (the second row), especially the root part in the red box due to the low contrast value there. Instead, our method is not restricted to the open or close condition, even we do not include any open bite condition teeth data in our training dataset because a tooth CT data captured under an open bite condition is generally invalid in orthodontic diagnosis. To further compare the statistical segmentation accuracy with their method, we first capture two sets of teeth data in an open bite condition and then conduct the comparison using them. Specifically, the DSC scores are 87.12% (theirs) and 92.64% (ours), and the average symmetric surface distance errors are 0.32mm (theirs) and 0.14mm (ours) respectively. We outperform them both visually and statistically.

## 4.3. Discussion

**Failure case.** There are two failure cases as shown in Fig. 8 (a) and (b). The segmentation will fail when there is extreme gray scale value in CT image, such as the metal ar-

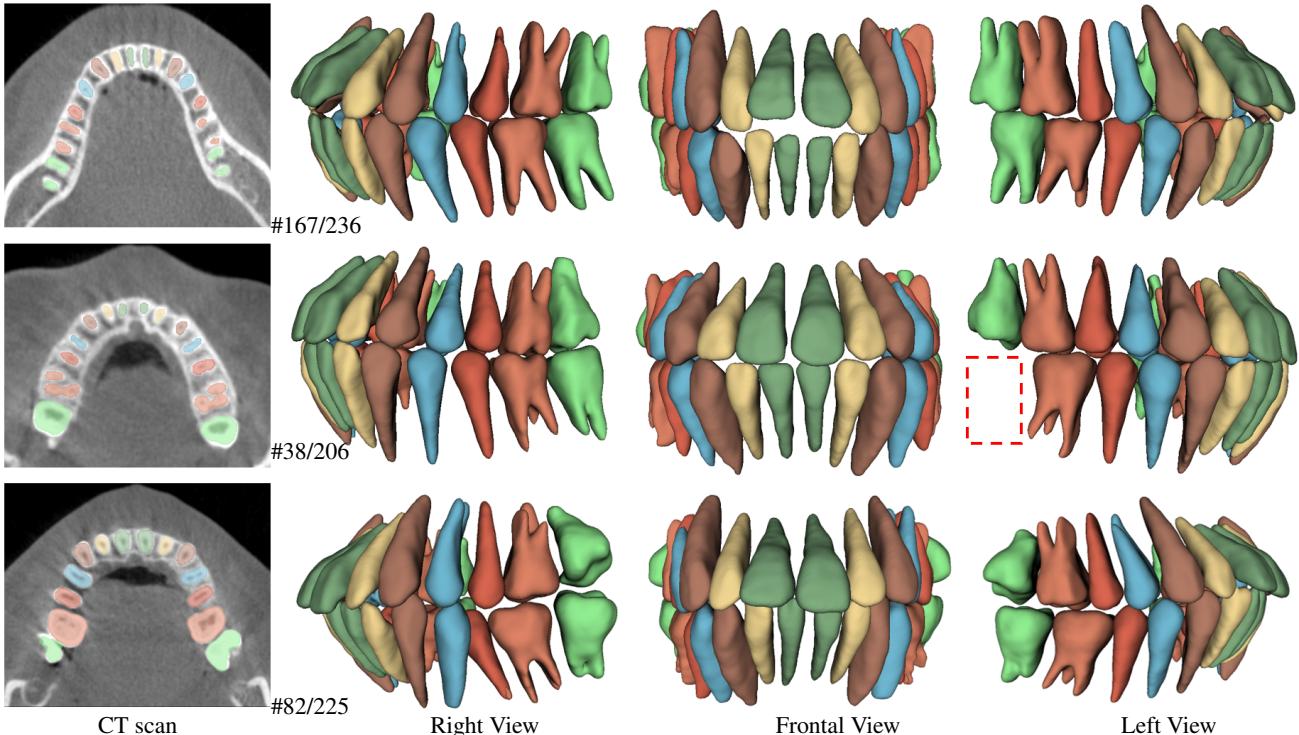


Figure 7. The results gallery of the tooth segmentation and identification. Different CT scans with segmentation results are shown in the first column, and the reconstructed 3D teeth models from three different views are shown in the following three columns. The numbers illustrate the scan indices and different colors illustrate different teeth as defined in Sec. 3.2.3. In addition, the second example contains a removed molar tooth, whose position is marked by the red dashed box.

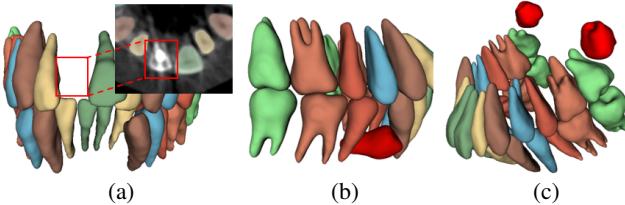


Figure 8. Failure cases and wisdom tooth detection. (a) Extreme gray scale value appears on CT image, e.g., metal artifact of dental implants. (b) Tooth with wrong orientation. (c) Correctly detected wisdom teeth with wrong labels.

tifart of dental implants (Fig. 8 (a)). And the identification will fail if the tooth has the wrong orientation (Fig. 8 (b)), since our network did not see this kind of data during the training process.

**Wisdom tooth.** Wisdom tooth is a special case for human since only a few people have this kind of tooth. Hence, we remove these teeth from CBCT images when preparing the training data. But when we feed this tooth to our network, it detects and segments it successfully as shown in Fig. 8 (c). We never add extra label for this tooth, therefore the tooth label is wrong, visualized with red color.

**Incomplete teeth.** Our testing dataset includes data with incomplete teeth. One example result is shown in the second row of Fig. 7, where one tooth has been removed from

the jaw. We could successfully segment all existing teeth with correct labels.

## 5. Conclusion

In this paper, we propose the first deep learning solution for accurate tooth instance segmentation and identification from CBCT images. Our method is fully automatic without any user annotation and post-processing step. It produces superior results by exploiting the novel learned edge map, similarity matrix and the spatial relations between different teeth. As illustrated, the proposed method significantly outperforms all other existing methods both qualitatively and quantitatively. Our newly proposed components make the popular RPN-based framework suitable for 3D applications with lower GPU memory and less training time requirements, and it can be generalized to other medical image processing tasks in the future.

**Acknowledgement** We thank the reviewers for the suggestions, Dr. Daniel Lee for collecting the teeth data, Dr. Lei Yang for proofreading, and Dr. Jian Shi for the valuable discussions. This work is supported by Hong Kong INNOVATION AND TECHNOLOGY FUND (ITF) (ITS/411/17FX).

## References

- [1] H. Akhoondali, R. Zoroofi, and G. Shirani. Rapid automatic segmentation and visualization of teeth in ct-scan data. *Journal of Applied Sciences*, 9(11):2031–2044, 2009. 2
- [2] S. Barone, A. Paoli, and A. V. Razionale. Ct segmentation of dental shapes by anatomy-driven reformation imaging and b-spline modelling. *International journal for numerical methods in biomedical engineering*, 32(6):e02747, 2016. 2, 7
- [3] H. Chen, Q. Dou, X. Wang, J. Qin, J. C. Cheng, and P.-A. Heng. 3d fully convolutional networks for intervertebral disc localization and segmentation. In *International Conference on Medical Imaging and Virtual Reality*, pages 375–382. Springer, 2016. 2
- [4] H. Chen, L. Yu, Q. Dou, L. Shi, V. C. Mok, and P. A. Heng. Automatic detection of cerebral microbleeds via deep learning based 3d feature representation. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 764–767. IEEE, 2015. 2
- [5] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In *IEEE CVPR*, volume 1, page 3, 2017. 2
- [6] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-sensitive fully convolutional networks. In *European Conference on Computer Vision*, pages 534–549. Springer, 2016. 2
- [7] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016. 2
- [8] Z. Deng and L. J. Latecki. Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in rgb-depth images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 2, 2017. 2
- [9] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng. 3d deeply supervised network for automated segmentation of volumetric medical images. *Medical image analysis*, 41:40–54, 2017. 2, 3
- [10] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012. 5
- [11] Y. Gan, Z. Xia, J. Xiong, G. Li, and Q. Zhao. Tooth and alveolar bone segmentation from dental computed tomography images. *IEEE journal of biomedical and health informatics*, 22(1):196–204, 2018. 2, 7
- [12] Y. Gan, Z. Xia, J. Xiong, Q. Zhao, Y. Hu, and J. Zhang. Toward accurate tooth segmentation from computed tomography images using a hybrid level set model. *Medical physics*, 42(1):14–27, 2015. 2
- [13] H. Gao and O. Chae. Individual tooth segmentation from ct images using level set method with shape and intensity prior. *Pattern Recognition*, 43(7):2406–2417, 2010. 2
- [14] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran. Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 350–359, 2018. 2, 3
- [15] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2
- [16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2
- [17] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 2, 3
- [18] M. Hosntalab, R. A. Zoroofi, A. A. Tehrani-Fard, and G. Shirani. Segmentation of teeth in ct volumetric dataset by panoramic projection and variational level set. *International Journal of Computer Assisted Radiology and Surgery*, 3(3-4):257–265, 2008. 1, 2
- [19] M. Hosntalab, R. A. Zoroofi, A. A. Tehrani-Fard, and G. Shirani. Classification and numbering of teeth in multi-slice ct images using wavelet-fourier descriptor. *International journal of computer assisted radiology and surgery*, 5(3):237–249, 2010. 7
- [20] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2018. 4
- [21] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick. Learning to segment every thing. *Cornell University arXiv Institution: Ithaca, NY, USA*, 2017. 2
- [22] D. X. Ji, S. H. Ong, and K. W. C. Foong. A level-set based approach for anterior teeth segmentation in cone beam computed tomography images. *Computers in biology and medicine*, 50:116–128, 2014. 1, 2
- [23] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017. 2
- [24] S. Keyhaninejad, R. Zoroofi, S. Setarehdan, and G. Shirani. Automated segmentation of teeth in multi-slice ct images. 2006. 2
- [25] L. Lechuga and G. A. Weidlich. Cone beam ct vs. fan beam ct: a comparison of image quality and dose delivered between two differing ct imaging modalities. *Cureus*, 8(9), 2016. 1
- [26] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015. 3
- [27] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. *arXiv preprint arXiv:1611.07709*, 2016. 2
- [28] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017. 2
- [29] Y. Miki, C. Muramatsu, T. Hayashi, X. Zhou, T. Hara, A. Katsumata, and H. Fujita. Classification of teeth in cone-beam ct using deep convolutional neural network. *Computers in biology and medicine*, 80:24–29, 2017. 7

- [30] Y. Pei, X. Ai, H. Zha, T. Xu, and G. Ma. 3d exemplar-based random walks for tooth segmentation from cone-beam computed tomography images. *Medical physics*, 43(9):5040–5050, 2016. [2](#), [7](#)
- [31] P. O. Pinheiro, R. Collobert, and P. Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015. [2](#)
- [32] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. In *European Conference on Computer Vision*, pages 75–91. Springer, 2016. [2](#)
- [33] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. [2](#)
- [34] Z. Ren and E. B. Sudderth. Three-dimensional object detection and layout prediction using clouds of oriented gradients. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1525–1533, 2016. [2](#)
- [35] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016. [2](#)
- [36] S. Song and J. Xiao. Sliding shapes for 3d object detection in depth images. In *European conference on computer vision*, pages 634–651. Springer, 2014. [2](#)
- [37] S. Song and J. Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 808–816, 2016. [2](#)
- [38] K. Yan, X. Wang, L. Lu, and R. M. Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3):036501, 2018. [2](#)
- [39] H.-T. Yau, T.-J. Yang, and Y.-C. Chen. Tooth model reconstruction based upon data fusion for orthodontic treatment simulation. *Computers in biology and medicine*, 48:8–16, 2014. [2](#)
- [40] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. *arXiv preprint arXiv:1709.04518*, 2017. [2](#)
- [41] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang. Md-net: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017. [2](#)
- [42] Z. Zhang, L. Yang, and Y. Zheng. Translating and segmenting multimodal medical volumes with cycle-and shapeconsistency generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9242–9251, 2018. [2](#)
- [43] X. Zhou, Y. Gan, J. Xiong, D. Zhang, Q. Zhao, and Z. Xia. A method for tooth model reconstruction based on integration of multimodal images. *Journal of Healthcare Engineering*, 2018, 2018. [2](#)