Computers in Biology and Medicine 152 (2023) 106296

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Teeth U-Net: A segmentation model of dental panoramic X-ray images for context semantics and contrast enhancement☆

Senbao Hou [a], Tao Zhou [a,c,*], Yuncan Liu [a], Pei Dang [a], Huiling Lu [b,**], Hongbin Shi [d]

[a] School of Computer Science and Engineering, North Minzu University, Yinchuan, China
[b] School of Science, Ningxia Medical University, Yinchuan, China
[c] Key Laboratory of Image and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan, China
[d] Urinary Surgery, General Hospital of Ningxia Medical University, Yinchuan, China

ARTICLE INFO

Keywords:
Dental panoramic X-ray images
Medical auxiliary diagnosis
Deep learning
Context semantics
Contrast enhancement

ABSTRACT

Background and objective: It is very significant in orthodontics and restorative dentistry that the teeth are segmented from dental panoramic X-ray images. Nevertheless, there are some problems in panoramic X-ray images of teeth, such as blurred interdental boundaries, low contrast between teeth and alveolar bone.
Methods: In this paper, The Teeth U-Net model is proposed in this paper to resolve these problems. This paper makes the following contributions: Firstly, a Squeeze-Excitation Module is utilized in the encoder and the decoder. And proposing a dense skip connection between encoder and decoder to reduce the semantic gap. Secondly, due to the irregular shape of the teeth and the low contrast of the dental panoramic X-ray images. A Multi-scale Aggregation attention Block (MAB) in the bottleneck layer is designed to resolve this problem, which can effectively extract teeth shape features and fuse multi-scale features adaptively. Thirdly, in order to capture dental feature information in a larger field of perception, this paper designs a Dilated Hybrid self-Attentive Block (DHAB) at the bottleneck layer. This module effectively suppresses the task-irrelevant background region information without increasing the network parameters. Finally, the effectiveness of the algorithm is validated using a clinical dental panoramic X-ray image datasets.
Results: The results of the three comparison experiments are shown that Accuracy, Precision, Recall, Dice, Volumetric Overlap Error and Relative Volume Difference for dental panoramic X-ray teeth segmentation are 98.53%, 95.62%, 94.51%, 94.28%, 88.92% and 95.97% by the proposed model respectively.
Conclusion: The proposed modules complement each other in processing every detail of the dental panoramic X-ray images, which can effectively improve the efficiency of preoperative preparation and postoperative evaluation, and promote the application of dental panoramic X-ray in medical image segmentation. There are more accuracy about Teeth U-Net than others model in dental panoramic X-ray teeth segmentation. That is very important to clinical doctors to cure in orthodontics and restorative dentistry.

1. Introduction

Each tooth in the human mouth is composed of crown, root and neck. The tooth is surrounded by a fibrous periodontal membrane, which holds the tooth in place. In medical images, the periodontal membrane soft tissue is also the key anatomical structure for the human eye to judge the tooth contour. Normally, the human eye can only see the crown area. For the diseases in the tooth neck and root area, doctors must analyze the disease with the help of medical images, so the dental image of preoperative examination is particularly important [1]. At present, there is no shortage of people began to excessively pursue high-quality diet. However, due to the unreasonable diet habits and incomplete dental cleaning work, the health status of residents is not optimistic, the incidence of oral diseases is increasing.

Due to the frequent use of oral teeth, dental care and restoration work is carried out throughout a person's life. According to the World

☆ This paper is the results of the research which is funded in collaboration with National Science Foundation.
* Corresponding author. Key Laboratory of Image and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan, China.
** Corresponding author. School of Science, Ningxia Medical University, Yinchuan 750004, China.
E-mail addresses: zhoutaonxmu@126.com (T. Zhou), lu_huiling@163.com (H. Lu).

https://doi.org/10.1016/j.compbiomed.2022.106296
Received 29 August 2022; Received in revised form 17 October 2022; Accepted 6 November 2022
Available online 12 November 2022

Health Organization, as of 2021, the number of elderly people over the age of 60 in the world has exceeded the number of children (under the age of 5), which means that the trend of aging is inevitable. Almost all older people suffer from receding gums and tooth loss. The elderly need more doctors and drug resources in dental care. However, the development rate of medical treatment cannot match the speed of population aging, which means that people's dental problems are often not diagnosed and treated [2]. Therefore, for some common dental problems, it is urgent to use computer-aided diagnosis (CAD) and treatment to improve the efficiency of dentists' diagnosis before surgery. For dentists, advanced training is required before they can be certified in dental surgery, which requires at least two years of study and more than four years of clinical practice. In clinical dental implant surgery, some novice doctors may lack clinical experience, resulting in the contact of the micro implant with the adjacent tooth root, resulting in the failure of the operation. Therefore, in the training stage of dentists, computer-aided diagnosis and treatment can be used to visualize the basic dental morphology and structure, which is convenient for interns to learn intuitively and shorten the training period of dentists. The imbalance of the number of doctors and patients and the high training cost of dentists have promoted the development and application of computer intelligence (segmentation, recognition and other algorithms) in the field of dental medicine [3]. At present, computer-assisted medical methods have become feasible and safe, and panoramic dental X-ray image can assist doctors to locate lesions, determine the extent of resection and protect adjacent structures, which has significant clinical significance and application value.

Panoramic dental X-ray image is a complete and clear display of the whole picture of the upper and lower jaw, the dentition and alveolar bone of the upper and lower jaw on a single image, and to provide help for the diagnosis of diseases around the jaw. Panoramic image can accurately determine teeth tilt angle, periodontal soft tissue, teeth root and alveolar bone conditions. Accurate measurement of anatomical images of internal dental organs can provide technical support for doctors' preoperative diagnosis and analysis, staging, formulation of surgical plans, and postoperative evaluation. In the field of dental image analysis, teeth image segmentation technology is the key technology to analyze teeth status, and also the first step of computer-aided dental diagnosis. The segmented images can provide effective information for the assessment of missing teeth, the judgment of teeth development, the location of buried teeth and the judgment of adjacent relationships. It marks teeth through panoramic X-ray images of teeth to obtain the Ground Truth of teeth. Then the segmentation model is used to obtain the segmentation results of teeth, which ultimately provides a direct or indirect basis for clinical diagnosis.

In teeth segmentation of dental panoramic X-ray images, the accurate teeth root segmentation plays a very crucial role in the evaluation of impacted teeth, superordinate teeth, and missing dentition, as well as the review and evaluation of correction results. However, due to the blurred boundary between teeth and the low contrast between teeth and alveolar bone in panoramic dental X-ray images, accurate tooth segmentation is always a difficult problem. Aiming at the above problems, this paper proposes the dental panoramic X-ray image Teeth U-net model with context semantics and enhanced tooth boundaries. Its main contributions are as follows:

1) Between the encoder and decoder, a dense skip connection mechanism is used to store teeth detail information from different encoders to solve the problem of blurred teeth boundaries.
2) Aiming at the problem of teeth contour, this paper designs a multi-scale aggregate attention block, which can adaptively learn multi-scale features of different encoders. So that its input into the bottleneck layer contained sufficient teeth context feature information.
3) A dilated hybrid self-attentive block is proposed, where the image features after dilated convolution are input into a parallel self-attentive module to extract channel and spatial information to make a visible distinction between teeth and alveolar bone.
4) 1500 clinical data sets are divided into 12 categories. The latest segmentation model is used to compare with Teeth U-Net.

This paper is organized as follows: we present the Teeth U-Net models design in Section 3. In Section 4, we present the experiments part. Eventually, Summary and future prospects are drawn in Section 5.

2. Related work

When doctors face massive data, it is the premise and guarantee of maximizing the application value of dental images to effectively use the dental segmentation technology of computer-aided diagnosis to locate, extract and quantitatively analyze the human dental lesions. In the task of medical image segmentation, due to the successful application of deep learning in computer vision tasks, some teeth segmentation methods based on deep learning have been proposed. As shown in Table 1, Tekin [4] et al. used mask region-based convolutional neural network (Mask R–CNN) to segment and number teeth in occlusal radiography images and obtained high-quality segmentation masks. Yang [5] et al. developed an automated and simplified dental image analysis method, which integrates dental image diagnosis knowledge, thus saving a lot of manual work in data preparation. Xia [6] et al. put the upper and lower mandibles of computed tomography (CT) images into the proposed model, successfully segmented a single tooth from the CT images of natural contact scanning of upper and lower teeth, and obtained a complete single tooth model. The U-Net based network provides a new research idea for recent teeth segmentation tasks. The U-Net network provides a new research idea for the recent tooth segmentation task. Koch et al. [7] realized the semantic segmentation task of teeth panoramic images based on the U-Net network, which can perform better sample segmentation and relies on a smaller and simpler network architecture. The Efficient Encoder-Decoder Network (EED-Net) architecture was constructed by Kong et al. [8] and others to segment maxillofacial quickly and accurately. The network consists of residual encoder, multipath feature extractor and object-oriented decoder. Zhao et al. [9] proposed a two-stage two-staged attention segmentation network (TSASNet) model for tooth localization and segmentation in dental panoramic X-ray images. This model can automatically aggregate pixel-level context information and capture those fuzzy teeth areas. Cui et al. [10] proposed tooth segmentation network (TSegNet) for 3D

Table 1
Summary of research status.

Work	Year	Approach	Objective
Tekin et al. [4]	2022	Mask R–CNN	Enhanced tooth segmentation and numbering
Yang et al. [5]	2018	Vgg 16	Age estimation
Xia et al. [6]	2017	Traditional method	Single tooth segmentation
Koch et al. [7]	2019	FCNs	Teeth segmentation
Kong et al. [8]	2020	EED-Net	Tooth contour segmentation
Zhao et al. [9]	2020	TSASNet	Teeth segmentation
Cui et al. [10]	2021	TSegNet	Three dimensional tooth segmentation
Feng et al. [11]	2021	URNet	Image dehazing
Liu et al. [12]	2020	Res-Unet	Nailfold capillaries
Wang et al. [13]	2021	HDA-ResUNet	Liver tumor
Alom et al. [14]	2018	RCL	Nuclei
Jin et al. [15]	2020	RAUNet	Liver tumor
Liu et al. [16]	2020	DRANet	ischemic stroke
Wang et al. [17]	2021	CLCU-Net	brain tumor
Jose et al. [18]	2019	IVD-Net	intervertebral disc
Zhang et al. [19]	2018	MDU-Net	Carcinoma of colon
Wang et al. [20]	2020	AFD-UNet	liver cancer
Mohammad et al. [21]	2020	Dense-Unet	Lung contour

scanning points of dental models to achieve tooth segmentation.

U-Net networks have achieved good results in medical image segmentation tasks. Many scholars have continued their research on U-Net structure, such as the improvement of encoder and decoder, the improvement of convolutional layer and the improvement of skip connection. The convolutional layer of U-Net can be added with residual unit, which can train the network effectively, solve the degradation problem well, deepen the number of network layers and improve the model performance. Feng et al. [11] proposed a U-Net based residual network (URNet), where the network embeds residual units in the bottleneck structure, significantly improving the image dehazing effect. Liu et al. [12] proposed a Res-Unet structure with two residual units in the coding and decoding parts, which deepens the number of layers in the network and has good performance for capillary segmentation. Wang et al. [13] proposed hybrid dilation and attention residual U-Net (HDA-ResUNet), which incorporates residual connections in each convolutional layer to obtain high-level features, and this model has fewer parameters and better segmentation compared to U-Net; Alom et al. [14] proposed R2Unet, which uses recurrent convolutional layer (RCL) and RCL with residual units in the coding and decoding units instead of regular convolutional layers, which helps to develop more efficient and deeper models.

The addition of an attention mechanism to the skip connection allows rescaling of the extracted features. Jin et al. [15] proposed the residual attention U-Net (RAUNet) for liver tumor segmentation. The attentional residual mechanism in this network contains a backbone branch and a soft mask branch. The backbone branch learns the original features, while the soft mask branch focuses on reducing noise and enhancing good features, and the method achieves good results in liver tumor segmentation with high scalability and generalization capability in brain tumor segmentation. Liu et al. [16] proposed deep residual attention network deep residual attention network (DRANet), in which the attention mechanism is composed of residual blocks and inflated convolution, and the attention mechanism improves the feature processing between the encoder and decoder of the network, allowing the model to better distinguish between the two lesion types. Wang et al. [17] proposed a cross-level connected U-Net (CLCU-Net). An attention mechanism consisting of a segmentation of the channel attention module is added to the encoded path and the skip connected path to extract useful information from the connected features and eliminate redundant information.

The incorporation of dense connectivity between encoders and decoders can better structure the relationship between different modules of the encoders and decoders. Jose et al. [18] proposed the intervertebral disc network (IVD-Net), a model that uses a dense mechanism to connect encoders layer by layer, with each encoder processing a different image pattern, leaving the model free to understand where and how different patterns should be processed and combined. Zhang et al. [19] proposed the multi-scale densely connected U-Net (MDU-Net), which uses dense connections between encoders and decoders to directly fuse adjacent feature maps of different scales at high and low levels, which improves the encoder, decoder and skip connection, reducing the overfitting from the dense connection. Wang et al. [20] proposed the adaptive fully dense UNet (AFD-UNet), a network based on Unet++ that adaptively and efficiently utilizes shallow and deep features by densely connecting the features of each layer of Unet++ through horizontal dense connections. Mohammad et al. [21] proposed the Dense-Unet, this network connects each layer of the encoder downwards layer by layer with each layer of the decoder upwards layer by layer to form a dense connection effect so that different levels of image combinations can be utilized.

Above all, this paper proposes a Teeth U-Net model for the problem of accurate segmentation of dental teeth, and uses a dense connection mechanism between the encoder and decoder to solve the teeth boundary blurring problem; In the encoder we use a multi-scale aggregated attention block that can adaptively learn multi-scale features from different encoder layers, so that its input to the bottleneck layer contains sufficient information of dental contextual features; In order to make a clear distinction between teeth and alveolar bone, a dilation hybrid self-attentive block is proposed, and the image features after dilated convolution are input to a parallel self-attentive block to extract channel and spatial information.

3. Teeth U-Net model design

3.1. Technical background

This paper proposes a Teeth U-Net network for medical images based on U-Net with contextual semantics and enhanced dental boundaries. Firstly, the dataset is cleaned from the overall image dataset and 1500 useable datasets are collated. Subsequently, the clinician outlines the labels of the segmented teeth data and identifies the segmented locations. The dental data set is then placed into the encoder of the proposed network, using the Squeeze-and-Excitation (SE) module with a convolutional layer to focus on the image features to pass through the bottleneck layer. Using a feature aggregation attention block and a dilated hybrid self-attention block at the bottleneck layer. Re-passing image features into the decoder. Finally, a dense skip connection of the different layers is used in the decoder to receive the passed information of the encoding layer and the feature information of the bottleneck layer. The decoder reduces the image semantic features of the teeth in the dental panoramic X-ray image to obtain the final segmentation result for each teeth.

3.2. U-net model

With the rapid development of deep learning technology, the application of deep learning in the field of medical images has attracted extensive research and attention. Among them, how to automatically identify and segment the lesions in medical images is one of the most concerned problems. To solve this problem, in 2015, Ronneberger et al. published U-Net [22] in MICCAI conference, which is a breakthrough progress of deep learning in medical image segmentation. U-Net is improved based on fully convolutional network (FCN), which consists of encoder, bottleneck module and decoder. Due to its U-shaped structure combining context information and fast training speed, the amount of data used is small. It satisfies the demands of medical image segmentation and is widely used in medical image segmentation. The structure of U-Net is shown in Fig. 1. Due to the diversity of lesion shapes and the difference of different organ structures, only using U-Net structure to segment dental teeth cannot meet the needs of accuracy and speed.

3.3. Teeth U-Net network architecture

The U-Net medical image segmentation model with contextual semantics and augmented tooth boundaries is shown in Fig. 2. The dental data set is passed into the proposed network. After encoder, bottleneck layer, skip connection and decoder. Eventually all the teeth are clearly segmented. The encoder consists of Same convolutional layers, SE modules, which are utilized to extract the contextual semantic information of the teeth. The bottleneck layer includes the multi-scale feature aggregation attention block (MAB), and the dilated hybrid self-attention block (DHAB). These modules will collect local features, extract multi-scale features and enhance perceptual representation performance. Dense skip connect module (DSM) is to apply convolution and pooling operations to images between different layers to reach the same image scale transmitted by the same layer, and then concatenate them with the decoder of the corresponding layer. The decoder is used to aggregate feature information from the different layers and can concatenate the features extracted by the encoder with the upsampled feature map. The decoder also includes the Same convolution layer and SE module, which receives the information of bottleneck layer and encoder to restore the

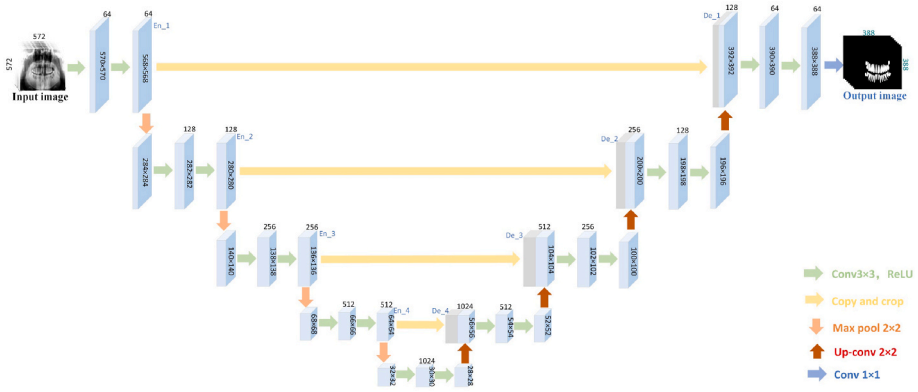

Fig. 1. U-Net segmentation model.

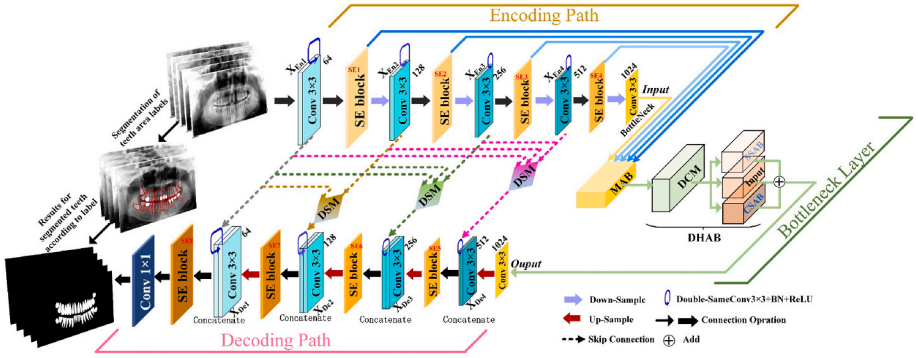

Fig. 2. Teeth U-Net segmentation model.

original image size and output the final segmentation result.

The difficulty of dental segmentation is significantly increased by the fact that the dental X-ray data set presents vague and difficult information and that an adult usually has 32 teeth, which means that there are 32 teeth positions that can be segmented. Although convolutional manipulation can be used to extract rich information while expanding the field of perception by stacking more layers, deeper convolutional layer structures do not provide good access to global information [23]. Whereas, the attention module can adjust the global information, and each point in the image will calculate the correlation with other points. The weight of pixels in the image is adjusted by the correlation information obtained from the attention feature map to highlight the important part of teeth in the panoramic dental X-ray image. Hence, in this paper, the channel attention SE module is applied to each layer of the convolutional layer to highlight vital information about the teeth. The U-Net architecture connects only the encoder and decoder layers that are symmetrical on the same layer when the decoder recovers the feature map spatial information. However, the simple skip connection ignores important information such as teeth position and boundaries in the lower-level semantic feature map. To facilitate the flow of information among feature maps, this paper proposes the Dense Skip Connection Module. It can exploit the potential of the network to correctly locate dental boundaries through feature reuse. For a particular X-ray image of a dental, feature maps at different scales in the encoder have different correlations with the object. If the encoder can automatically establish the weights of the scale at which the image conveys information, so that the network can adaptively recalibrate the channel features to highlight the most relevant feature channels. In addition, the roots of teeth are complex and uneven in shape. In this paper, we propose DHAB to absorb the root structure of teeth for better segmentation performance.

3.3.1. *Dense skip connection module*

In the U-Net structure, encoder and decoder paths are connected layer by layer at corresponding locations. These connections help in the localization of pixel classes. However, due to the variety and shape of teeth in the data set, down sampling in the encoder will lead to the loss of spatial structure information. In order to enable the encoder to save more details of the pre-feature mapping information, inspired by document [24], this paper proposes a dense skip connection between the encoder and the decoder. Specifically, each decoder layer of Teeth U-Net combines details from both the symmetric encoder layer and all the upper encoder layers. This model retains more spatial contextual information and facilitates the use of multi-level features to recover images.

In this paper, the feature mapping of the i-th encoder layer is denoted as X_{Eni}, and in a similar way the feature mapping of the i-th decoder layer is denoted as X_{Dei}. Taking the lowest level information of X_{De4} as an example, the decoder feature map X_{De4} fuses the low-level details of the encoder feature maps X_{En1}, X_{En2} and X_{En3}. The high-level semantic information of the symmetric encoder feature map X_{En4} is fused by pixel-level addition, and Fig. 3 illustrates how the feature map of X_{De4} is constructed. In order to achieve a pixel-level overlay of the different encoder layer feature maps, this paper needs to unify the size and number of channels of the different feature maps through maximum pooling and convolution operations. Specifically, this paper uses max pooling on X_{En1} to unify the size of X_{En1} and X_{En4}, and a 1×1 convolution operation on X_{En1} to unify the number of channels in X_{En1} and X_{En4}. Similarly, the image size and number of channels of X_{En2} and X_{En4} are unified using max pooling and a 1×1 convolution operation on X_{En2}. Finally, a 1×1 convolution operation with a max pooling operation is used on X_{En3} to unify the size of the images and the number of channels of X_{En3} and X_{En4}. As shown in Table 2. DSM4, all dimensions and number of channels are resized to the size of X_{De4} and then

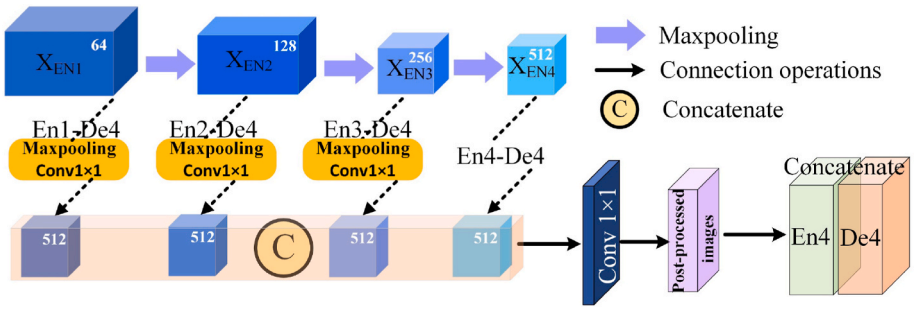

Fig. 3. Dense skip connection module.

Table 2
DSM structural operations.

DSM	Path	Operation
DSM4	En1-De4	MaxPool 8 × 8512, Conv1 × 1
	En2-De4	MaxPool 4 × 4 512, Conv1 × 1
	En3-De4	MaxPool 2 × 2 512, Conv1 × 1
	En4-De4	–
DSM3	En1-De3	MaxPool 4 × 4 256, Conv1 × 1
	En2-De3	MaxPool 2 × 2 256, Conv1 × 1
	En3-De3	–
DSM2	En1-De2	MaxPool 2 × 2 128, Conv1 × 1
	En2-De2	–
DSM1	En1-De1	–

concatenated with the corresponding X_{De4} using feature concatenation and Conv 1 × 1, resizing the number of channels to the size of X_{De4}. In this way, feature mappings containing different levels of semantic information are obtained and then they are concatenated with the features.

In Figs. 2 and 3, the operations are performed in the following order: a dense skip connection is performed, followed by feature concatenation, and finally two 3 × 3 Same convolution operations are performed. Each convolution operation is followed by a batch normalisation and a ReLU activation function. The expression for the dense skip concatenation is as follows.

$$X_{Dei} = \begin{cases} C(X_{Eni} \vee U(X_{Dei}+1)), i = 1 \\ C\left(C_1\left(\left(\bigvee_{k=1}^{i-1} C_1(P(X_{Enk})) \vee X_{Eni}\right)\right) \vee U(X_{Dei}+1)\right), 1 < i < 4 \\ C\left(C_1\left(\left(\bigvee_{k=1}^{i-1} C_1(P(X_{Enk})) \vee X_{Eni}\right)\right) \vee U(X_{ma})\right), i = 4 \end{cases} \quad (1)$$

where the function *C(.)* represents two consecutive 3 × 3 Same convolution operations. *C1(.)* represents a 1 × 1 convolution operation, each followed by a batch normalisation and ReLU activation function. *P(.)* represents the max pooling operation with a convolution kernel size of 2^{i-k}. *U(.)* indicates an operation with an upsampling coefficient of 2, and $\vee$ is a concatenating operation. X_{ma} represents the feature map after going through the hybrid attention module. The details of the architecture and the operation of each pathway are described in Table 2.

3.3.2. *Multi-scale aggregated attention blocks*

In order to enable the network to sufficiently improve the low contrast of the X-ray image as well as increase the teeth boundary information in the encoding stage, multiscale aggregation attention block (MAB) is proposed in this paper as shown in Fig. 4. In designing the MAB module this paper takes a number of factors into account. It is possible to go through a single convolution operation, set the step size to a large number and rapidly get the same size in the bottleneck layer. However, through experiments, it has been found that the step size, if done this way, will have cracks in the adjacent receptive fields. Therefore, in this paper, the convolution step is set to 2, which means that the adjacent sensory fields will not be duplicated and will cover a wider range. Each level after processing by SE module is defined as the input of MAB. After the SE module, the step size of each convolution operation is set to 2, and the filling is 1. After the operation of convolution kernel 3 × 3, it becomes the same size of the bottleneck layer for concatenating. The number of channels is then changed to 1024 after a 1 × 1 convolution operation. In the multi-scale aggregated attention block makes the module better able to handle features at different scales, and it is reasonable to combine these features for prediction. However, for different scales the feature map may have different relevance to that object. A model that automatically determines the scale weights for each pixel will allow the network to adapt itself to the corresponding scale for a given input. Therefore, we incorporate an attention module to learn image-specific weights for each scale to calibrate the features at different scales. This attention setup is used at the end of the network, as shown in Fig. 4. After concatenating the image information from different layers and resizing them to 32 × 64 × 1024, followed by averaging pooling and MLP respectively to obtain the channel coefficients α, as shown in Equation (2).

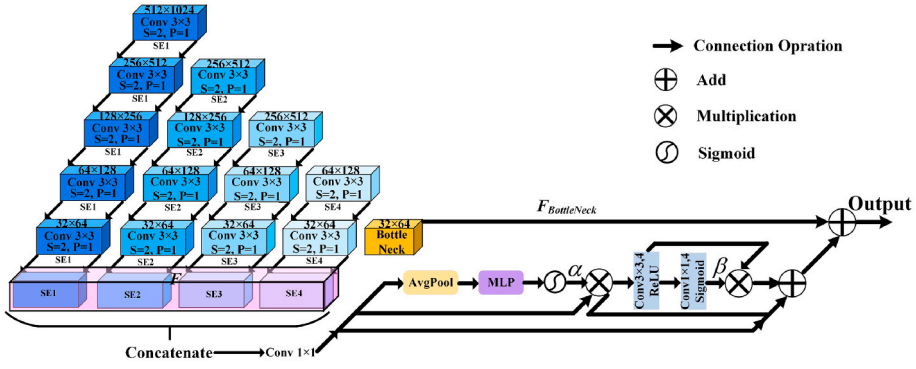

Fig. 4. MAB module.

$$\alpha = \sigma\{MLP[AvgPool(F)]\} \tag{2}$$

As shown in Equation (3). The channel coefficients are multiplied with the concatenated feature maps following 3 × 3 convolution, ReLU and 1 × 1 convolution, Sigmoid, to obtain the channel coefficients. As shown in Equation (3).

$$\beta = \sigma\left\{Conv_{1\times1}\left[Conv_{3\times3}, ReLU(F\times\alpha)\right]\right\} \tag{3}$$

The features are then joined together using residual connections. Finally the feature maps for each weight are obtained and the dimensions of the bottleneck layer input are summed. The process is shown in Equation (4):

$$F_{MAB} = F\times\alpha\times\beta + F\times\alpha + F_{BottleNeck} \tag{4}$$

3.3.3. Dilated hybrid self-attention block

In order to capture contextual information in a larger perceptual field without increasing the network parameters, while suppressing the ability to segment background regions that are irrelevant to the task, this paper designs a dilated hybrid self-attentive block. This block is made up of two stages, the first of which is a serial and parallel merged dilated convolution. In the proposes dilated convolution module, this paper uses different convolution operations in the four branches to extract features at different scales, which allows the network to learn more contextual information by fusing spatial information at different granularities. The former three branches operate in parallel, the fourth branch is a serial branch, and the dilation rate is set to 1, 2 and 3 respectively, while the size of the original convolution kernel is 3 × 3. The size of the dilation receptive field can be obtained according to Equation (5), where k denotes the size of the convolution kernel, d denotes the dilation rate, and R denotes the receptive field size of the dilation convolution. Each branch expands the receptive fields by enlarging the size of the original convolution kernel at a different dilatation rate. By using multiple parallel dilated convolutions with different receptive fields in the dilated convolution block, this paper can obtain multiple scales of contextual information from the feature maps of the dental panoramic X-ray image.

$$R = k + (k-1)(d-1) \tag{5}$$

The second stage takes as input a three-branch path with dilated convolution, channel self-attention and spatial self-attention in parallel. The Spatial Self-Attention block aims to extract essential information about the image and calculate the importance of each pixel feature in the spatial domain. The Channel Self-Attention block is used to inhibit useless features while highlighting useful features by using the learned global information to achieve feature recalibration. However, spatial self-attention ignores the discrepancies between channels. In contrast, channel self-attention aggregates global information directly and lacks a local representation of each channel. Consequently, the bottleneck layer of the architecture can continuously execute channel self-attention and spatial self-attention, which can better generate attention maps by further utilizing channel and spatial relations between different feature maps. In contrast to many complex natural images, medical images contain rare shapes and fixed structures. With this aspect in mind, this paper transposes the self-attentive module to a framework following dilated convolution.

Based on the module proposed by dual attention network (DA-Net) [25] at the bottleneck layer, this paper introduces it following the dilated convolution. This module consists of three subsequent block: the image features of the dilated convolutional input, the channels and the spatial self-attention block. As shown in Fig. 5, the encoder module extracts the features of the input image data and puts them into parallel dilated convolution with different dilation rates. These features are subsequently fed into two parallel attention modules, the Channel Self Attention Block (CSAB) and the Spatial Self Attention Block (SSAB), to generate the expressivity of the channel self-attention and spatial self-attention.

The SSAB is shown in Fig. 5, given a local feature $Input \in R^{C\times H\times W}$, it is initially fed into the convolution layer to generate two new feature maps Q and K, respectively. Where $\{Q, K\} \in R^{C\times H\times W}$. Then, the SSAB block inputs the feature *reshape* as $R^{C\times N}$, where $N = H \times W$ is the number of pixels. After that, matrix multiplication is performed between the transpose of Q and K. After that, matrix multiplication is performed between the transpose of Q and K. Finally the softmax layer is applied to calculate the spatial attention map $S \in R^{N\times N}$. S_{ji} measures the influence of the *i-th* location on the *j-th* location. The more similar the feature representation of two locations, the more it contributes to their correlation.

$$S_{ji} = \frac{exp\left(Q_i^T \bullet K_j\right)}{\sum_{i=1}^{N} exp\left(Q_i^T \bullet K_j\right)} \tag{6}$$

At the same time, the feature *input* is fed into the convolution layer to generate a new feature map $\in R^{C\times H\times W}$, again *reshape* it to $R^{C\times N}$. Then, matrix multiplication is performed between the transpose of V and S to *reshape* the result into $R^{C\times H\times W}$. Finally the input features *input* are summed element by element to obtain the final output $B \in R^{C\times H\times W}$, as shown below.

$$B_j = \sum_{i=1}^{N}\left(S_{ji}D_i\right) + input_j \tag{7}$$

The CSAB is shown in the bottom right of Fig. 5, this module is different from the SSAB. Input *reshape* of *input* to $R^{C\times N}$, then matrix multiplication between *input* and the transpose of *input*. Finally, the softmax layer is applied to obtain the channel attention map $P \in R^{C\times C}$, as illustrated below.

$$P_{ji} = \frac{\exp\left(input_i^T \bullet input_j\right)}{\sum_{i=1}^{C}\exp\left(input_i^T \bullet input_j\right)} \tag{8}$$

where P_{ji} is a measure of the impact of the i-th channel on the j-th channel. In addition, we perform matrix multiplication between *input* and the transpose of *input* and reshape the result to $R^{C\times H\times W}$. Then, an element-by-element summation operation is performed with *input* to obtain the final output X, giving the final output $X \in R^{C\times H\times W}$.

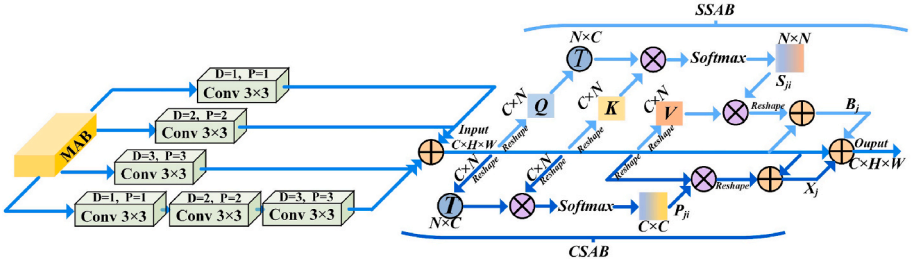

Fig. 5. Dilated Hybrid self-Attention Block.

$$X_j = \sum_{i=1}^{C} \left(P_{ji} input_j\right) + input_j \tag{9}$$

SSAB selectively aggregates features at each spatial location by weighting all spatial locations. This allows the model to capture the long-term dependency of features, where similar features will be correlated with each other. At the same time CSAB ensures that the full space is used for representation and normalisation, thus enhancing the contrast of features in different channels and allowing the model to have better discrimination.

The overall Dilated Hybrid self-Attention Block (DHAB) module is then shown below , D_1, D_2 and, D_3 are denoted as dilated convolutions with dilation rates of 1, 2, and 3, respectively. While ∪ denotes a parallel dilated convolution, and ∩ denotes the dilated convolution in series. The front two brackets indicate the series and parallel expansion convolution doing addition, followed by the spatial and channel self-attention. And *input* is the dilated attention as input to do addition with the two attention modules as shown below.

$$Output = [D_1 \cup D_2 \cup D_3] + [D_1 \cap D_2 \cap D_3] + B_j + X_j + input \tag{10}$$

3.3.4. Squeeze and Excitation module

In order that the network can learn to focus on the discrepancies among different viewpoints during the input of dental data images, the significance of the characteristics of each perspective is learned automatically. The SE module is introduced in the encoder and decoder of the model [26] The SE module essentially implements a self-attentive function of the channel, which recalibrates the channel feature response. It also learns global information by suppressing non-useful features and emphasising informative features, while feature recalibration can be accumulated through the SE module.

Although convolutional structures can expand the field of perception and extract rich information by stacking more layers, deeper convolutional layer structures do not provide sufficient access to global information. In turn, the SE Attention module provides access to global information. In this paper, the SE module for channel attention is placed after each convolutional layer, allowing the more accurate feature information extracted by the encoder to be passed to the underlying layer with the decoder. And the decoder can fuse the features extracted by the encoder with the upsampled feature map for enhancement. As shown in Fig. 6, feature map *X1* represents the encoder features and feature map *X2* represents the decoder features. The downsampling structure of the SE module is shown in Fig. 6 a. After two 3 × 3 Same convolutions to obtain the features concatenated with the corresponding layer, the SE module is output to obtain the output feature map *Y*. Fig. 6 b shows the upsampling structure of the SE module. After upsampling, feature map *X2* is concatenated with feature map *X1* from the channel dimension.

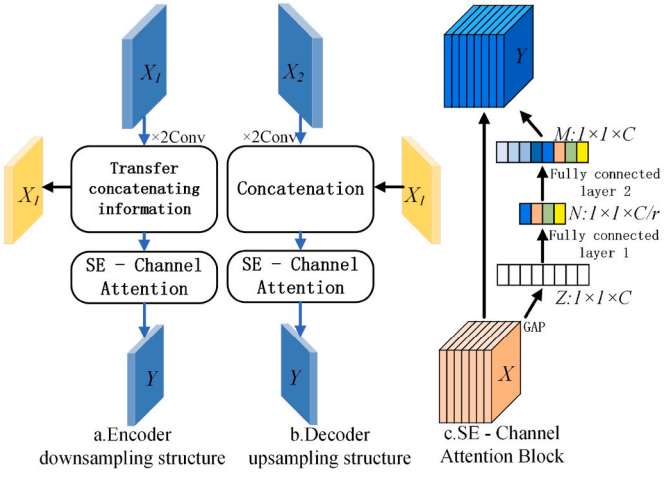

Fig. 6. Channel attention SE module.

The concatenated feature map is input to the channel attention SE module, and finally the feature map *Y* is obtained. The details structure of the channel attention module is shown in Fig. 6 c. The feature map *X* is extracted to the channel weight information *Z* by global average pooling (GAP). *Z* passes through two fully connected (FC) layers, and the obtained channel information is multiplied with the feature map *X* for channel weight adjustment to obtain the feature map *Y*.

4. Experiments

4.1. Experimental platforms and their data sets

All the networks in the experiments followed the same training strategy:

Hardware environment: Computer with 256 GB of RAM, graphics card NVIDIA TITAN V, processor Intel(R) Xeon(R) Gold 6154 CPU @ 3.00 GHz.

Software environment: Windows Server 2019 Datacenter 64-bit operating system, Python 3.7.10, Pytorch 1.7.0 deep learning framework, CUDA 11.3.58. In this paper, the RMSprop optimizer is used, which has the effect that a more gentle direction in the parameter space will result in improved network performance.

Parameter setting: The batchsize is 3, the number of iterations is 200, and the learning rate is 1e-5. It is adjusted using the callback function ReduceLROnPlateau strategy. The training speed is increased using the Gradient Scaler module.

The design of the loss function usually considers the characteristics of the data set, which is divided into 12 categories in this paper, but the distribution of each oral dental category in the data set is uneven. In order to accelerate the convergence speed of the model, the loss function in this paper is the cross-entropy loss function, and the smaller the loss function is, the smaller the difference between the predicted and true values of the model. As shown in Equation (11), y' represents the output of the model.

$$L = -y \log y' - (1-y)\log(1-y') \tag{11}$$

The data set is selected from 1500 clinical patients who undergo dental examination in the dental department of a tertiary hospital in Ningxia during the period of January 2020–June 2022. The data for this experiment was collected by a digital panoramic X-ray machine (Ingram Micro OP200D). The dental data set was taken by instructing the patient to remove any metal objects above the neck, such as hairpins, necklaces, earrings, with removable dentures, as well as caps and glasses, etc. The patient must stand in front of the machine in a correct position, stand upright, hold the handrails on both sides with both hands, adjust the height, let the patient bite the support, and then close the door of the radiation protection room. The data set was collected by each patient in close collaboration with the dentist. The labelled data was stored in CSV files by the dentist using the VIA tool and after cleaning and sorting the data, the format was converted to the png format required for the segmentation model. The data file contains the contour position of each teeth, and the original data is collated with the corresponding binary labelled data. The images in this dataset have grade differences and can be classified into 12 different grades, which are described in Table 3 and Fig. 7. The proposed method is trained on a dataset of clinical dental panoramic X-ray images containing all morphological and structural conditions of the teeth so that the method can be effectively applied in clinical practice. For the dataset setup, the original size of the X-ray was 1536 × 2976. However, in this paper, the dataset is adjusted to 500 × 1024 by cropping and scaling to place the network input, and the output is judged to be better after experiments. In addition, for each teeth category, the dataset is randomly divided into a training set, a validation set and a test set, with a segmentation ratio of 8:1:1. Finally, the training set contains 1200 images, while the validation and test sets each contain 150 images.

Table 3
Presentation of dental X-ray panoramic data sets.

Category	Types of teeth	Number
a.	Images of all teeth, including those with restorations and orthodontic appliances	75
b.	Images of all teeth, including teeth with restorations but excluding teeth with orthodontic appliances	62
c.	Images of all teeth, including teeth with dental orthodontic appliances but excluding teeth with restorations	30
d.	Images of all teeth, excluding teeth with restorations and teeth with dental orthodontic appliances	80
e.	Images of teeth with dental implants	26
f.	Images containing more than 32 teeth	1
g.	Image of missing teeth, including teeth with prostheses and dental appliances	61
h.	Image of missing teeth, including teeth with restorations but not teeth with orthodontic appliances	42
i.	Image of missing teeth, including teeth with dental orthodontic appliances but not teeth with restorations	31
j.	Image of missing teeth, excluding teeth with restorations and teeth with dental appliances	58
k.	Image of all teeth, only interrupted teeth	608
l.	Images of all teeth, only of decayed teeth	426

4.2. Evaluation indicators and data sets

In order to objectively and comprehensively evaluate the diagnostic performance of the network, and to facilitate comparison with other algorithms. This paper uses Accuracy (Acc), Precision (Pre), Dice coefficient, Recall, Volumetric Overlap Error (Voe) and Relative Volume Difference (Rvd) to evaluate Teeth U-Net. To standardise the four evaluation metrics, the values of Voe and Rvd are taken as 1 minus the difference between these two metrics respectively. A larger difference between them indicates a better result. Conversely a smaller value indicates a poorer effect. Table 4 shows the definitions of the evaluation indicators.

A teeth area that is correctly segmented as True Positive (TP), a normal area that is correctly segmented as True Negative (TN), a normal tissue area that is segmented as a teeth area as False Positive (FP) and a teeth area that is segmented as a normal area as False Negative (FN) are defined as True Positive. P denotes the target pixel predicted by the model and G denotes the target pixel in the label value (ground truth). Six evaluation metrics are used to demonstrate the advantages of the proposed network. The correctness of the predictions is assessed using Acc, by dividing the number of correctly segmented samples by the number of all samples. The accuracy of the teeth segmented by the system is assessed using Pre, by calculating the proportion of correct predictions that are positive as a percentage of all predictions that are positive. Dice is used to assess the similarity of the model predictions to the target pixels of the label values. Recall is utilized to assess the efficiency of the model in successfully segmenting the teeth by calculating the percentage of correct predictions that are positive as a percentage of all actual positive predictions. Voe is employed to assess the error rate of the predicted model with respect to the labelled values. Rvd in order to calculate the volumetric difference between the model predicted target pixels and the labelled target pixels. A lower volume overlap error Voe and relative volume difference Rvd indicates less segmentation error and better segmentation.

4.3. Experimental design and analysis

Three sets of experiments are conducted for panoramic dental X-ray image slices to demonstrate the superiority of Teeth U-Net, the model proposed in this paper. The first group of experiments demonstrates the impact of contextual semantic relevance on the segmentation network. The second set of experiments is a comparison with some advanced networks. Since this network proposes MAB and DHAB in the bottleneck layer, the comparison with advanced networks is done with proposed modules that have been proposed in the last five years and have similar proposed modules in the bottleneck layer. The third set of experiments is the ablation experiment, which is designed to demonstrate the effectiveness of each module in this paper for the segmentation of teeth.

4.3.1. Segmentation networks for contextual semantic relevance

The experiment of context semantic relevance proves the advantages of the proposed network through five groups of experiments. The training set is 1300 X-ray images, and the validation and test sets are both 150 X-ray images respectively. The method proposed in this paper includes the following aspects in terms of contextual semantic correlation: dense skip connections between encoder and decoder, SE modules. Therefore, this selected experiment of contextual semantic relevance

Table 4
Definitions of evaluation indicators.

Evaluation indicators	Definition	Evaluation indicators	Definition
Acc	$Acc = \frac{TP+TN}{TP+FP+FN+TN}$	*Pre*	$Pre = \frac{TP}{TP+FP}$
Dice	$Dice = \frac{2\times\lvert P\cap G\rvert}{\lvert P\rvert+\lvert G\rvert}$	*Recall*	$Recall = \frac{TP}{TP+FN}$
Voe	$Voe = abs\left(1-\left\lvert\frac{P\bigcap G}{P\bigcup G}\right\rvert\right)$	*Rvd*	$Rvd = \frac{abs(P-G)}{G}$

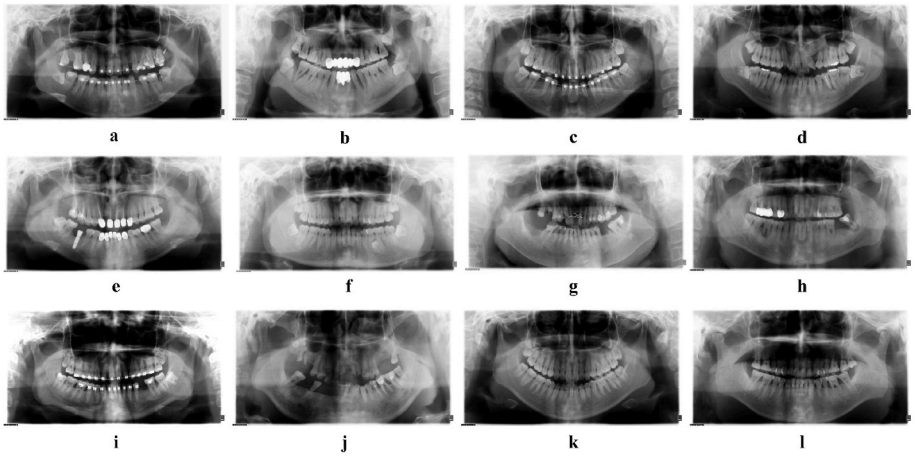

Fig. 7. Dental X-ray panoramic data set 12 classification map.

needs to include the SE module of SEResUNet [27] and the densely connected UNet++ [28] in the segmentation network to verify the effectiveness of the module. While U-Net [22] is a classical medical image segmentation network, U-Net is used as the benchmark in the experimental comparison of contextual semantic relevance. Since the proposed method in this paper adds the SE module after the Same convolutional layer, and ResUNet [29] replaces the convolutional layer of U-Net with a residual connected block, this experiment of contextual semantics is to compare the effectiveness of the SE module in the proposed method.

The six evaluation indexes in Table 5 show that the ResUNet model is the least effective in segmentation and the segmentation is poor. As can also be seen in Fig. 9, the ResUNet network shows an over-segmentation phenomenon in the e-column image, with a redundant white spot. The tooth on the lower left side of the column c image is not segmented to the root. The upper rightmost tooth in the d-column image is under-segmented and the tooth contour segmentation effect has a visible disadvantage to the naked eye. In the SEResUNet network, only the Rvd evaluation index is 3.24% higher than that of the U-Net, despite the addition of the squeeze incentive mechanism. The rest of the indices did not exceed the classical medical image segmentation network U-Net. It is also clear from the segmentation visualization results in Fig. 9 that U-Net shows a broken continuum in the processed image of the tooth root in the upper right corner of column d. In the upper left corner U-Net does not segment the tooth boundary. In contrast, SEResUNet is intact, which shows that U-Net is less effective in processing the tooth root boundaries. However, the U-Net segmented the root of the tooth in the lower left-hand side of the image in column c. No segmentation has occurred in the third tooth from the top in the column a image. The SEResUNet segmentation appears as a sporadic burr in the upper right side of the blocked tooth in the c-column image. This shows that U-Net does outperform SEResUNet in some indices in the dataset targeting the oral teeth. And the U-Net++ network can capture features at different levels and integrate them at different levels by means of feature overlay. The sensitivity to target objects of different sizes is different, and features with large sensory fields can be easily identified for large objects. However, in practical segmentation, information about the edges of large objects and the small objects themselves are easily lost to the deep network in one downsampling and one upsampling. This is where features with small sensory fields may be needed to help. And U-Net++ just has different sizes of perceptual fields. So it works better than the basic U-Net network in visualising the result graphs and various evaluation indices. However, the network proposed in this paper adds the idea of dense skip connection and applies the SE module to each layer of the encoder and decoder espectively. Moreover, the bottleneck layer gradually expands the perceptual field by dialated convolution and adds a hybrid attention module, presenting the ideas of dense connection mechanism and SE attention module in the network. The six evaluation indicators obtained for Acc, Pre, Recall, Dice, Voe and Rvd all exceeded U-Net++ by 6.13%, 3.25%, 0.69%, 0.76%, 1.03% and 2.81% respectively. As can be seen in the visualization results in Fig. 9, U-Net++ segmented only a white line at the root of the tooth in the upper left corner of the d-column segmentation map. In contrast, the network proposed in this paper clearly segments out the upper left third molar. The information fusion between contextual semantics is effectively achieved and the tooth boundary region is enhanced to improve segmentation performance. In Radar Fig. 8, it can be presented that the coverage of the proposed method in this paper can reach the maximum of the contextual semantic experiment. It shows that the model has better segmentation performance.

Table 5
Comparison of contextual semantic experiments.

Architecture	Acc%	Pre%	Recall%	Dice%	Voe%	Rvd%
U-Net [22]	92.27	93.82	91.94	92.78	86.61	88.02
U-Net++ [28]	92.40	92.37	93.82	93.52	87.89	93.16
SEResUNet [27]	92.22	93.51	91.89	92.62	86.29	91.26
ResUNet [29]	90.89	89.96	90.71	91.24	83.94	88.97
Ours(Teeth U-Net)	**98.53**	**95.62**	**94.51**	**94.28**	**88.92**	**95.97**

4.3.2. *Comparison with advanced segmentation networks*

The method proposed in this paper includes multiscale aggregated attention blocks, dense skip connections and DHAB to solve the segmentation problem of oral teeth. To demonstrate the effectiveness of the present model, this experiment is compared with parallel network models involving attentional mechanism, networks with dilated convolution and recently proposed advanced network models.(see Table 6).

To further validate the effectiveness of this approach in the teeth segmentation task, the segmentation results of this model on the dental dataset is compared with other state-of-the-art methods and the visualization of the segmentation is shown in Fig. 10. DA-Net [25], although proposing a parallel attention network, lacks the judgement of the boundaries of each teeth in the dataset targeting oral teeth. Moreover, the adhesions of each teeth are more severe. The position of the teeth and the boundaries of the roots are blurred, making it difficult to distinguish the status of the relationship among the teeth. U-shape hybrid Transformer Network (UTNet) [30] applies a self-attentive module in both the encoder and decoder, which is slightly higher than DA-Net in terms of the final evaluation metrics. However, the visualization of the segmentation shows that the roots of the mandibular third molar on the right side of column a appear under-segmented. Also in this network, the first molar portion of the column d image appears over-segmented. The maxilla and mandible bone in the e-column image also appear to be adherent. TransUnet [31] uses the Transformer mechanism in the encoder section. The evaluation index of the segmentation is slightly higher than that of UTNet. The following problem remains, this network shows over-segmentation in the maxillary first molar on the right side of the column b image, while the original image does not have this tooth. Also in the d-column images there is over-segmentation of the central incisor, the upper left first molar and the lower left third molar resulting in a slight lack of segmentation performance of TransUnet in the treatment of oral teeth. Tumor attention networks (TA-Net) [32] does not process the dental data set as perfectly as the original does the liver. On the lower left third molar in the d-column diagram of this model, it can be seen that the teeth segmented by TA-Net are enlarged at the boundary of the range, which is quite different from the original image. The DeeplabV3+ [33] model uses depth-separable convolution with dilated convolution though. But similarly to TA-Net in the d-column diagram the lower left third molar contour border is large and it is difficult to distinguish among the outer

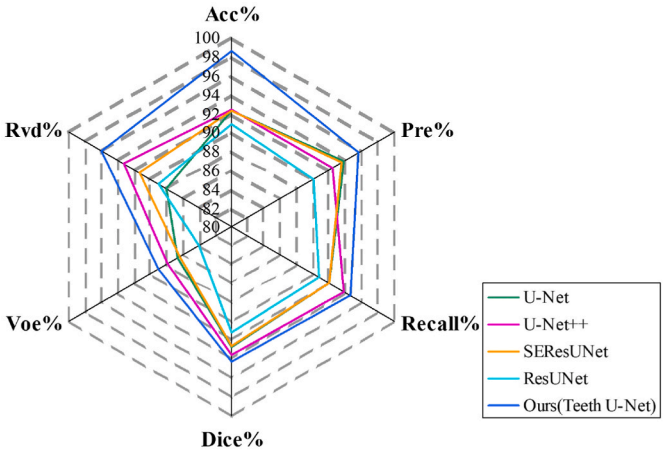

Fig. 8. Contextual semantic relevance experimental radar plot results.

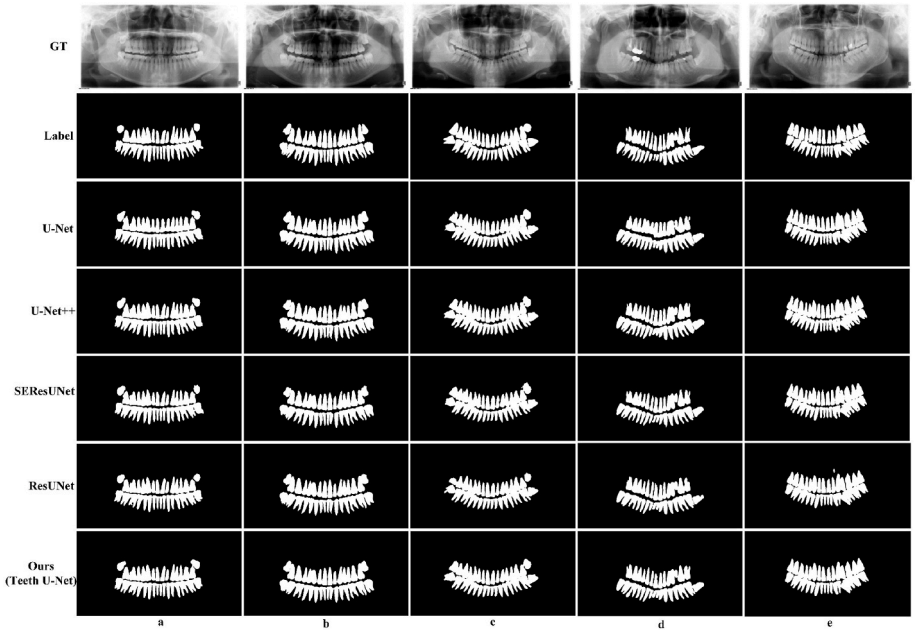

Fig. 9. Contextual semantic relevance experimental visualization results.

Table 6
Comparison of advanced networks.

Architecture	Acc%	Pre%	Recall%	Dice%	Voe%	Rvd%
Attention UNet [34]	95.08	90.68	90.97	92.22	85.72	94.52
DeeplabV3+ [33]	90.31	90.68	93.97	92.22	85.61	93.52
UTNet [30]	96.48	94.72	92.96	90.24	84.39	94.05
TranUNet [31]	97.96	91.78	91.60	91.60	85.54	95.28
TA-Net [32]	97.83	90.62	91.72	91.09	83.68	95.32
DANet [25]	85.66	83.93	89.89	84.36	76.61	82.12
Ours(Teeth U-Net)	**98.53**	**95.62**	**94.51**	**94.28**	**88.92**	**95.97**

borders of the teeth. The Attention UNet [34] architecture applies attention mechanisms to skip connections, allowing for better focus on salient regions and suppression of background regions. From the complex situation of the root boundary and teeth in the oral teeth, the segmentation result of Attention UNet is better. However, from a visualization point of view Attention UNet shows root splitting at the third molar on the lower right side. In contrast, the root of the third molar in the original image and label is in an inverted triangular shape. A visualization of the segmentation effect is shown in the table with the evaluation indices for a comparison of the different methods. It can be seen from the data that the proposed segmentation network model not only has a good segmentation effect, but also has higher indices than the other networks. The proposed network not only involves parallel hybrid attention in DA-Net, but also dilated convolution in conjunction with it. In addition, the dense skip connection to each layer of the SE module application also plays an effective role. It is proved that the proposed network is feasible to extract the tooth position information from dental X-ray images. The context information is effectively focused on the teeth in the image, and the segmentation visualization result is closer to the tag value. The proposed model can also be more intuitively perceived in the radar plot(Fig. 11), where it achieves the largest of all segmentation indices and accounts for a significantly larger proportion than the other networks.

4.3.3. Comparison of ablation experiments

To verify the validity of the model in this paper using each module, six groups of ablation experiments have been conducted on the same data set. The training set consists of 1300 X-ray images, and the validation set and test set consist of 150 X-ray images respectively. For Experiment 1, only the SE module is added to the model noted as Squeeze and Excitation U-Net# (SEUNet#) in order to distinguish it from the network that has emerged. In Experiment 2, only the dense skip connection UNet (DUNet) appears in the segmentation network model. Experiment 3, adding dense skip connections with SE modules to the network is called Squeeze-and-Excitation Dense UNet (SEDUNet). Experiment 4, in U-Net network, only the MAB module is marked as MUNet (MAB). In the fifth experiment, the multi-scale feature aggregation attention block and DHAB in the bottleneck layer are added to the U-Net network. Experiment 6, on the basis of experiment 5, a dense skip connection is added (DBUNet). The seventh experiment is the network proposed in this paper, called Teeth U-Net. Fig. 13 shows the visual segmentation results, Fig. 12 shows the comparison of radar charts of ablation experiments, Table 7 shows the comparison of evaluation indexes of various experimental data, and Fig. 14 shows the comparison of loss functions of ablation experiments.

As can be seen in Table 7, the ablation experiments of the proposed module have obtained good results in all evaluation index tables. The SEUNet# is very similar to the DUNet in comparison to all indices. In Acc, Dice, Recall and Voe DUNet is 0.1%, 0.44%, 0.96% and 0.78% higher than SEUNet#, while in Pre and Rvd, SEUNet# is 0.06% and 0.19% higher than DUNet. In the SEUNet# visualization Fig. 13, the lower left third molar in the a-column plot has a smaller range of segmented teeth, and in column b the upper right third molar shows multiple segmentation. In the network with dense skip connection, only the third molar on the upper right of column b is over segmented when two pieces of data are processed in the same place. However, the SEDUNet model when combining the two networks shows no defects in the a-column image and a smaller sporadic over-segmentation in the upper right third molar in the b-column image. The rest of the images are largely consistent with the labelled information, indicating that the combined use of the two models is indeed effective in segmenting the data set of teeth. There is over-segmentation on the right side of the upper right third molar in column b in MUNet model. But in BUNet model, the right side of the upper right third molar in the same column b is over-segmented. Other images are consistent with the original label. Pre and Voe are very close to the proposed network after adding dense connectivity to the bottleneck layer. The four metrics of Acc, recall, dice and RVD are 1.21%, 3.01%, 0.68% and 0.34% lower than the proposed

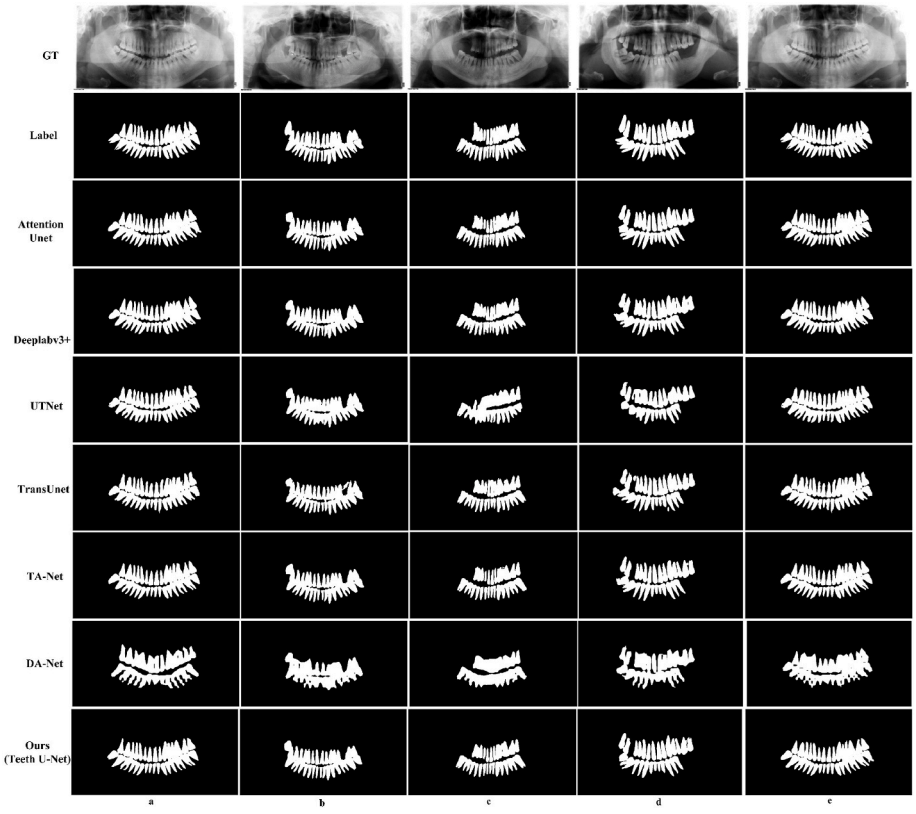

Fig. 10. Comparison results with advanced network visualization.

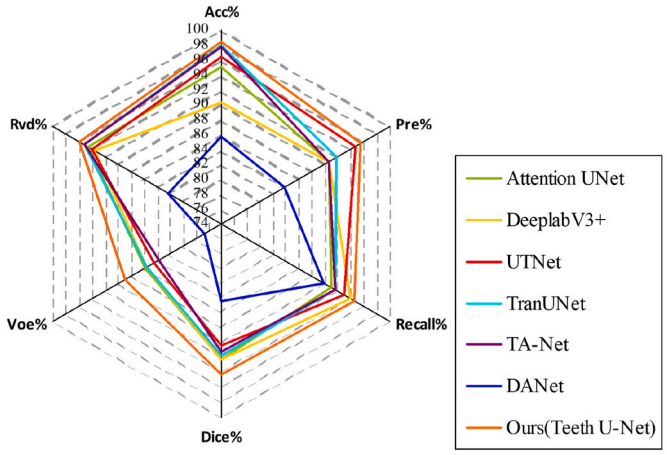

Fig. 11. Comparison results with advanced network radar maps.

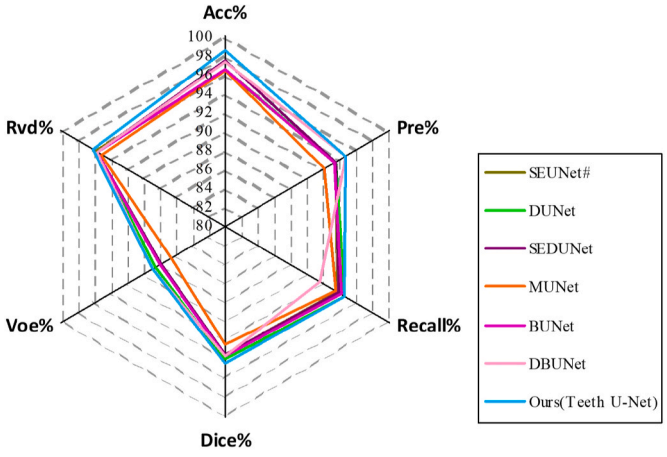

Fig. 12. Comparison results of radar maps of ablation experiments.

network respectively. The advantage of the dense skip connection in conjunction with the bottleneck layer module can also be seen in Fig. 13. And this model is very similar to the present network and label values, but the following problems still exist in the b-column images. For illustration, sporadic over-segmentation at the upper right third molar and under-segmentation at the left third molar. Nevertheless, the network as proposed in this paper presents very promising results for each of the evaluation indices in Table 7. Similarly the segmentation visualization results plotted in Fig. 13 are extremely close to the labelled values. The proposed dense skip connection, multi-scale aggregated attention blocks in the bottleneck layer and dilated hybrid self-attentive block are shown to play an invaluable role through six sets of ablation experiments. The advantages of the network proposed in this paper for segmentation of dental datasets are even more evident. Fig. 14 also shows that in the ablation experiment, the loss value of each module gradually decreases with the increase of epoch, and finally achieves the fitting. The dice coefficient value gradually increases with the increase of epoch, which also proves that the model proposed in this paper can achieve high accuracy regardless of the overall effect or the effect of individual design modules.

5. Conclusion and future work

In order to solve the problem of accurate segmentation of all teeth in the dental panoramic image and clear judgment of teeth root boundary, this paper proposes the Teeth U-net model for dental panoramic X-ray image segmentation. In the network, the dense skip connection between the encoder and the decoder is proposed to recover image features by

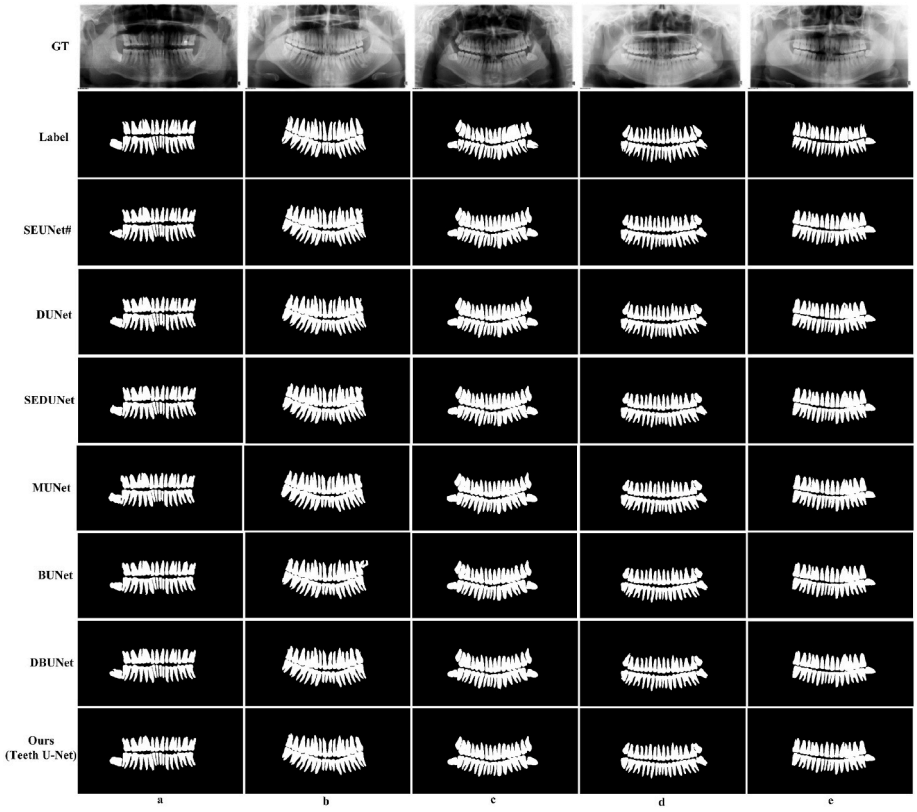

Fig. 13. Comparison of the results of the visualization of the ablation experiment.

Table 7
Comparison of ablation experiments.

Architecture	Acc%	Pre%	Recall%	Dice%	Voe%	Rvd%
SEUNet#	96.40	93.46	93.53	93.43	87.72	95.77
DUNet	96.50	93.40	94.49	93.87	88.50	95.58
SEDUNet	97.42	93.39	93.75	93.50	87.83	95.64
MUNet	96.38	92.09	93.43	92.36	86.58	95.03
BUNet	96.43	93.32	93.99	93.58	87.98	95.57
DBUNet	97.32	94.60	91.50	93.60	88.85	95.63
Ours(Teeth U-Net)	**98.53**	**94.62**	**94.51**	**94.28**	**88.92**	**95.97**

using multi-level connections. The MAB is designed in the bottleneck layer. At the same time, DHAB is proposed to enlarge the receptive field of the dental X-ray image and enhance the detailed feature information, so as to obtain richer semantic information, effectively solve the problem of teeth boundary ambiguity, and make the model as far as possible without loss of resolution. In order to verify the effectiveness of the Teeth U-Net model, the contextual semantic correlation experiments, the comparison with advanced segmentation networks and the ablation experiments are performed on the X-ray image dataset of clinical dental panoramic X-ray image. Through the visual results, the analysis of data tables and the image comparison of radar images, it can be proved that the indices Acc, Pre, Recall, Dice, Voe and Rvd of the Teeth U-Net model for dental panoramic segmentation are 98.53%, 95.62%, 94.51%, 94.28%, 88.92% and 95.97%, respectively. In order to verify the effectiveness of the internal modules of Teeth U-net model, this paper compares the Dice value of each module with the loss function. The results show that each module is complementary to each other in processing every detail of the dental X-ray image, which can effectively improve the efficiency of preoperative preparation and postoperative evaluation. To promote the application of dental teeth in medical image segmentation.

There are three main research directions for dental segmentation in the future: 1. There are low contrast and unnecessary interference information in the imaging methods of X-ray images. However, the improved structure of U-Net network generally uses single modality input, which often ignores the segmentation error of dental images caused by different imaging methods. In the future, the encoder in U-Net will be changed into the complementary mode of multiple modes, so that the model can converge more quickly and have higher accuracy. 2. This paper mainly studies the dental panoramic X-ray image segmentation. Labeling a large number of training data is a time-consuming and costly work, and the quality of the labeling is subjective and depends too much on the experience of the annotator. Compared with complex pixel level labels, weak supervised learning only needs weak label information such as point labels, boundary box labels or image level labels to obtain more accurate segmentation results. Therefore, subsequent research will attempt to apply weak segmentation to 2D and 3D segmentation of teeth images. 3. Most of the current segmentation algorithms only process two-dimensional images. With the continuous improvement of computer hardware performance and processing capacity, they can be extended to the research of segmentation algorithms for three-dimensional images. Because 3D images are not only closer to the actual situation, but also provide more abundant spatial information, it is expected to have a positive impact on the image segmentation effect.

CRediT authorship contribution statement

Senbao Hou: Conceptualization, Methodology, Investigation, Writing, Visualization, Supervision, Software, Results analysis. **Tao Zhou:** Writing-review & editing, Validation. **Yuncan Liu:** Data curation, Writing, Investigation. **Pei Dang:** Writing, Resources, Review. **Huiling Lu:** Investigation. **Hongbin Shi:** Data curation, Writing – review.

Declaration of competing interest

The authors declare that they have no known competing financial

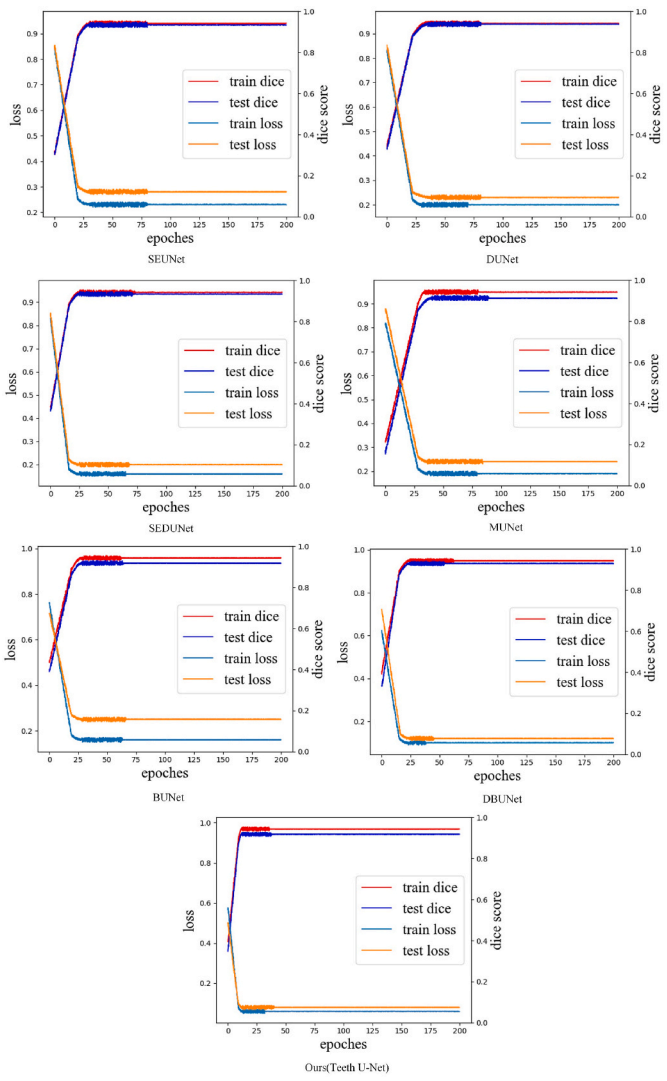

Fig. 14. Training loss, testing loss, training Dice and testing Dice change with epoch.

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant No. 62062003, Natural Science Foundation of Ningxia under Grant No. 2022AAC03149, North Minzu University Research Project of Talent Introduction under Grant No. 2020KYQD08.

References

[1] P.L. Lin, P.Y. Huang, P.W. Huang, H.C. Hsu, C.C. Chen, Teeth segmentation of dental periapical radiographs based on local singularity analysis, Comput. Methods Progr. Biomed. 113 (2) (2014) 433–445, https://doi.org/10.1016/j.cmpb.2013.10.015.

[2] P. K.J, K.C. Kwak, A trends analysis of dental image processing, in: 2019 17th International Conference on ICT and Knowledge Engineering, IEEE Press, 2019, pp. 1–5, https://doi.org/10.1109/ICTKE47035.2019.8966853.

[3] G. Silva, L. Oliveira, M. Pithon, Automatic segmenting teeth in X-ray images: trends, a novel data set, benchmarking and future perspectives, Expert Syst. Appl. 107 (2018) 15–31, https://doi.org/10.1016/j.eswa.2018.04.001.

[4] B.Y. Tekin, C. Ozcan, A. Pekince, Y. Yasa, An enhanced tooth segmentation and numbering according to FDI notation in bitewing radiographs, Comput. Biol. Med. 146 (2022), 105547, https://doi.org/10.1016/j.compbiomed.2022.105547.

[5] J. Yang, Y. Xie, L. Liu, B. Xia, Z. Cao, C. Guo, Automated dental image analysis by deep learning on small dataset, in: 2018 IEEE 42nd Annual Computer Software and Applications Conference, vol. 1, 2018, pp. 492–497, https://doi.org/10.1109/COMPSAC.2018.00076. IEEE(COMPSAC), IEEE.

[6] Z. Xia, Y. Gan, L. Chang, Individual tooth segmentation from CT images scanned with contacts of maxillary and mandible teeth, Comput. Methods Progr. Biomed. 138 (2017) 1–12, https://doi.org/10.1016/j.cmpb.2016.10.002.

[7] T.L. Koch, M. Perslev, C. Igel, S.S. Brandt, Accurate segmentation of dental panoramic radiographs with u-net, in: IEEE 16th International Symposium on Biomedical Imaging, 2019, pp. 15–19, https://doi.org/10.1109/ISBI.2019.8759563.

[8] Z. Kong, F. Xiong, C. Zhang, Z. Fu, Automated maxillofacial segmentation in panoramic dental X-ray images using an efficient encoder-decoder network, IEEE Access 8 (2020) 207822–207833, https://doi.org/10.1109/ACCESS.2020.3037677.

[9] Y. Zhao, P. Li, C. Gao, Y. Liu, TSASNet: tooth segmentation on dental panoramic X-ray images by two-stage attention segmentation network, Knowl. Base Syst. 206 (2020), 106338, https://doi.org/10.1016/j.knosys.2020.106338.

[10] Z. Cui, C. Li, N. Chen, G. Wei, R. Chen, Y. Zhou, D. Shen, W. Wang, TSegNet: an efficient and accurate tooth segmentation network on 3D dental model, Med. Image Anal. 69 (2021), 101949, https://doi.org/10.1016/j.media.2020.101949.

[11] T. Feng, C. Wang, X. Chen, H. Fan, K. Zeng, Z. Li, URNet: a u-net based residual network for image dehazing, Appl. Soft Comput. 102 (2021), 106884, https://doi.org/10.1016/j.asoc.2020.106884.

[12] S. Liu, Y. Li, J. Zhou, J. Hu, N. Chen, Y. Shang, Z. Chen, T. Li, Segmenting nailfold capillaries using an improved U-net network, Microvasc. Res. 130 (2020), 104011, https://doi.org/10.1016/j.mvr.2020.104011.

[13] Z. Wang, Y. Zou, P.X. Liu, Hybrid dilation and attention residual U-net for medical image segmentation, Comput. Biol. Med. 134 (12) (2021), 104449, https://doi.org/10.1016/j.compbiomed.2021.104449.

[14] M. Alom, C. Yakopcic, T.M. Taha, V.K. Asari, Nuclei segmentation with recurrent residual convolutional neural networks based u-net (R2U-Net), in: NAECON 2018 IEEE National Aerospace and Electronics Conference, 2018, pp. 228–233, https://doi.org/10.1109/NAECON.2018.8556686.

[15] Q. Jin, Z. Meng, C. Sun, H. Cui, R. Su, Ra-Unet, A hybrid deep attention-aware network to extract liver and tumor in CT scans, Front. Bioeng. Biotechnol. 8 (2020), 605132, https://doi.org/10.3389/fbioe.2020.605132.

[16] L. Liu, L. Kurgan, F. Wu, J. Wang, Attention convolutional neural network for accurate segmentation and quantification of lesions in ischemic stroke disease, Med. Image Anal. 65 (2020), 101791, https://doi.org/10.1016/j.media.2020.101791.

[17] Y.L. Wang, Z.J. Zhao, S.Y. Hu, F.L. Chang, CLCU-Net: cross-level connected U-shaped network with selective feature aggregation attention module for brain tumor segmentation, Comput. Methods Progr. Biomed. 207 (2021), 106154, https://doi.org/10.1016/j.cmpb.2021.106154.

[18] J. Dolz, C. Desrosiers, I.B. Ayed, IVD-net: intervertebral disc localization and segmentation in MRI with a multi-modal UNet, Lect. Notes Comput. Sci. 11397 (2019) 130–143, https://doi.org/10.1007/978-3-030-13736-6_11.

[19] J. Zhang, Y. Jin, J. Xu, X. Xu, Y. Zhang. MDU-net: Multi-Scale Densely Connected U-Net for Biomedical Image Segmentation, arXiv preprint arXiv: 1812.00352vol. 2.

[20] K. Eric, C. Chen, M.M. Hassan, A. Almogren, A deep learning based medical image segmentation technique in Internet-of-Medical-Things domain, Future Generat. Comput. Syst. 108 (2020) 135–144, https://doi.org/10.1016/j.future.2020.02.054.

[21] Y. Mohammad, J. Philippe, C. Farida, Dense-Unet: a light model for lung fields segmentation in Chest X-Ray images, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC), 2020, pp. 1242–1245, https://doi.org/10.1109/EMBC44109.2020.9176033.

[22] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional Networks for Biomedical Image Segmentation, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28. Medical Image Computing and Computer-assisted Intervention.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid Scene Parsing Network, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 6230–6239, https://doi.org/10.1109/CVPR.2017.660.

[24] X. Wang, Z. Li, Y. Huang, Y. Jiao, Multimodal medical image segmentation using multi-scale context-aware network, Neurocomputing 486 (2022) 135–146, https://doi.org/10.1016/j.neucom.2021.11.017.

[25] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3141–3149, https://doi.org/10.1109/CVPR.2019.00326.

[26] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, in: 2018 IEEE, CVF Conference on Computer Vision and Pattern Recognition, 7132-7141, doi: 10.1109/TPAMI.2019.2913372.

[27] Z. Cao, B. Yu, B. Lei, H. Ying, X. Zhang, D.Z. Chen, J. Wu, Cascaded SE-ResUnet for segmentation of thoracic organs at risk, Neurocomputing 453 (2021) 357–368, https://doi.org/10.1016/j.neucom.2020.08.086.

[28] Z. Zhou, M. Siddiquee, N. Tajbakhsh, UNet++: A Nested U-Net Architecture for Medicalimage Segmentation, 4th Deep Learningin Medical Image Analysis Workshop, 2018, pp. 3–11, https://doi.org/10.1007/978-3-030-00889-5_1.

[29] J. Liu, Y. Kang, J. Qiang, Y. Wang, D. Hu, Y. Chen, Low-dose CT imaging via cascaded ResUnet with spectrum loss, Methods 202 (2021) 10, https://doi.org/10.1016/j.ymeth.2021.05.005.

[30] Y. Gao, M. Zhou, D. Metaxas, UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation, arXiv preprint arXiv: 2107.00781vol. 2.

[31] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, Trans UNet: transformers make strong encoders for medical image segmentation, arXiv preprint arXiv: 2102.04306.

[32] S. Pang, A. Du, M.A. Orgun, Y. Wang, Z. Yu, Tumor attention networks: better feature selection, better tumor segmentation, Neural Network. 140 (2021) 203–222, https://doi.org/10.1016/j.neunet.2021.03.006.
[33] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deep Lab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848, https://doi.org/10.1109/TPAMI.2017.2699184.
[34] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention U-Net: Learning where to Look for the Pancreas, arXiv preprint arXiv: 1804.03999.