IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 29, 2020, accepted November 8, 2020, date of publication November 16, 2020,
date of current version November 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037677

Automated Maxillofacial Segmentation in Panoramic Dental X-Ray Images Using an Efficient Encoder-Decoder Network

ZHENGMIN KONG[1], (Member, IEEE), FENG XIONG[1], CHENGGANG ZHANG[1], ZHUOLIN FU[1], MAOQI ZHANG[2,3,4], JINGXIN WENG[2,3,4], AND MINGZHE FAN[2,3,4]

[1]School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China
[2]Department of Periodontology, School of Stomatology, Wuhan University, Wuhan 430072, China
[3]State Key Laboratory Breeding Base of Basic Science of Stomatology (Hubei-MOST), Wuhan University, Wuhan 430072, China
[4]Key Laboratory for Oral Biomedicine of Ministry of Education (KLOBM), School and Hospital of Stomatology, Wuhan University, Wuhan 430072, China

Corresponding author: Feng Xiong (xiongfeng@whu.edu.cn)

ABSTRACT The panoramic dental X-ray images are an essential diagnostic tool used by dentists to detect the symptoms in an early stage and develop appropriate treatment plans. In recent years, deep learning methods have been applied to achieve tooth segmentation of dental X-rays, which aims to assist dentists in making clinical decisions. Because the original images contain plenty of useless information, it is necessary to extract the region-of-interest (ROI) to obtain more accurate results by focusing on the maxillofacial region. However, a fast and accurate maxillofacial segmentation without hand-crafted features is challenging due to the poor image quality. In this study, we create a large maxillofacial dataset and propose an efficient encoder-decoder network model named EED-Net to solve this problem. This dataset consists of 2602 panoramic dental X-ray images and corresponding segmentation masks annotated by the trained experts. Based on the original structure of U-Net, our model structure contains three major modules: a feature encoder, a corresponding decoder, and a multipath feature extractor that connects the encoding path and the decoding path. In order to obtain more semantic features from the depth and breadth, we replace the convolution layer with the residual block in the encoder and adopt Inception-ResNet block in the multipath feature extractor. Inspired by the skip connection in FCN-8s, the lightweight decoder has the same channel dimension as the number of segmented objects. Besides, a weighted loss function is used to enhance segmentation accuracy. The comprehensive experimental results on the new dataset demonstrate that our model achieves better accuracy and speed trade-offs for maxillofacial segmentation than the latest methods.

INDEX TERMS Maxillofacial segmentation, panoramic radiographs, deep learning, encoder-decoder network.

I. INTRODUCTION

Dental radiographs are widely used in dentistry for clinical diagnosis, treatment, and surgery. These images enable dentists to find hidden dental structures, bone loss, and cavities [1], which are hard or impossible to detect in visual inspection. Hence, dentists can detect symptoms at an early stage and develop appropriate treatment plans [2]. In the dental examination, there are three types of conventional radiographs: bitewing, periapical, and panoramic [3]. The bitewing X-rays show portions of the upper and lower teeth to detect changes in bone thickness caused by gum disease. The periapical X-rays show the whole teeth in part of the upper jaw or lower jaw to evaluate the root area and surrounding bone structures. While the bitewing and periapical X-rays focus on the details of individual teeth or parts of teeth, the panoramic X-rays capture the entire mouth area, including all the teeth, gums, jaws, and bone structure [4] to provide more diagnostic evidence. Since the panoramic X-rays are filmed outside the mouth, they have better patient acceptance with faster shooting, less radiation exposure, and lower infection rate [5]. During diagnosis and treatment, such as root canal treatment, caries diagnosis and tooth orthodontics, dentists are required to analyze panoramic radiographs and record

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

specific symptoms of diseased teeth in the electronic medical record. However, it takes a long training time for a young doctor to read dental films accurately [6]. Therefore, considerable attention has been paid to automatic panoramic X-rays analysis.

Many studies have been carried out to explore tooth segmentation since the diagnosis is based on the analysis of teeth and surrounding tissues. Traditional image segmentation methods [7]–[10] can quickly extract dental information from a single radiograph type. However, these methods rely on well-designed manual features and lack sufficient generalization capability. The emergence of deep learning methods [11], [12] dramatically improves the accuracy of tooth segmentation under different tooth distributions. The image segmentation models based on deep learning mostly adopt an end-to-end learning framework [13], where both training and inference are performed by dense feedforward computation and backpropagation to learn the entire image at once. The training process is designed to optimize the overall parameters according to the error between predicted results and true labels. As the deep learning models need to focus more on teeth and surrounding areas to enhance pixel recognition, their typical inputs are often the region-of-interest (ROI) derived from the original image [14], [15]. In panoramic X-rays, the ROI is the maxillofacial region [16], [17], which contains all the teeth and discards most of the irrelevant information. Therefore, real-time extraction of the maxillofacial region is an essential preprocessing step in the panoramic dental X-ray analysis.

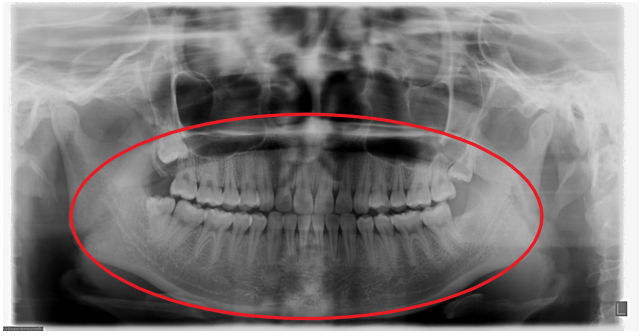


FIGURE 1. An example of the panoramic dental X-rays. The inner area of the red circle is the approximate outline of maxillofacial region that contains all the teeth. The pixels and illumination of the image are not uniformly distributed, and the high-noise outer area accounts for more than half of the pixels. Due to low-contrast and overlapping anatomical structures, the maxillofacial region has no strong edges.

However, an accurate and rapid maxillofacial segmentation is challenging due to the poor image quality of panoramic dental X-rays. As shown in Fig. 1, the low quality of the image comes from three aspects [18]: 1) unavoidable noise, 2) varying illumination, and 3) low pixel contrast of the tissue. A major dilemma is that panoramic radiographs show overlapping anatomical structures [19]. The overlap among teeth, jaws and surrounding bones then causes complex variations in grayscale levels of panoramic X-rays. In this case, static and dynamic methods have been developed to extract the ROI from panoramic X-rays. The static method defines a fixed rectangular window for all images. Through a statistical analysis of the maxillofacial morphology [20], [21], the anchor point (often the image center) and the rectangle size are determined, and this window is cropped from the image as the maxillofacial region. This method is theoretically effective for samples with similar tissue morphology. Nevertheless, the shape and position of the maxillofacial region are considerably variable in practice, which limits the application of the static window. The dynamic method employs manual threshold selection [22], morphological transformation [23], or wavelet variation [24] to extract the separation line of the upper and lower jaws for correcting the anchor point. Then, the modified anchor point and rectangular, trapezoidal [25] or oval window [26] are combined to capture the maxillofacial region. Although this method improves the segmentation accuracy and has higher flexibility to some extent, it requires a fine selection of pixel thresholds and fails to accurately locate the separation line when teeth are missing or overlapping. To circumvent the limitations of the above methods, an approach with greater feature extraction and generalization capability is needed. Fortunately, the deep learning approach [27]–[30] is competent for this task because of its remarkable success in medical image segmentation [31], [32]. However, the performance gain of this approach usually comes at the cost of high computation [33] and long processing time. Therefore, ensuring both the accuracy and the real-time performance of maxillofacial segmentation still remains a significant challenge.

In this study, we propose an efficient encoder and decoder network named EED-Net to address automatic maxillofacial segmentation using a newly built dataset. Since there is no open maxillofacial dataset in relevant research fields, we elaborately collect 2602 panoramic dental X-rays, excluding the hypoplasia and tooth decay. A group of dental experts are trained to implement the maxillofacial annotations. On this basis, we evaluate the performance of FCN-8s [34] and U-Net [35], which are commonly used as the baseline models in image segmentation. Despite that both models have their own advantages in aspect of speed and accuracy, they cannot balance these two performances. Against this background, we combine the structure of U-Net with the decoding method of FCN-8s to design a new maxillofacial segmentation model. The proposed model is mainly composed of three modules: a feature encoder, a corresponding decoder, and a multipath feature extractor that connects the encoder and the decoder in the last layer. Specifically, the common convolution connection in the encoder is replaced with a residual structure [36] to obtain deeper features. Besides, we construct a modified Inception-ResNet block [37] in the multipath feature extractor to increase the width of the feature search. The skip connection between the encoder-decoder pairs is reserved for passing the low-resolution texture and location information. To reduce parameters, we employ the category number of segmented objects to identify the decoder's channel dimension,

which is far less than the encoder. The extensive experiment results on the maxillofacial dataset demonstrate that our proposed model achieves better overall performance than the latest real-time semantic segmentation models, which verifies the efficiency of EED-Net.

The remainder of this paper is organized as follows. Section II introduces the related works on the maxillofacial segmentation methods and some latest deep learning semantic segmentation models. Section III describes the new dataset and the proposed model in detail. Section IV presents the experimental results. In Section V, our conclusion and discussion are provided.

II. RELATED WORKS

In the literature of maxillofacial segmentation methods for panoramic dental X-ray images, two strategies have been developed: 1) static method: observing the distribution of teeth and surrounding tissues in the image samples and artificially selecting a static window to crop the maxillofacial region; 2) dynamic method: enhancing the contrast of the original image, extracting the separation line of upper and lower jaws, and correcting the position and shape of the dynamic window.

To implement the first strategy, Oliveira and Proença [20] defined a rough maxillofacial region that took out the nasal and chin bones as the first stage of teeth segmentation. They measured the boundary distance of the maxillofacial region and fitted each distance variable to a corresponding Gaussian distribution. The shortest distance with 95% certainty was adopted as the window size to extract the ROI. According to the statistical analysis of image morphology, Fares and Feghali [21] positioned the anchor point at the maxillofacial center to reduce the distance variables.

To improve the adaptability and accuracy of segmentation algorithms, the dynamic methods were proposed. Jain and Chen [22] used transverse scanning to find the lower pixel intensity in the y-axis projection histogram. Combined with the Gaussian distribution assumption and a user-assisted initialization, this method found the gap valley between the upper and lower teeth. By locating these gaps, they divided the maxillofacial region into vertical strips for subsequent tooth segmentation. Wanat and Frejlichowski [26] utilized the areas between necks of teeth to determine the separation line, which did not depend on the gap between adjacent teeth. Harandi *et al.* [23] applied morphology transform and modified geodesic active contour on panoramic X-rays to achieve better separation and subtraction. Gumus [24] employed the discrete wavelet transform for better location of the ROI and adopted polynomial regression to form a smooth separation line with absent teeth.

The above-mentioned dynamic methods aim to separate the upper and lower jaws, whereas the deep learning method requires a complete maxillofacial image to achieve the training and inference of end-to-end models. Fares and Feghali [21] searched for the occlusal gap to locate the maxillofacial center and extracted the entire ROI with a fixed rectangle. However, their method cannot find the accurate maxillofacial region when the upper and lower jaws are connected. Hasan *et al.* [25] combined the gradient vector flow(GVF) snakes with K-means clustering to accomplish maxillofacial segmentation with missing gaps between upper and lower jaws. Although this method has been proved to work with more types of panoramic radiographs, it is highly dependent on manual threshold selection to extract the maxillofacial edges.

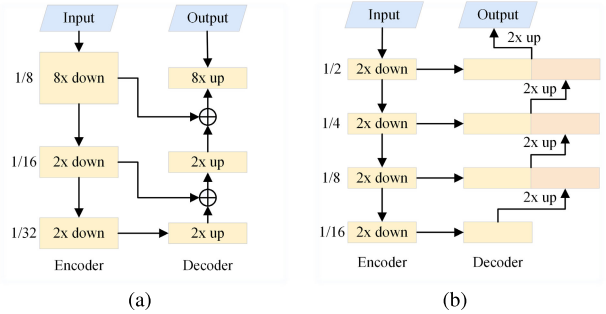


FIGURE 2. The network architectures of (a) FCN-8s and (b) U-Net. For both models, 2x down module downsamples the input feature map to half the size, 2x up module upsamples the input feature map to twice the size, and 1/8 denotes that the feature map size of the module is 1/8 of the input image size. Low-resolution and high-resolution features are fused by addition operations in FCN-8s and concatenation operations in U-Net.

In recent years, significant progress has been made to improve the performance of low-contrast medical image segmentation using deep learning algorithms. Among these works, FCN [34] and U-Net [35] are the popular classical models. FCN lays the foundation for most modern segmentation architectures. As shown in Fig. 2(a), the removal of the full connected layer allows the model to predict the dense output from an arbitrary-sized image theoretically. Specifically, VGG [38] is employed as the encoder to extract high-resolution semantic features. For the decoder, the transposed convolution and skip connection are integrated to restore the object spatial. In this context, three versions of the model: FCN-32s, FCN-16s, and FCN-8s are presented with different depths of the connection architecture. Since the frequent use of pooling layers in FCN results in a loss of low-resolution features, U-Net builds dense skip connections to further explore the recovery of image details. Fig. 2(b) shows that U-Net has a symmetrical encoder-decoder structure. Due to the concentration operation for multi-scale feature integration, low-resolution context and high-resolution features can be fused without loss. As a result, the abundant feature information effectively improves model performance at the cost of increased computation.

More recently, many lightweight models have been proposed for real-time semantic segmentation. RefineNet [39] constructed an encoder-decoder network in the full residual form [36] and leveraged multiple-level abstract features to perform high-resolution semantic segmentation. The multi-branch systems ICNet [40] and BiSeNet [41] learned a global context with reduced-resolution input in a deep

branch, while boundaries were learned in a shallow branch at full resolution. Fast-SCNN [42] merged the two-branch pattern with the encoder-decoder framework and introduced a novel initial layer to share the computations of both branches. LEDNet [43] proposed an asymmetric encoder-encoder architecture with channel split and attention pyramid to lighten the model complexity.

However, real-time maxillofacial segmentation with high accuracy remains challenging. Inspired by the above works, the proposed EED-Net incorporates the multipath feature extractor and the simplified decoder in the residual encoder-decoder framework to guarantee both speed and accuracy of maxillofacial segmentation.

III. METHODOLOGY

In this section, we describe our research methodology in detail from the following aspects: the establishment of the dataset, the structure of EED-Net, and the weighted loss function.

A. DATASET

1) THE SOURCE AND ATTRIBUTES OF THE IMAGES

In the early stage of our research, we found it difficult to find a maxillofacial segmentation dataset in the community. The existing open dental datasets contain no maxillofacial results, and their quantity is not enough to support the training of deep learning models. Therefore, we cooperate with the experts from the Hospital of Stomatology Wuhan University to build a maxillofacial dataset. To gather data quickly, we choose to screen panoramic dental X-rays from electronic medical records rather than waiting for new patients. It is worth noting that these images do not contain any personal privacy, so there is no ethical issue involved. Considering the maxillofacial growth and deformity caused by age, the experts remove the samples that are too young or too old. Moreover, every image is labeled ''non-periodontal disease'' or ''periodontal disease'' for the follow-up research. In total, our dataset is composed of 2602 panoramic X-rays, including 1146 ''non-periodontal disease'' samples and 1456 ''periodontal disease'' samples.

2) THE ANNOTATION PRINCIPLES OF THE MAXILLOFACIAL REGION

In medicine, the maxillofacial region is located at the beginning of the digestive tract and respiratory tract to communicate with the outside world through the oral cavity and nasal cavity. This description aims to delineate the specific functional area, but our research is designed for maxillofacial segmentation. Therefore, a clear maxillofacial definition is required to accommodate the panoramic X-rays. After analyzing the correlation between teeth and maxillofacial region in different types of the images, we develop the following three principles for maxillofacial annotation: a) the segmented region should contain all the teeth; b) there should be a significant difference between the segmented region and the unsegmented region; c) The boundary of the segmented region should be as smooth as possible. According to these rules, the new maxillofacial region is precisely outlined between the nose and chin, forming a smooth surface with the left and right jaws.

3) THE ANNOTATIONS OF THE PANORAMIC DENTAL X-RAYS

The maxillofacial annotations are achieved by anchoring and connecting the points in the images. In general, 30-50 points are required in an image to outline the target region, which makes the annotations labor-intensive and time-consuming. To improve annotation efficiency, we test some common annotation methods and finally choose VGG Image Annotator (VIA) [44]. This tool is easy-to-use and compatible with different operating systems, which is more suitable for collaborative annotations than other tools. Based on the proposed segmentation principles, our experts are trained to outline the maxillofacial region in the image using VIA, where the annotation files are obtained in JSON form. By converting the files into images in python, the maxillofacial masks are generated in grayscale. The segmentation results are verified by our experts and unqualified images are relabeled. Finally, we establish a dataset of 2602 panoramic X-rays and corresponding maxillofacial masks for deep learning models. Fig. 3 shows the typical images and their masks.

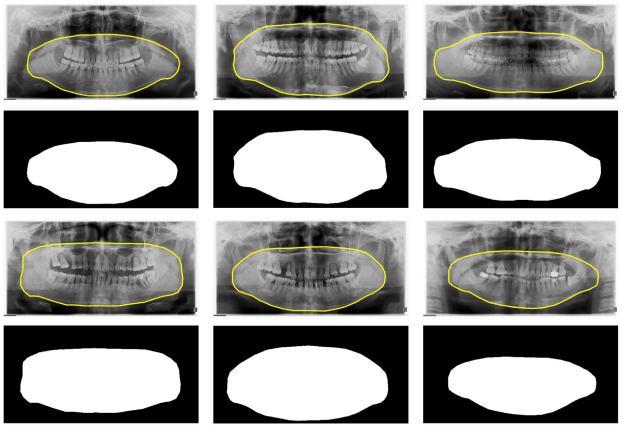


FIGURE 3. The panoramic X-rays in the first row are ''non-periodontal disease'', and the images in the third row are ''periodontal disease''. The yellow curves in these images are the maxillofacial contour outlined by the experts. The other rows are the corresponding masks generated by the annotations.

TABLE 1. The sample composition of different categories in the dataset.

Type	Non-periodontal	Periodontal	Total images
Training	916	1166	2082
Test	230	290	520
Total	1146	1456	2602

For the performance assessment of the models, we divide the whole dataset into a training set and a test set with a ratio of 4:1. Table 1 describes the distribution of different

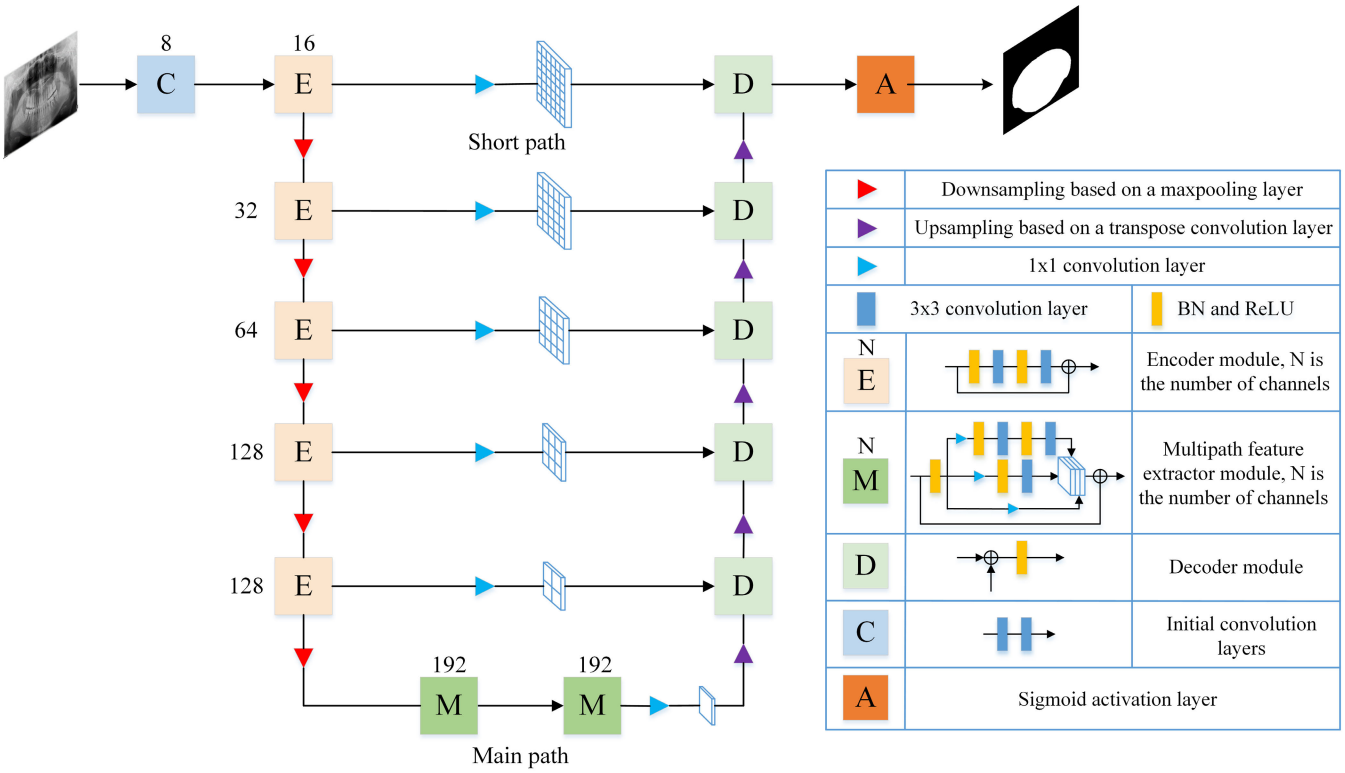


FIGURE 4. The structure of EED-Net. In addition to the initial convolution layers and the last sigmoid activation layer, the model is mainly composed of three parts: 1) the residual encoder modules, 2) the multipath feature extractors based on the simplified Inception-ResNet block, 3) the object-oriented decoder modules. The main path realizes the breadth extraction of the encoder output features, and passes high-resolution features to the decoder after channel compression. In order to recover image details accurately, the dense short paths are utilized to achieve the gradient flow in shallow layers. Since there is only one target type in maxillofacial segmentation, the corresponding decoder has a single feature channel, converted from a 1 × 1 convolution layer.

categories of images in the dateset. In order to maintain category balance, both types of samples are randomly assigned to the homologous datasets in the same ratio of 4:1. Since the original panoramic X-rays have a resolution of 2976 × 1536 pixels, our GPU will run out of memory during the calculation of such a large image. Therefore, we resize them to a resolution of 744 × 384 (which is a quarter of the original size) by the nearest interpolation. This dataset serves the purpose of supporting the verification and optimization of maxillofacial segmentation models.

B. EED-NET

Before developing our model, we evaluated the common segmentation models, FCN-8s and U-Net, on our maxillofacial dataset. The comparison results indicate that U-Net has higher accuracy, while FCN-8s uses less computing time with more parameters. It is worthwhile to design a new network model to inherit the advantages of these two baseline models. In this case, both accuracy and efficiency can be guaranteed.

Based on the structural analysis of existing models, we propose a novel efficient encoder-decoder network named EED-Net for maxillofacial segmentation. As illustrated in Fig. 4, the main architecture of EED-Net consists of three components: an encoder module, a decoder module, and a multipath feature extractor module. Before entering the serial encoder modules, a 3 × 3 convolution layer is used for the preliminary feature extraction and channel transformation of the input image. The encoder module adopts the residual structure to obtain effective high-resolution features and speed up the model calculation. In the encoding path, its channel increases by multiples to compensate for the information loss caused by maxpooling layers. Then, two concatenated multipath feature extractor modules (which we call the main path) are utilized to connect the encoding path and the decoding path. Since high-resolution features have a decisive influence on the results, the multiple feature extractor is constructed on the simplified Inception-ResNet structure. Due to the combination of receptive fields with different widths, the main path can obtain multi-dimensional deep features to enhance the discriminative capacity of the segmentation target. Next, our decoder module realizes the addition of input feature maps after the channel transformation using a 1 × 1 convolution layer. Most importantly, the number of segmented objects, not the encoder, determines the channel number in the decoders. Since each panoramic X-ray has only one maxillofacial region, the object-oriented decoders dramatically reduce the model complexity. Moreover, skip connections (which we call the short path) between

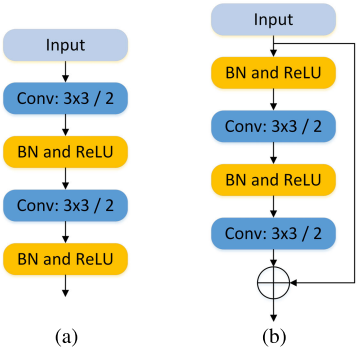


FIGURE 5. The structure of the encoder in (a) U-Net and (b) EED-Net. The full pre-activation residual structure is adopted in EED-Net to make training easier and improve generalization.

the encoder-decoder pairs are reserved to achieve the gradient flow in shallow layers, allowing the training of a very deep network. Finally, the segmentation results are obtained through the last sigmoid activation layer.

1) FEATURE ENCODER MODULE

The primary function of the encoder is to gradually reduce the spatial dimension of feature maps and capture more high-level semantic features. Since the proposed model is deeper than U-Net, we replace the ordinary convolution block with the residual block to avoid gradient degradation. Fig. 5 indicates the encoder structures of these two models. Both encoders are composed of two 3×3 convolution layers and activation layers. The activation layer contains a batch normalization layer [45] and a rectified linear unit (BN and ReLU). Compared with U-Net, our residual encoder has a shortcut connection and an addition operation that enable the model to learn residual features rather than entire features, which is easier to optimize. By learning existing various residual structures, we choose the pre-activation residual block [46] with activation layers in front of the weight layers to accelerate the training and improve the generalization ability. Besides, the convolution layers within the encoder have the same channel number. Since there is a maxpooling layer between adjacent encoders to reduce the feature map, the channel number of the next encoder is twice that of the previous encoder to compensate for information loss.

2) MULTIPATH FEATURE EXTRACTOR

After passing through five encoders, the input feature map of the main path is $1/32$ of the original image size. However, the high-resolution information plays a decisive role in the restoration of the segmented region. Since the excessive increase of convolution kernel and channel number will bring too many parameters, we construct a modified Inception-ResNet block to obtain deep features under multiple receptive fields. As shown in Fig. 6, the extractor has three separate paths, whose convolution kernels are 1×1, 3×3, and 5×5. The 5×5 path is composed of two 3×3 convolution layers. The 1×1 convolution layers are used to adjust the

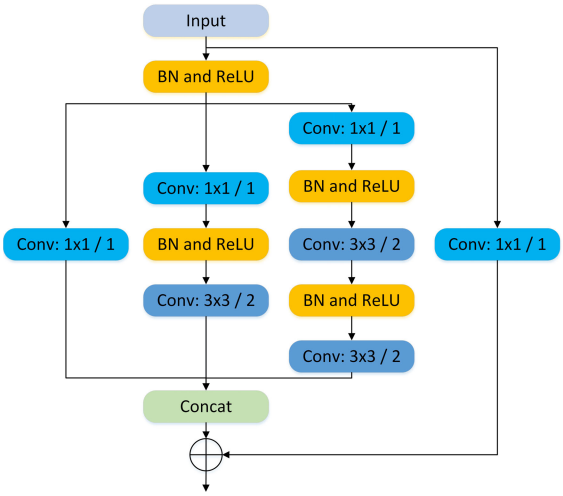


FIGURE 6. The structure of the multipath feature extractor.

input dimension to reduce the extra computation. In addition, the original Inception-ResNet structure is converted to the pre-activation form to keep consistent with the encoder. Then we use a concatenation operation to combine the multipath outputs with the same dimensions. Meanwhile, the feature concatenation causes no information loss. Finally, we use two 192-channel modules instead of a 384-channel module to build the main path.

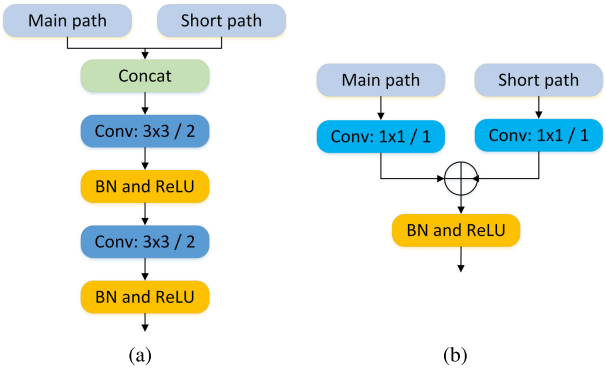


FIGURE 7. The structure of the feature decoder module in (a) U-Net and (b) EED-Net.

3) FEATURE DECODER MODULE

In the maxillofacial segmentation experiments, we find that U-Net is more accurate and FCN-8s is faster with some performance loss. Based on the analysis of two models, it can be inferred that the symmetric decoders of U-Net bring plenty of information redundancy, which slows down the calculation. To reduce the complexity, we refer to FCN-8s and adopt 1×1 convolution layers to recover the object details and spatial dimensions before information fusion. In Fig. 7, our decoder has no common weighted layers, and an addition operation completes the feature combination of the two paths. Unlike the single-channel decoder in FCN-8s, the channel

number in our decoder is determined by the number of the segmented objects to enhance the segmentation quality. Since each panoramic dental X-ray has only one target maxillofacial region, the object-oriented decoder is much lighter than U-Net. The activation layer is placed at the end of the decoder to ensure the pre-activated form.

C. WEIGHTED LOSS FUNCTION

As an end-to-end model, EED-Net is trained to predict whether each pixel belongs to the foreground or the background to extract the target region. Before the training, it is necessary to specify the task type and expected performance for choosing the loss function. Since regression functions are mainly used to predict dynamic variables, maxillofacial segmentation should be more a pixel classification problem than a regression problem. In general, the entropy loss function [47] is applied to train the multiple classification models by calculating the loss value of each pixel type. Considering that there are only two types of pixels in our task, we adopt the binary entropy loss function to estimate each pixel category. Equation (1) describes the loss calculation of a single image:

$$L_{bce}(y, p) = -\frac{1}{n}\sum_{i}^{N}\left(y_i \log p_i + (1 - y_i)\log(1 - p_i)\right) \quad (1)$$

where y is the ground truth of the input image and $y_i \in \{0, 1\}$, p is the predicted values of the pixel classification and $p_i \in [0, 1]$, and N denotes the total number of pixels. The final loss value L_{bce} is obtained by calculating the average loss of all pixels.

Actually, maxillofacial segmentation is not only a pixel classification problem, but also a region segmentation problem, that takes a connected whole in the image as the target area. Moreover, the small difference in pixel intensity between the bone and tooth structures on panoramic radiographs increases the difficulty in recognizing the maxillofacial edges. Considering the boundary connectivity, we adopt the dice coefficient loss [48] to focus on the overlap between the ground truth and the predicted maps. In image segmentation, the ground truth and prediction can be viewed as two sets, and the dice coefficient is used to gauge their similarity. In this manner, the predicted map approaches the ground truth by minimizing the dice coefficient loss in (2).

$$L_{dice}(y, p) = 1 - \frac{2\omega\sum_{i}^{N} y_i p_i}{\sum_{i}^{N} y_i^2 + \sum_{i}^{N} p_i^2} \quad (2)$$

where $\sum_{i}^{N} y_i p_i$ denotes the intersection of the ground truth y and the predicted results p, $\sum_{i}^{N} y_i^2$ and $\sum_{i}^{N} p_i^2$ are the quadratic sums of the pixel category of two samples. We set $\omega = 1$ in this study, and L_{dice} will come to zero when the predicted result is the same as the ground truth.

To balance the performance of EED-Net in pixel prediction and overall segmentation, we adopt the weighted loss function composed of binary entropy loss and dice coefficient loss. The final loss function is defined as

$$L_{loss}(y, p) = (1 - \alpha)L_{bce}(y, p) + \alpha L_{dice}(y, p) \quad (3)$$

where $\alpha \in [0, 1]$ is the weighted coefficient. This weighted loss function can be transformed into a single loss function by setting $\alpha = 0$ or $\alpha = 1$. The integrality of segmentation is emphasized with a large α. In our experiments, we set $\alpha = 0.8$ to get better performance.

IV. EXPERIMENTS AND RESULTS

In this section, the extensive experiments are conducted to validate the effectiveness of our approach. Firstly, the experimental environments and the implementation details of model training are introduced. Then, the evaluation metrics are presented to quantify the performance of the models in various aspects. Finally, the proposed model is optimized from the structure and loss function, and compared with other methods.

A. EXPERIMENT ENVIRONMENTS AND IMPLEMENTATION DETAILS

The training and testing are implemented on a desktop server with an Inter(R) i9-9900K CPU, 64 GB memory, and two NVIDIA GeForce RTX 2080 Ti graphics cards. These GPUs are configured with CUDA 10.0 and cuDNN 7.6 to realize the fast parallel computation of convolutions. The software environment is built on the Keras 2.3.1 with Tensorflow 1.15.0 in Python language. In implementation, the pixel values of the input image are normalized to 0 to 1 before entering the model. The convolution kernel parameters are initialized using Tikhonov regularization method (more often known as L2 regularization). Moreover, an Adam optimizer [49] is employed to minimize the loss value with a multi-staged learning rate. The training epoch is 100, and the batch input size is 4. During the training, the learning rate is $5e-4$ in 1 to 10 epochs, $2e-4$ in 11 to 50 epochs, and $1e-4$ in 51 to 100 epochs.

B. EVALUATION METRICS

In order to make a comprehensive evaluation, it is necessary to measure model performance in terms of segmentation integrity. To this end, we add some widely accepted objective criteria to evaluate these segmentation methods.

1) JACCARD SIMILARITY COEFFICIENT

Similar to the dice coefficient, the Jaccard index (also referred to as IoU) is used to measure the similarity of two sets. Defined as the size of the intersection divided by the size of the union, the Jaccard index can quantify the overlap between the predicted map and the ground truth. In our experiments, its value ranges from 0 to 1, and a larger value indicates a better segmentation result. The Jaccard index is defined as:

$$Jaccard = \frac{|X \cap Y|}{|X \cup Y|} \quad (4)$$

where X denotes the image array composed of the approximation of predicted results, and Y represents the image array of the ground truth. The elements in both X and Y are

either 0 or 1. In practice, the intersection and the union are realized by logical matrix operations.

2) HAUSDORFF DISTANCE

The hausdorff distance (HD) [50] is widely employed to evaluate medical image segmentation methods as an indicator to measure the maximum boundary of the segmentation surface. Since maxillofacial segmentation is a preprocessing step of tooth analysis in panoramic radiographs, HD provides a good measure of the usefulness of the segmentation results for the intended task. For overlapping binary images, HD can be obtained by calculating the distances from each point in one image to the nearest point in the other image, and then taking the largest distance. The bidirectional HD between two images is defined as follows:

$$hd(X, Y) = \max_{x \in X} \min_{y \in Y} \|x - y\|_2 \tag{5}$$

$$hd(Y, X) = \max_{y \in Y} \min_{x \in X} \|y - x\|_2 \tag{6}$$

$$HD(X, Y) = \max\left(hd(X, Y), hd(Y, X)\right) \tag{7}$$

where X denotes the binary predicted map and Y denotes the ground truth. We employ Euclidean distance to compute the distance between the pixels of the same class. While $hd(X, Y)$ and $hd(Y, X)$ are the one-sided HD between the binary images, the maximum $HD(X, Y)$ represents the longest segmentation error.

C. EXPERIMENTS

To validate the effectiveness of the proposed model, we design a series of comparative experiments on our maxillofacial dataset. We first explore the performance of EED-Net with different depths and investigate the impact of the weighted loss function on segmentation accuracy, then demonstrate the differences between the proposed model and the baseline models by visualizing the predicted maps, and finally confirm the performance in comparison with the latest real-time segmentation models.

1) EVALUATION OF MODEL DEPTH

Based on the flexible structure, we construct the corresponding models according to the number of encode-decoder pairs. These models have the same multipath extractor. Using the same binary cross entropy (BCE) loss and gradient learning rate, we test four different versions of the network: EED-Net-3, EED-Net-4, EED-Net-5, and EED-Net-6.

To obtain a comprehensive evaluation, the frame per second (fps) is augmented to describe the real-time performance of the models. In particular, the output rate measurement is performed on the whole dataset with all 2602 images to enhance the credibility. The overall performance of EED-Nets is presented in Table 2. As the network gets deeper, the spatial and time complexity of EED-Net gradually increases, indicating that the model parameters are mainly derived from the bottom encoder-decoder pairs. Among these models, EED-Net-3 with the fewest parameters achieves the highest output rate of 42.3 fps. In contrast, EED-Net-3 has the lowest segmentation accuracy, where the jaccard is 0.9790 and the hausdorff distance is 16.31. In general, the segmentation accuracy of the underfitting model benefits from additional parameters. However, the experimental results in Table 2 show that EED-Net-5 instead of EED-Net-6 yields the highest accuracy. As shown in Fig. 8, when the model depth exceeds three layers, the convergence trend of the loss values is roughly similar. For EED-Net-6, the increased parameters bring no improvement in segmentation accuracy, but lead to slight overfitting. Due to the small difference for EED-Nets in output rate, we can infer that a deeper network may not lead to better performance. Since both segmentation accuracy and speed must be considered in maxillofacial segmentation, the most accurate EED-Net-5 is selected for subsequent optimization.

TABLE 2. The performance comparison for EED-Nets of different depths using BCE loss.

Models	Acc	Jaccard	HD	FPS	Para(M)
EED-Net-3	0.9911	0.9790	16.31	**42.3**	**0.36**
EED-Net-4	0.9917	0.9804	10.45	41.9	0.61
EED-Net-5	**0.9921**	**0.9813**	**9.69**	41.0	0.92
EED-Net-6	0.9920	0.9810	9.97	40.2	1.23

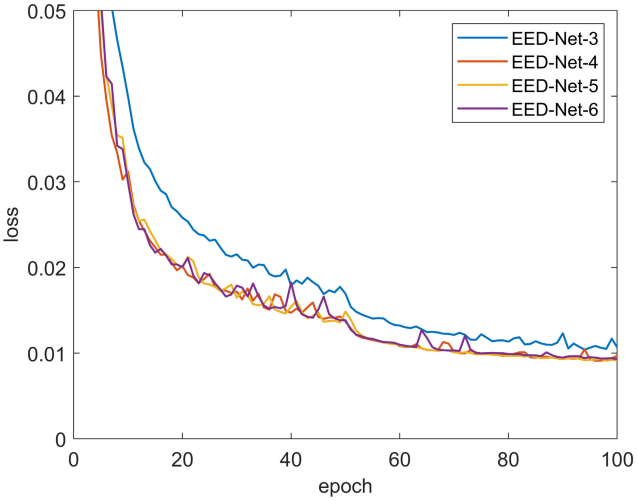


FIGURE 8. The training loss curves of EED-Nets.

2) EVALUATION OF THE LOSS FUNCTIONS

To verify the validity of the weighted loss function, we train EED-Net-5 with the other two classical losses under the same conditions. Specifically, the model performance of three typical weighted coefficients α are reported in Table 3. Besides, the loss curves of EED-Net-5 with different configurations are plotted in Fig. 9, where the weighted loss curves are located between the BCE curve and the Dice curve. With the increase of α, the curve gets closer to Dice, otherwise, it rises near BCE, which conforms to its definition. Table 3 shows that the weighted loss has a higher segmentation accuracy

TABLE 3. The comparison of segmentation accuracy of EED-Net-5 trained by different loss functions.

Loss fucntion	Acc	Jaccard	HD
EED-Net-5-BCE	0.9921	0.9814	9.69
EED-Net-5-Dice	0.9922	0.9815	9.13
EED-Net-5(α=0.2)	0.9925	0.9821	8.37
EED-Net-5(α=0.5)	0.9924	0.9819	8.97
EED-Net-5(α=0.8)	**0.9928**	**0.9829**	**8.32**

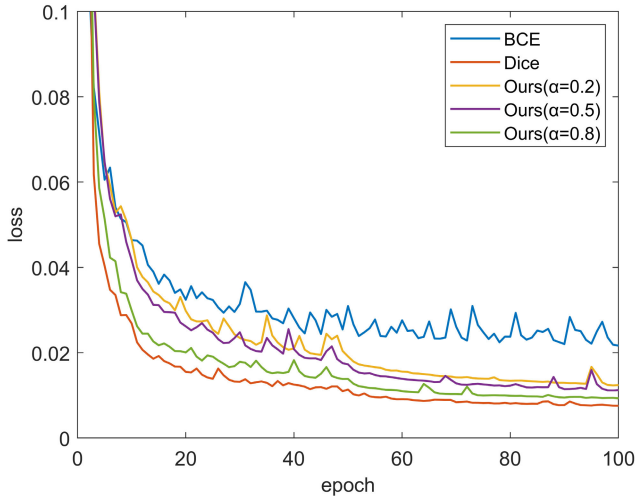


FIGURE 9. The loss curves of EED-Net-5 with different loss functions. BCE and Dice are the classical loss functions, and ours is the weighted realization of these two losses, where α is the weighted coefficient of Dice.

than the classical losses, especially in the hausdorff distance. However, the performance improvement of the weighted coefficient is nonlinear. According to the experimental results of the typical weights, the boundary weights bring greater performance gain than the middle weights, approximating a U-shaped distribution. Compared with BCE and Dice, the model receives the biggest boost with a α of 0.8, where the jaccard increases by 0.15% and 0.14%, and the hausdorff distance reduces by 14.14% and 8.87%. As a result, this version of EED-Net-5 is chosen as our final model.

3) COMPARISON WITH THE BASELINE MODELS

Since the proposed model is built on the basis of FCN-8s and U-Net, it is necessary to analyze the changes brought about by our adjustments. Fig. 11 displays the visual segmentation results of FCN-8s, U-Net, and EED-Net-5. The corresponding accuracy values are presented in Table 4. The input images are the representative dental X-rays selected from the test images. On the whole, all three models are able to extract the basic contours of the maxillofacial region. However, there are significant differences in the details of the contour edge. FCN-8s has continuous pixel recognition errors in the maxillofacial edges, resulting in a jagged boundary. Compared with FCN-8s, the boundaries of U-Net are more distinct and smoother. Based on the structural analysis,

TABLE 4. The accuracy performance of EED-Net-5 and the baseline models on the specified images. Image1 to Image5 are the test samples listed from top to bottom in X-ray column of Fig. 11.

Methods	Acc				
	Image1	Image2	Image3	Image4	Image5
FCN-8s	0.989	0.9932	0.9943	0.9908	0.9927
U-Net	0.9906	0.9941	0.9938	0.9913	0.9933
Ours	**0.9921**	**0.9956**	**0.9951**	**0.9919**	**0.9952**

Methods	Jaccard				
	Image1	Image2	Image3	Image4	Image5
FCN-8s	0.9727	0.9843	0.9847	0.9737	0.9762
U-Net	0.9769	0.9863	0.9836	0.9749	0.9783
Ours	**0.9805**	**0.9897**	**0.987**	**0.9767**	**0.9844**

Methods	HD				
	Image1	Image2	Image3	Image4	Image5
FCN-8s	16.28	7.21	7.28	8.94	7.81
U-Net	**12.37**	10.0	56.44	20.81	8.48
Ours	14.42	**4.0**	**7.07**	**7.62**	**5.39**

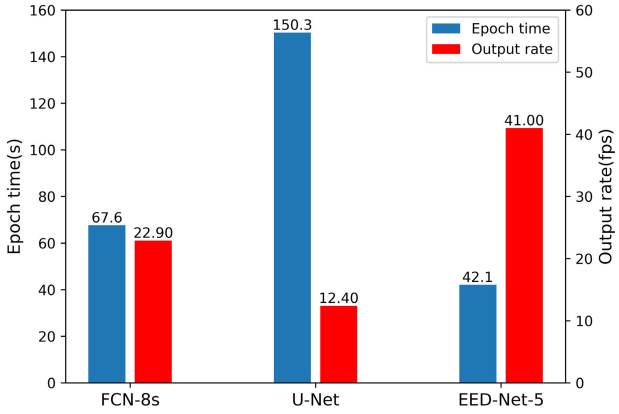


FIGURE 10. The comparison of output rate and average one-epoch training time of EED-Net-5 and the baseline models.

we consider that the encoder-decoder network form contributes to the fine-grained segmentation. However, there are still some regional pixel recognition errors around the edges, which looks like some bubbles. As for EED-Net-5, the residual encoders and the multipath feature extraction enable it to obtain a higher pixel accuracy and regional connectivity. Therefore, the results of EED-Net-5 are closer to the ground truth with smoother boundaries and fewer bubbles.

In addition to the recovery of image details, we illustrate the speed performance of the three models in Fig. 10. The concatenation operation of U-Net brings high computation while obtaining fine-grained segmentation, resulting in the longest training time and lowest output rate. The output form of FCN-8s enables it to lead U-Net 10 fps in speed with decreased accuracy. Due to the combination of channel pruning and decoder simplification, EED-Net-5 reduces the epoch time by 37.7% from 67.6 to 42.1 and increases the output rate by 79.1% from 22.9 fps to 41.0 fps compared to FCN-8s. In summary, the proposed model achieves significant improvements in accuracy and speed over the baseline models.

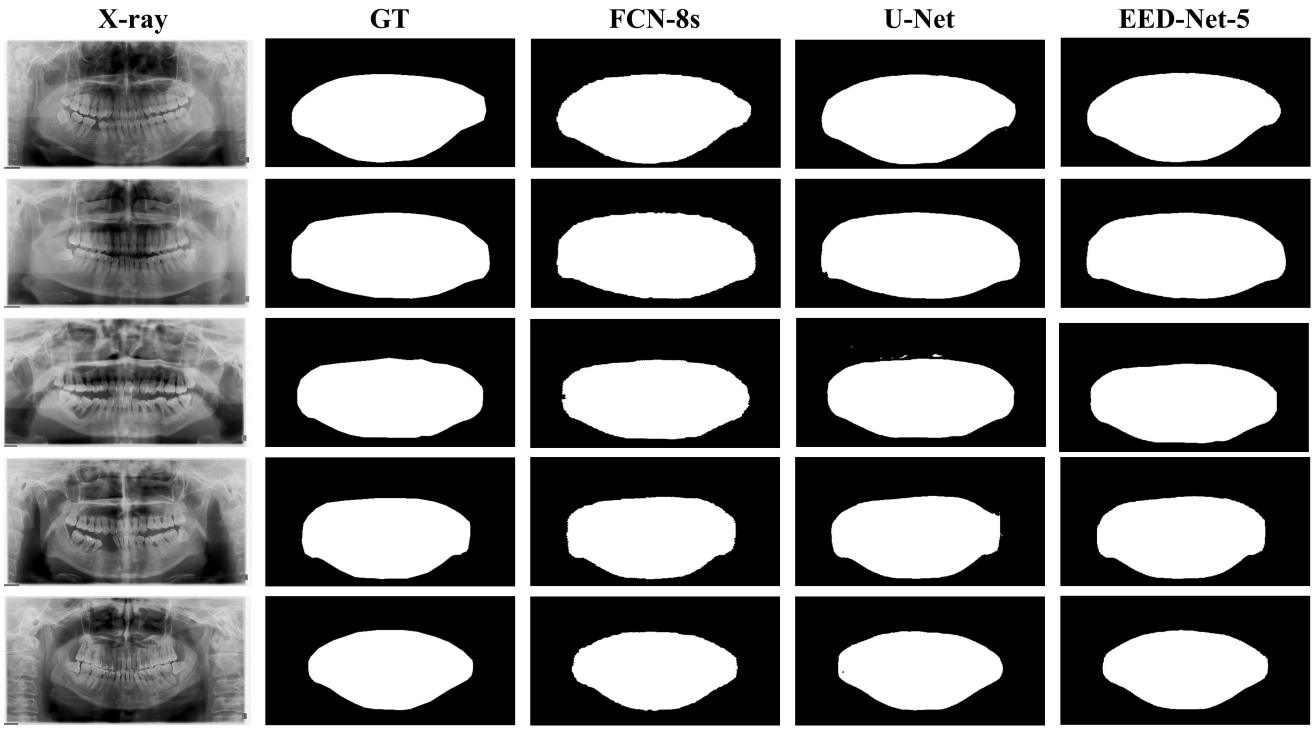


FIGURE 11. The sample results of the three models. The first column is five different dental X-rays where the first two samples are "non-periodontal" and the others are "periodontal". The "GT" is the ground truth generated by manually annotation. The other columns denote the visual segmentation results of FCN, U-Net, and EED-Net-5.

4) COMPARISON WITH THE LATEST METHODS

To further confirm the effectiveness of the proposed model, we design comparison experiments with the latest real-time segmentation models on the maxillofacial dataset. The weighted loss and Adam optimizer with stepped learning rate are employed for all models. It is worth noting that the model performance and properties may be slightly different from the original method because we have migrated the code of authors to our environment. From Table 5, it can be seen that except for the classical models, RefineNet and Fast-SCNN are representative of the other models. RefineNet achieves 0.9926, 0.9827, and 8.93 in the accuracy, the jaccard, and the hausdorff distance respectively, but this improvement is derived from the increase of model parameters, and its segmentation speed is unsatisfactory. Fast-SCNN achieves the output rate of 66.2 fps, but this real-time performance comes at the expense of segmentation delicacy, which is detrimental to subsequent image processing. The performance of other models on maxillofacial segmentation falls between these two models, and they fail to find a balance between speed and accuracy.

The results in Table 5 indicate that our model is a better implementation of maxillofacial segmentation. The proposed EED-Net-5 achieves 0.9928 in the accuracy, 0.9829 in the jaccard and 8.32 in the hausdorff distance with the fewest parameters, which outperforms the other methods. In the aspect of segmentation speed, our model can maintain an output rate of 41.0 fps, which meets the requirements of real-time segmentation. Since the environmental configuration of the models is identical, the highest accuracy and the increased speed further validate the efficiency of our network architecture. This is mainly due to the optimization of the encoder-decoder structure and the addition of the multipath feature extractor, which enables our model to focus on image details with fewer parameters. Considering both accuracy and speed performance, the conclusion can be reached that the proposed model achieves better overall performance than the other real-time segmentation methods for maxillofacial segmentation.

TABLE 5. The performance comparison between EED-Net-5 and the latest segmentation methods on the maxillofacial dataset.

Methods	Acc	Jaccard	HD	FPS	Para(M)
FCN-8s [34]	0.9916	0.9802	9.04	22.9	21.48
U-Net [35]	**0.9926**	0.9824	11.95	12.4	8.63
RefineNet [39]	**0.9926**	**0.9827**	**8.93**	13.4	26.50
ICNet [40]	0.9907	0.9791	10.32	40.6	6.74
BiSeNet [41]	0.9842	0.9638	21.39	34.3	5.07
Fast-SCNN [42]	0.9861	0.9691	12.49	**66.2**	**1.56**
LEDNet [43]	0.9916	0.9810	9.61	34.8	2.77
Ours	**0.9928**	**0.9829**	**8.32**	**41.0**	**0.92**

V. CONCLUSION AND DISCUSSION

An accurate and automatic extraction of the ROI is essential for the researches of panoramic dental X-rays. In this study, we establish a maxillofacial dataset and propose an efficient encoder and decoder network model named EED-Net to achieve maxillofacial segmentation. The dataset consists of 2602 panoramic X-ray images and the corresponding maxillofacial mask results. Based on the structure of U-Net and the decoding form of FCN-8s, the proposed EED-Net is composed of the residual encoders, the multipath feature extractors, and the object-oriented decoders. The encoders and extractors are used to capture deep features, and the channel number of the decoders is simplified to reduce the parameters. Moreover, we adopt a weighted loss function to further improve segmentation accuracy. The extensive experiments on the dataset demonstrate that EED-Net outperforms other methods in segmentation accuracy. In addition, our model can output an average of 41.0 fps segmented images per second, which is highly feasible to realize the preprocessing of panoramic X-rays. Based on the extensibility of the baseline models, it can be inferred that our method can be applied to the X-ray images with a single target type. For the multi-class segmentation, the effectiveness of the proposed method requires further study. Since we have to compress the original images to avoid insufficient computer memory in this work, how to directly extract ROI using high-resolution images also deserves further exploration.

REFERENCES

[1] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger, P. Fischer, T. F. Cootes, and C. Lindner, "A benchmark for comparison of dental radiography analysis algorithms," *Med. Image Anal.*, vol. 31, pp. 63–76, Jul. 2016.

[2] K.-J. Park and K.-C. Kwak, "A trends analysis of dental image processing," in *Proc. 17th Int. Conf. ICT Knowl. Eng. (ICT&KE)*, Nov. 2019, pp. 1–5.

[3] G. Silva, L. Oliveira, and M. Pithon, "Automatic segmenting teeth in X-ray images: Trends, a novel data set, benchmarking and future perspectives," *Expert Syst. Appl.*, vol. 107, pp. 15–31, Oct. 2018.

[4] R. Kaur, R. S. Sandhu, A. Gera, and T. Kaur, "Edge detection in digital panoramic dental radiograph using improved morphological gradient and MATLAB," in *Proc. Int. Conf. Smart Technol. For Smart Nation (SmartTechCon)*, Aug. 2017, pp. 793–797.

[5] V. Rushton, K. Horner, and H. Worthington, "The quality of panoramic radiographs in a sample of general dental practices," *Brit. Dental J.*, vol. 186, no. 12, pp. 630–633, Jun. 1999.

[6] A. Wirtz, S. G. Mirashi, and S. Wesarg, "Automatic teeth segmentation in panoramic X-ray images using a coupled shape model in combination with a neural network," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 712–719.

[7] Y. Y. Amer and M. J. Aqel, "An efficient segmentation algorithm for panoramic dental images," *Procedia Comput. Sci.*, vol. 65, pp. 718–725, Jan. 2015.

[8] M. K. Alsmadi, "A hybrid fuzzy C-means and neutrosophic for jaw lesions segmentation," *Ain Shams Eng. J.*, vol. 9, no. 4, pp. 697–706, Dec. 2018.

[9] C. K. Modi and N. P. Desai, "A simple and novel algorithm for automatic selection of ROI for dental radiograph segmentation," in *Proc. 24th Can. Conf. Electr. Comput. Eng. (CCECE)*, May 2011, pp. 000504–000507.

[10] L. Gráfová, M. Kašparová, S. Kakawand, A. Procházka, and T. Dostálová, "Study of edge detection task in dental panoramic radiographs," *Dentomaxillofacial Radiol.*, vol. 42, no. 7, Jul. 2013, Art. no. 20120391.

[11] A. Betul Oktay, "Tooth detection with convolutional neural networks," in *Proc. Med. Technol. Nat. Congr. (TIPTEKNO)*, Oct. 2017, pp. 1–4.

[12] T. L. Koch, M. Perslev, C. Igel, and S. S. Brandt, "Accurate segmentation of dental panoramic radiographs with U-NETS," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 15–19.

[13] L. Song, J. Lin, Z. J. Wang, and H. Wang, "An end-to-end multi-task deep learning framework for skin lesion analysis," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2912–2921, Oct. 2020.

[14] G. Jader, J. Fontineli, M. Ruiz, K. Abdalla, M. Pithon, and L. Oliveira, "Deep instance segmentation of teeth in panoramic X-ray images," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 400–407.

[15] D. Frejlichowski and R. Wanat, "Automatic segmentation of digital orthopantomograms for forensic human identification," in *Proc. Int. Conf. Image Anal. Process.* Berlin, Germany: Springer, 2011, pp. 294–302.

[16] C. Buchart, G. S. Vicente, A. Amundarain, and D. Borro, "Hybrid visualization for maxillofacial surgery planning and simulation," in *Proc. 13th Int. Conf. Inf. Vis.*, Jul. 2009, pp. 266–273.

[17] M. H. Bozkurt and S. Karagol, "Jaw and teeth segmentation on the panoramic X-ray images for dental human identification," *J. Digit. Imag.*, early access, Aug. 2020.

[18] P.-L. Lin, P.-Y. Huang, and P.-W. Huang, "An automatic lesion detection method for dental X-ray images by segmentation using variational level set," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 5, Jul. 2012, pp. 1821–1825.

[19] J.-H. Lee, S.-S. Han, Y. H. Kim, C. Lee, and I. Kim, "Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs," *Oral Surgery, Oral Med., Oral Pathol. Oral Radiol.*, vol. 129, no. 6, pp. 635–642, Jun. 2020.

[20] J. Oliveira and H. Proença, "Caries detection in panoramic dental X-ray images," in *Computational Vision and Medical Image Processing*. Dordrecht, The Netherlands: Springer, 2011, pp. 175–190.

[21] C. Fares and M. Feghali, "Tooth-based identification of individuals," *Int. J. New Comput. Archit. Appl.*, vol. 3, no. 1, pp. 22–34, 2013.

[22] A. K. Jain and H. Chen, "Matching of dental X-ray images for human identification," *Pattern Recognit.*, vol. 37, no. 7, pp. 1519–1532, Jul. 2004.

[23] A. A. Harandi, H. Pourghassem, and H. Mahmoodian, "Upper and lower jaw segmentation in dental X-ray image using modified active contour," in *Proc. Int. Conf. Intell. Comput. Bio-Med. Instrum.*, Dec. 2011, pp. 124–127.

[24] E. Gumus, "Segmentation and root localization for analysis of dental radiographs," *Signal, Image Video Process.*, vol. 10, no. 6, pp. 1073–1079, Sep. 2016.

[25] M. M. Hasan, W. Ismail, R. Hassan, and A. Yoshitaka, "Automatic segmentation of jaw from panoramic dental X-ray images using GVF snakes," in *Proc. World Automat. Congr. (WAC)*, Jul. 2016, pp. 1–6.

[26] R. Wanat and D. Frejlichowski, "A problem of automatic segmentation of digital dental panoramic X-ray images for forensic human identification," in *Proc. CESCG*, 2011, pp. 1–8.

[27] X. Xu, C. Liu, and Y. Zheng, "3D tooth segmentation and labeling using deep convolutional neural networks," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 7, pp. 2336–2348, Jul. 2019.

[28] Y. Rao, Y. Wang, F. Meng, J. Pu, J. Sun, and Q. Wang, "A symmetric fully convolutional residual network with DCRF for accurate tooth segmentation," *IEEE Access*, vol. 8, pp. 92028–92038, 2020.

[29] S. Tian, N. Dai, B. Zhang, F. Yuan, Q. Yu, and X. Cheng, "Automatic classification and segmentation of teeth on 3D dental model using hierarchical deep learning networks," *IEEE Access*, vol. 7, pp. 84817–84828, 2019.

[30] C. Lian, L. Wang, T.-H. Wu, F. Wang, P.-T. Yap, C.-C. Ko, and D. Shen, "Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3D intraoral scanners," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2440–2450, Jul. 2020.

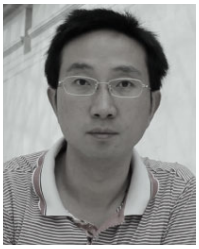
[31] C. Lyu, G. Hu, and D. Wang, "HRED-Net: High-resolution encoder-decoder network for fine-grained image segmentation," *IEEE Access*, vol. 8, pp. 38210–38220, 2020.

[32] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder–decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2020.

[33] J. Cheng, P.-S. Wang, G. Li, Q.-H. Hu, and H.-Q. Lu, "Recent advances in efficient computation of deep convolutional neural networks," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 64–77, Jan. 2018.

[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2016, *arXiv:1602.07261*. [Online]. Available: http://arxiv.org/abs/1602.07261
[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[39] V. Nekrasov, C. Shen, and I. Reid, "Light-weight RefineNet for real-time semantic segmentation," 2018, *arXiv:1810.03272*. [Online]. Available: http://arxiv.org/abs/1810.03272
[40] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 405–420.
[41] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.
[42] R. P K Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast semantic segmentation network," 2019, *arXiv:1902.04502*. [Online]. Available: http://arxiv.org/abs/1902.04502
[43] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, and L. J. Latecki, "Lednet: A lightweight encoder-decoder network for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1860–1864.
[44] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proc. 27th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2019, pp. 2276–2279.
[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 630–645.
[47] Y. Zhou, X. Wang, M. Zhang, J. Zhu, R. Zheng, and Q. Wu, "MPCE: A maximum probability based cross entropy loss function for neural network classification," *IEEE Access*, vol. 7, pp. 146331–146341, 2019.
[48] D. Kang, S. Park, and J. Paik, "SdBAN: Salient object detection using bilateral attention network with dice coefficient loss," *IEEE Access*, vol. 8, pp. 104357–104370, 2020.
[49] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 499–513, Feb. 2020.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980



ZHENGMIN KONG (Member, IEEE) received the B.Eng. and Ph.D. degrees from the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China, in 2003 and 2011, respectively. From 2005 to 2011, he was with the Wuhan National Laboratory for Optoelectronics as a member of the Research Staff, and was involved in beyond-3G and UWB system design. From 2014 to 2015, he was with the University of Southampton, U.K., as an Academic Visitor, and investigated physical-layer security and artificial intelligence. He is currently an Associate Professor with the School of Electrical Engineering and Automation, Wuhan University. His current research interests include physical-layer security, signal processing, computer vision, and artificial intelligence.

FENG XIONG received the B.Eng. degree from the Department of Artificial Intelligence and Automation, Wuhan University, Wuhan, China, in 2018, where he is currently pursuing the master's degree with the School of Electrical Engineering and Automation. His research interests include the Internet of Things, artificial intelligence, and computer vision.

CHENGGANG ZHANG received the B.Eng. degree from the Department of Artificial Intelligence and Automation, Wuhan University, Wuhan, China, in 2018, where he is currently pursuing the master's degree with the School of Electrical Engineering and Automation. His research interests include electricity load forecasting, data mining, and artificial intelligence.

ZHUOLIN FU received the B.Eng. degree from the Department of Artificial Intelligence and Automation, Wuhan University, Wuhan, China, in 2018, where he is currently pursuing the master's degree with the School of Electrical Engineering and Automation. His research interests include software development, few-shot learning, and artificial intelligence.

MAOQI ZHANG received the bachelor's degree in dental medicine from the School of Stomatology, Nanchang University, Nanchang, China, in 2019. She is currently pursuing the master's degree in periodontal science with the School of Stomatology, Wuhan University. Her main research interests include medical services, medical diagnostic imaging, medical diagnosis, medical treatment, and medical tests.

JINGXIN WENG received the bachelor's degree in medicine from the Guanghua School of Stomatology, Sun Yat-sen University, Guangzhou, China, in 2019. She is currently pursuing the master's degree with the School of Stomatology, Wuhan University. Her research interests include the medical diagnosis, medical diagnosis imaging, and medical services.

MINGZHE FAN received the bachelor's degree in oral medicine from the School of Stomatology, Shandong University, Jinan, China, in 2009. He is currently pursuing the master's degree with the School of Stomatology, Wuhan University. His research interests include the medical treatment, medical diagnosis, and medical services.