Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062400

ARTICLE

Teeth YOLACT: An Instance Segmentation Model Based on Impacted Tooth Panoramic X-Ray Images

Tao Zhou[1,2], **Yaxing Wang**[1,2,*], **Huiling Lu**[3], **Wenwen Chai**[1,2], **Yunfeng Pan**[1,2] **and Zhe Zhang**[1,2]

[1]School of Computer Science and Engineering, North Minzu University, Yinchuan, 750021, China
[2]Key Laboratory of Image and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan, 750021, China
[3]School of Medical Information & Engineering, Ningxia Medical University, Yinchuan, 750004, China

*Corresponding Author: Yaxing Wang. Email: wang_yaxing2024@163.com

Received: 17 December 2024; Accepted: 06 March 2025; Published: 19 May 2025

ABSTRACT: The instance segmentation of impacted teeth in the oral panoramic X-ray images is hotly researched. However, due to the complex structure, low contrast, and complex background of teeth in panoramic X-ray images, the task of instance segmentation is technically tricky. In this study, the contrast between impacted Teeth and periodontal tissues such as gingiva, periodontal membrane, and alveolar bone is low, resulting in fuzzy boundaries of impacted teeth. A model based on Teeth YOLACT is proposed to provide a more efficient and accurate solution for the segmentation of impacted teeth in oral panoramic X-ray films. Firstly, a Multi-scale Res-Transformer Module (MRTM) is designed. In the module, depthwise separable convolutions with different receptive fields are used to enhance the sensitivity of the model to lesion size. Additionally, the Vision Transformer is integrated to improve the model's ability to perceive global features. Secondly, the Context Interaction-awareness Module (CIaM) is designed to fuse deep and shallow features. The deep semantic features guide the shallow spatial features. Then, the shallow spatial features are embedded into the deep semantic features, and the cross-weighted attention mechanism is used to aggregate the deep and shallow features efficiently, and richer context information is obtained. Thirdly, the Edge-preserving perception Module (E2PM) is designed to enhance the teeth edge features. The first-order differential operator is used to get the tooth edge weight, and the perception ability of tooth edge features is improved. The shallow spatial feature is fused by linear mapping, weight concatenation, and matrix multiplication operations to preserve the tooth edge information. Finally, comparison experiments and ablation experiments are conducted on the oral panoramic X-ray image datasets. The results show that the APdet, APseg, ARdet, ARseg, mAPdet, and mAPseg indicators of the proposed model are 89.9%, 91.9%, 77.4%, 77.6%, 72.8%, and 73.5%, respectively. This study further verifies the application potential of the method combining multi-scale feature extraction, multi-scale feature fusion, and edge perception enhancement in medical image segmentation, which provides a valuable reference for future related research.

KEYWORDS: The oral panoramic X-ray; instance segmentation; impacted teeth; vision transformer; the edge-preserving

1 Introduction

Human teeth are composed of the crown, root, and neck [1]. With the improvement of human living standards and the increase in fine food, the total amount of human oral chewing activity has decreased. The chewing function is degenerated. The jawbone is insufficient to grow. The teeth eruption is limited. These lead to a high incidence of impacted teeth [2]. According to the relationship between the long-axis

impacted tooth and the long-axis second molar, the impacted tooth can be divided into the vertical impacted tooth, the horizontal impacted tooth, the mesial impacted tooth, the distal impacted tooth, the inverted impacted tooth, and the buccal impacted tooth. The representative diagram is shown in Fig. 1, and the specific description is shown in Table 1. Impacted teeth not only cause a variety of oral diseases, such as caries, gingivitis, periodontitis, and odontogenic tumors, but also are often accompanied by dental diseases, such as dentition defects, tooth loss, misaligned dentition, adjacent tooth displacement, and tooth shelter each other. These have a negative impact on the life and mental health of the patient [3]. The representative figure is shown in Fig. 2, and the specific details are shown in Table 2.

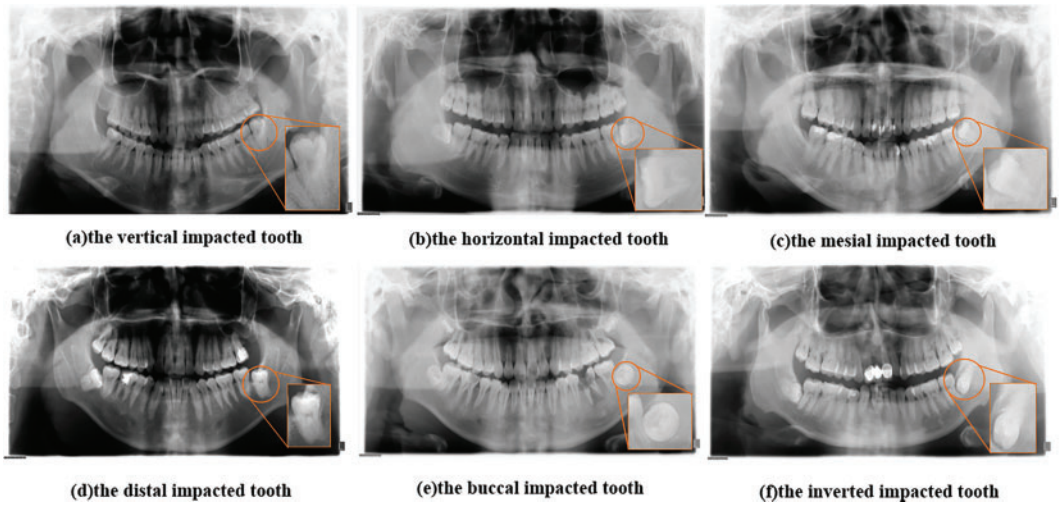

Figure 1: A typical example of the long-axis impacted tooth and the long-axis second molar

Table 1: Specific explanation about the long-axis impacted tooth

Num.	Type	Explanation
(a)	The vertical impacted tooth	The impacted tooth root is toward the jawbone base, and the crown is toward the vertical direction
(b)	The horizontal impacted tooth	The impacted tooth root is toward forward or backward, and the crown is toward horizontal direction
(c)	The mesial impacted tooth	The impacted tooth root is toward the midline or mesial line, and the crown is toward the mesial line of the adjacent tooth
(d)	The distal impacted tooth	The impacted tooth root is toward the midline or distal line, and the crown is toward the distal line of the adjacent tooth
(e)	The buccal impacted tooth	The impacted tooth root is toward the buccal side, and the crown is toward the oral interior
(f)	The inverted impacted tooth	The impacted tooth root is toward the top, and the crown is toward the bottom

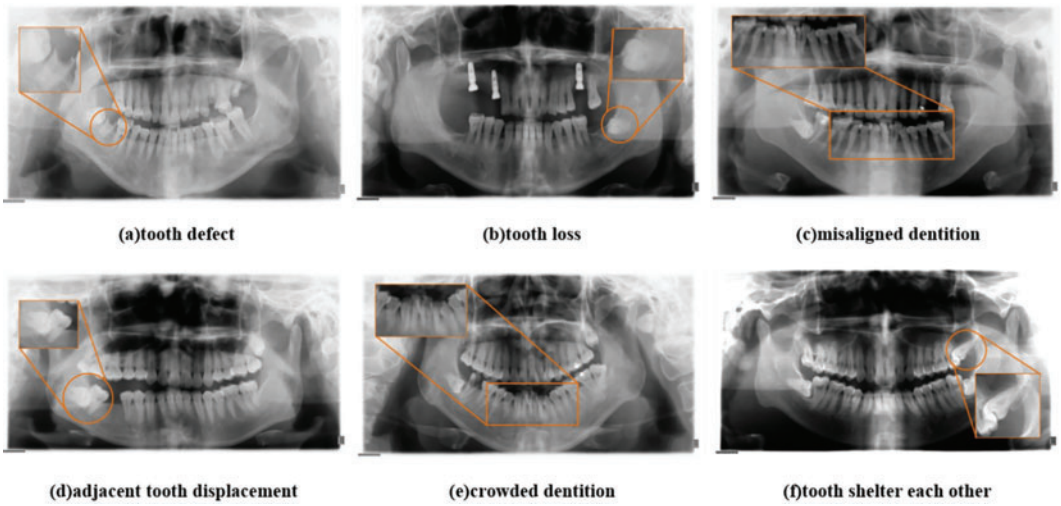

Figure 2: A typical example of accompanying symptoms of the impacted tooth

Table 2: A specific description of the accompanying symptoms of impact

Num.	Type	Explanation
(a)	Tooth defect	The tooth surface is damaged or has a partial tooth defect
(b)	Tooth loss	Lose one or more teeth
(c)	Misaligned dentition	Teeth are not aligned or irregular
(d)	Adjacent tooth displacement	The adjacent teeth of the impacted teeth are missing
(e)	Crowded dentition	There is not enough space among the teeth, resulting in tight alignment or overlapping
(f)	Tooth shelter each other	The teeth are overlapped or misaligned so that they cannot bite each other

The X-ray is widely used in the medical field, such as pneumonia X-ray images [4], dental X-ray images and so on. The oral panoramic X-ray images can show dentition, alveolar bone, tooth root, dental periodontal, and jawbone joints [5]. Therefore, the panoramic X-ray image is the preferred medical image for dentists to examine Teeth. Firstly, Dentists analyze panoramic X-ray dental images to check the status of teeth, such as position, number, arrangement, and morphology, in order to assess the teeth's state and the existing problems. Secondly, Dentists assess the state of periodontal tissue by observing the gingival, periodontal membrane, alveolar bone, and other periodontal tissues. Thirdly, the morphology, length, curvature of the root, and condition of the root canal are detected to evaluate whether the tooth apex is infected or diseased. Fourthly, adjacent structures such as temporomandibular joint, temporalis muscle, and masseter muscle are examined to assess whether there is dysfunction. Finally, based on the above indicators, combined with the patient's clinical symptoms and oral examination results, the corresponding therapeutic schedule is formulated. This process not only takes a lot of time and energy but also requires a lot of professional knowledge and experience from the doctor. With the successful application of artificial intelligence technology in Computer-Aided Diagnosis [6], the doctor's burden is greatly reduced. At present, how to apply artificial intelligence technology to improve the precision and efficiency of dental disease diagnosis is becoming a research hotspot.

In recent years, many important advances have been made in deep learning-based approaches, especially in the field of medical image segmentation. A tooth segmentation network based on double auxiliary information is proposed by Lu et al. [7]. The network innovatively combines multi-modal feature fusion and an adaptive weighting mechanism, which significantly optimizes the extraction ability of tooth boundaries and achieves excellent segmentation results in complex backgrounds. An automatic tooth segmentation method based on a convolutional neural network was proposed by Wang et al. [8]. This method can efficiently extract tooth features in intraoral scanning images and perform well in capturing tooth boundaries by designing a multi-layer convolutional network. Jader et al. [9] used Mask Region-based Convolutional Neural Network (Mask R-CNN) [10] for instance segmentation in panoramic X-ray images. This experiment proved the effectiveness and potential application of instance segmentation for impacted tooth segmentation. Graham et al. [11], through the combination of a dilated convolution layer and other neural network structures, realized the segmentation of gland instances in colon tissue images while minimizing the loss of information. Research shows that the traditional neural network with extended convolution can obtain more accurate segmentation results. Bhatti et al. [12] proposed a Mask R-CNN model embedded in a feature pyramid network for instance segmentation of breast X-ray images. This method replaces the feature recommendation network of the Mask R-CNN model with the Feature Pyramid Network (FPN) [13] to generate higher-quality feature maps, effectively improving the performance of lesion detection. Wang et al. [14] proposed the Anchor-Free Polyp Mask (AFP Mask) model, which proposes a method that does not rely on predefined anchor boxes to achieve local localization and instance segmentation of polyps and can quickly and accurately segment lesion areas. The above studies show that instance segmentation provides important support for accurate lesion localization and segmentation, auxiliary diagnosis, and so on. However, there are still some issues with instance segmentation about impacted tooth in oral panoramic X-ray images: (1) Due to the different position and inclination Angle of the impacted tooth in the dental arch, it can lead to dentition defect, tooth defect, malocclusion, adjacent tooth displacement, and tooth overlap. The spatial relationship and connection structure between impacted teeth and normal teeth are complex. (2) Due to the high density and component similarity of oral tissues, the contrast between the impacted tooth and periodontal tissues, such as gingiva, periodontal membrane, and alveolar bone, is low in oral panoramic X-ray images, resulting in unclear boundaries of the impacted tooth.

Aiming to the above problems, this paper proposes the Teeth You Only Look at Coefficients (YOLACT) model for instance segmentation of panoramic X-ray tooth images. The main contributions are as follows:

1. The MRTM is designed, in the module, the depth-wise separable convolution with different perceptive fields and Residual Block (Res_block) is used to improve the model sensitivity to the lesion size, the Vision Transformer are used to improve the model perception ability about global features, it makes the model more effective in extracting the different position and inclination angle of impacted tooth feature in the dental arch.
2. The CIaM is designed to fuse deep and shallow features. The shallow spatial features are guided by the deep semantic features. Then, the shallow spatial features are embedded into the deep semantic features, and the cross-weighted attention mechanism is used to efficiently aggregate the deep and shallow features, and richer context information is obtained. The model can better understand the association between impacted teeth and periodontal tissue and improve the precision and efficiency of impacted teeth segmentation and detection.
3. The E2PM is designed to enhance the teeth edge features. The first order differential operator is used to get the tooth edge weight, perception ability of tooth edge features is improved. The shallow spatial feature is fused by linear mapping, weight concatenation and matrix multiplication operations to

preserve the tooth edge information. The discrimination between impacted teeth and periodontal tissue is improved to achieve more precise and fine segmentation of impacted teeth.

2 Related Work

Compared with the two-stage instance segmentation model, YOLACT, as a single-stage case segmentation model, is characterized by fast reasoning speed, high detection precision, and accurate segmentation. Therefore, the YOLACT model is the preferred scheme in this paper. Due to the excellent performance and wide range of applications of the Transformer model, it has achieved good results especially in the field of instance segmentation.

2.1 Instance Segmentation Method Based on YOLACT

The YOLACT model achieves good results in the medical field. Many researchers have improved the structure of YOLACT to adapt it to more medical tasks. Better Real-time Instance Segmentation (YOLACT++) is proposed by Bolya et al. [15]. Deformable convolution is applied to the feature extraction network. The ability of the model to extract instances with different scales, aspect ratios, and rotations is enhanced. The YOLACT model of the two-tower structure was designed by Jia et al. [16]. This network introduced the bottom-up path enhancement module. There are feature pyramid network modules in YOLACT that are paired with this module. The aim is to shorten the transmission path, and the low-level and high-level features extracted from pathology images are fused. The Model performance is improved. A double attention network based on YOLACT++ architecture was introduced by Roy et al. [17]. Brain tumor scans MRI (Magnetic resonance imaging) images are used for detection and segmentation. To sum up, there are good prospects for the development of the YOLACT model in the medical field. In this paper, the Teeth YOLACT model is designed to segment impacted teeth. The structure is shown in Fig. 3.

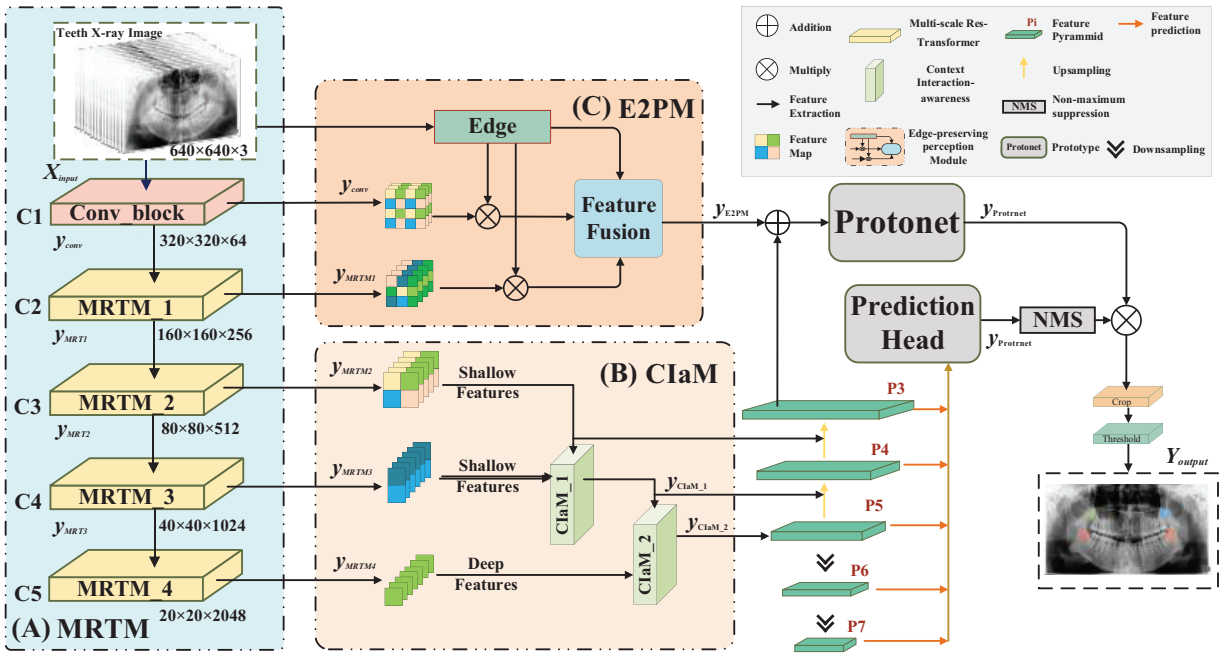

Figure 3: Overall structure of Teeth YOLACT model

2.2 Instance Segmentation Method Based on Transformer

The multi-head self-attention is exploited by the Transformer [18]. The ability to extract global features is enhanced, and an important role is played in the instance segmentation. The Transformer module is applied in instance segmentation by Guo et al. [19]. The global information in the image is better captured by the method. The global information in the image is better captured by the method, and there is a stronger feature representation ability. The performance of object detection and instance segmentation is improved. The Transformer module dealing with masking attention is designed by Cheng et al. [20]. The model uses masking attention in the Transformer decoder to limit attention to local features centered on the predicted segment. Shi et al. [21] proposed Robust Foveal Visual Perception for Vision Transformers (TransNeXt), which introduces a biomimetic-designed Token mixer that simulates the foveal vision and continuous eye movement mechanism in human biology, enabling global perception for each Token on the feature map. To sum up, Transformer possesses the advantages of global information capture ability, rich feature representation learning, and context awareness. In many application scenarios, Transformer-based instance segmentation models perform very well. Therefore, the multi-scale Res-Transformer module is designed in this paper. The structure is shown in Fig. 4.

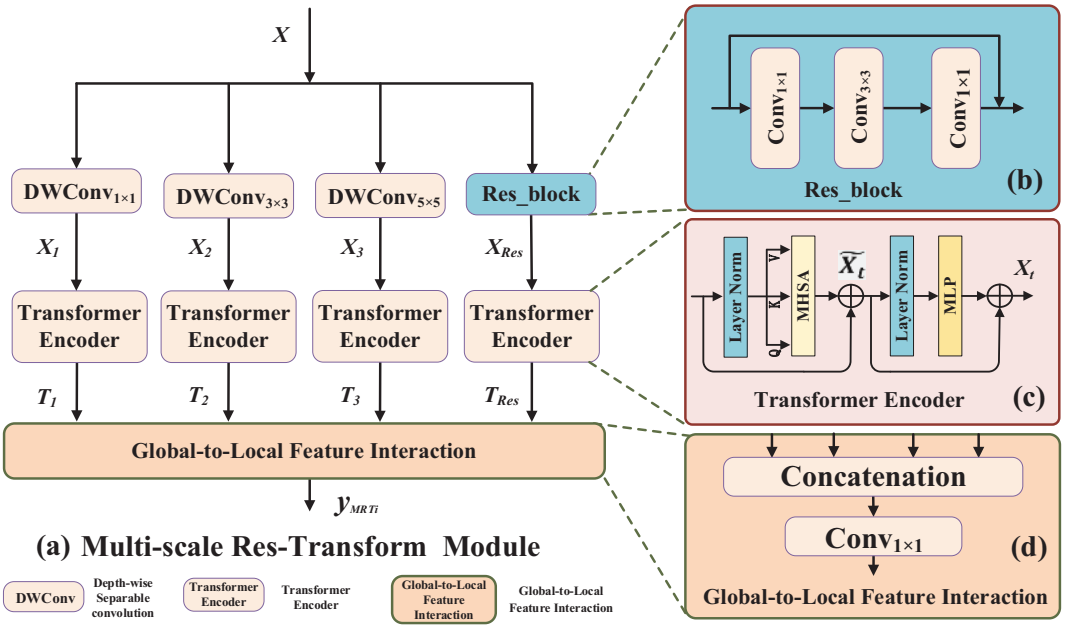

Figure 4: Structure of multi-scale Res-Transformer module

3 Method

It is a challenging task to achieve precise instance segmentation of impacted teeth in oral panoramic X-ray images. Therefore, the Teeth YOLACT model is proposed in this paper. The model is based on the YOLACT model. In order to fully extract effective features and make the model more precise in recognizing and location-impacted teeth, the MRTM is designed. In order to obtain more rich context information and make the model better understand the relationship between impacted tooth and periodontal tissue, the CIaM is designed. In order to enhance tooth contour information and make the model distinguish impacted teeth from periodontal tissue more accurately, the E2PM is designed. The model structure is shown in Fig. 3. The Teeth YOLACT pseudo-code is shown in Algorithm 1. There are five stages from C1 to C5 in the feature

extraction network, as shown in Fig. 3A. In the network, the output of C1 is the input of C2, the output of C2 is the input of C3, and so on until the end of C5. There are 7 × 7 convolution, batch normalization, ReLU (Rectified Linear Unit) activation functions, and Max-pooling in the C1 convolution module. Three parts are included, from C2 to C5. In the first part, multi-scale local features are extracted by the depth-wise separable convolution with different sizes and Res_block [22]. In the second part, global features are extracted in the Transformer encoder. In the third part, the global feature and local features are interacted in the channel dimension. Among these, the number of stacks of Res_block in C2 to C5 is 3, 4, 6, and 3.

Algorithm 1: Teeth YOLACT pseudo-code

Input: X_{input}

Output: Y_{output}

Multi-Scale Res-Transformer:

$y_{Conv} = Res\left(X_{input}\right); y_{MRT1} = MRT_1\left(y_{Conv}\right); y_{MRT2} = MRT_2\left(y_{MRT1}\right);$
$y_{MRT3} = MRT_3(y_{MRT2}); y_{MRT4} = MRT_4(y_{MRT3});$

Context Interaction-awareness Module:

$y_{CIaM1} = CIaM_1(y_{MRT3} \otimes y_{MRT2})); y_{CIaM2} = CIaM_2(CIaM_1 \otimes y_{MRT4}));$

Edge-preserving perception Module:

$y_{E2PM} = Fusion((SA(Canny(X_{input}) \otimes y_{MRT1}) \otimes SA(Canny(X_{input})) \otimes y_{Conv}) \otimes SA(Canny(X_{input})));$

Feature Pyramid Networks:

$P5 = y_{CIaM2}; P6 = UpSampling\left(P5\right); P7 = UpSampling\left(P6\right); P4 = \text{Down}Sampling\left(P5\right);$
$P3 = \text{Down}Sampling\left(P4\right);$

Feature Prediction Networks:

$y_{prediction} = Conv\left(P3\right) \oplus Conv\left(P4\right) \oplus Conv\left(P5\right) \oplus Conv\left(P6\right) \oplus Conv\left(P7\right);$
$y_{protonet} = y_{E2PM} \oplus y_{P3}; Y_{output} = Threshold\left(Crop\left(NMS\left(y_{prediction}\right) \otimes y_{protonet}\right)\right);$

END

The CIaM module consists of CIaM_1 and CIaM_2 in series, as shown in Fig. 3B. Firstly, the y_{MRTM2} and y_{MRTM3} generated by the C3 layer and C4 layer are input into CIaM _1, and the deep feature and the shallow feature are interacted by the cross-attention mechanism. Secondly, the y_{MRTM3} and y_{MRTM4} produced by CIaM_1 and C5 are input into CIaM_2 to perform interactions between the deep feature and the shallow feature. Finally, the output of y_{CIaM1} and y_{CIaM2} are transmitted to FPN.

The E2PM module is mainly composed of an edge extraction module and a feature fusion module, as shown in Fig. 3C. Firstly, the edge of the original image is extracted by the Canny operator, and the obtained edges are used to enhance the tooth edges. Secondly, these weighted feature maps are fused to retain the tooth edge information. Finally, the y_{E2PM} and P3 are added by element-by-element, the results are input into Protonet to obtain $y_{protonet}$.

3.1 Multi-Scale Res-Transformer Module (MRTM)

In panoramic X-ray images, due to the spatial relationship and connection structure being complex between the impacted teeth and normal teeth, it is difficult to extract sufficient lesion features. Therefore, the depth-wise separable convolution of different perceptive fields and the Res_block with translation invariance are used to effectively extract impacted tooth local features with different shapes and sizes. The global association between the impacted teeth and periodontal tissue is established by the self-attention mechanism in the Transformer, and global features are obtained by capturing long-term dependencies. The MRTM

is designed to compose a feature extraction network in this paper, as shown in Fig. 4a. The module is a four-branch structure, which consists of depth-wise separable convolutions with three different convolution kernels and Res_block, where Res_block is shown in Fig. 4b. The Transformer encoder is connected to each branch. X_i ($i \in \{1, 2, 3, Res\}$) is processed by the Transformer encoder to obtain T_i ($i \in \{1, 2, 3, Res\}$), where the Transformer encoder is shown in Fig. 4c. T_i ($i \in \{1, 2, 3, Res\}$) are inputted into the global-local feature interaction part. In this part, local features and global features are fused by channel-level concatenation, where the global-local feature interaction module is shown in Fig. 4d.

$$X_1 = DWconv_{1\times1}(X) \tag{1}$$

$$X_2 = DWconv_{3\times3}(X) \tag{2}$$

$$X_3 = DWconv_{5\times5}(X) \tag{3}$$

$$X_{Res} = Conv_{1\times1}(X) \tag{4}$$

where X is the input feature map and X_{Res}, X_1, X_2, and X_3 are the output vectors.

Step 1: The panoramic X-ray image is inputted into the C1 layer to extract features. In C2–C5, the depth-wise separable convolution with 3 different convolution kernels of 1 × 1, 3 × 3, and 5 × 5, Res_block, is included to extract local features at different scales.1 × 1 convolution is used to extract lesion features with small size, 3 × 3 convolution is used to extract lesion features with middle size, and 5 × 5 convolution is used to extract lesion features with large size. The local feature extraction capability is further improved by the parallel Res_block so that the module can be used more fully.

Step 2: The obtained X_i ($i \in \{1, 2, 3, Res\}$) are used to input into the corresponding Transformer encoder. The Transformer encoder models the feature correlation at different locations by self-attention mechanism and multi-layer perceptron. Operation steps: Firstly, input vectors X_i ($i \in \{1, 2, 3, Res\}$), W_q, W_k, and W_v are converted into three different vectors: query matrix (Q), key matrix (K), and value matrix (V). They are divided into multiple heads; Attention weights are calculated by an activation function. Secondly, the attention weights are multiplied with parameter matrices. The feature attention is computed for each head output. Finally, the feature attention output of MHSA is obtained by concatenating the feature attention output of each head.

To sum up, Formula (5) represents the process by which the input vector is processed by Layer Normalization (LN), Multi-Head Self-Attention (MHSA), and residual. The Formula (6) represents the process by which the input vector is processed by LN, Multilayer Perceptron (MLP), and residual.

$$\widetilde{X}_t = MHSA(LN(X_{t-1})) + X_{t-1} \tag{5}$$

$$X_t = MLP\left(LN\left(\tilde{X}\right)\right) + \widetilde{X}_t \tag{6}$$

Step 3: T_i ($i \in \{1, 2, 3, Res\}$) are inputted into the global local feature interaction module, perform feature Concatenation through concatenation operation, and the channel is adjusted by 1 × 1 convolution operation to obtain y_{MRTi} ($i \in \{1, 2, 3, Res\}$), as shown in Formula (7).

$$y_{MRTi} = Conv_{1\times1}([T_1; T_2; T_3; T_{Res}]) \tag{7}$$

3.2 Context Interaction-Awareness Module (CIaM)

During network training, shallow features contain low-level spatial information, such as contour and edge details, but lack sufficient semantic information. With the increment of the network depth, the model can extract richer semantic features, such as whether the tooth is an impacted tooth or the connection

between the impacted tooth and the periodontal tissue. However, spatial information is not enough. In order to obtain richer contextual features and better understand the association between impacted teeth and periodontal tissues, it is necessary to efficiently aggregate deep semantic information and shallow spatial details. Therefore, this paper designs the CIaM, as shown in Fig. 5. In this module, the shallow spatial features are guided by the deep semantic features. Then, the shallow spatial features are embedded into the deep semantic features. The cross-weighted attention mechanism is used to aggregate the deep and shallow features efficiently, and richer context information is obtained. The model can improve precision and efficiency.

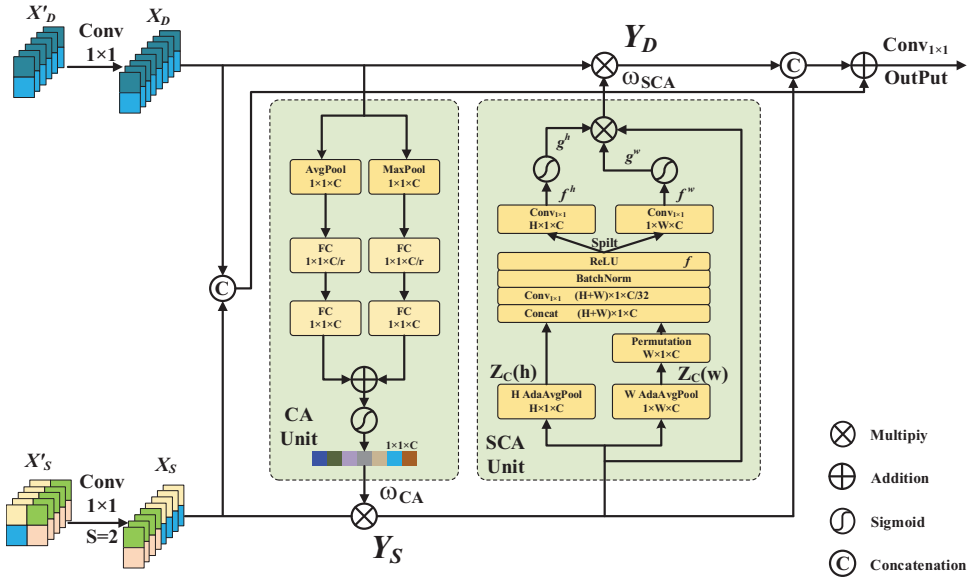

Figure 5: Structure of context interaction-awareness module

The Channel Attention (CA) Unit is adopted in this module. The compressed channel attention in deep features is used to reflect the channel's importance and focuses on the global context to provide semantic information. Firstly, X_D uses max pooling and average pooling to generate C-dimension vectors, which capture channel correlation through two consecutive Fully Connected (FC) layers, respectively. The first FC layer reduces the feature dimension to C/r (r = 16). After the ReLU activation function, the second FC layer is used to restore the channel dimension. Secondly, the two are added element by element, and the Sigmoid function normalizes the feature values to obtain the channel attention weight ω_{CA}. Finally, the shallow features are weighted by multiplying X along the channel dimension, and the calculation process of ω_{CA} is shown in Formula (8).

$$\omega_{CA} = \sigma\left\{FC\left(\delta\left(FC\left(MaxPool\left(X_D\right)\right)\right)\right) \bigoplus FC\left(\delta\left(FC\left(AvgPool\left(X_D\right)\right)\right)\right)\right\} \tag{8}$$

For the Spatial Coordinate Attention (SCA) Unit, the spatial coordinate attention map is generated from the shallow feature map, the goal of which is to enable the network to focus on the lesion region by calculating the spatial location weight. In horizontal and vertical directions, the shallow feature maps are pooled using the pooling kernel of (H, 1) and (1, W), and the feature maps $Z_C(h)$ and $Z_C(w)$ are obtained.

The transpose operation of $Z_C(w)$ is concatenated with $Z_C(h)$. Then, the channels are converted using 1×1 convolutions. The calculation process is shown in Formula (9).

$$f = \sigma(BN(Conv_{1\times1}([Z_C(w); Z_C(h)])) \tag{9}$$

where $f \in R^{\frac{C}{r}\times(H+W)}$ is the intermediate feature map, r is the down-sampling ratio.

In the horizontal direction, a 1×1 convolution is performed on f to recover the channel dimension and obtain the tensor $f^h \in R^{\frac{C}{r}\times H}$; after the Sigmoid function, the spatial attention weight is obtained g^w; in the vertical direction, a 1×1 convolution is applied to f to recover the channel dimension and obtain the tensor $f^w \in R^{\frac{C}{r}\times W}$, After the Sigmoid function, the spatial attention weight is obtained g^w. Finally, multiply g^h, g^w, and Y_S point-wise to get Y_D. The calculation process of g^h and g^w is shown in Formulas (10) and (11). The Y_D calculation process is shown in Formula (12).

$$g^h = \sigma\left(Conv_{1\times1}\left(f^h\right)\right) \tag{10}$$

$$g^w = \sigma(Conv_{1\times1}(f^w)) \tag{11}$$

$$Y_D = Y_S \otimes g^h \otimes g^w \tag{12}$$

3.3 Edge-Preserving Perception Module (E2PM)

The degree of contrast between the impacted teeth and periodontal tissues, such as gingiva, periodontal membrane, and alveolar bone, is low in oral panoramic X-ray images. As a result, the boundary of the impacted tooth is not clear. To improve the perceptual ability about tooth contour and discrimination ability between impacted teeth and periodontal tissue, realizing more accurate and finer impacted tooth segmentation. Therefore, E2PM is designed in this paper, as shown in Fig. 6. In the module, the first-order differential operator is used to obtain the tooth edge weight, and the perception ability of tooth edge features is improved. The shallow spatial feature is fused by linear mapping, weight concatenation, and matrix multiplication operations, which are used to preserve the tooth edge information.

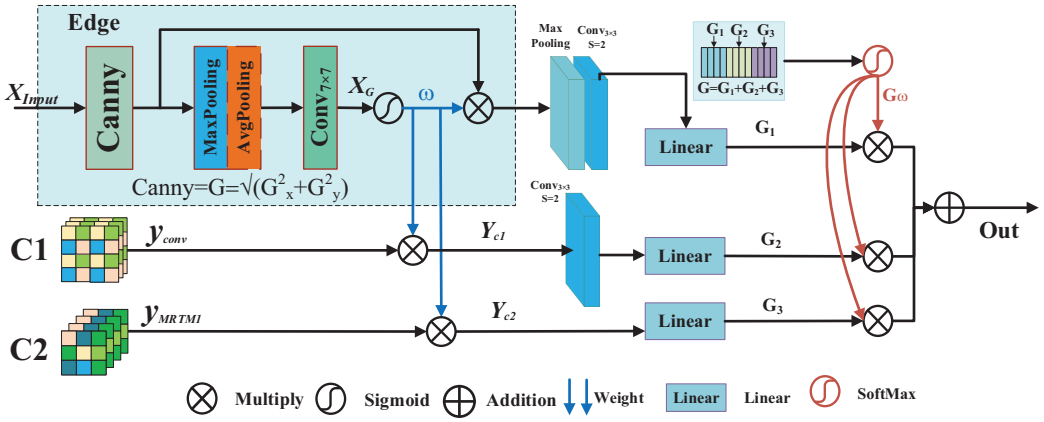

Figure 6: Structure of edge-preserving perception module

Step 1: Firstly, the first-order differential operator Canny is used to extract the tooth edge features of the original image (X_{input}). Then, the feature map X_G is obtained by Max pooling, average pooling, and 7×7 convolution operation. The calculation process is shown in Formula (13).

$$X_G = Conv_{7\times7}\left(Avgpool\left(Maxpool\left(Canny\left(X_{input}\right)\right)\right)\right) \tag{13}$$

Step 2: The tooth edge weights ω are obtained using the Sigmoid function. Then, the ω is used to enhance the tooth edge information in the shallow feature maps (y_G, y_{C1}, y_{C2}), the feature maps Y_G, Y_{C1}, Y_{C2} are obtained. The calculation process of Y_G, Y_{C1} and Y_{C2} is shown in Formulas (14)–(16).

$$Y_G = \omega \bigotimes (Canny\,(X_G)) \tag{14}$$

$$Y_{C1} = \omega \bigotimes y_{C1} \tag{15}$$

$$Y_{C2} = \omega \bigotimes y_{C2} \tag{16}$$

Step 3: Scaling the feature maps Y_G and Y_{C1} to the same size as Y_{C2}. Firstly, mapping Y_G to G1 by Max pooling, 3 × 3 convolutions with step size 2 and Linear mapping; Mapping Y_{C1} to G2 via a 3 × 3 convolution with step size 2 and Linear; mapping Y_{C2} to G3 by Linear, the calculation process of G1, G2, and G3 is shown in Formulas (17)–(19). Secondly, when concatenating G1, G2, and G3, the SoftMax function is used to convert them into a probability distribution between 0 and 1. The adaptive attention weight G_ω is obtained. Thirdly, the G_ω are multiplied with G1, G2, and G3 to obtain 3 feature maps. Finally, the three feature maps are fused element-by-element to obtain Out. The calculation process of Out is shown in Formula (20).

$$G1 = Conv_{3\times3,s=2}\,(MaxPool\,(Y_G)) \tag{17}$$

$$G2 = Conv_{3\times3,s=2}\,(I_{c1}) \tag{18}$$

$$G3 = Conv_{3\times3}\,(I_{c1}) \tag{19}$$

$$Out = (G_1 \bigotimes G_\omega) \bigoplus (G_2 \bigotimes G_\omega) \bigoplus (G_3 \bigotimes G_\omega) \tag{20}$$

4 Experiments and Discussion

4.1 Dataset and Preprocessing

From January 2020 to June 2022, 1500 clinical patients who underwent oral examination in the Department of Stomatology of a $Class_{iii}$ Grade A hospital in Ningxia were selected as the datasets. There are 1500 oral panoramic X-ray images in the datasets, including 1200 images in the training set and 300 images in the test set. In the datasets, the image format is the Digital Imaging and Communications in Medicine (DICOM) format; this paper converts the DICOM format image to PNG format. Then, under the guidance of professional doctors, Labelme software is used to mark the contour of the lesion. The corresponding JSON file is generated, which includes the category label of the lesion, the coordinate value of the marking point, the width and height of the image, and the image path.

4.2 Experimental Environment and Parameter Environment

This paper adopts the PyTorch deep learning framework to implement the Teeth YOLACT model. The configure ratio of the hardware environment is as follows: processor: Intel(R) Xeon(R) Gold 6154 CPU @ 3.00 GHz, memory: 256 GB, graphics card NVIDIA TITAN V. Software environment: Windows Server 2019 Datacenter64-bit operating system, Pytrch1.12.1, Python version 3.7.12. Cuda version: 11.3.58. In the process of network training, the batch size of training is 4, the learning rate is 0.001, and the stochastic gradient descent algorithm is used as the optimizer to optimize the model, where the parameter momentum is 0.9, and the weight attenuation coefficient is 0.0005.

4.3 Evaluation Index

In order to evaluate the performance of the Teeth YOLACT network. In this paper, Intersection over Union (IoU), Average Precision (AP), Average Recall (AR), and Mean Average Precision (mAP) are used as the evaluation criterion. The specific calculation formulas are given in (21)–(24).

$$IOU = \frac{TP}{(TP + FP + FN)} \tag{21}$$

$$AP = \frac{1}{|t|}\frac{1}{|th|}\sum_{t}\frac{TP(t)}{TP(t) + FN(t)} \tag{22}$$

$$AR = \frac{1}{|c|}\frac{1}{|th|}\sum_{t}\frac{TP(t)}{TP(t) + FN(t)} \tag{23}$$

$$mAP = \frac{AP_{0.5} + AP_{0.55} + \ldots + AP_{0.95}}{10} \tag{24}$$

where *th* represents the threshold of each category, *t* is the number of detected samples, *c* is the detection category.

In order to verify the instance segmentation performance of the Teeth YOLACT model, two experiments are set in this paper. The first is the ablation experiment. This group of experiments is used to verify the influence of each module on the detection and segmentation performance in this model group. The second is comparative experiments; Teeth YOLACT is compared with different instance segmentation networks to illustrate advancement. Both experiments are evaluated using the same datasets of panoramic X-ray images. Among them, AP_{det}, AR_{det}, and mAP_{det} represent evaluation indicators for detection, while AR_{seg}, AP_{seg}, and mAP_{seg} represent evaluation indicators for segmentation.

4.4 Ablation Experiments

In ablation experiments, there are 5 experiments to verify MRT, CIaM, and E2PM. It is shown in Table 3. Experiment 1, YOLACT, Resnet50 is used as the YOLACT feature extraction network. Experiment 2, MRTM-YOLACT, MRTM is used as the YOLACT feature extraction network. Experiment 3, YOLACT+MRTM+CIaM, MRTM is used as the YOLACT feature extraction network, and the CIaM module is added to the deep layer of the feature extraction network. Experiment 4, YOLACT+MRTM+E2PM, MRTM is used as the YOLACT feature extraction network, and the E2PM module is added to the shallow layer of the feature extraction network. Experiment 5, Teeth YOLACT, the network proposed in this paper.

Table 3: Ablation experiment design

Model	Resnet50	FPN	Head	MRTM	CIaM	E2PM
Experiment 1	✓	✓	✓	✗	✗	✗
Experiment 2	✗	✓	✓	✓	✗	✗
Experiment 3	✗	✓	✓	✓	✓	✗
Experiment 4	✗	✓	✓	✓	✗	✓
Experiment 5	✗	✓	✓	✓	✓	✓

The experimental results are shown in Table 4. The indicators of Experiment 2 are higher than those of Experiment 1. The evaluation indicators AP_{det}, AP_{seg}, AR_{det}, AR_{seg}, mAP_{det}, and mAP_{seg} are increased by 1.6%, 1.1%,0.6%, 0.5%, 0.2%, and 0.7% separately. The experimental results show that the combination of depth-wise separable convolution and Res_block with the Transformer encoder can fully extract the effective features of impacted teeth. Experiment 3 Compared with Experiment 2, the evaluation indicators AP_{det}, AP_{seg}, AR_{det}, AR_{seg}, mAP_{det}, and mAP_{seg} are increased by 0.6%, 2.1%, 0.9%, 0.7%, 0.8% and 0.9% separately. The experimental results showed that efficient aggregation of deep semantic information and shallow spatial details can obtain richer context information and improve the accuracy of impacted teeth segmentation and detection. Experiment 4 Compared with Experiment 2, the evaluation indicators AP_{det}, AP_{seg}, AR_{det}, AR_{seg}, mAP_{det}, and mAP_{seg} are increased by 1.2%, 1.1%, 0.9%, 1.1%, 0.7%, and 1.7% separately. The experimental results showed that the discrimination is enhanced between impacted teeth and periodontal tissue. Achieve more accurate segmentation of impacted tooth instances. In Experiment 5, Teeth YOLACT, the evaluation indicators AP_{det}, AP_{seg}, AR_{det}, AR_{seg}, mAP_{det}, and mAP_{seg} reached 89.9%, 91.9%, 77.4%, 77.6%, 72.8%, and 73.5%. That is the best result.

Table 4: Results of ablation experiments (IoU = %)

MODEL	AP_{det} (0.75)	AP_{det} (0.75)	AR_{det} (0.50:0.95)	AR_{det} (0.50:0.95)	mAP_{det} (0.50:0.95)	mAP_{det} (0.50:0.95)
Experiment 1	86.9%	88.1%	76.0%	75.8%	71.6%	71.0%
Experiment 2	88.5%	89.2%	76.6%	76.3%	71.8%	71.7%
Experiment 3	89.1%	91.3%	77.5%	77.0%	72.6%	72.6%
Experiment 4	89.7%	90.3%	77.5%	77.4%	72.5%	73.4%
Experiment 5	89.9%	91.9%	77.4%	77.6%	72.8%	73.5%

In order to further show the instance segmentation performance of each sub-module in the proposed model. The radar chart compares the evaluation indicators results of different improved modules. This can be seen in Fig. 7.

The visualization result of the panoramic X-ray image in the Teeth YOLACT is shown in Fig. 8. It can be seen that the model can accurately detect and segment impacted teeth.

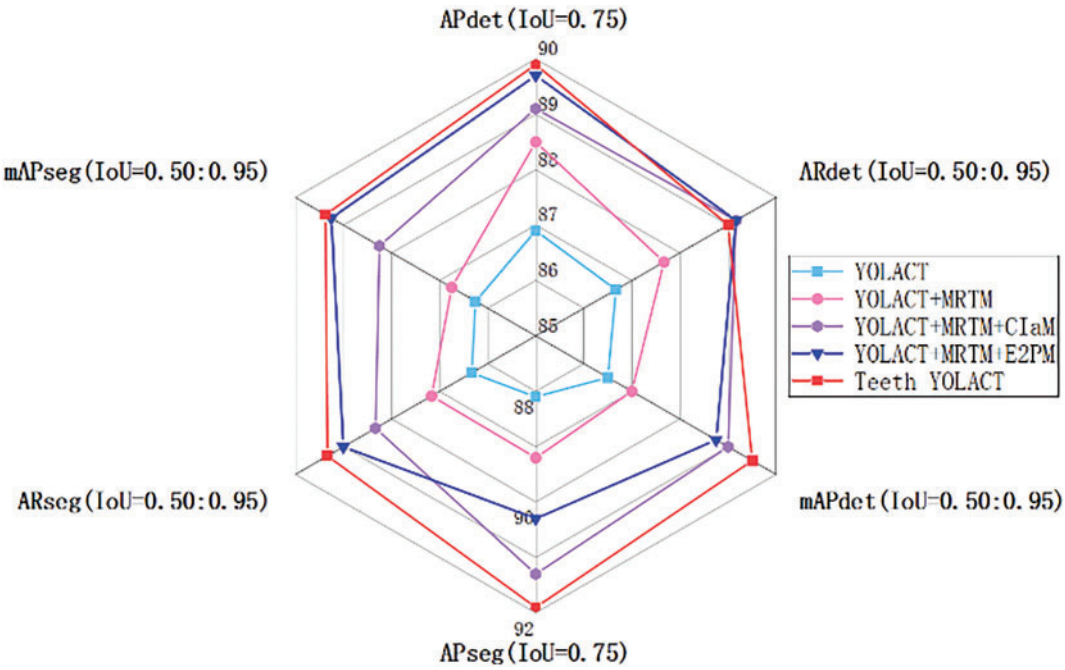

Figure 7: Radar chart comparison of different module instance segmentation results of teeth

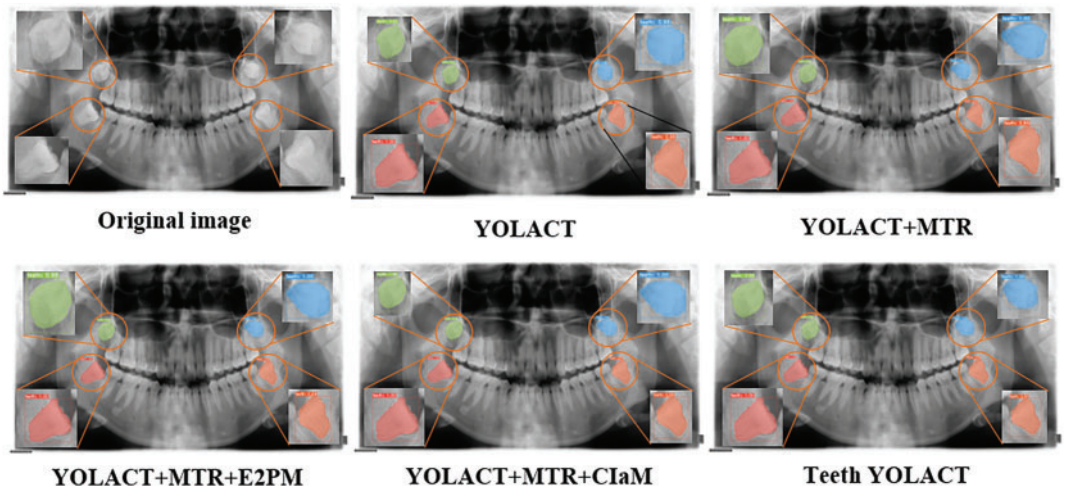

Figure 8: Visualization results of the model

4.5 Comparison Experiments

In Comparative Experiment 6, comparative experiments are designed to verify the model advancement. Experiment 1 Segmenting Objects by Locations (SOLOv2) [23]: Comparison with the SOLOV2 model, Experiment 2 Conditional Convolutions for Instance Segmentation (ConInst) [24]: Comparison with the ConInst model, Experiment 3 BoxInst: High-Performance Instance Segmentation with Box Annotations (BoxInst) [25]: Comparison with the BoxInst model, Experiment 4 Top-Down Meets Bottom-Up for Instance Segmentation (BlendMask) [26]: Comparison with the BlendMask model, Experiment 5 YOLACT: Comparison with the YOLACT model, Experiment 6 YOLOv8seg: Comparison with the YOLOv8seg [27]

model, Experiment 7 Teeth YOLACT: the network proposed in this paper. The experimental results are shown in Table 5.

Table 5: Comparison results of segmentation models for different instances (IoU = %)

MODEL	AP_{det} (0.75)	AP_{det} (0.75)	AR_{det} (0.50:0.95)	AR_{det} (0.50:0.95)	mAP_{det} (0.50:0.95)	mAP_{det} (0.50:0.95)
SOLOV2	/	87.2%	/	73.1%	/	68.3%
Conlnst	87.3%	91%	76.7%	76.4%	70.9%	71.5%
Boxlnst	85.7%	69.2%	76.3%	62.5%	70.5%	56.7%
BlendMask	85.1%	89.9%	76.4%	77.0%	71.0%	72.3%
YOLACT	86.9%	88.1%	76.0%	75.8%	71.6%	71.0%
YOLOv8seg	90.7%	90.2%	77.3%	77.1%	74.5%	70.1%
OUR	89.9%	91.9%	77.4%	77.6%	72.8%	73.5%

Note: SOLOV2 has no detection head and cannot give the value of the detection index. The AP_{det}, AR_{det} and Map_{det} are null. "/" represent null.

The experimental results are shown in Table 5. The Teeth YOLACT instance segmentation model is proposed in this paper. In terms of detection and segmentation, the AP_{det} of Teeth YOLACT model reaches 89.9%, which is 3% and 2.6% higher than YOLACT and ConInst. It reaches 91.9% on AP_{seg}, which is 0.9% and 2% higher than BlendMask and ConInst. In terms of Average Recall, Teeth YOLACT achieved 77.4% and 77.6% in AR_{det} and AR_{seg}, respectively, which are 1.4% and 1.8% higher than those of YOLACT. In terms of comprehensive evaluation indicators, Teeth YOLACT reaches 72.8% and 73.5% in mAP_{det} and mAP_{seg}, respectively, which are 1.2% and 2.5% higher than those of YOLACT. In summary, Teeth YOLACT outperforms the existing mainstream methods in terms of detection and segmentation accuracy, recall rate, and comprehensive performance, especially in complex tooth instance segmentation scenarios, showing higher application value and performance potential. These advantages are mainly due to the innovative design of model multi-scale feature extraction, deep and shallow feature fusion, and edge perception enhancement.

In order to show the performance of the Teeth YOLACT model more intuitively. This chart compares the evaluation metrics of different improved modules in order to more intuitively show the advantages of the model on complex tooth edge feature extraction. It can be seen from the figure that the Teeth YOLACT method performs well in multiple indicators, especially in the four evaluation indicators AP_{seg} (0.75), AR_{det} (0.50:0.95), mAP_{det} (0.50:0.95), and mAP_{seg} (0.50:0.95). Compared with other methods, the detection and segmentation accuracy of Teeth YOLACT achieves higher values in most scenarios. The experimental results show that Teeth YOLACT has stronger generalization ability and robustness in the task of impacted tooth instance segmentation as shown in Fig. 9.

In order to further verify the effectiveness of the proposed method, a visual comparison with the existing mainstream instance segmentation methods is shown in Fig. 10. By showing the segmentation effects of different methods, the advantages of the proposed method in complex panoramic X-ray image tasks are more intuitively reflected.

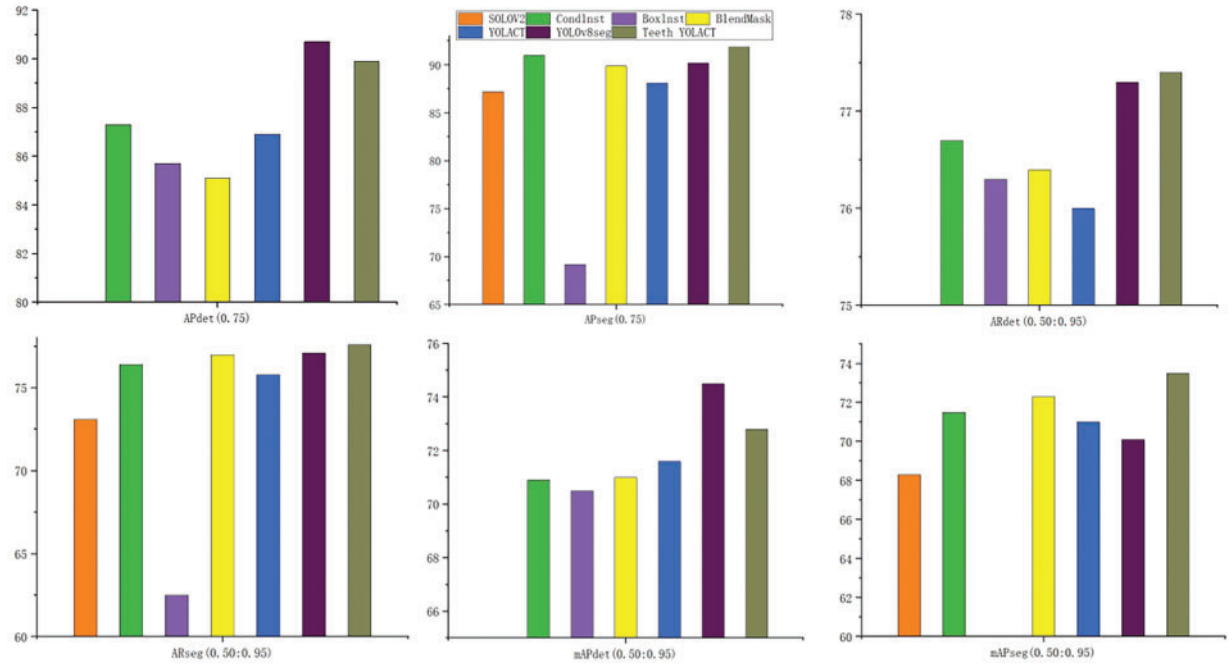

Figure 9: Teeth YOLACT compared to other examples

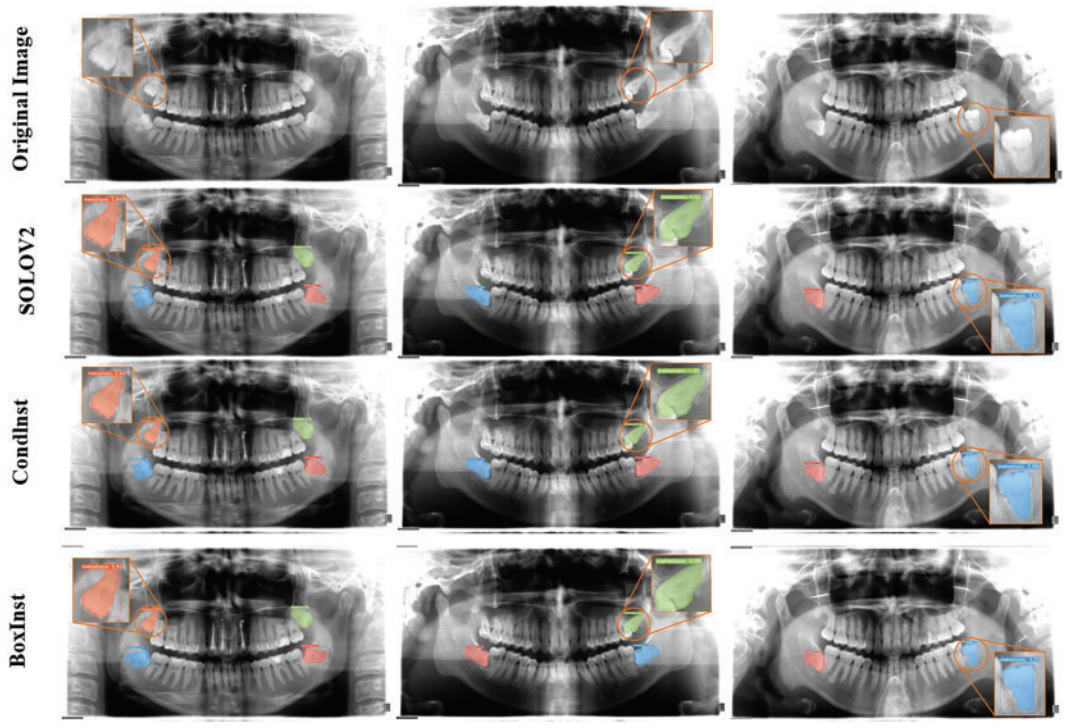

Figure 10: (Continued)

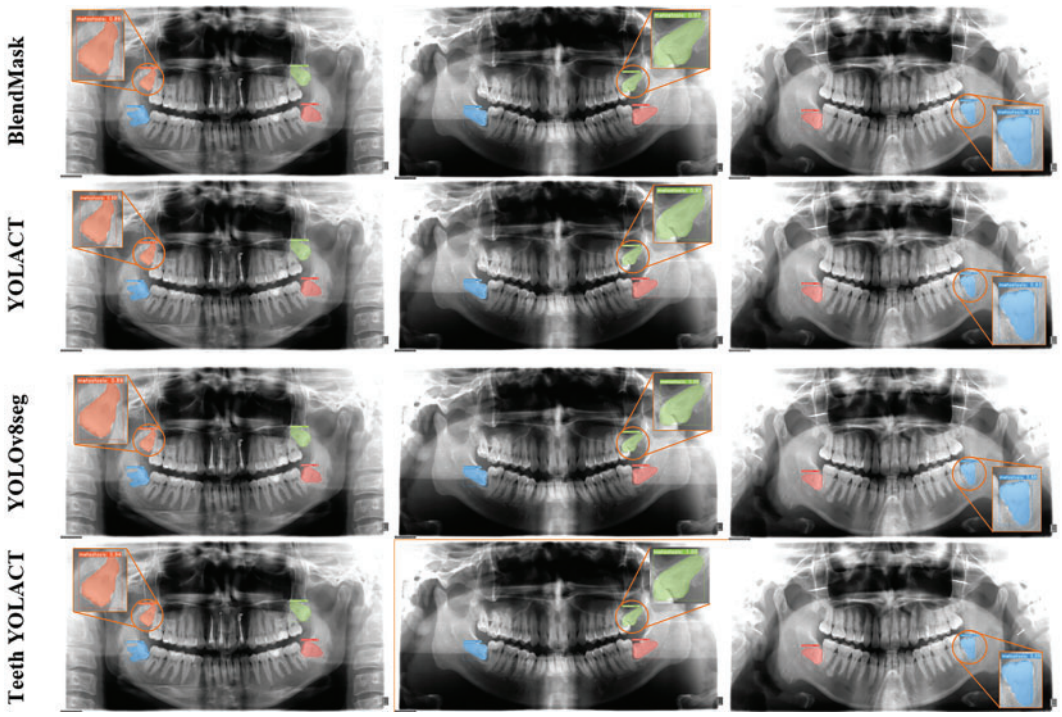

Figure 10: Visualization of existing mainstream instance segmentation methods

5 Conclusion and Future Work

To address the challenges of impacted teeth in panoramic X-ray images, such as dental deformities and low contrast with periodontal tissues, this study proposes the Teeth YOLACT instance segmentation model for detecting and segmenting impacted teeth. The MRTM is designed to effectively extract features, enabling more accurate recognition and localization of impacted teeth. The CIaM is designed to fuse deep and shallow features. In addition, the E2PM is designed to enhance the teeth edge features and improve the discrimination between the model-impacted teeth and the periodontal tissue. The proposed model realizes more accurate segmentation of impacted teeth. In order to verify the effectiveness of the Teeth YOLACT model. The comparative experiment and ablation experiment are carried out on the oral panoramic X-ray image dataset. The AP_{det}, AP_{seg}, AR_{det}, AR_{seg}, mAP_{det}, and mAP_{seg} indicators of tooth instance segmentation by the Teeth YOLACT model are 89.9%, 91.9%, 77.4%, 77.6%, 72.8%, and 73.5%, respectively. The results show that the Teeth YOLACT model can effectively improve the detection and segmentation of impacted teeth in oral panoramic X-ray images. There is an important reference value for the auxiliary diagnosis of dental lesions based on oral panoramic X-ray images.

In this paper, the Teeth YOLACT model is proposed and compared with other instance segmentation models, and the indicators are significantly improved. However, there are still some limitations of the model: (1) At present, the number of samples in the dataset is not large and the types of teeth are not rich. (2) The current model learning is a strong supervised learning. Weak supervised learning and the generalization ability about instance segmentation of impacted teeth are essential. (3) Although the model tries some strategies to improve the accuracy of instance segmentation, the accuracy of instance segmentation is still not very high, and the consumption of computing and storage resources is still significant.

In oral panoramic X-ray images, the instance segmentation of impacted teeth is a challenge. In the future, research can be carried out from the following aspects:

1. Increasing dataset size and diversity: In cooperation with more medical institutions, diverse oral X-ray image datasets containing different age groups, pathological types, and image quality are collected, and automated or semi-automated annotation tools are developed to reduce annotation costs.
2. Explore weakly supervised and semi-supervised learning methods: Advanced self-supervised learning and contrastive learning techniques are used to improve the performance of the model in scenarios with insufficient labeled data by combining a small amount of labeled data and a large amount of unlabeled data.
3. Optimization model structure: This paper studies the lightweight neural network structure, reduces the computational complexity and storage requirements through parameter pruning and model quantization technology, and makes the model more suitable for deployment in resource-constrained medical devices.

In summary, in the field of medical image instance segmentation, these strategies are helpful to guarantee the instance segmentation models' innovation and improve the model performance in practical application scenarios. Such as rich medical image data sets, improving the generalization ability, and supervision methods of the model. Through these measures, doctors can be provided with more accurate diagnostic information, improve medical quality, improve patient treatment experience, promote the progress of medical research, and promote the development of the medical field.

Acknowledgement: The authors would like to thank the anonymous editors and reviewers for their critical and constructive comments and suggestions.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China (Grant No. 62062003), Natural Science Foundation of Ningxia (Grant No. 2023AAC03293).

Author Contributions: Tao Zhou: Data curation, Project administration, Writing—review & editing, Funding acquisition, Supervision. Yaxing Wang: Methodology, Writing–original draft, Validation, Visualization, Software. Wenwen Chai: Investigation, Writing—review & editing, Formal analysis. Yunfeng Pan: Data curation, Investigation, Validation. Zhe Zhang: Visualization, Validation. Huiling Lu: Writing—review & editing, Resources. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Due to ethical and privacy concerns, it is not convenient to share this data and material.

Ethics Approval: All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. This research was conducted with human volunteers, and all participating human volunteers provided informed consent. The research received approval from the North Minzu University and Ningxia Medical University. The research ethics certificate numbers are No. 2024-17 and No. 2024-G253, respectively.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Hou SB, Zhou T, Liu YC, Dang P, Lu HL, Shi HB. Teeth U-Net: a segmentation model of dental panoramic X-ray images for context semantics and contrast enhancement. Comput Biol Med. 2023;152(2):106296. doi:10.1016/j.compbiomed.2022.106296.
2. Guarnieri RCC, Vernucci R, Vichi M, Leonardi R, Barbato E. Impacted maxillary canines and root resorption of adjacent teeth: a retrospective observational study. Med Oral Patol Oral Cir Buccal. 2016;21(6):e743. doi:10.4317/medoral.21337.

3. Ness GM, Blakey GH, Hechler BL, Miloro M, Ghali GE, Larsen PE, et al. Impacted teeth. In: Miloro M, Ghali GE, Larsen PE, Waite P, editors. Peterson's principles of oral and maxillofacial surgery. Berlin/Heidelberg, Germany: Springer; 2022. p. 131–169.
4. Zhou T, Peng CY, Guo YJ, Wang HX, Niu YX, Lu HL. Identity-mapping ResFormer: a computer-aided diagnosis model for pneumonia X-ray images. IEEE Trans Instrum Meas. 2025;74:1–12. doi:10.1109/TIM.2025.3534218.
5. Zhao Y, Li JC, Cheng BD, Niu NJ, Wang LG, Gao WG, et al. Applications and challenges of deep learning in dental imaging. Image Graph. 2024;29(3):0586–607. doi:10.11834/jig.230062.
6. Chen Y, Guo X, Xia Y, Yuan Y. Disentangle then calibrate with gradient guidance: a unified framework for common and rare disease diagnosis. IEEE Trans Med Imaging. 2024;43(5):1816–27. doi:10.1109/TMI.2023.3349284.
7. Lu JF, Huang XY, Song CH, Li CJ, Hu YY, Xin RL, et al. CISA-UNet: dual auxiliary information for tooth segmentation from CBCT images. Alex Eng J. 2025;114(7):543–55. doi:10.1016/j.aej.2024.11.103.
8. Wang X, Alqahtani KA, Van den Bogaert T, Shujaat S, Jacobs R, Shaheen E. Convolutional neural network for automated tooth segmentation on intraoral scans. BMC Oral Health. 2024;24(1):804. doi:10.1186/s12903-024-04582-2.
9. Jader G, Fontineli J, Ruiz M, Abdalla K, Pithon M, Oliveira L. Deep instance segmentation of teeth in panoramic X-ray images. In: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images; 2018 Oct 29–Nov 1; Parana, Brazil. p. 400–7.
10. He KM, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy. p. 2961–9.
11. Graham S, Chen H, Gamper J, Dou Q, Heng PA, Snead D, et al. MILD-Net: minimal information loss dilated network for gland instance segmentation in colon histology images. Med Image Anal. 2019;52(5):199–211. doi:10.1016/j.media.2018.12.001.
12. Bhatti HMA, Li J, Siddeeq S, Rehman A, Manzoor A. Multi-detection and segmentation of breast lesions based on mask RCNN-FPN. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine; 2020 Dec 16–19; Seoul, Republic of Korea. p. 2698–704.
13. Kirillov A, Girshick R, M He K, Dollar P. Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 6399–408.
14. Wang D, Chen S, Sun X, Chen Q, Cao Y, Liu B, et al. AFP-mask: anchor-free polyp instance segmentation in colonoscopy. IEEE J Biomed Health Inform. 2022;26(7):2995–3006. doi:10.1109/JBHI.2022.3147686.
15. Bolya D, Zhou C, Xiao F, Lee YJ. YOLACT++ better real-time instance segmentation. IEEE Trans Pattern Anal Mach Intell. 2022;44(2):1108–21. doi:10.1109/TPAMI.2020.3014297.
16. Jia Y, Lu C, Li X, Ma M, Pei Z, Sun Z, et al. Nuclei instance segmentation and classification in histopathological images using a DT-Yolact. In: 2021 20th International Conference on Ubiquitous Computing and Communications; 2021 Dec 20–22; London, UK. p. 414–20.
17. Roy N, Roy SK, Sharan P. Effective brain tumor segmentation for MRI image analysis using dual attention network based YOLACT++. In: 10th International Conference on Computing for Sustainable Global Development; 2023 Mar 15–17; New Delhi, India. p. 10–5.
18. Zhou T, Niu YX, Lu HL, Peng CY, Guo YJ, Zhou HY. Vision transformer: to discover the four secrets of image patches. Inf Fusion. 2024;105:102248. doi:10.1016/j.inffus.2024.102248.
19. Guo R, Niu D, Qu L, Li Z. SOTR: segmenting objects with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, BC, Canada. p. 7157–66.
20. Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 1290–9.
21. Shi D. Robust foveal visual perception for vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. p. 17773–83.
22. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA.

23. Wang XL, Zhang RF, Kong T, Li L, Shen CH. SOLOv2: dynamic and fast instance segmentation. Adv Neural Inf Process Syst. 2020;33:17721–32.
24. Tian Z, Shen C, Chen H. Conditional convolutions for instance segmentation. In: Computer Vision-ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. Berlin/Heidelberg, Germany: Springer.
25. Tian Z, Shen C, Wang X, Chen H. BoxInst: high-performance instance segmentation with box annotations. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 5439–48.
26. Chen H, Sun K, Tian Z, Shen C, Huang Y, Yan Y. BlendMask: top-down meets bottom-up for instance segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 8570–8.
27. Jocher G, Chaurasia A, Qiu J. YOLO by Ultralytics. [cited 2025 Jan 1]. Available from: https://github.com/ultralytics/ultralytics.