Knowledge-Based Systems 206 (2020) 106338

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

TSASNet: Tooth segmentation on dental panoramic X-ray images by Two-Stage Attention Segmentation Network

Yue Zhao [a,b,*,1], Pengcheng Li [a,b,1], Chenqiang Gao [a,b], Yang Liu [c], Qiaoyi Chen [a,b], Feng Yang [a,b], Deyu Meng [d,e]

[a] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[b] Chongqing Key Laboratory of Signal and Information Processing, Chongqing 400065, China
[c] Stomatological Hospital of Chongqing Medical University, Chongqing 401147, China
[d] Macau Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macau
[e] School of Mathematics and Statistics and Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 12 May 2020
Received in revised form 6 July 2020
Accepted 28 July 2020
Available online 5 August 2020

Keywords:
Panoramic X-ray image
Attention model
Tooth segmentation

ABSTRACT

Tooth segmentation acts as a crucial and fundamental role in dentistry for doctors to make diagnosis and treatment plans. In this paper, we propose a Two-Stage Attention Segmentation Network (TSASNet) on dental panoramic X-ray images to address the issues suffered in the tooth boundary and tooth root segmentation task which are caused by the low contrast and uneven intensity distribution. We firstly adopt an attention model which is embedded with global and local attention modules to roughly localize the tooth region in the first stage. Without any interactive operator, the attention model so constructed can automatically aggregate pixel-wise contextual information and identify coarse tooth boundaries. To better obtain final boundary information, we use a fully convolutional network as the second stage to further segment the real tooth area from the attention maps obtained from the first stage. The effectiveness of TSASNet is substantiated on the benchmark dataset containing 1,500 dental panoramic X-ray images, our proposed method achieves 96.94% of accuracy, 92.72% of dice and 93.77% of recall, significantly superior to the current state-of-the-art methods.

1. Introduction

Accurate tooth segmentation is significant in the oral field for guiding clinical diagnosis and developing appropriate surgical plans. In orthodontics, dentists urgently need information on the patient's tooth movement and tooth root absorption during treatment, which is used to evaluate the state of the teeth, correct the bad relationship between the teeth and shorten the orthodontic treatment cycle. The prerequisites for this process are accurately segmenting teeth from the dental panoramic X-ray images. In addition, accurate tooth segmentation can also be used for age identification, forensic identification and to find the hidden dental structures, benign or malignant masses.

Nevertheless, on the one hand, tooth segmentation mostly relies on manual or semi-automatic interactive segmentation by dentists, which is tedious and relies heavily on the dentists' prior knowledge [1–3]. For a large number of images taken with relatively low image qualities, it is expensive and laborious, and even infeasible for some tooth regions to let professional dentists correctly recognize. On the other hand, only small benchmark datasets are available with dental radiography up to now (39 cases in [4]), to address this problem, Silva et al. [5] conducted a dataset which consists of 1,500 dental X-ray panoramic images.

However, tooth segmentation is still a challenging task due to the following aspects: (1) The dental panoramic X-ray images are generally noisy and have low contrast due to the radiation limitation, which inclines to make the image with poor or even lack of boundaries. Lin et al. [6,7] first enhanced the tooth images to separate teeth and gums, and then edge extraction was used to segment the tooth region. Chandran et al. [8] employed CLAHE to enhance the dental images, and then the segmentation results were obtained by the Otsus threshold method. (2) The number of tooth roots may be different, and the contrast of the tooth roots are poor, so it is difficult to segment the topological shapes of several roots of the same molars continuously. Researchers employed level set methods [9,10] on the single slice of 3D Cone beam CT to obtain the tooth root segmentation. There also are some researchers trying to employ horizontal and vertical integral projection to solve this problem [11,12], but the segmentation performance is not satisfied. (3) Different patients

* Corresponding author.
E-mail address: zhaoyue@cqupt.edu.cn (Y. Zhao).
[1] Both authors contributed equally.

https://doi.org/10.1016/j.knosys.2020.106338
0950-7051/

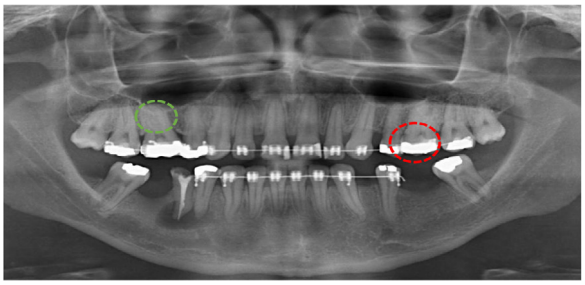


Fig. 1. An example for challenges of the tooth segmentation on a panoramic X-ray image. The green circle denotes the almost invisible boundaries and topological structures. The red circle indicates the dental instruments, and one can see that the teeth and jaw bone appear with low contrast features.

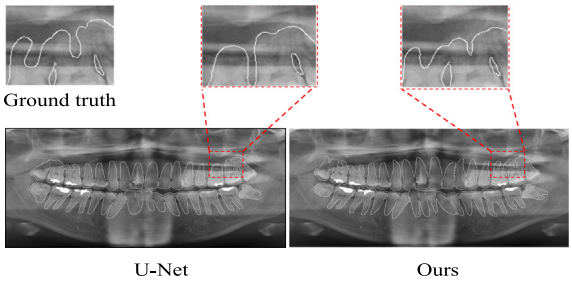


Fig. 2. An example for qualitative comparison with the U-Net. The white lines indicate the segmentation contours. Top row: the close-up view on the segmentation contours inside the red rectangle. Bottom row: Teeth segmentation results from the U-net model (left) and the proposed method (right).

have various teeth statuses and some dental instruments such as dental implants, root canal and metal rack are obstacles of accurate tooth segmentation, as shown as in Fig. 1.

Inspired by the great success of the attention models on various computer vision tasks [13–17] and the two-stage strategy on the object detection [18], in this paper we propose a two-stage segmentation strategy to handle the great challenges suffered in various tooth segmentation scenarios on low-contrast dental panoramic X-ray images. We first adopt global and local attention modules to roughly localize the dental regions in the first stage and then use a fully convolutional network to further search for the exact dental region in the second stage.

In Fig. 2, we illustrate an example for qualitative comparison with the U-net model [19] on teeth segmentation. In this experiment, the main difficulty lies in the detection of weak-defined boundaries. One can point out that the segmentation contours from the U-Net model fail to seek some weakly-defined boundaries, as depicted in the middle of the top row. In contrast, the segmentation result obtained from the proposed model can correctly detect these boundaries, as shown in the right picture of the top row.

In this paper, the main contributions can be summarized as follows:

- Firstly, we propose a new two-stage attention segmentation network for tooth detection and segmentation. To our best knowledge, the proposed model is the first one which exploits a two-stage strategy for tooth localization and segmentation in dental panoramic X-ray images. Our method is capable of automatically encoding richer information and extract more discriminative representations via the hierarchical architecture.
- Next, the proposed attention model focuses on automatically catching the real tooth region, thus independent of user-provided intervention. Furthermore, the proposed method can effectively alleviate the inhomogeneous intensity distribution problem which may yield strong unexpected effects in the tooth segmentation task. Moreover, the tooth structures and the tooth root topologies can be obtained simultaneously by the proposed model.
- Finally, the proposed model achieves 92.72% and 96.94% in terms of dice score and accuracy measure respectively on the benchmark dataset of dental panoramic X-ray images, which indeed significantly outperforms state-of-the-art methods.

The remaining of this paper is organized as follows. In Section 2, the related tooth segmentation methods in the literature are reviewed. In Section 3, we present the proposed method, i.e. the TSASNet and the corresponding implementation details. The experimental results and the discussion are reported in Section 4. The conclusion is drawn in Section 5.

2. Related work

In recent years, the medical image segmentation tasks, including but not limited to the tooth segmentation task on X-ray images, have given rise to great attention to solving various challenging issues. These methods can be roughly divided into two research lines: the traditional methods where prior knowledge and image features are considered and the deep learning-based methods which are driven by data.

Traditional Methods. Traditional methods very often rely on the prior knowledge and image features such as image gradients and region-based similarity measure [20–22]. The locally adaptive threshold method was used to segment teeth in some panoramic X-ray images [22], for which the local image statistical features can be efficiently and effectively exploited. A template matching method with the Otsus threshold and Mahalanobis distance technique was later proposed in [23]. However, it is difficult for this model to identify the tooth roots by using a thresholding strategy. Region growing methods [24,25] were used to segment panoramic and bitewing X-ray images. These methods are known to be time-consuming and also be sensitive to noise. Histogram-based methods were used to choose a gray level threshold to identify enamel caries and proximal caries [26]. Active contour models [27,28] based on the level set formulation [29–32] were further constructed by using a variation of the boundary models to perform tooth segmentation. These methods can handle sharp corners and cusps, as well as topological changes, but suffers from demanding requirement on initialization and computation complexity. Also, such models are often tracked into the unexpected local minima. The mathematical morphology method was also exploited to obtain segmentation in dental images [33]. This method with high accuracy rate and low time complexity. However, this method focuses only on dealing with bitewing and periapical dental images. Lira et al. [34,35] combined the active contour model and the shape analysis for tooth segmentation. The shape models were also used in [35] for tooth recognition. The morphological transform, subtracting images and the active contour evolution scheme were considered in [36] for tooth segmentation. However, the performance of this model heavily depends on initialization. In summary, these traditional methods heavily rely on the matching between the prior and the image features and are usually only suitable for specific scenes. Moreover, it is not a trivial matter to generalize them to similar scenes, since these methods contain expertise knowledge.

Deep Learning-based Methods. The deep learning-based methods have attracted extensive attention in recent years. Recently, it is confirmed that encoding global features and local features are powerful to feature representation and visual feature learning [37,38], and some works devoted to use self-spaced learning

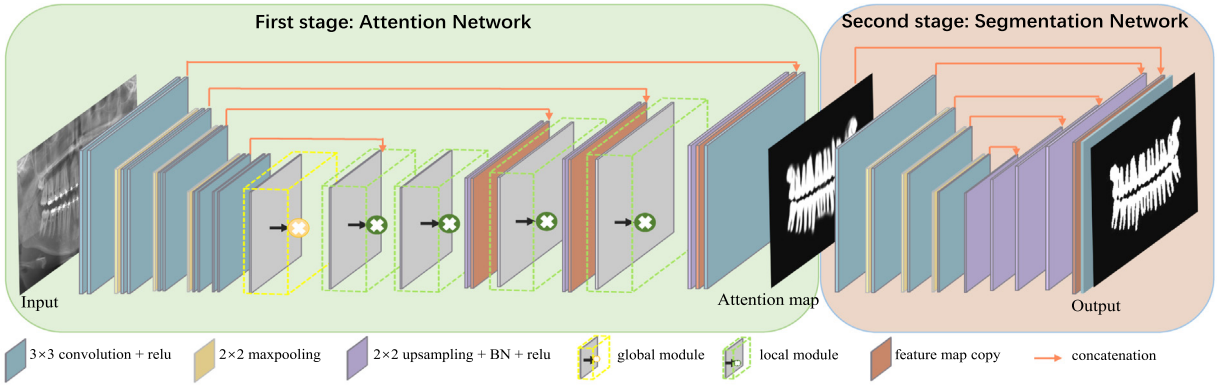


Fig. 3. The proposed TSASNet pipeline involving two stages. The first stage adopts a pixel-wise contextual attention network to get attention map, while the second one segments the attention map to achieve the final segmentation results.

and background subtraction to process images [39,40]. But there still exist only few approaches devoted to addressing the tooth segmentation problem on dental panoramic X-ray images [41–44]. Specifically, Yang et al. [41] proposed a streamlined dental image analysis approach on a small labeled dataset which is comprised of only 196 cases. However, this approach only supports CNNs for simple tooth disease diagnosis and classification. Chen et al. [42] used a Mask R-CNN architecture [45] to segment the tooth area by modifying the label to instance-level to make predictions. This method also proved that the instance segmentation methods can be used in the tooth segmentation task. Wirtz et al. [43] used a coupled shape model in conjunction with a neural network. However, wisdom teeth segmentation was neglected by this method, and the matching of the shape model depends heavily on the tooth structure of the input image.

3. Methodology

3.1. Overview

The proposed TSASNet mainly concentrates on automatic localization and segmentation of the whole and real teeth from a dental panoramic X-ray image. The proposed TSASNet pipeline is comprised of two stages, as depicted in Fig. 3. Specifically, we exploit, in the first stage, the attention network which encodes both the global and local attention modules to derive pixel-wise contextual information. As a consequence, the dental regions of interest can be roughly localized from the attention maps. In the second stage, more accurate tooth segmentation results can be obtained after the fully convolutional network. The proposed TSASNet network can be trained in an end-to-end fashion.

3.2. Attention network

To obtain a more accurate attention map of a panoramic X-ray image for subsequent tooth segmentation task, we take both global attention and local attention into consideration to construct our attention network.

The first stage network in Fig. 3 shows our attention network architecture. The global and local attention modules are embedded into a unified encoder–decoder network, where the encoder tends to extract the dental panoramic X-ray image features, and the decoder inclines to firstly recover the resolution and image size, and then output the attention maps by a sigmoid activation function. The attention network is able to capture both high-level global contexts and low-level details, simultaneously.

Fig. 4(a) depicts the details of our designed global and local attention modules. Inspired by the ReNet [46] and the PiCANet [47], in order to get the global attention, the receptive field of each pixel should be the whole feature map. To this end, the global attention module is obtained by four long short-term memory (LSTM) networks of vertical and horizontal alternation [48]. We use a flatten operator on the vertical and horizontal direction of the image, and then use recurrent neural networks to process those flattened sequential data. Specifically, the outputs of encoder features are parallelly converted into four spatial features by using four different directional scannings (left-to-right, right-to-left, top-to-bottom, and bottom-to-top), as shown in Fig. 4(a). We focus on learning the local spatial dependencies of the sequential image regions, by features obtained from four directions, each pixel can remember its contextual information of all four directions, which can then naturally integrate the global context.

We denote the obtained feature maps as $\mathbf{F}$, with channels C, width W, and height H, respectively. $\mathbf{y}^{w,h}$ means the feature vector obtained in each pixel of location (w, h). To obtain the attention weight vector of each pixel in the feature maps which can see the information of four directions, we first adopt a 1×1 convolution operator to transform the feature map channels to D, where $D = W \times H$. For each pixel location, its attention weight vector $\boldsymbol{\alpha}^{w,h}$ can be calculated by a softmax function:

$$\boldsymbol{\alpha}_i^{w,h} = \frac{\exp(\mathbf{y}_i^{w,h})}{\sum_{j=1}^{D} \exp(\mathbf{y}_j^{w,h})}, \tag{1}$$

where $i \in \{1, \ldots, D\}$, and $\mathbf{y}^{w,h}, \boldsymbol{\alpha}^{w,h} \in \mathbb{R}^D$.

Then the attention weight vector in each location (w, h) can be reshaped as a 2D matrix of size $H \times W$. Finally, we can get the global attention maps by calculating the weighted average value based on the reshaped attention matrix and the feature maps $\mathbf{F}$:

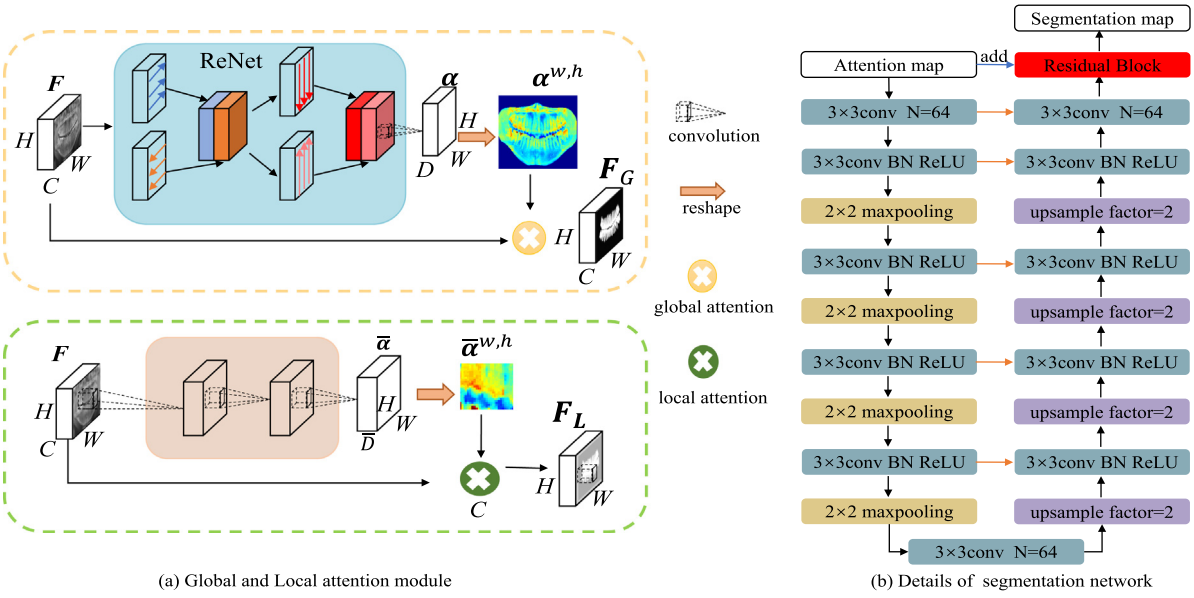
$$\mathbf{F}_G = \sum_{i=1}^{D} \boldsymbol{\alpha}_i^{w,h} \mathbf{F}, \tag{2}$$

where $\mathbf{F}_G \in \mathbb{R}^{W \times H \times C}$.

The local attention module is composed of two continuous convolution layers. Specifically, we transform encoder's feature maps channel to $\bar{D} = \bar{W} \times \bar{H}$, where $\bar{W}$ and $\bar{H}$ denote local width and height, respectively. We make continuous convolution in the local areas $\bar{W} \times \bar{H}$. Furthermore, we also adopt a softmax activation function to get the local attention weight vector $\bar{\boldsymbol{\alpha}} \in \mathbb{R}^{\bar{D}}$. Finally, the local attention feature maps can be obtained by:

$$\mathbf{F}_L = \sum_{i=1}^{\bar{D}} \bar{\boldsymbol{\alpha}}_i^{w,h} \mathbf{F}, \tag{3}$$

where $\mathbf{F}_L \in \mathbb{R}^{W \times H \times C}$.



(a) Global and Local attention module (b) Details of segmentation network

Fig. 4. Details of the attention modules and segmentation network: (a) Global (orange box) and Local (green box) attention module, the variables C, H and W stand for the channels, the image height and weight, respectively. (b) Parameter details of the segmentation network.

3.3. Segmentation network

The segmentation network is constructed based on the fully convolutional networks. Inspired by encoder–decoder architecture like U-Net [19], which includes five convolutional layers and four pooling layers in the encoder, and deconvolution is used in the decoder to recover resolution. Our segmentation network has a contracting path and an expansive path. The difference of the architecture between the U-Net model and the proposed one is that our segmentation network is developed by the residual connection to refine the image segmentation with the attention map. Specifically, we fuse the attention map in the attention network with the last feature map in the segmentation network and generate the final results to further refine both regions and boundaries.

The resolution of the feature maps decreases gradually and the number of channels increases when performing pooling operators. In the expansive path, we use a deconvolution operator to replace the bi-linear interpolation to upsample and recover resolutions, and then shortcut connections from the contracting path to the expansive path are exploited to supplement essential high-resolution features. At the final layer, a 1×1 convolution and a softmax function are used to map each feature vector to the desired classes. In addition, we add batch normalization and the ReLU activation function into the convolution layers as conventional, as shown in Fig. 4(b).

3.4. Hybrid loss function

Inspired by a hybrid loss function strategy [49,50], two loss functions are used to train the proposed TSASNet. We jointly learn the two-stage task together with an end-to-end fashion. Specifically, we jointly supervise attention and segmentation map prediction during training.

We denote by $\hat{y}$ the predicted attention or segmentation probability which is expressed as follows:

$$\hat{y} = g(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{x})}, \tag{4}$$

where g is the sigmoid activation function, the vector of weights $\mathbf{w}$ is optimized through adaptive moment estimation (Adam), and $\mathbf{x}$ represents the input feature maps. The first loss function considered is a binary cross-entropy (BCE) loss applied to ground truth with the prediction of the attention map and the tooth segmentation map respectively, which can be denoted by:

$$\mathcal{L}_{\text{bce}} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})), \tag{5}$$

where $y \in \{0, 1\}$ is the true label.

We take into account the SSIM loss as the second choice for the loss function of TSASNet. Note that the SSIM loss was first considered for image quality assessment as introduced in [51]. In our method, we make use of the emphasis of the structural similarity between the ground truth and the segmentation maps, rather than calculating the differences between two images in a pixel-by-pixel manner. We denote by x the predict attention map or segmentation map and by y the corresponding binary ground truth mask. The SSIM loss can be formulated as follows:

$$\mathcal{L}_{\text{ssim}} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{6}$$

where μ_x, μ_y and σ_x, σ_y are the mean and standard deviations of x and y respectively, σ_{xy} is their co-variance, $C_1 = 0.01^2$ and $C_2 = 0.03^2$ are used to avoid dividing by zero.

Finally, our training loss can be defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{bce}} + \lambda_2 \mathcal{L}_{\text{ssim}}, \tag{7}$$

where λ_1 and λ_2 are positive constants that dominate the related importance for the corresponding loss. In this paper, we empirically set $\lambda_1 = 20$ and $\lambda_2 = 1$ to guarantee gradients returned by these losses are of same order of magnitude.

3.5. Implementation details

The tooth segmentation challenge lies in the low contrast ratio between the bone area and the tooth area in panoramic X-ray images, leading to extremely fuzzy boundaries which may yield great segmentation errors. As shown in Fig. 3, we adopt a two-stage network to joint training. In the first stage, we use the pixel-wise contextual attention network to get the attention maps which contains all tooth area, and after that, we directly set the attention maps as input in the second stage. The ground truth

Table 1
The detailed information about the benchmark dataset.

Category	Missing teeth	Restoration	Dental appliance	Dental implant	Average teeth	Mean intensity	Images
1		✓	✓		32	108.30	73
2		✓			32	108.29	220
3			✓		32	107.25	45
4					32	107.31	140
5	✓			✓	18	109.36	120
6					37	100.43	170
7	✓	✓	✓		27	108.50	115
8	✓	✓			29	106.72	457
9	✓		✓		28	107.33	45
10	✓				28	105.94	115

is used again to guide the segmentation network to conduct fine segmentation on the attention maps.

The training and testing phases of our model are implemented on 4 NVIDIA GeForce GTX 1080Ti GPUs with PyTorch with an end-to-end fashion. We set the batch size and epoch to 8 and 50 to get the gradient update respectively. The initial learning rate is set to 0.001, and we decay the learning rate by a factor o 0.1 at per 600 steps. Finally, we choose Adam optimizer to minimize the hybrid loss function.

4. Experimental results

4.1. Dataset and evaluation criteria

Dataset. We use the benchmark public dataset [5] to evaluate our proposed TSASNet. This dataset includes a total of 1,500 panoramic X-ray images of teeth which has 10 categories, and the detailed information is listed in Table 1. We randomly select 1,200 images from 10 categories as the training set, and select 150 images as the validation and testing set from the remaining data respectively.

Evaluation Metrics. In medical image segmentation community, the dice coefficient [52–54] is extensively adopted to evaluate the performance of an image segmentation model, and can be defined as follows:

$$Dice = \frac{2|P \bigcap G|}{|P| + |G|}, \tag{8}$$

where P and G are the predicted segmented region and the ground truth region, respectively. The maximum of the dice coefficient is 1, which indicates the highest segmentation accuracy, while the minimum is 0, which corresponds to the lowest segmentation accuracy.

Actually, the image segmentation can also be considered as a pixel classification task, so we also use the commonly used metrics of classification tasks to evaluate the segmentation performance of the proposed TSASNet as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{9}$$

$$Specificity = \frac{TN}{FP + TN}, \tag{10}$$

$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

$$Recall = \frac{TP}{TP + FN}, \tag{12}$$

where TP means the pixel number of true prediction of tooth area, FP means the pixel number of false prediction of tooth area, TN is the pixel number of true prediction of background area, and FN denotes the pixel number of false prediction of background area.

Table 2
Performance comparison with state-of-the-art methods.

	Acc.	Spec.	Pre.	Recall	Dice	Params(M)
U-Net [19]	96.04	97.68	89.89	90.18	89.33	31.04
BiseNet [55]	95.05	95.98	85.53	92.48	87.80	12.2
DenseASPP [56]	95.50	97.76	90.09	86.88	88.13	46.16
SegNet [57]	96.38	98.32	92.26	89.05	90.15	29.44
BASNet [58]	96.77	**98.64**	94.56	90.11	92.12	87.06
Ours	**96.94**	97.81	**94.97**	**93.77**	**92.72**	**78.27**

4.2. Comparison with state-of-the-art models

We compare our proposed TSASNet with recent state-of-the-art learning and non-learning methods.

Learning-based Methods. Since there is rarely specific deep learning based method for tooth segmentation, we compared with the U-Net [19], BiseNet [55], DenseASPP [56], SegNet [57], and a two-stage network BASNet [58], which are all can be used in dental panoramic X-ray image segmentation task. For a fair comparison, we use the same parameter settings and report the performance on the test dataset. Table 2 shows the quantitative comparison results and the number of parameters of different methods. It can be obviously observed that TSASNet outperforms all state-of-the-art methods on all metrics, except unsubstantially lower in specificity metrics compared with BASNet, where the result further demonstrates that our attention module is more focused on foreground localization, and we reduce more than 10% parameter numbers compared with the two-stage network BASNet. Our segmentation dice performance has 0.6% to around 5% gains, which is a significant improvement in tooth segmentation task.

Fig. 5 also shows the segmentation results derived from the methods mentioned above. One can clearly see that the U-Net, BiseNet, DenseASPP and SegNet are prone to misclassify the jaw bone as teeth, while the proposed method can correctly find the segmentation. The BiseNet and DenseASPP methods have slightly weaker ability to handle the low contrast images, and the segmentation results appear noisy. As shown in Fig. 5, those state-of-the-art methods mentioned above suffer from the issues of either over-segmentation or under-segmentation when dealing with the dental panoramic X-ray images, while our method can accurately preserve the boundaries of teeth and the complete tooth structures. We further evaluate the effect of our attention modules on distinguishing the dental and background regions. This can be interpreted by the fact that our global and local attention modules aggregate pixel-wise contextual information, leading to more accurate segmentation results and can capture subtle contrast ratio differences between real tooth areas and other mouth structures.

Non-learning based methods. We also exploit the differences between the proposed approach and the traditional non-learning methods [59–61], where the experiment results are shown in Table 3 and in Fig. 6.

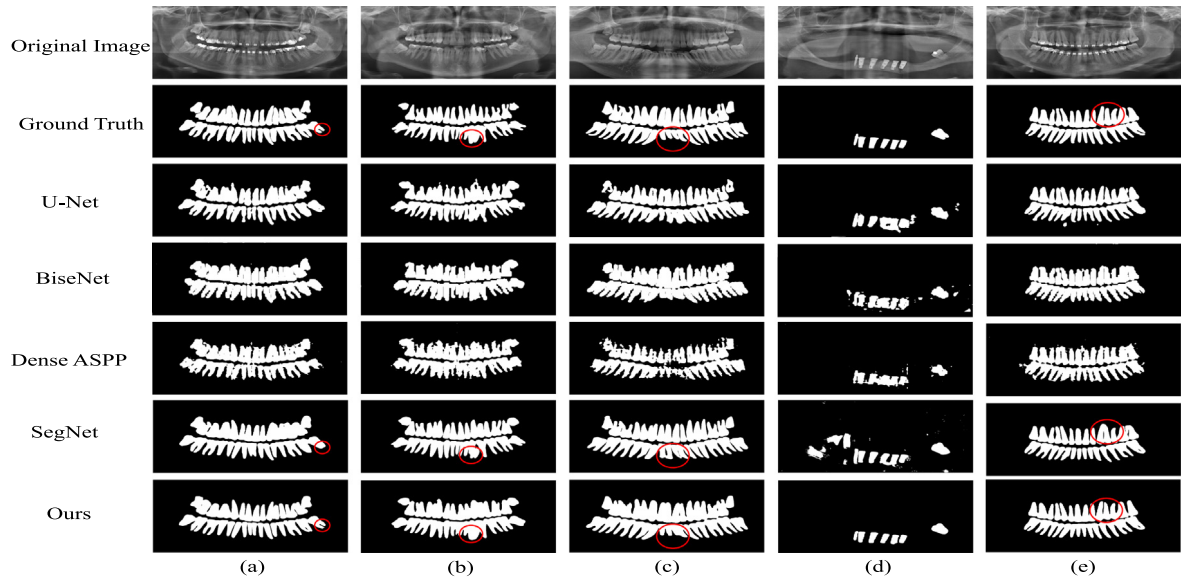


Fig. 5. Qualitative comparison results with the state-of-the-art image segmentation methods. The original images and the corresponding ground truth are shown in the first two rows. Rows 3 to 7 illustrate the segmentation results respectively derived from the U-Net, BiseNet, DenseASPP, SegNet and the proposed method. Red circle indicate the fine distinction between the ground truth, the SegNet and the proposed method.

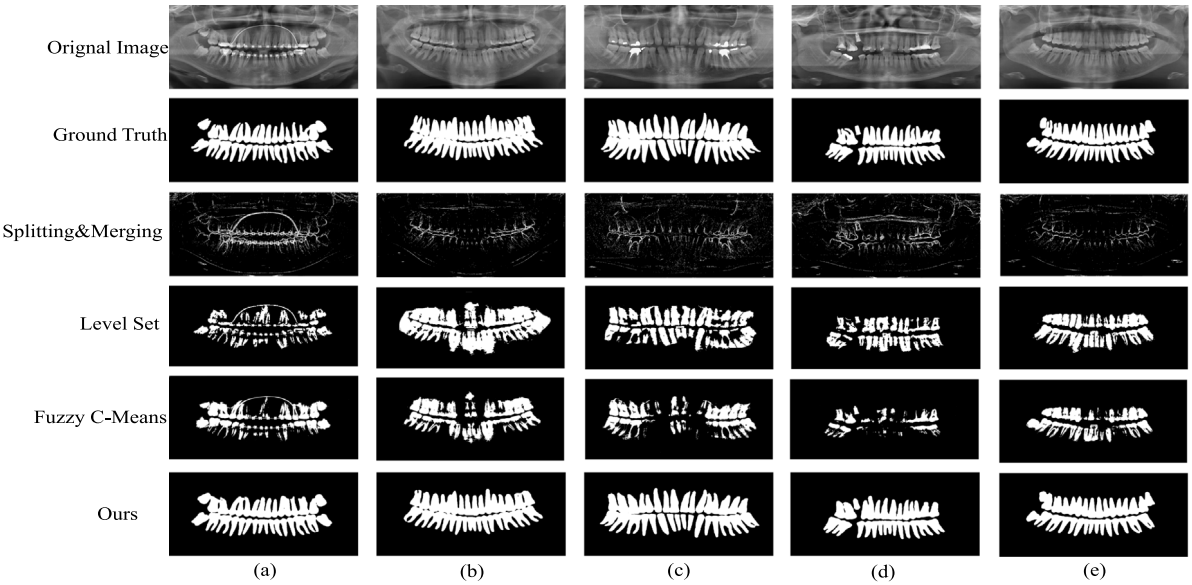


Fig. 6. Qualitative comparison results with the non-learning based methods. The original images and the corresponding ground truth are shown in the first two rows. Rows 3 to 6 illustrate the segmentation results respectively derived from Splitting and merging, Level set, Fuzzy C-Means and the proposed method.

Table 3
The average segmentation performance (%) of non-learning based methods.

Method	Accuracy	Specificity	Precision	Recall	Dice
FCM [59]	82.15	90.78	61.42	45.03	49.47
Level set [60]	75.50	78.45	48.37	68.49	59.17
Splitting&Merging [61]	81.33	**99.15**	81.24	8.14	14.48
Ours	**96.94**	97.81	**94.97**	**93.77**	**92.72**

The fuzzy C-Means (FCM) method [59] achieved accuracy score of 82.15% and specificity 90.78%, it indicates that this method is more likely to classify background pixels. Fig. 6 shows the segmentation results of the FCM method. From the results in the last two rows, one can observe that some tooth areas are not detected when applying the FCM method. In contrast, our method can find the tooth areas correctly.

The level set-based method [60] has the best performance in the sense of dice metric compared to the other traditional methods. It is known that reasonable initialization is helpful for evolving curves to detect sharp corners. From Fig. 6, we can find that this level set-based method has the poor ability in extracting the tooth roots from alveolar bones. Although some tooth areas are correctly identified, the alveolar bones are not properly classified as background.

At the same time, we also find that there are some holes within the segmented tooth regions derived from the level set-based method [60], which is unfavorable for clinical application. Although the segmentation performance can be improved through optimized parameters, it still belongs to manual tuning. In contrast, our method is fully automatic without manual intervention.

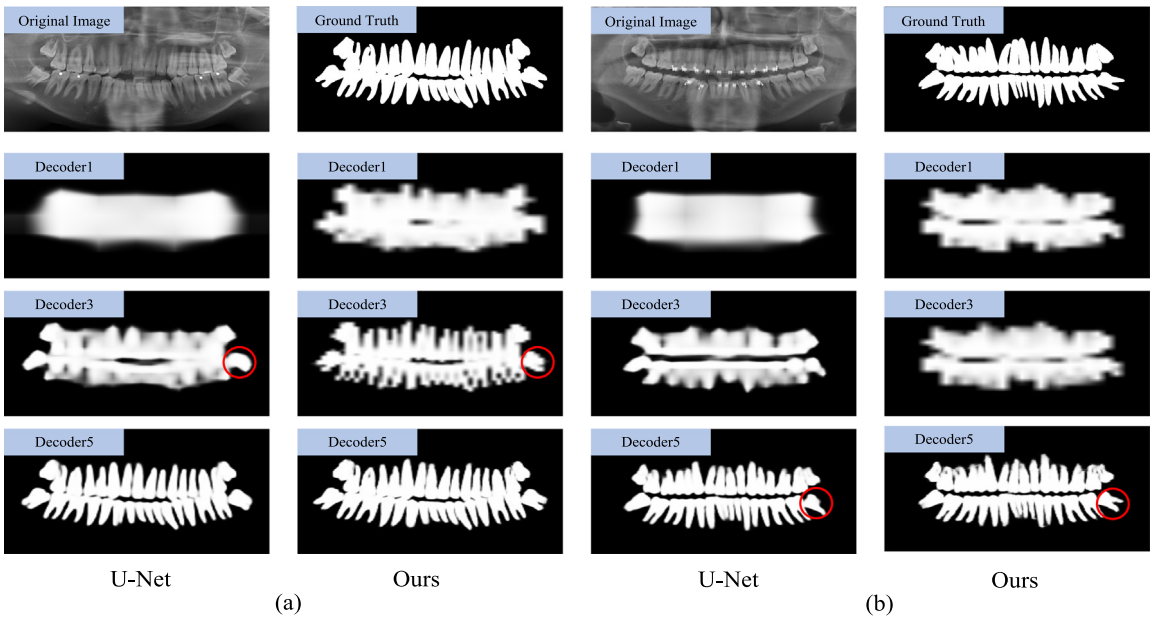


Fig. 7. Comparison of attention map visualized results: we compare the global attention module in decoder1 and compare local modules in decoder3 and decoder5. Our global module can localize the roughly tooth region in decoder1, and red circle in decoder3 and decoder5 indicate the superiority of the boundary maintaining with the local attention module.

From Table 3, we can observe that the specificity of the splitting-and-merging method [61] is high up to 99.15%, while the recall rate is only 8.14%. The reason can be that this method focuses on the segmentation of background pixels and slightly insufficient in the ability to correctly classify foreground pixels into tooth areas.

In general, our proposed method is superior to traditional prior knowledge-inspired methods in practical applications. This is mainly because our method can perform more automatic segmentation of dental panoramic X-ray images, which can effectively remove the interference of background information while ensuring the complete segmentation of the tooth area, so our proposed method is of great help to the clinical application of dentists.

4.3. Ablation study

We conduct ablation studies to validate the effects of our TSASNet components. Additional experiments are designed with our proposed novel components. All different settings are trained on the same dataset and we report the image-level dice coefficient score and the pixel-level metrics on our test dataset for comparison.

Effectiveness of the attention modules. To validate the effectiveness of our attention modules, we show the quantitative comparison results of our attention modules against baseline networks in Table 4, and we take U-Net [19] based network as our baseline network. '0G0L' means there is no global attention module or local attention module in two-stage U-shape architecture. '0G5L' denotes that five local attention modules are employed in our first stage attention network. '1G4L' and '2G3L' indicate that there are one global and four local attention modules and two global and three local attention modules in the attention network respectively. Follow the above settings, the comparison results show that the number of global and local attention modules can directly affect the segmentation result, and when we use different numbers of global and local attention modules to combine global and multi-scale local context features, we can effectively improve the model performance. From Table 4, we can see the setting of '1G4L' bring the most gains in precision, recall and dice, this is may due to the global attention module would focus on more background information, with the increasing number of the global attention modules, more background pixels would be recognized as foreground pixels. It is worth noting that we do not explore all the settings of the global attention module on the first stage decoding path, since there is a consensus that global context information performs better as the network deepens in the encoding path while the receptive field is diminished in the decoding process. We also demonstrate the effectiveness of the proposed attention network by showing visualized results, as Fig. 7 shows, for better visualization, we interpolate all feature maps to the same spatial resolution as original images. Fig. 7 shows two examples of dental panoramic X-ray images and their corresponding ground truth.

Table 4
Ablation study on the different number of global and local modules:'G' and 'L' denote global and local module respectively.

	Accuracy	Specificity	Precision	Recall	Dice
U-Net	96.04	97.68	89.98	90.18	89.33
0G0L	96.80	97.45	93.06	90.27	91.60
0G5L	96.27	97.49	90.80	91.56	91.01
2G3L	96.82	**98.85**	91.91	89.27	92.10
Ours	**96.83**	97.79	**93.34**	**93.10**	**92.36**

We denote the feature maps in the decoder path as F_D, G and L means global attention module and local attention module respectively. The second row in Fig. 7 shows the F_{D1} of the baseline U-Net network (left) and the F_{D1G} of our model (right). We can obviously see that the global attention module can not only locate the tooth region accurately but also have a better ability to distinguish different teeth. The last two rows in Fig. 7 shows the performance of the local attention modules, and we can see that the local attention module helps to refine the tooth region more sharpen.

Effectiveness of the loss functions. To evaluate the effectiveness of our proposed hybrid loss function, a set of experiments on different loss settings are conducted based on '1G4L' attention architecture. Table 5 shows the comparison results. We can observe that BCE loss and SSIM loss have similar performance in

Table 5
Ablation study on different loss function settings.

	Accuracy	Specificity	Precision	Recall	Dice
L_{bce}	96.81	97.79	91.91	93.10	92.36
L_{ssim}	96.80	**98.30**	93.42	91.27	92.21
$L_{bce} + L_{ssim}$	**96.94**	97.81	**94.97**	**93.77**	**92.72**

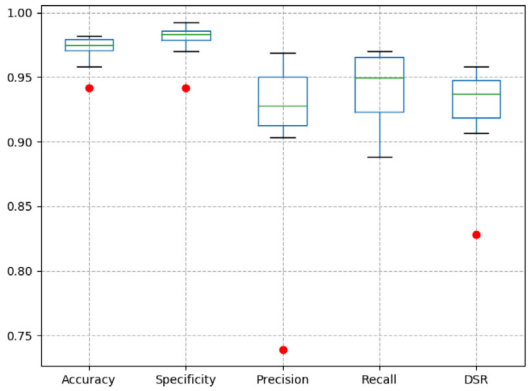


Fig. 8. Metrics comparison between 10 categories of the teeth situation. We calculated the metrics for each image category separately on the testing set. The green line denotes the median value of each metric, the red point denotes outlier.

segmentation accuracy and dice metrics, while SSIM loss has higher performance in Specificity, it is due to the number of background pixels are slightly more than the tooth pixels, and this metric is putting emphasis on the background pixels. When we mix two loss functions to form a hybrid loss function, the dice metric have 0.5% gains dramatically. The increase in overall performance could be attributed to the use of multi-aspect metrics, which provides a clear clue for the following researches of such methods.

4.4. Discussion

Label quality. Fig. 8 shows the box plot of the metrics between different categories of dental panoramic X-ray images. As can be seen from Fig. 8, the result shows that there is an outlier in all metrics expect Recall. It is noted that the outlier in the dental benchmark dataset is harmful to the performance of the proposed method. Actually, the categories of 7 and 8 in the benchmark dataset are labeled more coarse than other kinds of categories. Fig. 9 shows the example of different ground truth categories, from up to bottom, are category 7, category 8 and category 2 respectively. We can clearly observe that although the label in category 7 is sharper than other categories, our proposed method can handle the defect of ground truth, and our segmentation results could more reasonable than the poorly labeled data.

Difficult cases. There still have some difficult cases in our method and in Fig. 10 we give two examples for such cases. The first one is the segmentation of the supernumerary tooth situation. It is challenging because there is no distinct boundary between those teeth, and some teeth even overlap directly. Another one is the segmentation of teeth with fillers. In this case, the tooth roots are almost homogenized by the upper jaw, since our network rarely see this kind of data during the training process.

5. Conclusion

In this paper, we propose a two-stage attention segmentation network for the tooth segmentation task. Firstly, we adopt an attention model that embed global and local attention modules to coarsely localize the dental regions, and then we use a fully convolutional network to further segment the real dental regions. By doing so, this model is capable of automatically aggregating pixel-wise contextual information and catching those blurry dental regions without any interactive operations. Experimental results demonstrate that TSASNet can obtain superior segmentation performance on dental panoramic X-ray images over other state-of-the-art methods along this research line, and has highly competitive performance compared with the current state-of-the-art medical image segmentation methods.

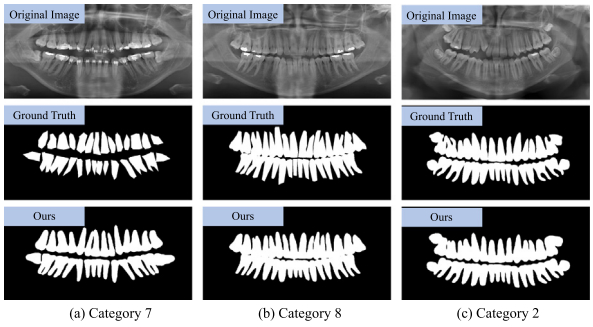


Fig. 9. Examples of different categories' visualized ground truth and segmentation results, where the ground truth are more sharpen in category 7 than the others.

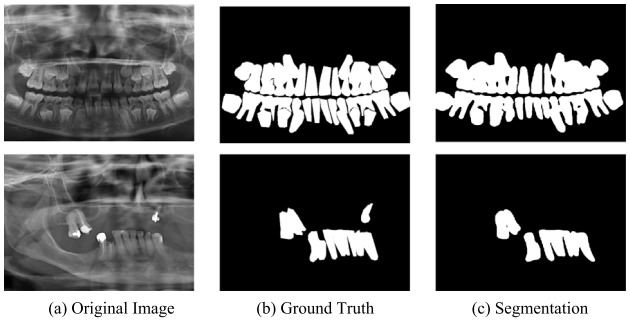


Fig. 10. Segmentation results for two challenging examples.

CRediT authorship contribution statement

Yue Zhao: Conceptualization, Methodology, Investigation, Resources, Writing - review & editing. **Pengcheng Li:** Methodology, Software, Data curation, Writing - original draft. **Chenqiang Gao:** Conceptualization, Validation, Supervision, Funding acquisition. **Yang Liu:** Resources, Validation. **Qiaoyi Chen:** Data curation. **Feng Yang:** Data curation, Formal analysis. **Deyu Meng:** Conceptualization, Investigation, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61571071, 61906025), Chongqing Research Program of Basic Research and Frontier Technology, China (No.

cstc2018jcyjAX0227), the Science and Technology Research Program of Chongqing Municipal Education Commission, China under Grant KJQN201900607, the Education Informatization Project of Chongqing University of Posts and Telecommunications, China under Grant xxhyf2019-01.

References

[1] Ching-Wei Wang, Cheng-Ta Huang, Jia-Hong Lee, Chung-Hsing Li, Sheng-Wei Chang, Ming-Jhih Siao, Tat-Ming Lai, Bulat Ibragimov, Tomaž Vrtovec, Olaf Ronneberger, et al., A benchmark for comparison of dental radiography analysis algorithms, Med. Image Anal. 31 (2016) 63–76.
[2] Ho Chul Kang, Chankyu Choi, Juneseuk Shin, Jeongjin Lee, Yeong-Gil Shin, Fast and accurate semiautomatic segmentation of individual teeth from dental CT images, Comput. Math. Methods Med. 2015 (2015).
[3] Zhongyi Li, Hao Wang, Interactive tooth separation from dental model using segmentation field, PLoS One 11 (8) (2016) e0161159.
[4] Olaf Ronneberger, Philipp Fischer, Thomas Brox, Dental X-ray Image Segmentation Using a U-Shaped Deep Convolutional Network, ISBI, 2015.
[5] Gil Silva, Luciano Oliveira, Matheus Pithon, Automatic segmenting teeth in x-ray images: Trends, a novel data set, benchmarking and future perspectives, Expert Syst. Appl. 107 (2018) 15–31.
[6] P.L. Lin, Y.H. Lai, P.W. Huang, An effective classification and numbering system for dental bitewing radiographs using teeth region and contour information, Pattern Recognit. 43 (4) (2010) 1380–1392.
[7] Phen-Lan Lin, Yan-Hao Lai, Po-Whei Huang, Dental biometrics: Human identification based on teeth and dental works in bitewing radiographs, Pattern Recognit. 45 (3) (2012) 934–946.
[8] Veena Chandran, Ginsi S. Nizar, Philomina Simon, Segmentation of dental radiograph images, in: Proceedings of the Third International Conference on Advanced Informatics for Computing Research, 2019, pp. 1–5.
[9] Seunghwan Shin, Yoonho Kim, A study on automatic tooth root segmentation for dental CT images, J. Soc. e-Bus. Stud. 19 (4) (2014).
[10] Yangzhou Gan, Zeyang Xia, Jing Xiong, Qunfei Zhao, Ying Hu, Jianwei Zhang, Toward accurate tooth segmentation from computed tomography images using a hybrid level set model, Med. Phys. 42 (1) (2015) 14–27.
[11] Omaima Nomir, Mohamed Abdel-Mottaleb, Fusion of matching algorithms for human identification using dental X-ray radiographs, IEEE Trans. Inf. Forensics Secur. 3 (2) (2008) 223–233.
[12] Robert Wanat, Dariusz Frejlichowski, A problem of automatic segmentation of digital dental panoramic x-ray images for forensic human identification, Proc. CESCG (2011) 165–172.
[13] Shihui Zhang, He Li, Weihang Kong, Xiaowei Zhang, Weidong Ren, An attention-guided and prior-embedded approach with multi-task learning for shadow detection, Knowl.-Based Syst. (2020) 105540.
[14] Degui Xiao, Xuefeng Yang, Jianfang Li, Merabtene Islam, Attention deep neural network for lane marking detection, Knowl.-Based Syst. (2020) 105584.
[15] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, Tat-Seng Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5659–5667.
[16] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, Shuicheng Yan, Diversified visual attention networks for fine-grained object classification, IEEE Trans. Multimed. 19 (6) (2017) 1245–1256.
[17] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[18] Ross Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[19] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[20] Chenqiang Gao, Deyu Meng, Yi Yang, Yongtao Wang, Xiaofang Zhou, Alexander G Hauptmann, Infrared patch-image model for small target detection in a single image, IEEE Trans. Image Process. 22 (12) (2013) 4996–5009.
[21] Chenqiang Gao, Lan Wang, Yongxing Xiao, Qian Zhao, Deyu Meng, Infrared small-dim target detection based on Markov random field guided noise modeling, Pattern Recognit. 76 (2018) 463–475.
[22] Rarasmaya Indraswari, Agus Zainal Arifin, Dini Adni Navastara, Naser Jawas, Teeth segmentation on dental panoramic radiographs using decimation-free directional filter bank thresholding and multistage adaptive thresholding, in: 2015 International Conference on Information & Communication Technology and Systems, ICTS, IEEE, 2015, pp. 49–54.
[23] Arisa Poonsri, Napapa Aimjirakul, Theekapun Charoenpong, Chamaiporn Sukjamsri, Teeth segmentation from dental x-ray image by template matching, in: 2016 9th Biomedical Engineering International Conference, BMEiCON, IEEE, 2016, pp. 1–4.
[24] Agus Zainal Arifin, Rarasmaya Indraswari, Nanik Suciati, Eha Renwi Astuti, Dini Adni Navastara, Region merging strategy using statistical analysis for interactive image segmentation on dental panoramic radiographs, Int. Rev. Comput. Softw. 12 (1) (2017) 63–74.
[25] Amelia Sahira Rahma, Eva Firdayanti Bisono, Agus Zainal Arifin, Dini Adni Navastara, Rarasmaya Indraswari, Generating automatic marker based on combined directional images from frequency domain for dental panoramic radiograph segmentation, in: 2017 Second International Conference on Informatics and Computing, ICIC, IEEE, 2017, pp. 1–6.
[26] Shubhangi Vinayak Tikhe, Anjali Milind Naik, Sadashiv D. Bhide, T. Saravanan, K.P. Kaliyamurthie, Algorithm to identify enamel caries and interproximal caries using dental digital radiographs, in: 2016 IEEE 6th International Conference on Advanced Computing, IACC, IEEE, 2016, pp. 225–228.
[27] Chunming Li, Rui Huang, Zhaohua Ding, J. Chris Gatenby, Dimitris N. Metaxas, John C. Gore, A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI, IEEE Trans. Image Process. 20 (7) (2011) 2007–2016.
[28] Pulkit Pandey, Anupama Bhan, Malay Kishore Dutta, Carlos M. Travieso, Automatic image processing based dental image analysis using automatic Gaussian fitting energy and level sets, in: 2017 International Conference and Workshop on Bioinspired Intelligence, IWOBI, IEEE, 2017, pp. 1–5.
[29] C. Li, C. Xu, C. Gui, M.D. Fox, Distance regularized level set evolution and its application to image segmentation, IEEE Trans. Image Process. 19 (12) (2010) 3243–3254.
[30] Stanley Osher, James A. Sethian, Fronts propagating with curvature-dependent speed: algorithms based on hamilton-Jacobi formulations, J. Comput. Phys. 79 (1) (1988) 12–49.
[31] Yue Zhao, Shuxu Guo, Min Luo, Yu Liu, Michel Bilello, Chunming Li, An energy minimization method for MS lesion segmentation from T1-w and FLAIR images, Magn. Reson. Imaging 39 (2017) 1–6.
[32] Yue Zhao, Shuxu Guo, Min Luo, Xue Shi, Michel Bilello, Shaoxiang Zhang, Chunming Li, A level set method for multiple sclerosis lesion segmentation, Magn. Reson. Imaging 49 (2018) 94–100.
[33] Eyad Haj Said, Diaa Eldin M. Nassar, Gamal Fahmy, Hany H. Ammar, Teeth segmentation in digitized dental X-ray films using mathematical morphology, IEEE Trans. Inf. Forensics Secur. 1 (2) (2006) 178–189.
[34] P.H.M. Lira, G.A. Giraldi, L.A.P. Neves, An automatic morphometrics data extraction method in dental x-ray image, in: Proceedings of the International Conference on Biodental Enginnering, Porto, Portugal, 2009, 77–82.
[35] Pedro H.M. Lira, Gilson A. Giraldi, Luiz A.P. Neves, Panoramic dental X-ray image segmentation and feature extraction, in: Proceedings of V Workshop of Computing Vision, Sao Paulo, Brazil, 2009.
[36] Azam Amini Harandi, Hossein Pourghassem, Hamid Mahmoodian, Upper and lower jaw segmentation in dental X-ray image using modified active contour, in: 2011 International Conference on Intelligent Computation and Bio-Medical Instrumentation, IEEE, 2011, pp. 124–127.
[37] Aming Wu, Linchao Zhu, Yahong Han, Yi Yang, Connective cognition network for directional visual commonsense reasoning, in: Advances in Neural Information Processing Systems, 2019, pp. 5669–5679.
[38] Linchao Zhu, Yi Yang, ActBERT: Learning global-local video-text representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8746–8755.
[39] Hongwei Yong, Deyu Meng, Wangmeng Zuo, Lei Zhang, Robust online matrix factorization for dynamic background subtraction, IEEE Trans. Pattern Anal. Mach. Intell. 40 (7) (2017) 1726–1740.
[40] Deyu Meng, Qian Zhao, Lu Jiang, A theoretical understanding of self-paced learning, Inform. Sci. 414 (2017) 319–328.
[41] Jie Yang, Yuchen Xie, Lin Liu, Bin Xia, Zhanqiang Cao, Chuanbin Guo, Automated dental image analysis by deep learning on small dataset, in: 2018 IEEE 42nd Annual Computer Software and Applications Conference, COMPSAC, vol. 1, IEEE, 2018, pp. 492–497.
[42] Hu Chen, Kailai Zhang, Peijun Lyu, Hong Li, Ludan Zhang, Ji Wu, Chin-Hui Lee, A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films, Sci. Rep. 9 (1) (2019) 3840.
[43] Andreas Wirtz, Sudesh Ganapati Mirashi, Stefan Wesarg, Automatic teeth segmentation in panoramic X-ray images using a coupled shape model in combination with a neural network, in: MICCAI, 2018.
[44] Shreyansh A. Prajapati, R. Nagaraj, Suman Mitra, Classification of dental diseases using CNN and transfer learning, in: 2017 5th International Symposium on Computational and Business Intelligence, ISCBI, IEEE, 2017, pp. 70–74.
[45] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
[46] Francesco Visin, Kyle Kastner, Kyunghyun Cho, Matteo Matteucci, Aaron Courville, Yoshua Bengio, Renet: A recurrent neural network based alternative to convolutional networks, 2015, arXiv preprint arXiv:1505.00393.

[47] Nian Liu, Junwei Han, Ming-Hsuan Yang, PiCANet: Learning pixel-wise contextual attention for saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3089–3098.
[48] Kai Sheng Tai, Richard Socher, Christopher D. Manning, Improved semantic representations from tree-structured long short-term memory networks, 2015, arXiv preprint arXiv:1503.00075.
[49] Kenta Moriwaki, Ryota Yoshihashi, Rei Kawakami, Shaodi You, Takeshi Naemura, Hybrid loss for learning single-image-based HDR reconstruction, 2018, arXiv preprint arXiv:1812.07134.
[50] Yanyan Wei, Zhao Zhang, Haijun Zhang, Richang Hong, Meng Wang, A coarse-to-fine multi-stream hybrid deraining network for single image deraining, 2019, arXiv preprint arXiv:1908.10521.
[51] Zhou Wang, Eero P. Simoncelli, Alan C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, vol. 2, IEEE, 2003, pp. 1398–1402.
[52] Dinggang Shen, Guorong Wu, Heung-Il Suk, Deep learning in medical image analysis, Ann. Rev. Biomed. Eng. 19 (2017) 221–248.
[53] Ming Yan, Jixiang Guo, Weidong Tian, Zhang Yi, Symmetric convolutional neural network for mandible segmentation, Knowl.-Based Syst. 159 (2018) 63–71.
[54] Li Wang, Dong Nie, Guannan Li, Élodie Puybareau, Jose Dolz, Qian Zhang, Fan Wang, Jing Xia, Zhengwang Wu, Jiawei Chen, et al., Benchmark on automatic 6-month-old infant brain segmentation algorithms: The iseg-2017 challenge, IEEE Trans. Med. Imaging (2019).
[55] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, Nong Sang, Bisenet: Bilateral segmentation network for real-time semantic segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 325–341.
[56] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, Kuiyuan Yang, Denseaspp for semantic segmentation in street scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3684–3692.
[57] Vijay Badrinarayanan, Ankur Handa, Roberto Cipolla, Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling, 2015, arXiv preprint arXiv:1505.07293.
[58] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, Martin Jagersand, Basnet: Boundary-aware salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7479–7489.
[59] Mutasem K. Alsmadi, A hybrid fuzzy c-means and neutrosophic for jaw lesions segmentation, Ain Shams Eng. J. (2016).
[60] Abdolvahab Ehsani Rad, Mohd Shafry Mohd Rahim, Alireza Norouzi, Digital dental x-ray image segmentation and feature extraction, TELKOMNIKA Indonesian J. Electr. Eng. 11 (6) (2013) 3109–3114.
[61] R. Bala Subramanyam, K. Purushotham Prasad, B. Anuradha, Different image segmentation techniques for dental image extraction, Int. J. Eng. Res. Appl. 4 (7) (2014) 173–177.