J Syst Sci Complex (2023) 36: 257–272

Transformer-Based Deep Learning Network for Tooth Segmentation on Panoramic Radiographs*

SHENG Chen · WANG Lin · HUANG Zhenhuan · WANG Tian · GUO Yalin · HOU Wenjie · XU Laiqing · WANG Jiazhu · YAN Xue

DOI: 10.1007/s11424-022-2057-9
Received: 24 January 2022 / Revised: 23 March 2022

Abstract Panoramic radiographs can assist dentist to quickly evaluate patients' overall oral health status. The accurate detection and localization of tooth tissue on panoramic radiographs is the first step to identify pathology, and also plays a key role in an automatic diagnosis system. However, the evaluation of panoramic radiographs depends on the clinical experience and knowledge of dentist, while the interpretation of panoramic radiographs might lead misdiagnosis. Therefore, it is of great significance to use artificial intelligence to segment teeth on panoramic radiographs. In this study, SWin-Unet, the transformer-based Ushaped encoder-decoder architecture with skip-connections, is introduced to perform panoramic radiograph segmentation. To well evaluate the tooth segmentation performance of SWin-Unet, the PLAGH-BH dataset is introduced for the research purpose. The performance is evaluated by F1 score, mean intersection and Union (IoU) and Acc, Compared with U-Net, LinkNet and FPN baselines, SWin-Unet performs much better in PLAGH-BH tooth segmentation dataset. These results indicate that SWin-Unet is more feasible on panoramic radiograph segmentation, and is valuable for the potential clinical application.

Keywords Deep convolutional neural network, panoramic radiograph, SWin-Unet, Tooth segmentation.

1 Introduction

Panoramic radiographs are a commonly used auxiliary diagnostic tool in oral clinical practice. They can help dentist make a comprehensive evaluation of patients oral health and help

SHENG Chen
Medical School of Chinese PLA, Beijing 100853*, China; Department of Stomatology, the first Medical Centre, Chinese PLA General Hospital, Beijing* 100853*, China*. Email: shengchen301@163.com.
WANG Lin (Corresponding author) · HUANG Zhenhuan · WANG Tian · GUO Yalin · HOU Wenjie · XU Laiqing · WANG Jiazhu · YAN Xue
Department of Stomatology, the First Medical Centre, Chinese PLA General Hospital, Beijing 100853*, China; Beihang University, Beijing* 100191*, China; Medical School of Chinese PLA, Beijing* 100853*, China.* Email: 13581891907@163.com; zhenhuan@buaa.edu.cn; wangtian@buaa.edu.cn; guoyalin301@163.com; houwenjie2022@163.com; xulaiqing@163.com; yuiyuan654@163.com; a137872648@163.com.
◇ *This paper was recommended for publication by Editor QI Hongsheng.*

dentist to quickly understand the overall tooth health status. Panoramic radiographs show the whole tooth information, i.e., tooth crown, tooth root, degree of tooth inclination, and jaw information. They have the advantages of simplicity, low cost and low radiation. However, panoramic radiographs still face following difficulties in the screening of oral diseases: 1) Complex and diverse lesions: Panoramic radiographs often reflect many different pathological images at the same time. The pathological characteristics are complex, the size and location of the lesions are different. Thus, it is difficult to make accurate and comprehensive diagnosis. 2) Dependence on professional assessment: The evaluation of panoramic radiographs depends seriously on clinical experience and knowledge of dentists. Moreover, there are often diagnostic differences among different dentists. 3) Human interpretation is prone to misdiagnosis and omission diagnosis: In the process of interpreting Panoramic radiographs, dentists tend to easily ignore the non principal complaint teeth abnormalities of hidden lesions, and only pay attention to the complaints of the patient, leading to the non principal complaint teeth missing the best treatment opportunity. 4) Image overlap: The panoramic radiograph is a two-dimensional planar image, which may cause the overlap and distortion of anatomical structures and the blur of tooth structure[1]. In addition, the unbalanced distribution of medical resources also aggravate the resistance of residents to receive oral health care services, which will face severe test. Therefore, the automatic diagnosis on panoramic radiographs is of great significance for the clinical work.

The precise detection and localization of specific anatomical structures on medical images, e.g., object segmentation, serves as the first step in identifying pathology and the key issue for an automated diagnostic system[2]. Image signal provides one potential way to analyze the situation[3]. Dental caries, periapical lesions, periodontal disease, odontogenic cysts and tumors all occur in the tooth tissue or around the tooth. The automated diagnosis and identification of these diseases on panoramic radiographs highly depended on accurate segmentation of the tooth tissue, including clear tooth boundaries. Furthermore, accurate segmentation of tooth tissue is also critical for making orthodontic treatment plans[4]. The localization and orientation evaluation of the teeth from the panoramic radiographs determines the orthodontic treatment procedure and treatment time. Thus, this task is subjective and sometimes inconsistent among experts. Furthermore, it is time consuming and tedious to sketch tooth profiles or markers by hand, and dentists often struggle to cope when the number of patients is large[5]. Therefore, the precise segmentation of the teeth on panoramic radiographs is of great significance in the oral clinic.

Traditional image segmentation and pattern recognition methods are categorized as threshold-based segmentation[6, 7], region-based segmentation[8, 9], energy functional-based segmentation[10]. In addition to the traditional image segmentation methods, the deep CNN is an efficient recognition and segmentation method developed in recent years[11, 12]. It has been widely used in biomedical field, and has achieved great achievements in medical image analysis.

CNN is one of the most useful artificial neural networks, facilitating extensive computer vision tasks, such as medical image classification[13, 14]. Computer vision techniques based on CNNs have been impressive for radiographic recognition of pathologies and have the potential to

meet the need for accurate tooth segmentation on panoramic radiographs[15, 16]. However, there are still many problems behind the rapid development of intelligent dental radiologic images: 1) Teeth vary from person to person. It is difficult to determine a set of parameters that can cover any of all teeth. In addition, image quality, tooth loss, caries, artifacts, presence of implants and bridges and variations of the teeth in between patients, and patients' age interfere with the segmentation task[17]. 2) Lack of high-quality labeled training samples. Supervised learning based method is wildly adopted in computer vision tasks. It requires a lot of precisely labeled data for learning. Labeling data depends on the doctor's professional knowledge, which is time-consuming. There are few publicly available dataset for dental radiologic images, which result in overfitting or poor robustness and less generalization. It is difficult to clinical transformation and commercial use[18]. 3) At present, the proposed deep learning-based model still needs to be optimized. Existing methods are less robustness and generalization for the practical applications[19].

This paper validates the feasibility of the deep CNN model as a dental image segmentation tool. This paper proposes a tooth segmentation method on panoramic radiographs, which can be adopted in the automatic diagnosis of clinical work. The tooth tissue boundaries on panoramic radiographs are extracted. The performance of different neural network are analyzed. In this study, we introduced SWin-Unet for panoramic radiographs segmentation, which is a pure transformer-based U-shaped Encoder-Decoder architecture. To evaluate its teeth segmentation performance, we collected 100 panoramic radiographs to establish PLAGH-BH dataset. We evaluated the model performance using F1 score, Mean intersection and union (IoU), and Acc. It is compared to three baseline algorithms: U-Net, Link-Net, and FPN. The results show that SWin-Unet performs better than other segmentation models on our introduced PLAGH-BH teeth segmentation dataset and the publicly available dataset. SWin-Unet is a transformer-based Ushaped encoder-decoder architecture with skip-connections. Traditional CNN shows good performance in image segmentation, but due to the limitation of convolutional operations, the global and remote semantic information interaction cannot be learned well. SWin-Unet builds a symmetric encoder-decoder structure with jump connections, implements a local to global self-attention mechanism, and develops a patch expanding layer without convolution and interpolation operations (patch expanding layer) to increases upsampling and feature dimension, which outperforms convolution based methods[20, 21].

2 Deep Learning-Based Tooth Segmentation

2.1 Tooth Segmentation

Image segmentation divides the global image into multiple regions (pixel sets) to make the image representation easier for analysis. Tooth segmentation on X-ray radiographs is essential for automatic diagnosis of department of stomatology. In dentistry, X-ray radiographs include two types[22]: Intraoral radiographs performed by placing film inside the mouth (bitewing radiographs and periapical radiographs) and extraloral radiographs by placing the patient between the radiograph and the X-ray source (panoramic radiographs). Most automatic diagnostic stud-

ies in dental radiology have focused on intraoral radiography, whose independent tooth images were easier for segmentation. In addition, due to the limitations of extraloral radiographs, such as topological imaging produces multi-lever noise. In some cases, the incisors image is covered, and shown in low contrast with complex morphological properties[23]. There are few studies on tooth segmentation on externaloral radiographs. However, extraoral radiograph are widely used in oral clinical practice, which can observe the basic tooth structure, tooth number, tooth development, periodontal and periapical situation. Thus, it is particularly important for tooth imaging segmentation of extraoral radiograph.

Deep learning is an algorithm with artificial neural network architecture. The deep learning methods, such as convolutional neural networks and recurrent neural networks, have achieved excellent results in different fields, such as computer vision, bioinformatics, natural language processing, and audio recognition, etc[24, 25]. The artificial neurons of CNN respond to local receptive filed, they perform well for image processing tasks. One or more convolutional layers and pooling layers process the two dimensional data with the feedforward mechanism. With the development of deep learning algorithms, X-ray images are increasingly used as the input, as shown in Figure 1. This paper inputs panoramic radiographs into four neural networks. The segmentation performance is compared. The feasibility of the method on panoramic dental

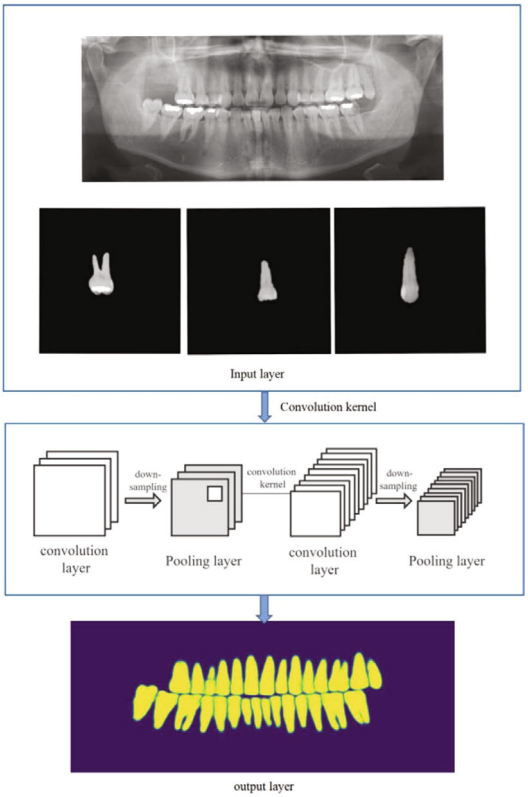

Figure 1 Tooth segmentation on panoramic radiograph based on deep learning

image segmentation are clarified, and its potential clinical application value are proved. The deep learning methods help to improve the diagnostic accuracy and efficiency of clinicians, and reduce the misdiagnosis and omission diagnosis.

2.2 Evaluation

We use x and w^l to represent the input and the corresponding convolutional filters, which introduces convolutional features into the forward propagation. We implement $f(x * w^l)$ based on a new layer, which is generic and can be independently used for any network, e.g., CNNs. The formal process is implemented as:

$$F^{l+1} = f_{CNN}(F^l, X), \tag{1}$$

where F^l stands for the feature map for the l-th layer. f_{CNN} denotes the convolutional operation implemented as a new layer or module. As shown in Figure 1, a pooling operation is also introduced on the output of each layer as:

$$\widehat{F}^{l+1} = Pooling(F^{l+1}), \tag{2}$$

where $\widehat{F}^{l+1}$ is the pooling feature which can introduce robustness for the object segmentation task. Correspondingly, the backward pass can be formulated as:

$$w_{t+1} = w_t + \alpha_t \frac{\partial \mathscr{L}}{\partial w}, \tag{3}$$

where $\mathscr{L}$ is the loss of network, and the layer index is omitted for easy presentation.

Proposition *Given that $\alpha_t \geq \alpha_{t+1}$, for loss function $\mathscr{L}$, the update rule shown in Equation* (3) *can lead to a learning converge of deep learning process defined as:*

$$\min_x \ \mathscr{L}(x), \tag{4}$$

where $x \in B$ is the variable, B is a finite set of input data, $\mathscr{L}$ is a convex function. The convergence can be also achieved based on the stochastic gradient descent (SGD) method, whose details are shown in [26].

We used the following criteria to evaluate the segmentation performance of four deep convolutional neural network models for 90 panoramic radiographs in the training dataset:

1) F1 score: F1 explains the degree of pixel overlap between ground truth and the predicted results, manifested as the balance between precision and recall.

$$Precesion = \frac{TP}{TP + FP}, \tag{5}$$

$$Recall = \frac{TP}{TP + FN}, \tag{6}$$

$$F_1 = \frac{2 \times (Recall \times Precision)}{Recall + Precision}. \tag{7}$$

Among them, TP (true positive) is true positive, FP (false positive) is false positive, FN (false negative) is false negative, and TN was true negative. TP is the region where the ground

truth overlaps with the predicted results. FP is the non-overlapping region between the ground truth and the predicted results. FN is the non-overlapping region in ground truth.

2) Mean intersection over union (IoU): IoU shows the area of overlap, between the results and the tooth segmentation in the ground truth region. The following is the formula for the IoU:

$$IoU = \frac{TP}{TP + FP + FN}. \tag{8}$$

The F1 score and mean IoU evaluate the performance of the segmentation from 0 to 1, and the closer to 1 indicates higher performance.

3) Accuracy (Acc): It indicates that the number of correctly predicted samples accounts for the percentage of all the samples. Its values range is 0 to 1. The closer the value of Acc is to 1, the higher the accuracy of segmentation. The formula is as follows:

$$Acc = \frac{TP + TN}{TP + FN + FP + TN}. \tag{9}$$

2.3 Swin-Unet: Unet-Like Pure Transformer for Tooth Segmentation

Transformer has achieved excellent application results in many visual tasks, but there is no related work in tooth image segmentation. ViT is based on a complete self-attention Transformer structure without using CNN. The image is divided into fixed-size patches, then the patches are fed into the linear projections along with their positions[27]. We apply the transformer-based image segmentation algorithm SWin-Unet for the tooth image segmentation task.

Tooth segmentation based on SWin-Unet is similar to U-Net's medical image segmentation, where the transformer model is the core modeling block. The tokenized image patches are sent to the transformer-based U-shaped encoding-decoding architecture via jump connections for local global semantic feature learning. The images are learnt for a global representation. The hierarchical Swin Transformer with an offset window as an encoder to represent the image spatial context features and design a decoder with a patch extended layer for upsampling operations to recover the spatial resolution of the feature graph[21]. SWin-Unet's architecture is shown in Figure 2. The bottleneck block, encoders, decoders, and jump connections are shown. The encoder splits the panorama into non-overlapping patches of size 4×4, converting the input images into sequence embedding. Then, the dimension of the feature is transformed to fit for the demand. The hierarchical features are learnt by the delicately designed blocks of the Swin transformer. The features are learnt in the transformer block, and the dimensions are changed in the patch merging layer. The decoder is relative to the encoder. The transformer block and patch expanding layer are designed correspondingly. Due to the encoder and downsampling precesses, the spatial information is lost. Thus, the multi-scale features are fused for the compensation. The upsampling, reshaping processes are conducted in the patch layer. The last patch expanding layer up-samples the feature the input resolution. The up-sampled features are adopted for segment prediction. We adopt SWin-Unet for tooth segmentation. The panoramic radiographs are split into several non-overlapping image patches, and then delivered to a transformer-based encoder to learn deep features. The decoder and encoder architectures

extract and fused the multiscale features to image representation, and make further segmentation.

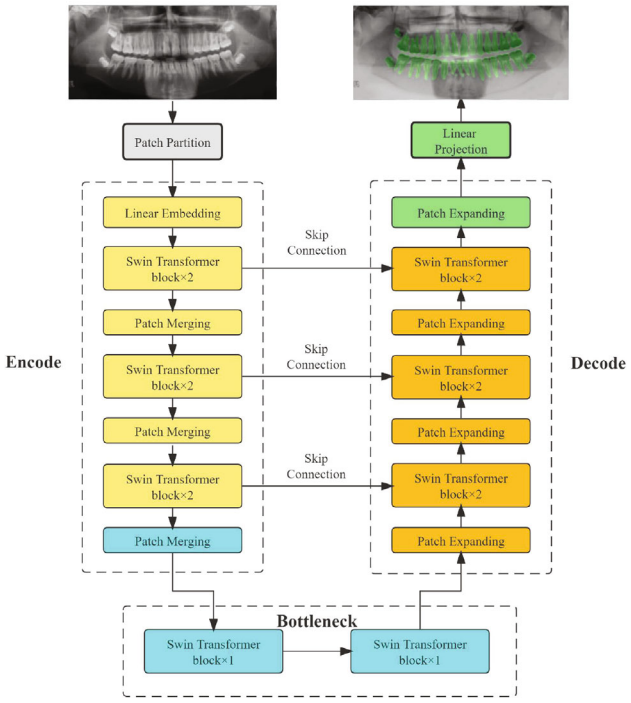

Figure 2 Tooth segmentation schematic diagram of SWin-Unet on panoramic radiographs

The SWin transformer block is different from the traditional multi-head self-attention (MSA) module in that it is constructed based on the shift window. In Figure 3, two consecutive Swin transformer blocks are introduced. Each Swin transformer block is composed of LayerNorm (LN) layer, multi-head self attention module, residual connection and 2-layer MLP with GELU non-linearity. The window based multi-head self attention (W-MSA) module and the shifted window-based multi-head self attention (SW-MSA) module are applied in the two successive transformer blocks, respectively. Consecutive Swin Transformer blocks are computed as:

$$\widehat{z}^{l} = W - MSA(LN(z^{l-1})) + z^{l-1}, \tag{10}$$

$$\widehat{z}^{l} = MLP(LN(z^{l})) + z^{l}, \tag{11}$$

$$\widehat{z}^{l+1} = SW - MSA(LN(z^{l})) + z^{l}, \tag{12}$$

$$\widehat{z}^{l+1} = MLP(LN(z^{l+1})) + z^{l+1}, \tag{13}$$

where $\widehat{z}^{l}$ and z^{l} denote the outputs features of the (S)W-MSA module and the MLP module for block l, respectively. W-MSA and SW-MSA denote window based multi-head self-attention using regular and shifted window partitioning configurations, respectively.

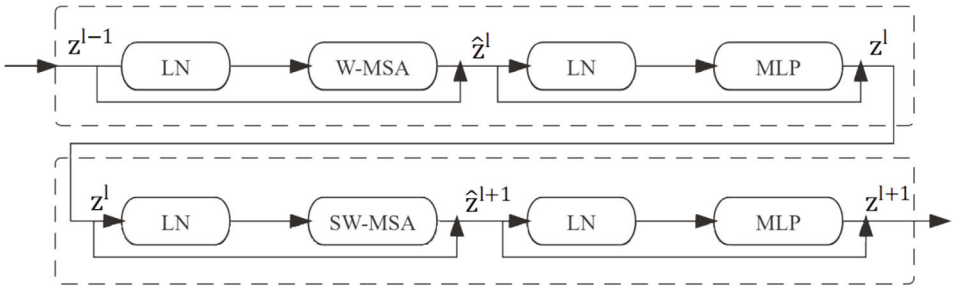

Figure 3 Swin transformer block architecture diagram

A standard transformer block includes two main modules: Self Attention (SA) module and Multi-Layer Perceptron (MLP) module. The input is $f_{in} \in R^{n \times d}$, the corresponding query, key, and value are calculated as:

$$f_Q = f_{in} W_Q, \quad f_K = f_{in} W_K, \quad f_V = f_{in} W_V, \tag{14}$$

where W_Q, W_K, W_V are queries, keys and values, respectively. The attention scores are computed as:

$$\boldsymbol{A} = softmax\left(\frac{f_Q f_K^T}{\sqrt{d}}\right). \tag{15}$$

Finally, the weighted sum of attention weights A and f_V are computed. The integrated features are:

$$f_{out} = \boldsymbol{A} f_V \cdot W_O, \tag{16}$$

where W_O are the projection matrix. The MLP module contains two linear layers parameterized by $W_1 \in R^{d \times (\varepsilon d)}$, $b_1 \in R^{\varepsilon d}$ and $W_2 \in R^{(\varepsilon d) \times d}$, $b_2 \in R^d$, respectively. ε is the expand ratio of MLP layers. Denote the input to MLP as $f_{in} \in R^n$, the output is:

$$f_{out} = GeLU(f_{in} W_1 + b_1) W_2 + b_2. \tag{17}$$

2.4 Deep Baseline Segmentation Model

Image segmentation is a hot topic in the field of deep learning. Representative baseline algorithms include U-Net, Link-Net, FPN, etc. The following briefly describes the architectural principles of these models.

U-type Convolutional Neural Network (U-Net)[28]: The method integrates shallow and deep feature information through upper and lower sampling and jump connection to expand the feature information. Thus, it reduces the training burden and make the edge information in the image more accurate. The U-Net network architecture is shown in Figure 4. It consists of encoder paths and decoder paths. The encoder path consists of the repeated application of two 3 × 3 convolutions. Each of the two convolutional layers was followed by a maximum pooling layer of 2 × 2 (stride 2), and a rectified linear unit (ReLU) for downsampling. In the upsampling of the decoding path, each step will have a 2 × 2 convolution layer and two 3 ×

3 convolution layers. Meanwhile, the upsampling of each step will add the feature graph from the relative strain contraction path. The last layer of the network is a 1 × 1 convolution layer. Through this operation, the feature vectors of 64 channels can be converted into the required number of classification results. Finally, the whole network of U-Net has 23 convolutional layers. U-Net can carry out convolution operation on pictures of any shape or size. Krois, et al.[29] used hyper-parameters search to adjust model architecture, and finally applied a seven-layer U-Net deep convolutional neural network (CNN) to detect periodontal bone loss (PBL) on panoramic radiographs. The results showed that CNN trained on a limited number of radiographic image segments showed at least similar discrimination to dentists assessing PBL on panoramic radiography. The amount of diagnostic work done by dentists using X-rays could be reduced by applying machine learn-based techniques.

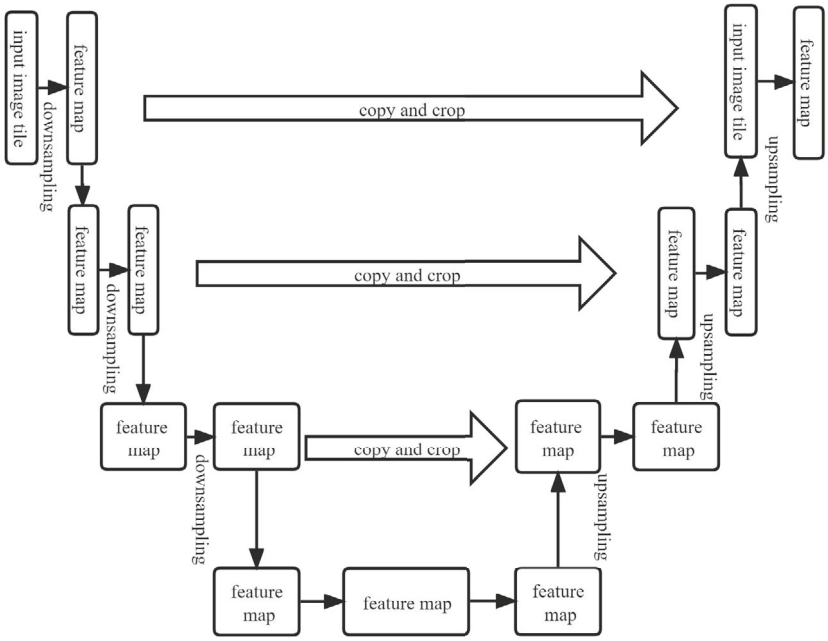

Figure 4 U-Net network architecture diagram

Link-Net: It is an efficient semantic segmented neural network with the advantages of jump connection, residual blocks, and encoder-decoder architecture. The original Link-Net uses the ResNet18 as its encoder. Link-Net showed high accuracy on multiple benchmarks and efficient computation. The Link-Net's network architecture is shown in Figure 5, consisting of an encoder and a decoder. The encoder starts with an initial block that convolves kernel input images of size 7 × 7, stride 2, or performs spatial maximum pooling (max-pooling) on regions of size 3 × 3, stride 2. The latter part of the encoder consists of a residual block, represented as the encoder block. The full convolution technique is used in the decoder[30]. At present, there are

few applications of Link-Net in tooth segmentation on panoramic radiography, but the studies for segmentation of chest radiographs[31].

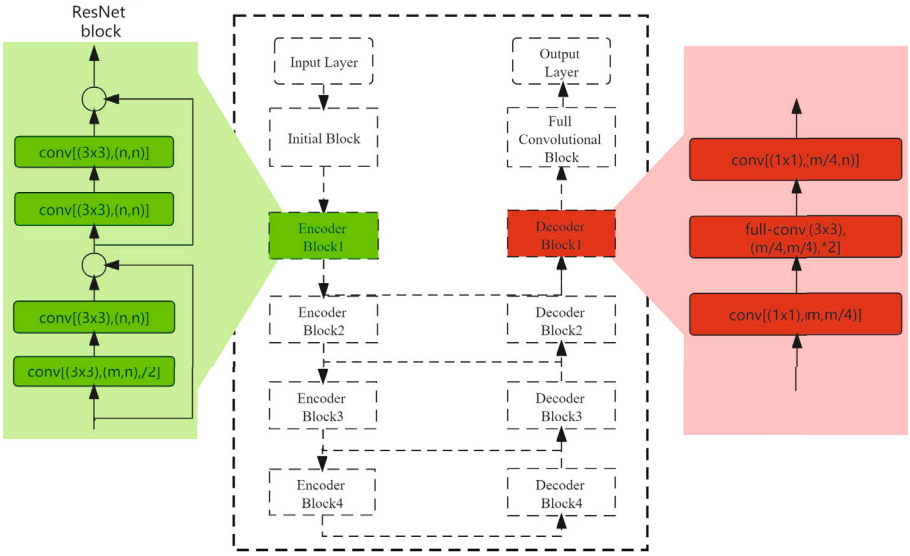

Figure 5 Link-Net network architecture diagram

Feature Pyramid network (FPN)[32]: It is a kind of high-level semantic feature map of various scales with top-down path and lateral connection proposed by utilizing the inherent multi-scale pyramid hierarchy structure of deep convolutional neural network, as shown in Figure 6. The structure of the feature pyramid mainly includes three parts: Bottom-up, top-down, and lateral connection. The process of bottom-up is the process of entering pictures into the backbone ConvNet to extract features. Some of the dimensions of the backbone output are unchanged, some are 2-fold. Take the layers with the output size as a stage, the output features of the last layer of each stage are extracted. The process of top-down is to sample the feature map and then pass it down. The purpose of upsampling is to enlarge the picture and insert the appropriate interpolation algorithm between the original image pixels. Lateral connection mainly includes three steps: 1×1 convolution reduces dimension for each stage output feature map, then fuses the obtained features with the feature maps sampled from the previous layer, and then convolute a 3×3 to obtain the feature output of this layer. Tooth segmentation based on cone-beam computed tomographic is a complex task, due to limited contrast and potential interference of various artifacts. Lahoud, et al.[33] proposed an artificial intelligence-driven tooth segmentation algorithm based on FPN, which realized automatic tooth detection and segmentation instead of artificial tooth contour localization.

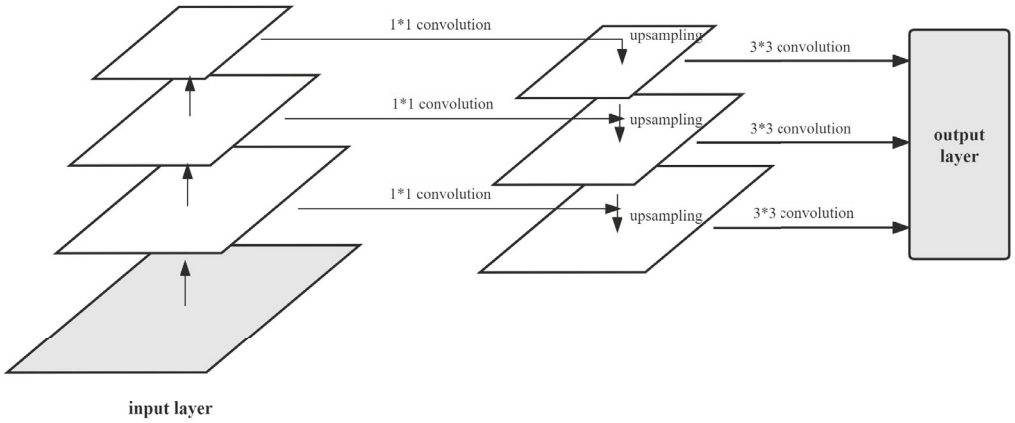

Figure 6 FPN network architecture diagram

3 Experiments

3.1 The PLAGH-BH Database

From January 2018 to January 2021, 100 panoramic radiographs were randomly selected from adult patients (aged 20-65 years) taken in the Department of Oral Radiology, PLA General Hospital, and a PLAGH-BH dataset was established. Among them, 90 panoramic radiographs were used for training and 10 for testing. Dental caries, periapical lesions, periodontal disease, various types of restorations (intracoronal, crowns, and bridges), orthodontic brackets and wires, and the third molars were included in the dataset. All panoramic radiographs were made using ORTHOPHOS XG 5 Ceph. Exposure conditions were set to 60 kV and 60 mA, according to the manufacturer's instructions. Exposure time was 14.1 seconds. All data are anonymized after the study was reviewed by the Ethics Review Committee of the PLA General Hospital.

3.2 PLAGH-BH Database Annotation

All raw images obtained by ORTHOPHOS XG 5 Ceph were 2440 × 1292 pixels in size. The tooth boundaries on panoramic radiographs were manually annotated by oral radiologists with more than 5 years of experience. Generate images with tooth labels based on the number of teeth on a panoramic radiographs. 90 panoramic radiographs in the dataset are used for training.

3.3 Implementation Details

The backbone of the three baseline algorithms was performed using the ResNet-50 network, while SWin-Unet used SWinNet. The input image size was set to 512 × 512, and the network used BCE (binary cross entropy) as a loss function, using the Adam optimization algorithm and a learning rate of 0.001 during learning, requiring approximately 160 epoch for network

convergence.

Convergence conclusion Based on our extensive experiments, we can observe that our deep method can converge to a local mimina with a better performance when compared with the sate-of-the-art methods on object segmentation.

4 Results

In the PLAGH-BH dataset, the performance evaluation of deep learning network models of U-Net, SWin-Unet, Link-Net and FPN in tooth image segmentation on the panoramic radiographs is shown in Table 1. Compared to the other three deep learning network models, SWin-Unet outperformed the others in terms of F1 score, IoU, and Acc. The segmentation results in the publicly available dataset panoramic dental X-rays with segmented mandibles (PDX)[34] are shown in Table 2. With the proposed segmented based method, we obtain the comparable results. Compared to U-Net and FPN, SWin-Unet outperforms both deep learning network models in F1 score, IoU, and Acc, but has a slightly lower F1 score, IoU, and Acc compared to Link-Net. The segmentation effect of SWin-Unet in the PLAGH-BH dataset, as shown in Figure 7.

Table 1 Results on the PLAGH-BH dataset

Method	F1-score	IoU	Acc
U-Net	0.5372	0.3674	0.8601
Link-Net	0.5731	0.4030	0.8718
FPN	0.3789	0.2407	0.8731
SWin-Unet	0.6372	0.4689	0.8852

Table 2 Segmentation results on the PDX dataset

Method	F1-score	IoU	Acc
U-Net	0.8542	0.7455	0.8837
Link-Net	0.8776	0.7820	0.8897
FPN	0.7948	0.6597	0.8695
SWin-Unet	0.8203	0.6956	0.8704

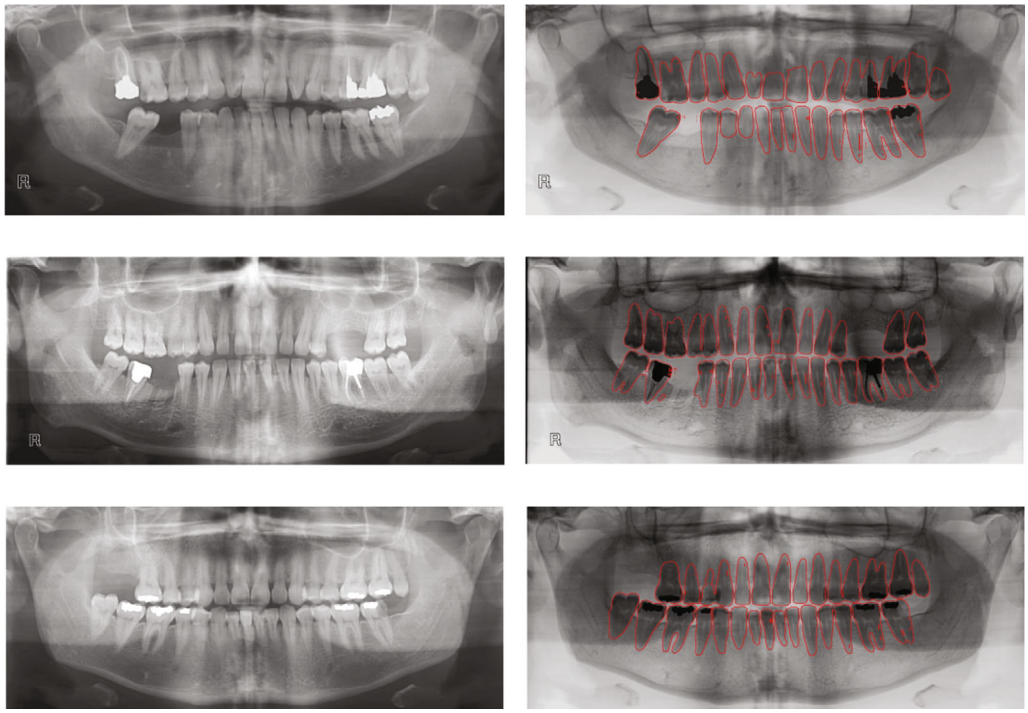

Figure 7 Visualization results of SWin-Unet on the PLAGH-BH dataset

5 Discussion

In the present work, four convolutional neural network-based segmentation models have made more accurate segmentation of tooth tissue images in the PLAGH-BH dataset. SWin-Unet model implemented by SwinNet backbone is used to train images with 512 × 512. According to the segmentation results, the performance evaluation F1 score is 0.6372, the mean IoU is 0.4689, and Acc is 0.8852. Compared with the performance evaluation indexes of three deep convolutional neural network learning models: U-Net, Link-Net, FPN, SWin-Unet obtains better performance. However, this differs from the results tested on the publicly available dataset, as shown in Table 2, where Link-Net U-Net performs better. In addition, the segmentation performance of the four deep convolutional neural network segmentation models in PLAGH-BH is worse compared with publicly available datasets, which may be related to randomly selected samples of the patients, the dentition presented by panoramic radiographs in different datasets were different, and there may be more factors affecting tooth segmentation of CNN in the PLAGH-BH dataset of our established dataset. Moreover, our dataset did not exclude wisdom teeth, whose position and shape vary from person to person, so there is not enough training data for the reliable shape model training. Therefore, its accurate segmentation is more difficult. Furthermore, tooth filling in the dataset also affects the accuracy of tooth segmentation, and also when the fillings do not suit the true profile of the teeth.

Panoramic radiographs are a commonly used auxiliary diagnostic tool in the oral clinic. The

dentist can diagnose the disease according to the image presented, and make corresponding treatment plans for the patients. In addition to showing the whole tooth structure in the oral cavity, the panoramic radiographs also shows the temporomandibular region, nose and facial bones. Because it is a two-dimensional image, it often causes overlapping anatomy and blurred the tooth edge, which poses a great obstacle to the diagnosis and work efficiency of dentists. Clinical reading depends on the dentists' experience, is often subjective, and the reading results vary from person to person. Therefore, the clinical application of automated diagnostic systems is particularly important. Accurate segmentation of tooth tissue has also become the basis of auxiliary diagnosis. Silva, et al.[35] applied traditional image segmentation methods to dental images, we evaluated the segmentation performance of different segmentation methods and the effect of X-ray types on the segmentation effect by the classification of segmentation methods and X-ray images. It is finally concluded that none of the studied algorithms enable a completely accurate segmentation of tooth tissue in the currently used dataset images, which may be associated with the low contrast of teeth and bone. The AI methods of deep learning show better performance in automated tooth segmentation compared to previous morphology methods. Wirtz, et al.[23] proposed an automated tooth segmentation method using a coupled shape model combined with a neural network model in panoramic radiographs. The results show that it performs better than traditional segmentation methods and it is able to handle difficulties in tooth loss, fracture, tooth filling. Even so, tooth segmentation on panoramic radiographs is still very difficult, because the panoramic image is not limited to a single portion. In addition, patients with tooth variation, missing, prosthetic artifacts, some other conditions of poor image quality will affect the segmentation of the image. Existing methods still have great room for improvement in the tooth segmentation of panoramic radiographs. Therefore, we still need to find a network model that is more suitable for the segmentation. The SWin-Unet and Link-Net models with the highest IoU scores are still of lower accuracy than most clinicians and are not suitable for separate analysis in the current state. A multi-source expansion of the training dataset is needed to improve the utility of the model. With the developement of deep CNN in recent years, improved inference accuracy is expected in the future through the adoption of updated architectures combined with hardware upgrades.

6 Conclusion

We successfully demonstrate the feasibility of the SWin-Unet deep learning convolutional neural network model for tooth segmentation on panoramic radiographs. We firstly introduce PLAGH-BH dataset. SWin-Unet shows superior segmentation performance. Our work highlights the potential of deep learning methods as an image segmentation tool for an objective, real-time diagnosis with minimal clinical labor requirements. Regarding the current shortcomings of deep learning in image segmentation, future work includes the use of a larger training image set to improve the segmentation accuracy and robustness.

References

[1] Perschbacher S, Interpretation of panoramic radiographs, *Australian Dental Journal*, 2012, **57**: 40–45.

[2] Kim J, Kim H, and Ro Y, Iterative deep convolutional encoder-decoder network for medical image segmentation, *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2017, 685–688.

[3] Zhao J, Ma Y, Pan Z, et al., Research on image signal identification based on adaptive array stochastic resonance, *Journal of Systems Science and Complexity*, 2022, **35**(1): 179–193.

[4] Wu C, Tsai W, Chen Y, et al., Model-based orthodontic assessments for dental panoramic radiographs, *IEEE Journal of Biomedical and Health Informatics*, 2017, **22**(2): 545–551.

[5] Ammar H, Ngan P, Crout R, et al., Three-dimensional modeling and finite element analysis in treatment planning for orthodontic tooth movement, *American Journal of Orthodontics and Dentofacial Orthopedics*, 2011, **139**(1): 59–71.

[6] Jiang Y, Qian J, Lu S, et al., LRVRG: A local region-based variational region growing algorithm for fast mandible segmentation from cbct images, *Oral Radiology*, 2021, **37**(4): 631–640.

[7] Wang T, Qiao M, Zhang M, et al., Data-driven prognostic method based on self-supervised learning approaches for fault detection, *Journal of Intelligent Manufacturing*, 2020, **31**(7): 1611–1619.

[8] Razali M, Ahmad N, Hassan R, et al., Sobel and canny edges segmentations for the dental age assessment, *Proceedings of International Conference on Computer Assisted System in Health*, 2014, 62–66.

[9] Pérez-Benito F, Signol F, Perez-Cortes J, et al., A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation, *Computer Methods and Programs in Biomedicine*, 2020, **195**: 105668.1–36.

[10] Bergeest J and Rohr K, Efficient globally optimal segmentation of cells in fluorescence microscopy images using level sets and convex energy functionals, *Medical Image Analysis*, 2012, **16**(7): 1436–1444.

[11] Gong X, Chen S, Zhang B, et al., Style consistent image generation for nuclei instance segmentation, *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2021, 3994–4003 .

[12] Mao M, Gao P, Zhang R, et al., Dual-stream network for visual recognition, *Proceeings of Advances in Neural Information Processing Systems*, 2021, 34–46.

[13] Esteva A, Kuprel B, Novoa R, et al., Dermatologist-level classification of skin cancer with deep neural networks, *Nature*, 2017, **542**(7639): 115–118.

[14] Wang T, Qiao M, Lin Z, et al., Generative neural networks for anomaly detection in crowded scenes, *IEEE Transactions on Information Forensics and Security*, 2018, **14**(5): 1390–1399.

[15] Leite A F, Van Gerven A, Willems H, et al., Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs, *Clinical Oral Investigations*, 2021, **25**(4): 2257–2267.

[16] Vinayahalingam S, Xi T, Bergé S, et al., Automated detection of third molars and mandibular nerve by deep learning, *Scientific Reports*, 2019, **9**(1): 1–7.

[17] Xu X, Liu C, and Zheng Y, 3D tooth segmentation and labeling using deep convolutional neural networks, *IEEE Transactions on Visualization and Computer Graphics*, 2018, **25**(7): 2336–2348.

[18] Van Eycke Y, Foucart A, and Decaestecker C, Strategies to reduce the expert supervision required for deep learning-based segmentation of histopathological images, *Frontiers in Medicine*, 2019, **6**: 222–231.

[19] Miotto R, Wang F, Wang S, et al., Deep learning for healthcare: Review, opportunities and challenges, *Briefings in Bioinformatics*, 2018, **19**(6): 1236–1246.

[20] Liu P, Song Y, Chai M, et al., Swin–unet++: A nested swin transformer architecture for location identification and morphology segmentation of dimples on 2.25 cr1mo0. 25v fractured surface, *Materials*, 2021, **14**(24): 7504.1–15.

[21] Luo C, Zhang J, Chen X, et al., UCATR: Based on CNN and transformer encoding and cross-attention decoding for lesion segmentation of acute ischemic stroke in non-contrast computed tomography images, *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021, **2021**: 3565–3568.

[22] Wang C, Huang C, Lee J, et al., A benchmark for comparison of dental radiography analysis algorithms, *Medical Image Analysis*, 2016, **31**(24): 63–76.

[23] Wirtz A, Mirashi S G, and Wesarg S, Automatic teeth segmentation in panoramic x-ray images using a coupled shape model in combination with a neural network, *Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention*, 2018, 712–719.

[24] Chan H, Samala R, Hadjiiski L, et al., Deep learning in medical image analysis, *Deep Learning in Medical Image Analysis*, 2020, **1213**: 3–21.

[25] Schwendicke F, Golla T, Dreher M, et al., Convolutional neural networks for dental image diagnostics: A scoping review, *Journal of Dentistry*, 2019, **91**: 103226.1–8.

[26] Goodfellow I, Bengio Y, and Courville A, *Deep Learning*, MIT Press, Cambridge, 2016.

[27] Zhang Y, Zhang S, Li Y, et al., Single- and cross-modality near duplicate image pairs detection via spatial transformer comparing CNN, *Sensors* (Basel), 2021, **21**(1): 255.

[28] Zhou Z, Siddiquee M M R, Tajbakhsh N, et al., Unet++: A nested u-net architecture for medical image segmentation, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, **11045**: 3–11.

[29] Krois J, Ekert T, Meinhold L, et al., Deep learning for the radiographic detection of periodontal bone loss, *Scientific Reports*, 2019, **9**(1): 1–6.

[30] Chaurasia A and Culurciello E, Linknet: Exploiting encoder representations for efficient semantic segmentation, *Proceedings of IEEE Visual Communications and Image Processing*, 2017, 1–4.

[31] Arora R, Saini I, and Sood N, Multi-label segmentation and detection of covid-19 abnormalities from chest radiographs using deep learning, *Optik*, 2021, **246**: 167780.1–18.

[32] Lin T, Dollár P, Girshick R, et al., Feature pyramid networks for object detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 2117–2125.

[33] Lahoud P, EzEldeen M, Beznik T, et al., Artificial intelligence for fast and accurate 3-dimensional tooth segmentation on cone-beam computed tomography, *Journal of Endodontics*, 2021, **47**(5): 827–835.

[34] Nishitani Y, Nakayama R, Hayashi D, et al., Segmentation of teeth in panoramic dental X-ray images using U-Net with a loss function weighted on the tooth edge, *Radiol Phys. Technol.*, 2021, **14**(1): 64–69.

[35] Silva G, Oliveira L, and Pithon M, Automatic segmenting teeth in x-ray images: Trends, a novel data set, benchmarking and future perspectives, *Expert Systems with Applications*, 2018, **107**15–31.