LABANET: LEAD-ASSISTING BACKBONE ATTENTION NETWORK FOR ORAL MULTI-PATHOLOGY SEGMENTATION

Huabao Chen[1]*, Xiaolong Huang*[2]*, Qiankun Li*[3]*, Jianqing Wang*[4]*, Bo Fang*[5]*, Junxin Chen*[6]

[1]College of Energy and Electrical Engineering, Hohai University, Nanjing, China
[2]School of Artificial Intelligence, Chongqing University of Technology, Chongqing, China
[3]Department of Automation, University of Science and Technology of China, Hefei, China
[4]School of Business and Management, Shanghai International Studies University, Shanghai, China
[5]College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
[6]School of Software, Dalian University of Technology, Dalian, China.

ABSTRACT

This paper presents a Lead-Assisting Backbone Attention Network (LABANet), which is able to perform multi-pathology instance segmentation of dental panoramic X-rays. A Lead-Assisting Attention Backbone (LAAB), containing two Swin-Transformers, is first developed for feature extraction. The following Region Proposal Network (RPN) and RoIAlign modules further convert the extracted features to a fixed-size feature map. Finally, an improved attention head with a Squeeze-and-Excitation (SE) block is constructed for object classification, bounding-box regression, and mask segmentation. By taking advantage of the global attention mechanism, the LABANet can better achieve multiple pathology segmentation. Experiment results demonstrate its effectiveness and advantages over state-of-the-art methods.

***Index Terms*—** Oral diseases, Multi-pathological segmentation, Lead-Assisting attention backbone, Global attention mechanism

1. INTRODUCTION

Oral diseases have been estimated to affect nearly half of the global population nowadays, among which 2.3 billion people suffer from permanent dental caries and the direct medical cost is 298 about billion dollars [1]. Oral panoramic radiographs (X-rays) play an important role in diagnosing dental disease that can be used by doctors to detect hidden tooth structures, bone loss, and other problems. However, the diagnosis of dental disease based on oral panoramic radiographs is an error-prone process as it is highly dependent on the experience of the dentist [2]. In addition, most of the patients have more than one oral disease because of the high correlation between oral diseases. Therefore, a reproducible and accurate multi-pathology segmentation method for oral panoramic radiographs is in a high demand.

In recent years, deep learning (DL) has achieved great success in medical imaging fields [3, 4]. Originating from the highly correlated nature of oral diseases, multi-pathological segmentation becomes a difficult but practically valuable topic. Some researchers tried to segment teeth and oral diseases with Convolutional Neural Networks (CNN), and promising results have been achieved. For example, Cantu $et\ al.$ [5] used U-Net to detect caries lesions on bitewings and found it significantly more accurate than dentists, and Chen $et\ al.$ [6] improved the accuracy of tooth segmentation from panoramic X-ray images by a novel multi-scale structural similarity (MS-SSIM) loss. To achieve accurate segmentation of caries lesions, Zhu $et\ al.$ [7] proposed a new CNN with a full-size axial attention module, called CariesNet. However, most existing works are based on single-pathology segmentation, and few of them studied the multi-pathology segmentation which is more clinically applicable.

This paper presents a Lead-Assisting Backbone Attention Network (LABANet), which enables multi-pathology instance segmentation of dental panoramic X-rays. Feature extraction of the input X-ray image is first performed by a Lead-Assisting Attention Backbone (LAAB), which is constructed by two Swin-Transformers [8]. The extracted features are then converted to a fixed-size feature map using the following Region Proposal Network (RPN) [9] and RoIAlign [9] modules. Finally, an improved attention head with a Squeeze-and-Excitation (SE) [10] block is constructed, which performs object classification, bounding-box regression, and mask segmentation with the fixed-size feature map. Because the LAAB is able to extract global features, the extracted feature map has rich contextual information. In addition, the improved attention head better echoes the global attention feature map extracted by the LAAB and renders the proposed LABANet great capability for accurate lesion segmentation. Validation experiments have been performed, and

Corresponding Author: Junxin Chen (junxinchen@ieee.org). This work is funded by the National Natural Science Foundation of China (No. 62171114) and the Fundamental Research Funds for the Central Universities (No. DUT22RC(3)099).

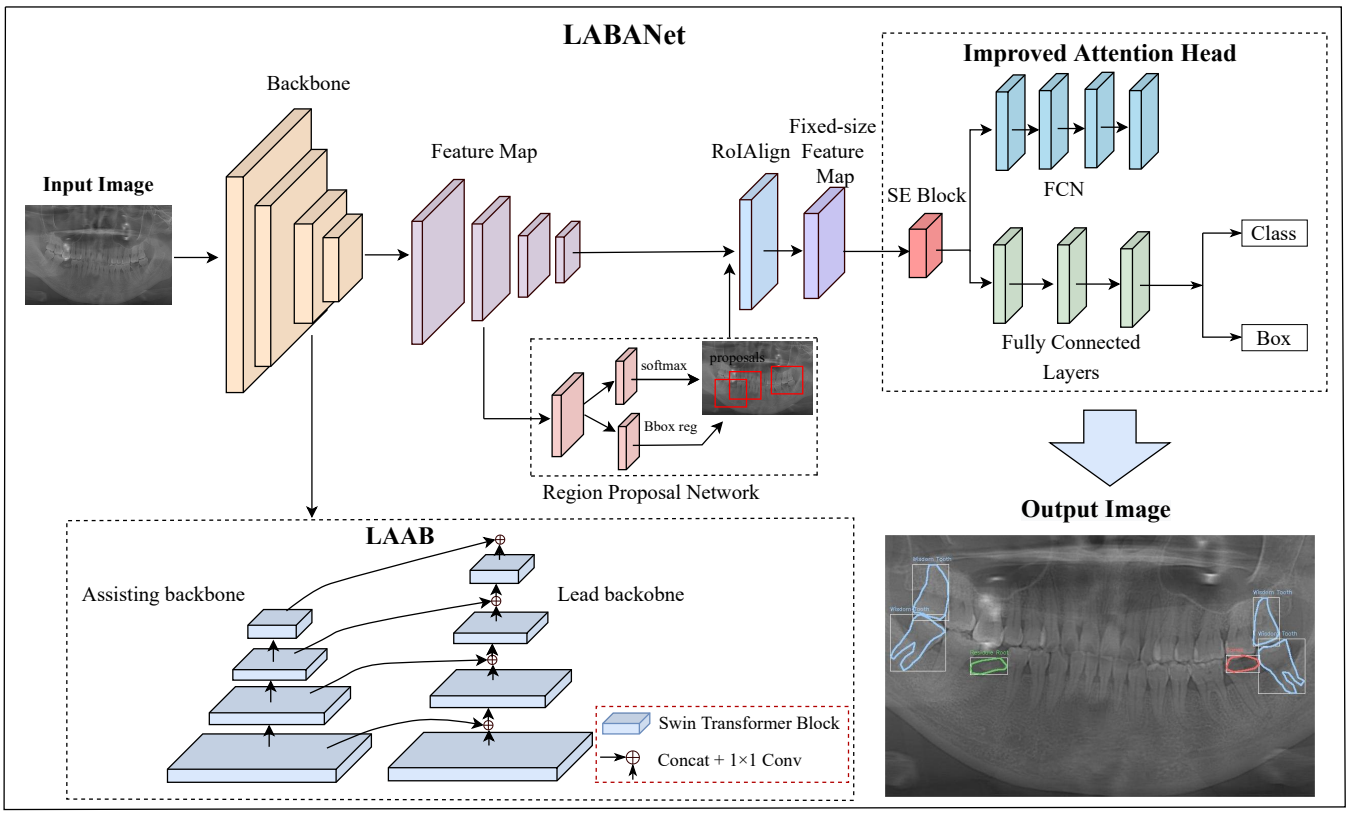

Fig. 1. The architecture of the LABANet.

the results well demonstrate the effectiveness and advantages of our method.

Our main contributions are three folds. 1) We propose the LABANet that has good contextual information extraction capability for multi-pathology dental panoramic images. 2) A SE block is added to the head for upgrading LABANet's global attention mechanism. 3) Validation experiments have been performed, and the results well demonstrate the performance and advantages of the LABANet on wisdom teeth, residual roots, and caries segmentation.

2. METHOD

2.1. Architecture of LABANet

The structure of the LABANet is demonstrated in Fig 1. It includes an LAAB for generating the feature map, RPN and RoIAlign for generating ROI on the fixed-size feature map, and an improved attention head for result prediction.

A dental panoramic X-ray image is first fed into the developed LAAB, which consists of two Swin Transformers, for feature extraction, and a feature map is generated in this step. Then, RPN and RoIAlign modules are employed to combine the feature map to generate ROI, and further convert it to a fixed-size feature map. Finally, the fixed-size feature map is input to the improved attention head with an SE block, recalibrating global features for object classification, bounding-box regression, and mask segmentation. In specific, fully connected layers are employed for object classification and bounding-box regression, while the Fully Convolutional Networks (FCN) is used for mask segmentation.

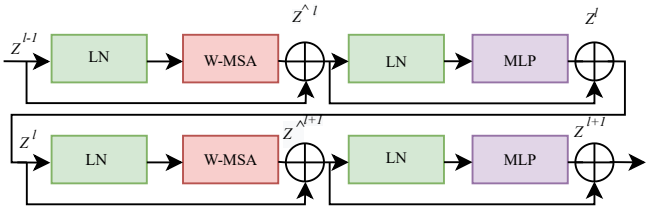

Fig. 2. The structure of the Swin Transformer.

2.2. LAAB

In CNN-based feature extraction, the introduction of the pooling layer [11] causes the loss of information. The Swin Transformer is a model based on the Encoder-Decoder framework without the pooling layer, it solves the inherent localization limitations of CNNs by introducing an attention mechanism. The structure of the Swin Transformer is illustrated in Fig. 2.

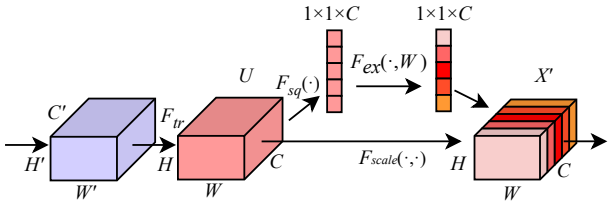

Fig. 3. The overview of the SE Block.

Inspired by [12], LAAB is developed in this paper to further enhance the backbone's global feature extraction capability. The LAAB includes a lead backbone and an assisting backbone, which are both based on Swin Transformer. Furthermore, a concatenate method is developed for connecting the lead and assisting backbones. Specifically, the two feature maps of size $H \times W \times C$ for the lead and assisting backbones in stage $(i-1)$ are connected on the channel. After that, the size of the combined feature map becomes $H \times W \times 2C$, and then the number of its channels becomes C by 1×1 convolution layer. The combined feature map is used as the input of stage i of the lead backbone. The operation is mathematically denoted as:

$$x_l^i = F_l^i(C(W(H(x_l^{i-1}, x_a^{i-1})))), 2 \leq i \leq 4, \quad (1)$$

where $H(\cdot)$ represents the concatenation, and function $W(\cdot)$ denotes a 1×1 convolution layer. $C(\cdot)$ denotes a convolution operation, $F(\cdot)$ realizes the feature aggregation mechanism with a convolution followed by a batch normalization and a ReLU activation function, and x_l^i represents the output at stage i of the lead backbone. Since Swin Transformer has 4 stages, $2 \leq i \leq 4$. For the subsequent segmentation process, only the output features of the lead backbone ($x_l^i, i = 2, 3, 4$) are fed into the next part, while the outputs of the assisting backbone are forwarded to the lead backbone.

2.3. Improved Attention Head

Considering the complexity of multi-pathology segmentation, the head of the segmentation network also needs to understand contextual information. Therefore, an improved attention header is proposed where the added SE block can perform global feature recalibration on the fix-sized feature map to accomplish object classification, bounding box regression, and mask segmentation. The structure of the SE block is shown in Fig. 3, which includes a Squeeze block and an Excitation block. The Squeeze block compresses global spatial information into channel descriptors. It generates channel-based statistics through global average pooling. Formally, a statistic $z \in \mathbb{R}^C$ is generated by shrinking U through its spatial dimensions $H \times W$. The c-th element of z is thus obtained by:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j). \quad (2)$$

To exploit the information aggregated in the squeeze operation, the excitation block uses a simple gating mechanism with sigmoid activation as:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \quad (3)$$

where δ refers to the ReLU function, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$.

The Excitation block parameterizes the gating mechanism by forming a bottleneck with two fully-connected (FC) layers around the non-linearity, i.e. a dimensionality-reduction layer with reduction ratio r, a ReLU, and then a dimensionality-increasing layer returning to the channel dimension of the transformation output U. The final output of the Excitation block is obtained by rescaling U with the activations s, that is

$$x_c^{'} = F_{scale}(u_c, s_c) = s_c u_c, \quad (4)$$

where $X^{'} = [x_1^{'}, x_2^{'}, ..., x_c^{'}]$ and $F_{scale}(u_c, s_c)$ refers to channel-wise multiplication between the scalar s_c and the feature map $u_c \in \mathbb{R}^{H \times W}$.

Through the SE block, the head part has attention mechanisms, which echo the global feature map extracted by the backbone. Therefore, fully connected layers and FCN can achieve better predictions when dealing with complex multi-pathological.

3. EXPERIMENTAL RESULT

3.1. Dataset and Implementation

The dataset contains 1000 panoramic images of the oral cavity. In these images, a total of 2177 areas were labeled as wisdom teeth, 716 areas were labeled as caries, and 822 areas were labeled as residual roots. The dataset is divided into a training set, a validation set, and a test set with a ratio of 8:1:1. The samples were ethically approved by the medical ethics committee of Pujiang Dental Hospital with the certification number 2022-0102.

The experiments are implemented using Nvidia Tesla A40 GPU with 48GB of memory. The code is implemented with MMDetection 2.14.0 in Ubuntu 16.04. The images are cropped to multiple sizes from 1200×2200 to 1600×2600 during the model training, using an Adam optimizer with the initial learning rate of $1e^{-4}$, batch size of 2, and 100 epochs.

3.2. Result and Comparison

The results of the LABANet and other state-of-the-art instance segmentation methods are listed in Table 1. Our method achieves a new state-of-the-art mean Dice value of 0.866. For wisdom teeth, residual roots, and caries, the proposed method achieves Dice values of 0.860, 0.903, and 0.836, respectively. It demonstrates that the proposed method has great segmentation accuracy in segmenting multi-lesion

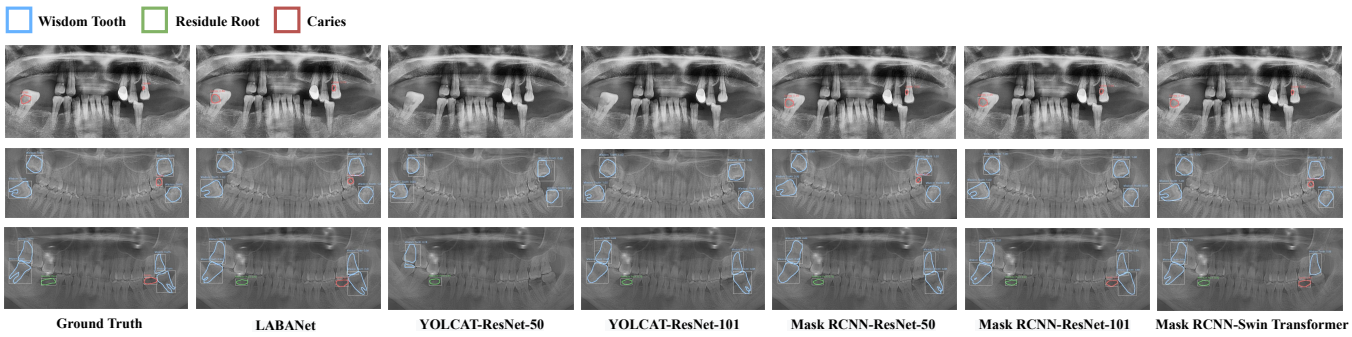

Fig. 4. Segmentation results of LABANet with other state-of-the-art methods on the oral panoramic radiograph dataset.

Table 1. Comparison of LABANet with state-of-the-art methods.

Model	Bbox mAP	Segm mAP	$Dice_{average}$	$Dice_{wt}$	$Dice_{rr}$	$Dice_{c}$
YOLACT [13]-ResNet-50	0.733	0.755	0.634	0.678	0.713	0.510
YOLACT [13]-ResNet-101	0.763	0.714	0.731	**0.891**	0.793	0.510
Mask RCNN [9]-ResNet-50	0.885	0.823	0.848	0.845	0.880	0.820
Mask RCNN [9]-ResNet-101	0.874	0.834	0.857	0.877	**0.904**	0.791
Mask RCNN [9]-Swin Transformer [8]	0.879	0.825	0.861	0.877	0.886	0.819
LABANet	**0.916**	**0.884**	**0.866**	0.860	0.903	**0.836**

regions. Regarding the bbox mAP and segm mAP metrics, the proposed method also achieves new state-of-the-art values of 0.916 and 0.884, respectively. The results well prove the performance advantages of the proposed LABANet on instance recognition and segmentation.

Fig. 4 shows the segmentation results. The LABANet can kindly find the small caries lesions from oral panoramic radiographs, whereas YOLCAT [13] performs poorly for the segmentation of small caries lesions. Compared with other methods, our proposal is able to find all the regions when there are multiple types of lesion regions, with relatively smooth segmentation contours and the most accurate segmentation results.

4. ABLATION STUDIES

Ablation studies have been conducted to verify the effectiveness of each proposed component, and the results are listed in Table 2. The ResNet-50 backbone with the original head is used as a baseline, which is based on the Mask RCNN [9] framework. The first row lists the results of the baseline, while the result of using the LAAB is listed in the second row. In the third row, the composite connection is further applied, and an improved attention head is added finally.

Comparing the first two rows, it can be seen that the LAAB's global feature extraction capability greatly improves the performance of segmentation (increasing Segm mAP by about 4%). Considering the second and third rows, we can see the positive effect of the composite connection. Moreover, the attention head has significantly improved the performance (increasing Segm mAP by about 2%), which can be indicated by comparing the last two rows.

Table 2. Ablation study of the LABANet.

Baseline	LAAB	Composite Connection	Attention Head	Segm mAP
✓				0.823
	✓			0.862
	✓	✓		0.867
	✓	✓	✓	**0.884**

5. CONCLUSION

In this paper, we proposed a Lead-Assisting Backbone Attention Network (LABANet), which enables multi-pathology instance segmentation of dental panoramic X-rays. In the proposed segmentation network, a Lead-Assisting Attention Backbone (LAAB) was first developed for global feature extraction. The following RPN and RoIAlign modules generated a fixed-size feature map with the extracted features, and the improved attention head with an SE block is constructed finally for object classification, bounding-box regression, and mask segmentation. Taking advantage of the global attention mechanism, the LABANet can better segment multi-pathology. Experimental results demonstrate its advantage over state-of-the-art methods.

6. REFERENCES

[1] GRF Collaborators et al., "A systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, 2018.

[2] Margrit-Ann Geibel, Steffen Carstens, Ulrike Braisch, Alexander Rahman, M Herz, and A Jablonski-Momeni, "Radiographic diagnosis of proximal caries—influence of experience and gender of the dental staff," *Clinical Oral Investigations*, vol. 21, no. 9, pp. 2761–2770, 2017.

[3] Parnian Afshar, Konstantinos N Plataniotis, and Arash Mohammadi, "Capsule Networks for brain tumor classification based on MRI images and coarse tumor boundaries," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1368–1372.

[4] Bo Fang, Junxin Chen, Wei Wang, and Yicong Zhou, "Combining multiple style transfer networks and transfer learning for lge-cmr segmentation," in *ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 1201–1205.

[5] Anselmo Garcia Cantu, Sascha Gehrung, Joachim Krois, Akhilanand Chaurasia, Jesus Gomez Rossi, Robert Gaudin, Karim Elhennawy, and Falk Schwendicke, "Detecting caries lesions of different radiographic extension on bitewings using deep learning," *Journal of Dentistry*, vol. 100, pp. 103425, 2020.

[6] Qiaoyi Chen, Yue Zhao, Yang Liu, Yongqing Sun, Chongshi Yang, Pengcheng Li, Lingming Zhang, and Chenqiang Gao, "MSLPNet: multi-scale location perception network for dental panoramic X-ray image segmentation," *Neural Computing and Applications*, vol. 33, no. 16, pp. 10277–10291, 2021.

[7] Haihua Zhu, Zheng Cao, Luya Lian, Guanchen Ye, Honghao Gao, and Jian Wu, "CariesNet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image," *Neural Computing and Applications*, pp. 1–9, 2022.

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin Transformer: hierarchical Vision Transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[10] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-Excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling, "CBNetV2: A composite backbone network architecture for object detection," *ArXiv preprint arXiv:2107.00420*, 2021.

[13] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee, "YOLACT: real-time instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9157–9166.