Submitted: 21 December, 2023 Accepted: 04 March, 2024 Published: 03 November, 2024 DOI:10.22514/jocpd.2024.136

ORIGINAL RESEARCH

Journal of Clinical Pediatric Dentistry

Co-Mask R-CNN: collaborative learning-based method for tooth instance segmentation

Chen Wang[1], Jingyu Yang[1], Hongzhi Liu[1], Peng Yu[2,*], Xijun Jiang[3,*], Ruijun Liu[4]

[1]The School of Computer and Artificial Intelligence, Beijing Technology and Business University, 100048 Beijing, China
[2]Department of Cariology and Endodontology, Peking University School and Hospital of Stomatology, 100081 Beijing, China
[3]Department of Pediatric Dentistry, Peking University School and Hospital of Stomatology, 100081 Beijing, China
[4]The School of Software, Beihang University, 100191 Beijing, China

***Correspondence**
yupeng@bjmu.edu.cn
(Peng Yu);
jiangxijun@bjmu.edu.cn
(Xijun Jiang)

Abstract

Traditional tooth image analysis methods primarily focus on feature extraction from individual images, often overlooking critical tooth shape and position information. This paper presents a novel computer-aided diagnosis method, Collaborative learning with Mask Region-based Convolutional Neural Network (Co-Mask R-CNN), designed to enhance tooth image analysis by leveraging the integration of complementary information. First, image enhancement is employed to generate an edge-enhanced tooth edge image. Then, a collaborative learning strategy combined with Mask R-CNN is introduced, where the original and edge images are input simultaneously, and a two-stream encoder extracts feature maps from complementary images. By utilizing an attention mechanism, the output features from the two branches are dynamically fused, quantifying the relative importance of the two complementary images at different spatial positions. Finally, the fused feature map is utilized for tooth instance segmentation. Extensive experiments are conducted using a proprietary dataset to evaluate the effectiveness of Co-Mask R-CNN, and the results are compared against those of an alternative segmentation network. The results demonstrate that Co-Mask R-CNN outperforms the other networks in terms of both segmentation accuracy and robustness. Consequently, this method holds considerable promise for providing medical professionals with precise tooth segmentation results, establishing a reliable foundation for subsequent tooth disease diagnosis and treatment.

Keywords

Deep learning; Two-stream collaborative network; Tooth instance segmentation; Bitewing radiograph

1. Introduction

The continuous integration of computer-aided technology and artificial intelligence has accelerated automation in various industries. Image understanding technology has received extensive attention in the field of medical image analysis. It can be used for segmentation, reconstruction and three-dimensional visualization of medical images, thus opening up new possibilities for advances in medical imaging and diagnostic processes [1]. X-ray imaging is very common in dental images commonly used by oral and maxillofacial surgeons. This type of imaging utilizes X-rays, which vary in intensity due to the density and thickness of different tissues, to penetrate the body. The resulting images show varying degrees of brightness or black-and-white contrast. In addition, X-rays are cost-effective and have a low radiation dose [2–4]. However, there are two major challenges in the use of X-rays for diagnosing oral diseases. First, the boundaries of tooth structures in images often show blurring due to factors such as missing and overlapping teeth, as well as significant differences in the distribution of tooth and pulp tissue. Second, due to the widespread use of restorative and implant materials, some dental metallic materials produce metal artefacts, which reduce contrast and blur structures and ultimately affect the dental image quality and disease diagnosis outcomes [5]. In traditional X-ray diagnostic methods, dentists rely mainly on personal experience and visual perception to analyse dental structures and formulate treatment plans, which can bring about a high degree of subjectivity. In contrast, deep learning-based medical image understanding techniques can help dentists achieve automated image interpretation. Accurate dental image segmentation and recognition can assist in clinical decision-making, thus improving clinical efficiency and reducing misdiagnosis rates [6].

Traditional tooth image segmentation methods include various approaches, such as threshold-based segmentation methods [7–9], edge detection-based segmentation methods [10, 11] and region-based image segmentation methods [12–14]. These methods have been applied to dental image segmentation, but they have several limitations, such as weak robustness to image noise and artefacts, which can easily cause degradation of segmentation performance.

With the rise of deep learning algorithms such as convolutional neural networks, the development prospects of medical image analysis automation are more objective. Compared with traditional dental instance segmentation methods that need complex rules for modelling, data-driven deep learning methods have stronger modelling and generalization capabilities. Among them, Mask R-CNN (Mask Region-based Convolutional Neural Network) [15] has attracted much attention for its excellent high accuracy, high scalability, multitask learning ability and migratory nature. Jader *et al*. [16] proposed the first system capable of detecting and segmenting each tooth in a panoramic radiograph image. Pinheiro *et al*. [17] investigated and compared two Mask R-CNN-based schemes to improve rough segmentation boundaries. In 2022, Chandrashekar *et al*. [18] introduced a collaborative learning model that improves learning performance by combining an independent tooth instance segmentation model, Mask R-CNN, with a recognition model, Faster R-CNN [19]. Lee *et al*. [20] differed from the approach of Jader *et al*. [16] in that each panoramic radiograph in the dental images used in this study generated multiple independently annotated mask images based on the number of teeth included. Zhao *et al*. [21], based on the instance segmentation model Mask R-CNN combined with the U-Net architecture, modified the segmentation branching to improve the segmentation effect. Chung *et al*. [22] proposed a pixel labelling-based neural network that is robust to metal artefacts. Silva *et al*. [23] analysed the performance of four classical network architectures (Mask R-CNN [15], PANet (Path Aggregation Network) [24], HTC (Hybrid Task Cascade for Instance Segmentation) [25], and ResNeSt (Split-Attention Networks) [26]) on standard panoramic X-ray film datasets. The results show that these architectures can be used for dental instance segmentation and numbering tasks, with PANet performing the best. Leite *et al*. [27] proposed a segmentation framework that combines two deep convolutional neural networks, DeepLab-v3 and FCN-ResNet101. However, this study used clear images of adolescent teeth and lacked consideration of the effects of the presence of artefacts, implants, and changes in teeth between patient and patient age. These methods have achieved good results in tooth instance segmentation; however, instance segmentation applied to the dental domain still has the following problems. (1) There is an insufficient quantity of data: the quantity of data in the dental domain is relatively small compared to that in other domains, as is the high cost of annotation, which makes it difficult to expand the data. (2) Morphological diversity: the diversity of tooth morphology makes the instance segmentation algorithm more complex. (3) Poor image quality: In the field of teeth, due to the interference of metal artefacts, which often leads to blurred or inconspicuous edges of the teeth and occlusion between the teeth, algorithms need to determine which pixels belong to which teeth and segment them, which puts a higher demand on the accuracy and robustness of the algorithms.

In conclusion, although the improved method based on Mask R-CNN achieved satisfactory experimental results, limitations such as the effect of clinical data noise and insufficient extraction of global contextual information from the feature extraction module still hinder the effectiveness of the dental instance segmentation method in dealing with region boundaries. However, collaborative learning integrates multiple pieces of input information, allows information exchange and sharing between original and edge images, and can comprehensively consider feature representations in different image spaces, thus improving the learning capability of the whole system and resulting in better model generalization performance and reducing the risk of overfitting. Therefore, in this paper, we propose an improved method for complementary image feature fusion based on the attention mechanism from the perspective of collaborative learning, which can extract more comprehensive contextual information. The contributions of our approach are worth mentioning and can be summarized as follows.

A Co-Mask R-CNN network is proposed that features image enhancement, dual-branch feature extraction, and dynamic feature fusion, allowing for reduced noise and sharper textural details in clinical images, resulting in more robust and correlated feature extraction with a strong global context.

A collaborative learning strategy is introduced, incorporating an image enhancement branch to acquire tooth edge images, which, combined with tooth and contextual background information from the original images, comprehensively considers tooth edge information to obtain more complete features.

An attention module is employed to dynamically determine the feature representations obtained from different branches, establishing long-range dependencies between different locations, which aids in better understanding the interrelationships between different positions and capturing semantic correlations more effectively.

2. Materials and methods

2.1 Network architecture

There are several issues with the current methods for improving Mask R-CNN. First, for the extracted regions of interest (RoI), the mask branch uses full convolutional operations for semantic segmentation. While fully convolutional operations have good sensitivity to local semantic information, they neglect contextual information. Second, Mask R-CNN utilizes a two-stage strategy involving detection followed by segmentation, with the segmentation results constrained by the detection outcomes. Finally, while the model performs relatively well for well-conditioned teeth, it is prone to errors when teeth are missing and lacks consideration of shape and position information. To address the aforementioned challenges, our proposed solution is a collaborative learning network model known as the Co-Mask R-CNN, as illustrated in Fig. 1.

The Co-Mask R-CNN method adopts an encoder-decoder structure consisting of three main modules: an image enhancement module, a collaborative learning module, and a feature fusion module. The image enhancement module aims to improve the contrast and accentuate the textural details present in the original image, thereby aiding in detecting tooth edge lines. The original and enhanced images are fed into the collaborative learning module, where the encoder extracts complementary features from the tooth-related images to obtain complementary information. The feature fusion module utilizes the attention mechanism proposed by Fu *et al*. [28] to dynamically

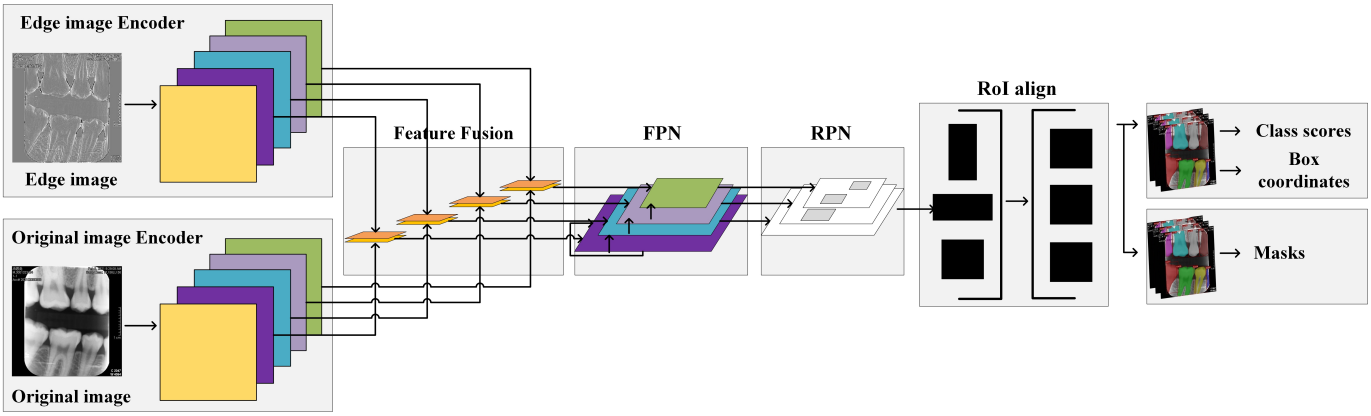


FIGURE 1. Co-Mask R-CNN Network Structure. FPN: Feature Pyramid Network; RPN: Region Proposal Network.

fuse the feature maps extracted by the collaborative learning module. This attention mechanism can incorporate contextual information and compensate for the lack of deep features in Mask R-CNN. Additionally, the weight distributions of the two input images in the model can be adaptively adjusted to achieve optimal weight ratios. Ultimately, the prediction branch produces the ultimate segmentation map. In the following sections, we explain the implementation details for each of these components.

2.1.1 Image enhancement

To reduce brightness, inconsistent exposure, and blurred edges. These factors significantly impact the effectiveness of segmentation networks, image enhancement is crucial in X-ray images. The present investigation introduces a data augmentation technique to address these challenges inspired by the methodology established by Zhou *et al*. [29]. To alleviate the issue of indistinct dental boundaries, we employ the Sobel operator to convolve the image and enhance the prominence of edge information, resulting in edge image generation. This approach reduces noise, enhances image contrast and sharpness, and standardizes image quality to sharpen textural details. Edge images enhance the edge-related information within the original image, emphasizing the outlines and forms of the teeth. In contrast, the original image presented a more comprehensive depiction of dental information, encompassing contextual background details as well.

In the enhancement process, the contrast-enhanced image $I_{ce}(x,y;\delta)$ is obtained through the following formula:

$$I_{ce}(x,y;\delta) = \eta I(x,y) + \theta G(x,y;\delta) * I(x,y) + S(x,y) * I(x,y) + \beta \quad (1)$$

Where $I(x,y)$ is the raw bitewing radiograph, * denotes the convolution operator, $G(x,y;\delta)$ represents the Gaussian filter with standard deviation δ, and $S(x,y)$ denotes the Sobel filter with a 3×3 kernel. The contrast-enhanced images are used as inputs for one branch of the collaborative learning network model. Fig. 2 shows the effect of applying image enhancement, where (1) is the original image, (2) is the image enhancement by Zhou *et al*. [29], and (3) is the edge image. As shown in the figure, the edge image provides higher contrast and brightness and sharper edges than does the original and Zhou's methods.

2.1.2 Complementary feature colearning

Although image enhancement techniques can reduce noise, improve image quality, and sharpen texture details, they may also alter pixel values and lead to the loss of some detailed features. To address this challenge, we introduce a collaborative learning framework inspired by the network architecture initially proposed by Kumar *et al*. [30], which aims to integrate complementary information from the original and edge images for better image analysis. Specifically, the Co-Mask R-CNN includes an original image encoder and an edge image encoder, which correspond to the CNN portion of the model and extract visual features from different images.

In addition, traditional CNNs often suffer from gradient vanishing and exploding problems when the network depth is increased; thus, regularization initialization and intermediate regularization layers are required to mitigate these issues. However, these methods may encounter the issue of network degradation, where the accuracy on the training set may reach a plateau or even decrease as the network depth increases. Therefore, we employ a specific residual learning structure. The ResNet101 [31] architecture consists of convolutional blocks and identity blocks, which serve as the fundamental residual blocks. The residual learning structure employs forward neural networks and shortcut connections. These connections enable straightforward identity mapping without introducing extra parameters or escalating computational complexity. The convolutional block alters network dimensionality by having distinct input and output dimensions, whereas the identity block deepens the network by maintaining the same input and output dimensions.

With these improvements, Co-Mask R-CNN can better learn the relationship between the original and enhanced images, thus improving the accuracy and effectiveness of image analysis.

2.1.3 Complementary feature fusion

This study incorporates a dual attention mechanism comprising a Channel Attention Module (CAM) and a Position Attention Module (PAM) to leverage the extracted features from both branches more effectively [32]. These modules dynamically assign weights to the two branches, effectively incorporate

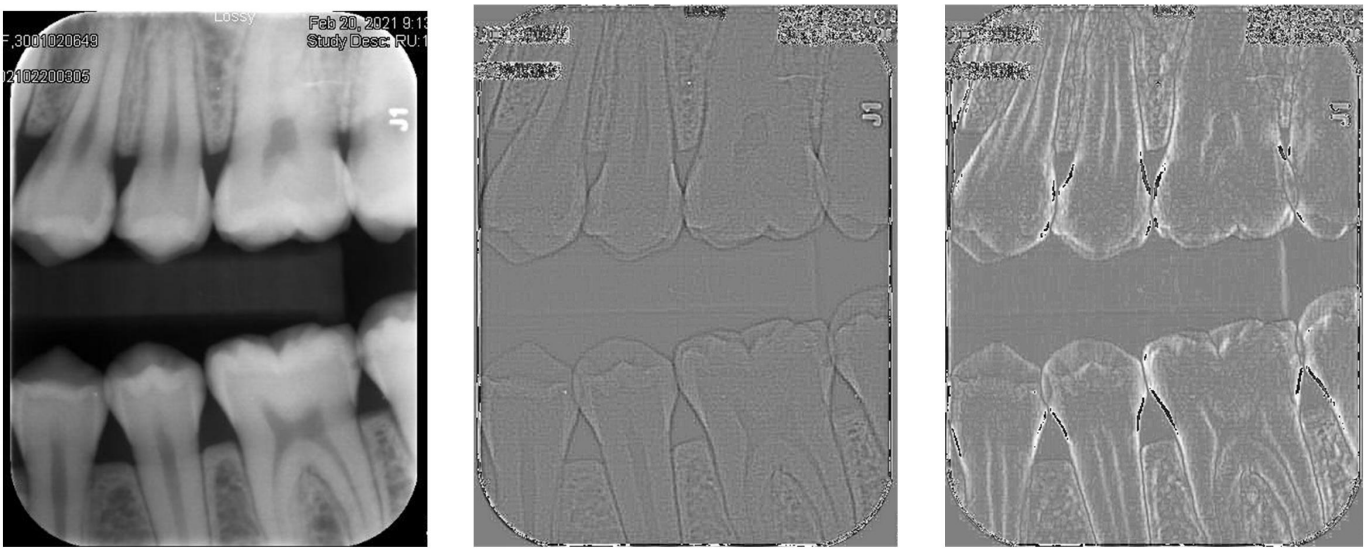


FIGURE 2. Image enhancement results.

contextual information, and achieve complementary feature aggregation. The outputs of the attention modules are transformed using convolutional layers and then combined through weighted summation to achieve feature fusion, as depicted in Fig. 3.

2.1.3.1 Position attention module

The PAM explores the spatial relationships within the feature maps obtained from the two branches. This is accomplished by updating the features at specific positions and assigning weights based on the similarities between corresponding features at those positions. As a result, positions with similar features can mutually enhance each other, irrespective of their spatial separation. By encoding extensive contextual information into the original features, the PAM enhances the representativeness significantly.

As illustrated in Fig. 4, $P\_enhancement$ and $P\_original$ represent the feature maps of the edge image extracted after the backbone network and the feature map of the local image, respectively. The feature map $P\_enhancement \in R^{C\times H\times W}$ acquired from the enhancement branch is initially fed into a convolutional layer, resulting in the generation of two novel feature maps, denoted as G and F, $\{G, F\} \in R^{C\times H\times W}$. Subsequently, the G, F matrix is transformed into an $R^{C\times N}$ matrix, where N = H × W repres matrients the overall pixel count. Through the multiplication of the transposed F and G matrices, the spatial attention map $M \in R^{N\times N}$ is derived. To ensure normalization, a softmax function is applied to this map, producing a normalized output.

$$M_{yx} = \frac{\exp\left(G_x \cdot F_y\right)}{\sum_{i=1}^{N} \exp\left(G_x \cdot F_y\right)} \tag{2}$$

Where M_{yx} ensures the influence of the x^{th} position on the y^{th} position, and the correlation between the feature representations of two positions increases when their corresponding M_{yx} values increase. Concurrently, feature $P\_enhancement$ undergoes a convolutional layer, generating a fresh feature map denoted as $O \in R^{C\times H\times W}$. This map is then reshaped into $R^{C\times N}$. Subsequently, the dot product of O and the transposed M matrix is computed, resulting in a reshaped output $R^{C\times H\times W}$. This output $Q \in R^{C\times H\times W}$ is subsequently scaled by the parameter λ and added elementwise to the original features $P\_original \in R^{C\times H\times W}$ of the input branch, as shown below:

$$Q_y = \lambda \sum_{x=1}^{N} \left(M_{yx} D_x\right) + P\_original_y \tag{3}$$

By initializing the weight parameter λ to 0 and progressively adapting it during the training phase, Eqn. 3 provides valuable insights. This approach implies that the resultant feature Q at each position is obtained by combining features from all positions and the original feature *via* weighted summation. This mechanism facilitates the integration of global contextual information and the selective incorporation of context guided by the spatial attention map. As a consequence, similar semantic features are mutually reinforced, leading to improved intraclass compactness and enhanced semantic consistency.

2.1.3.2 Channel attention module

The CAM treats each channel of a feature map as a distinctive response corresponding to a specific class, thus revealing the interrelated nature of diverse semantic responses. By delving into the interdependencies among channel maps, CAM has the ability to highlight interdependence between feature channels and enhance the specificity of semantics.

Fig. 5 illustrates the architecture of the CAM, where $P\_enhancement$ and $P\_original$ represent the feature maps of the edge image extracted after the backbone network and

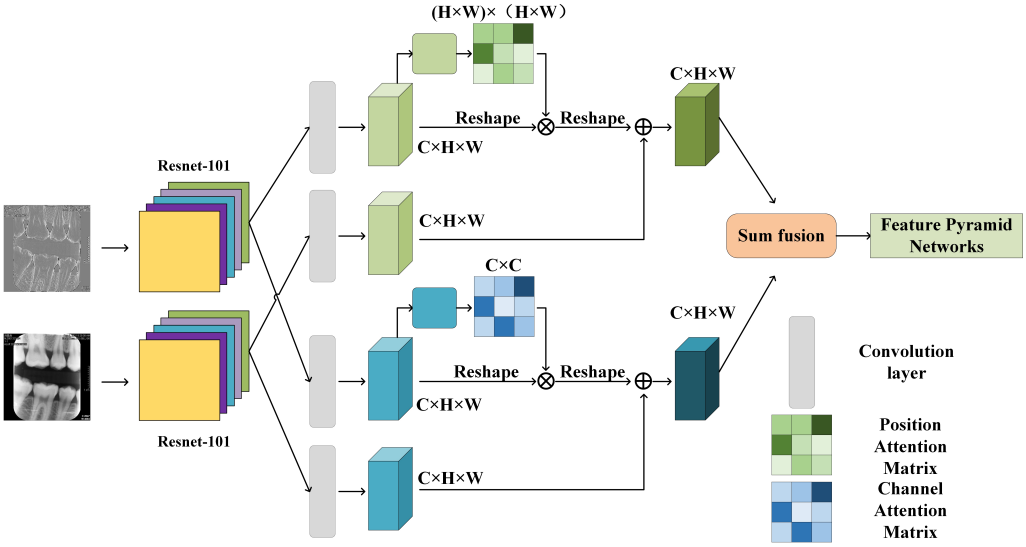


FIGURE 3. Dual attention network structure.

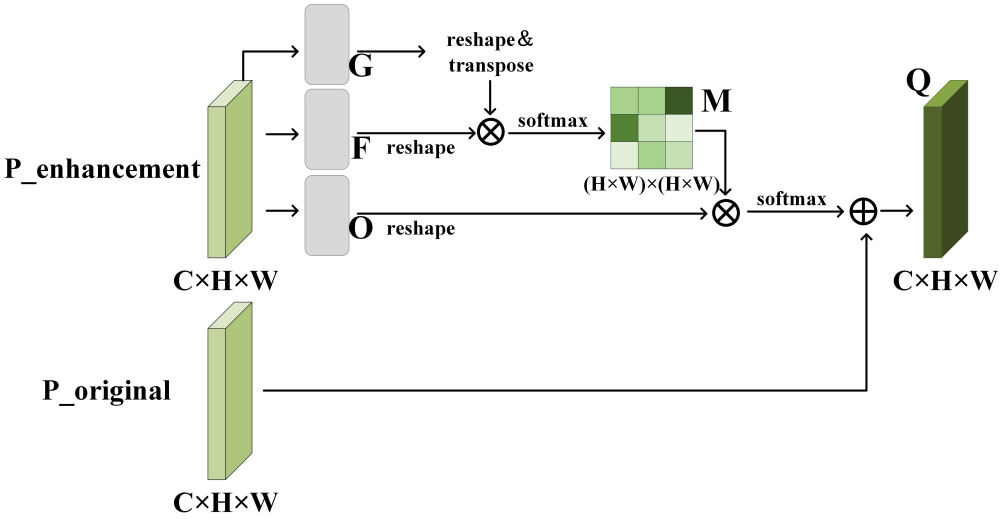


FIGURE 4. Position attention module.

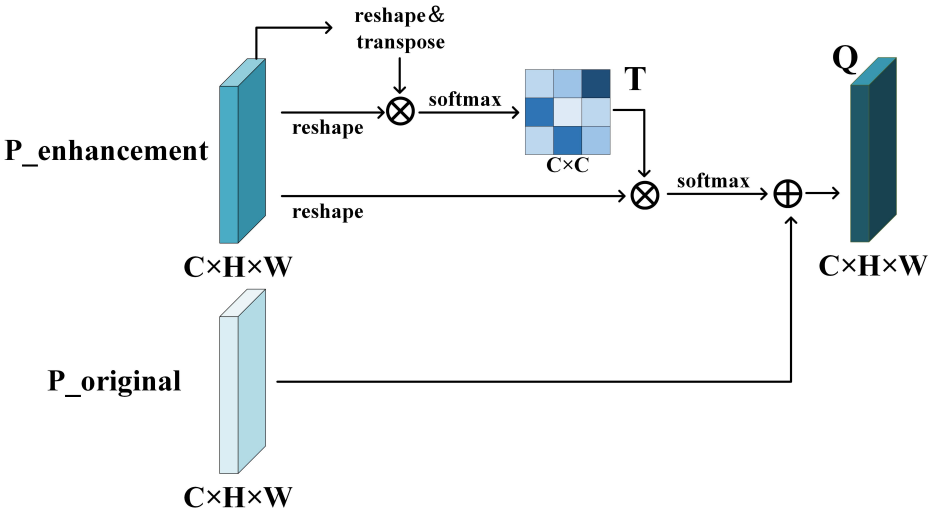


FIGURE 5. Channel attention module.

the feature map of the local image, respectively. Unlike the PAM, feature $P\_enhancement$ is directly reshaped into $R^{C\times N}$ and then subjected to matrix multiplication *via* transpose. The resulting feature maps are multiplied and passed through a softmax function, generating the channel feature map $T \in R^{C\times C}$, as shown below:

$$T_{yx} = \frac{\exp\left(P\_enhancement_x \cdot P\_enhancement_y\right)}{\sum_{x=1}^{N} \exp\left(P\_enhancement_x \cdot P\_enhancement_y\right)} \quad (4)$$

Where T_{yx} indicates the impact of the x^{th} channel on the y^{th} channel. Furthermore, matrix multiplication between T and the transpose of $P\_enhancement$ is performed, resulting in the reshaped feature $T \in R^{C\times H\times W}$. This outcome is then scaled by the parameter λ and added elementwise to the original feature $P\_original$ from the input branch, yielding the final output $Q \in R^{C\times H\times W}$, as shown below:

$$Q_y = \lambda \sum_{x=1}^{N} \left(T_{yx} P\_enhancement_x\right) + P\_original_y \quad (5)$$

Where the weight parameter λ is initialized to 0 and gradually assigned higher significance during the training phase. The equation illustrates that the ultimate feature representation of each channel is a combination of the features extracted from all channels, including the original features, with each channel's contribution weighted accordingly. By performing a weighted summation of all channels, the final feature representation can capture long-range semantic dependencies between feature maps, improving feature distinctiveness. To leverage the benefits of distant contextual information more effectively, the proposed approach combines the features extracted by two attention modules.

2.1.4 Prediction branch

The prediction branch integrates information by leveraging the network architecture of Mask R-CNN and employing the region proposal network (RPN) and ROI modules to extract and consolidate relevant features from the feature maps, as depicted in Fig. 6. Subsequently, the feature vectors are individually fed into two separate prediction network branches. The upper branch performs classification prediction and bounding box regression, while the lower branch is responsible for generating masks that correspond to the detected objects.

2.2 Datasets

We collected and curated the experimental dataset from Peking University School and Hospital of Stomatology. It encompasses a collection of 815 bitewing radiographs focusing on children's teeth. Each image was resized to 512 × 512, and the label data were annotated by a professional dentist with extensive clinical experience. After annotation, the researchers used LabelMe software to pixel-level annotate the tooth area drawn by the dentist on the bitewing radiographs. The generated annotation data were saved in JSON files, including the contour coordinates of each tooth. These annotated data were divided into training and testing sets for validation of the tooth instance segmentation network.

2.3 Metrics for statistical analysis

To evaluate the performance of Co-Mask R-CNN on tooth instance segmentation tasks, we employ metrics such as AP (Average Precision), AR (Average Recall), and IOU (Intersection Over Union). Precision serves as a metric for assessing the accuracy of a model when making positive predictions and represents the percentage of correctly identified positive instances. Recall quantifies the proportion of true positives in the testing dataset that the model correctly detects, reflecting the comprehensiveness of positive identification. The evaluation is conducted by employing IOU thresholds of 50 and 75, along with a range of IOU thresholds from 50 to 95 with increments of 5. The IOU = (50:95) signifies the mAP (mean Average Precision) calculated across various IOU thresholds, thereby offering a comprehensive measure of the model's performance across a spectrum of IOU values. The inclusion of IOU increment averaging aims to ensure that the model performs well not only at IOU = 50 but also at higher IOU thresholds. By calculating the AP of the model at various IOU values and taking the average, we obtain a comprehensive assessment of the model's accuracy.

2.4 Implementation details

In the experimental process, the dataset was divided into a training set and a test set at a ratio of 9:1. The training set was subsequently divided into five parts for fivefold cross-validation, after which the average value was taken as the final result. The hyperparameters, such as the initial learning rate, learning rate decay strategy and iteration number, were dynamically adjusted, the learning rate decay method used was the cosine function, and the initial learning rate was set to 0.0001 at the beginning of the training process. The same loss function as that of Mask R-CNN was used, and the cross-entropy function was used as the classification loss. Similarly, the SmoothL1 loss function was used as the bounding box. Similarly, the loss function as Mask R-CNN was used. Similarly, the cross-entropy function was used as the classification loss, the SmoothL1 loss function was used as the bounding box loss, the binary cross-entropy function was used as the mask loss, and the sum of each loss was obtained as the final loss value to optimize the model performance. A total of 100 iterations were carried out in the training phase of this experiment, and as shown in Fig. 7, the decline rate of the loss function tends to level off when the epoch is greater than 80, at which time the final model in the training phase is determined.

3. Results

The performance of the Co-Mask R-CNN network is demonstrated by fivefold cross-validation, and the results are shown in the table. This experiment counts the detection and segmentation results of the models separately to fully represent the performance of the Co-Mask R-CNN network.

As shown in the table, Co-Mask R-CNN achieves better performance than the other models; however, as the IOU

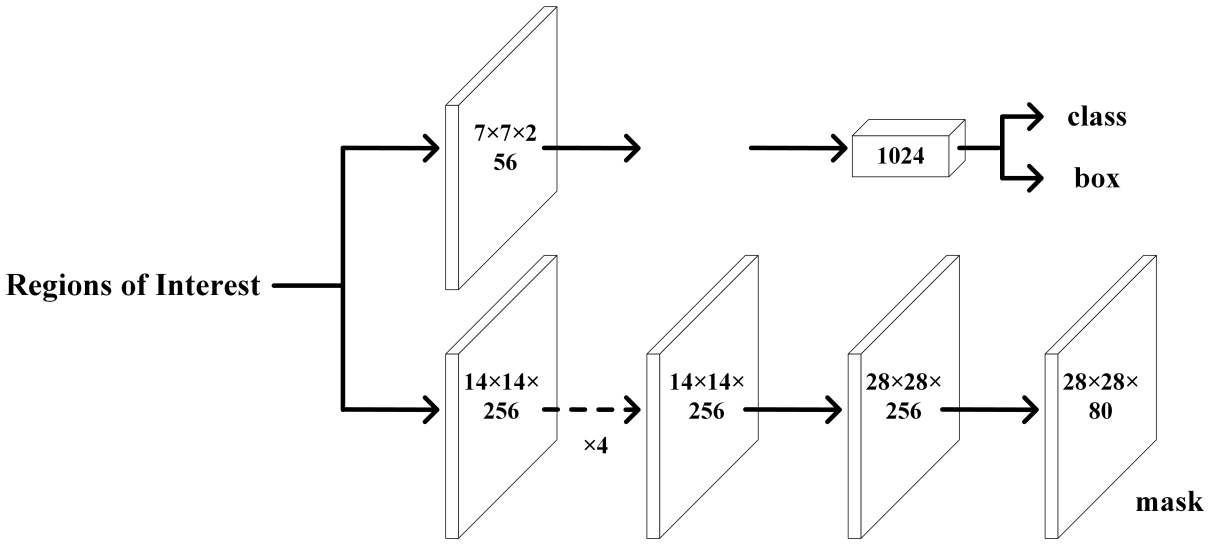


FIGURE 6. Prediction branch.

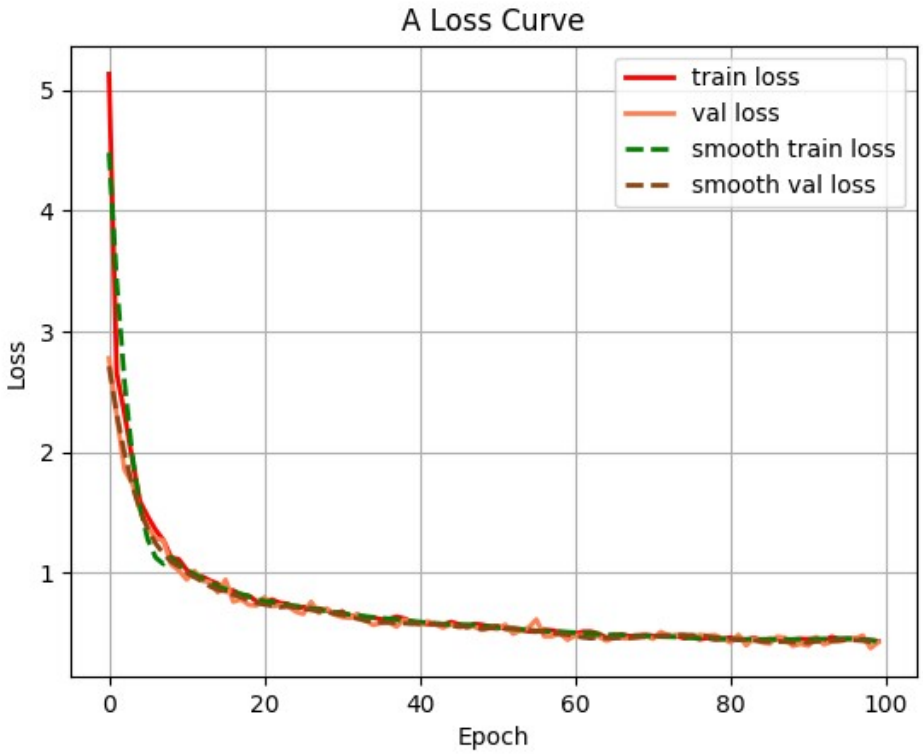


FIGURE 7. Loss curves.

confidence increases, both the detection and segmentation performances decrease. According to the segmentation results, as shown in Table 1, the best results in terms of precision for AP_{50}, AP_{75} and $AP_{50:95}$ are 0.988, 0.942 and 0.768, respectively, and the recall is 0.814. Table 2 shows that the best detection results in terms of precision for AP_{50}, AP_{75} and $AP_{50:95}$ are 0.988, 0.955 and 0.808, respectively, and the best result in terms of recall is 0.844. It can be seen that the performance of segmentation is lower than the performance of detection because segmentation belongs to the pixel-level task, which is more difficult compared to the detection result, and thus the performance is lower. performance is lower.

4. Discussion

4.1 Comparisons with state-of-the-art models

To assess the superior performance of Co-Mask R-CNN, we performed comparative experimental analysis with popular instance segmentation networks such as SSD (Single Shot MultiBox Detector) [32], YOLOv3 (You Only Look Once v3) [33], YOLOv7 (You Only Look Once v7) [34], and Mask R-CNN [15] and evaluated the performance based on their recall rate and accuracy at various thresholds. Notably, all

the models were trained without utilizing pretrained weights. Table 3 provides a comprehensive performance comparison of the tooth image dataset. Our Co-Mask R-CNN outperforms other classical models in terms of both precision and recall.

TABLE 1. Segmentation results.

	Average Precision			Average Recall
Methods	AP_{50}	AP_{75}	$AP_{50:95}$	$AR_{50:95}$
Fold 1	0.988	0.935	0.760	0.809
Fold 2	0.988	0.932	0.761	0.809
Fold 3	0.987	0.942	0.767	0.814
Fold 4	0.988	0.941	0.768	0.814
Fold 5	0.987	0.932	0.754	0.803
Average	0.987	0.936	0.762	0.809

TABLE 2. Detection results.

	Average Precision			Average Recall
Methods	AP_{50}	AP_{75}	$AP_{50:95}$	$AR_{50:95}$
Fold 1	0.988	0.955	0.805	0.840
Fold 2	0.988	0.954	0.808	0.843
Fold 3	0.987	0.952	0.807	0.844
Fold 4	0.988	0.945	0.800	0.840
Fold 5	0.988	0.951	0.805	0.841
Average	0.987	0.951	0.805	0.841

TABLE 3. Comparison of tooth segmentation results.

	Average Precision			Average Recall
Methods	AP_{50}	AP_{75}	$AP_{50:95}$	$AR_{50:95}$
SSD [32]	0.922	0.802	0.676	0.687
YOLOv3 [33]	0.948	0.830	0.688	0.654
YOLOv7 [34]	0.950	0.841	0.716	0.563
Mask R-CNN [15]	0.941	0.807	0.709	0.771
Co-Mask R-CNN	0.987	0.936	0.762	0.809

As shown in Table 3, at an IOU threshold of 50, the various models perform similarly in terms of precision. However, as the IOU threshold increases, Co-Mask R-CNN has more advantages, achieving a much higher precision rate than other classical models at IOU = (50:95), with a 5.3% improvement compared to Mask R-CNN. For recall, Co-Mask R-CNN still obtains the best results at high IOU thresholds, indicating that it has higher accuracy in detecting objects, and the generated detection boxes are closest to the true labels, resulting in more precise segmentation boundaries.

These experimental findings affirm the potential effectiveness of the proposed Co-Mask R-CNN as a viable solution for tooth instance segmentation tasks.

The experimental results of three images are visualized in Table 4, including a normal clear image, an image with metal artefact noise, and an image with uneven exposure. As shown in the first row of the normal clear image, SSD missed some teeth, while YOLOv3 and YOLOv7 made correct detections; however, the detection scores for each tooth box were much lower than those of the Co-Mask R-CNN. In the second row of the image with metal artefact noise, SSD and YOLOv7 missed some teeth, while Co-Mask R-CNN had higher detection scores than YOLOv3 and Mask R-CNN. In the third row, only YOLOv3 correctly detected teeth with uneven exposure in the image, while SSD and YOLOv7 missed some teeth. Although both Mask R-CNN and Co-Mask R-CNN exhibited instances of false detections, we demonstrated superior performance in terms of segmentation boundaries. The segmentation boundaries produced by the Co-Mask R-CNN were smoother and more accurate than those produced by the Mask R-CNN, resulting in a larger area of correct segmentation. Especially in scenarios where the object region tends to be larger than the background, the Co-Mask R-CNN excels and closely approximates the true label values. Its exceptional performance in accurately delineating tooth regions further establishes its efficacy in this specific segmentation task.

4.2 Ablation study

To validate the effectiveness of each module of the Co-Mask R-CNN, we performed four ablation experiments using the original image dataset and the enhanced image dataset. The first experiment involved training Mask R-CNN solely on the original image dataset, denoted as M_{s1}. The second experiment trained Mask R-CNN on the augmented image dataset, referred to as M_{s2}. The M_{s3} network model is a two-branch Mask R-CNN but does not use a feature fusion module. Simple feature fusion is performed by superposition of the two-branch feature map sampling channels, and the datasets are the original image and the enhanced image. Finally, the Co-Mask R-CNN model was trained on both the original and augmented image datasets. The instance segmentation results obtained from these experiments are presented in Table 5. In this study, the segmentation and classification results are separately discussed to facilitate comparison with the results of other methods.

4.3 Comparison of tooth segmentation results

A comparison of the results of the four sets of experiments is shown in Table 6. The quantitative evaluation shows that the collaborative learning model proposed in this paper improves the precision rate and recall by 5.3% and 3.8%, respectively, compared with the original Mask R-CNN model of M_{s1} at IOU = (50:95); however, M_{s2} has a decrease in the precision rate and recall by 2.1% and 2.7%, respectively, compared with M_{s1}. This is because the dataset used for M_{s2} is an enhanced image, and although the enhanced image makes the edge lines more obvious, the enhancement causes the image to lose more detail, worsening the segmentation results. Furthermore, M_{s3} has some performance improvement compared to M_{s1}, but the performance decreases compared to the model proposed in this experiment, thus proving the effectiveness of the proposed feature fusion module.

As shown in Table 5, we present the instance segmentation results of the ablation study on three representative dental

TABLE 4. Tooth instance segmentation results.

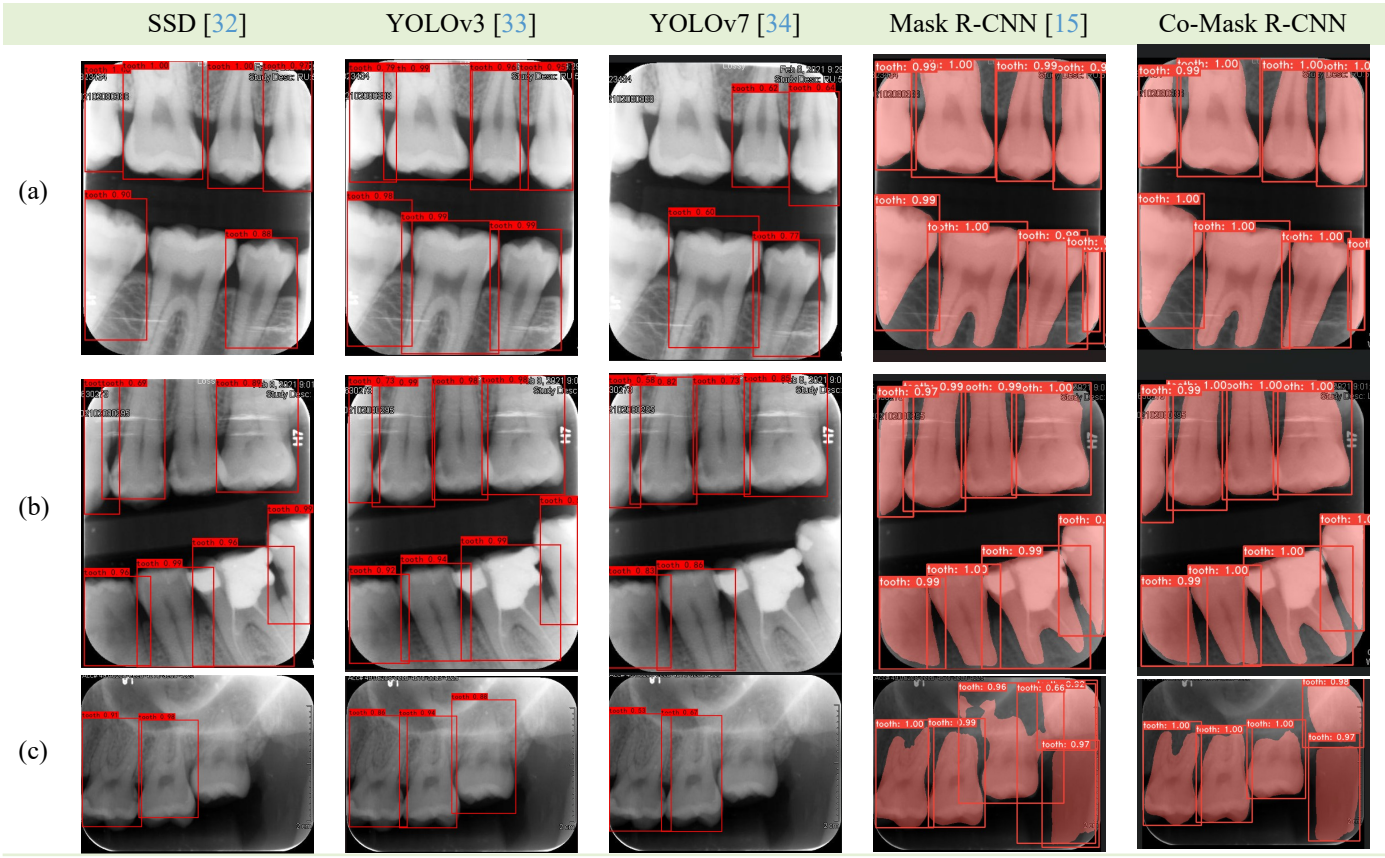
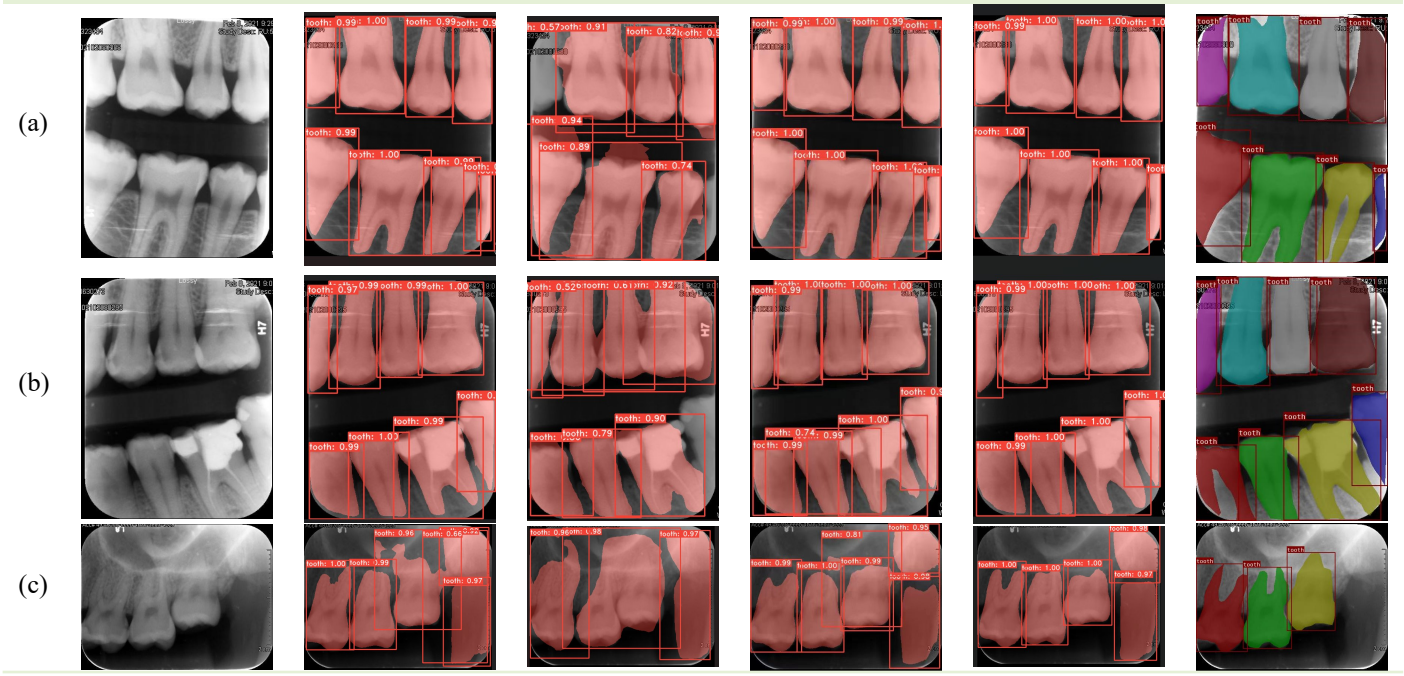


TABLE 5. Ablation study.



images: a clear normal image (a), an image with metal artefacts (b), and an image with tooth loss (c). In image (a), M_{s1} exhibited poor performance, with two missed teeth and unclear boundaries. In contrast, M_{s1}, M_{s3} and Co-Mask R-CNN accurately segmented each tooth, with Co-Mask R-CNN producing smoother and clearer boundaries. For image (b), M_{s2} also missed teeth and produced unclear boundaries, while Co-Mask R-CNN produced smoother and more accurate segmentation boundaries, indicating its robustness to metal artefacts. For image (c), all four experiments resulted in oversegmentation, highlighting the need for improved models for images with exposure interference and missing teeth. Compared with the original Mask R-CNN, the Co-Mask R-CNN shows an improvement in tooth segmentation accuracy, demonstrating the

TABLE 6. Comparison of tooth segmentation results.

	Average Precision			Average Recall
Method	AP_{50}	AP_{75}	$AP_{50:95}$	$AR_{50:95}$
M_{s1}	0.941	0.87	0.709	0.771
M_{s2}	0.946	0.857	0.688	0.744
M_{s3}	0.940	0.887	0.702	0.764
Co-Mask R-CNN	0.987	0.936	0.762	0.809

potential of attention to detail extraction for enhancing the segmentation performance of neural networks. Nonetheless, further research is required to address the challenges posed by dental images with complex backgrounds and missing or damaged teeth.

4.4 Comparison of tooth detection results

Table 7 shows the detection results of the four experiments, which demonstrate that the proposed collaborative instance segmentation model achieves the best performance in accurately identifying and detecting teeth. Compared to the Mask R-CNN M_{s1}, the Co-Mask R-CNN achieves 4.3% and 2.9% improvements in precision and recall, respectively, at IOU = (50:95), and the Co-Mask R-CNN also obtains better detection results than does M_{s3}. However, M_{s2} has the worst detection, which is in line with the segmentation results. This also proves the effectiveness of the proposed module in this study.

TABLE 7. Comparison of teeth detection results.

	Average Precision			Average Recall
Method	AP_{50}	AP_{75}	$AP_{50:95}$	$AR_{50:95}$
M_{s1}	0.934	0.844	0.762	0.812
M_{s2}	0.946	0.889	0.755	0.801
M_{s3}	0.958	0.906	0.765	0.817
Co-Mask R-CNN	0.987	0.951	0.805	0.841

In Table 5, the object detection labels and corresponding scores are presented above the bounding boxes. For image (a), Co-Mask R-CNN demonstrated the best detection performance, without any missed or false detections, with the detection boxes almost perfectly overlapping with the ground truth and with the highest detection score. M_{s1} and M_{s3} exhibited instances of multiple detections, while M_{s2} had the poorest detection results with missed and false detections. For image (b), Co-Mask R-CNN still demonstrated the best detection performance, with accurate detection boxes. M_{s1} detected the same number of boxes as did the Co-Mask R-CNN but with lower detection scores. M_{s2} exhibited the poorest detection results. For the image (c) with missing teeth, M_{s2} had the poorest detection results; M_{s1} and M_{s3} exhibited instances of false detections, detecting six teeth; and Co-Mask R-CNN also exhibited instances of false detections, detecting five teeth. Nevertheless, Co-Mask R-CNN still demonstrated the best detection performance, with higher detection scores for the correctly detected boxes than did M_{s1}. Therefore, the proposed instance segmentation model provides more accurate tooth detection, with higher recall, than does its original network.

Upon analysing the segmentation and detection results, we observed that the performance of the Co-Mask R-CNN surpassed that of the Mask R-CNN model (M_{s1}). Furthermore, as the IOU threshold increased, the differences among the four experiments became more prominent. The performance of the Co-Mask R-CNN model demonstrates its ability to produce results that closely align with the ground truth labels, providing further evidence of the effectiveness of the dual-branch fusion strategy proposed in this paper. Specifically, the attention-based feature fusion module prioritizes detailed information extraction, thereby integrating diverse multiscale contextual features. By employing an attention mechanism to fuse the weights of the original and enhanced images, the issues of inadequate global information extraction and potential segmentation errors caused by localization in the conventional Mask R-CNN were successfully resolved. This approach significantly enhances the model's generalization ability and results in more precise tooth instance segmentation. The fused model exhibits superior stability and surpasses the unfused results.

5. Conclusions

This paper proposes a tooth image segmentation network Co-Mask R-CNN, that integrates a self-attention mechanism and a collaborative learning strategy. To address the issues of low brightness and blurry boundaries in tooth images, we introduce an image enhancement method to reduce noise, unify image quality, and sharpen texture details.

The collaborative learning strategy is utilized to learn features in a joint manner from both the original and enhanced images. By incorporating an attention mechanism, the feature maps from the two branches are dynamically fused, facilitating the integration of complementary information. The attention mechanism addresses the challenge of inadequate extraction of low-level spatial information and global contextual information by the network. This approach effectively mitigates these limitations and enables the network to capture more comprehensive and meaningful spatial and contextual details. By incorporating the CAM and PAM, the proposed approach enhances the spatial relationships between pixels, resulting in improved tooth image segmentation performance. This method demonstrates robust applicability and can be effectively employed in various medical image segmentation tasks. It exhibits versatility and the potential to produce reliable results across different medical imaging scenarios. In future research, our focus will be on further optimizing the multiscale aspects of the method. It is worth mentioning that the present model is restricted by the segmentation of two-dimensional medical image slices. Nevertheless, our future endeavours will focus on addressing segmentation tasks for higher-dimensional medical images.

AVAILABILITY OF DATA AND MATERIALS

The data presented in this study are available on reasonable request from the corresponding author.

AUTHOR CONTRIBUTIONS

CW and JYY—designed the research study; performed the research. PY and XJJ—analyzed the data. CW, JYY, PY and XJJ—wrote the manuscript. HZL and RJL—revised the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the Ethics Committee of Peking University School and Hospital of Stomatology (ref no. PKUSSIRB-202272029).

ACKNOWLEDGMENT

Not applicable.

FUNDING

This research was funded by the Beijing Natural Science Foundation and Haidian Original Innovation Joint Fund (grant number: L222052), the National Natural Science Foundation of China (grant number: 62201018), the National Science and Technology Major Project (grant number: 2022ZD0119502), and the Youth Founding of Peking University School and Hospital of Stomatology (grant number: PKUSS20220109).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Kumar R, Saha P. A review on artificial intelligence and machine learning to improve cancer management and drug discovery. International Journal for Research in Applied Sciences and Biotechnology. 2022; 9: 149–156.

[2] Huang C, Wang J, Wang S, Zhang Y. A review of deep learning in dentistry. Neurocomputing. 2023; 554: 126629.

[3] Helli S, Hamamci A. Tooth instance segmentation on panoramic dental radiographs using u-nets and morphological processing. Duzce University Journal of Science and Technology. 2022; 10: 39–50.

[4] Martins M V, Baptista L, Luís H, Assunção V, Araújo MR, Realinho V. Machine learning in x-ray diagnosis for oral health: a review of recent progress. Computation. 2023; 11: 115.

[5] Zhang J, Jiang Y, Gao F, Zhao S, Yang F, Song L. A fast automatic reconstruction method for panoramic images based on cone beam computed tomography. Electronics. 2022; 11: 2404.

[6] Polizzi A, Quinzi V, Ronsivalle V, Venezia P, Santonocito S, Giudice A Lo, *et al*. Tooth automatic segmentation from CBCT images: a systematic review. Clinical Oral Investigations. 2023: 27: 3363–3378.

[7] Bruellmann D, Sander S, Schmidtmann I. The design of an fast fourier filter for enhancing diagnostically relevant structures—endodontic files. Computers in Biology and Medicine. 2016; 72: 212–217.

[8] Oltu B, Karaca BK, Uyar T, Uyar DS. Detection of occlusal plaque and caries using fuzzy C means based segmentation algorithm. 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). IEEE. 2021; 1–5.

[9] Wang C, Huang C, Lee J, Li C, Chang S, Siao M, *et al*. A benchmark for comparison of dental radiography analysis algorithms. Medical Image Analysis. 2016; 31: 63–76.

[10] Patil S, Kulkarni V, Bhise A. Algorithmic analysis for dental caries detection using an adaptive neural network architecture. Heliyon. 2019; 5: e01579.

[11] Yau H, Yang T, Chen Y. Tooth model reconstruction based upon data fusion for orthodontic treatment simulation. Computers in Biology and Medicine. 2014; 48: 8–16.

[12] Wang Y, Liu S, Wang G, Liu Y. Accurate tooth segmentation with improved hybrid active contour model. Physics in Medicine and Biology. 2018; 64: 015012.

[13] Salimzadeh S, Kandulu S. Teeth segmentation of bitewing X-ray images using wavelet transform. Informatica. 2020; 44: 421–426.

[14] Modi CK, Desai NP. A simple and novel algorithm for automatic selection of ROI for dental radiograph segmentation. 2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE. 2011; 504–507.

[15] He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. 2017 IEEE International Conference on Computer Vision. 2017; 2961–2969.

[16] Jader G, Fontineli J, Ruiz M, Abdalla K, Pithon M, Oliveira L. Deep instance segmentation of teeth in panoramic X-ray images. 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE. 2018; 400–407.

[17] Pinheiro L, Silva B, Sobrinho B, Lima F, Cury P, Oliveira L. Numbering permanent and deciduous teeth *via* deep instance segmentation in panoramic X-rays. 17th International Symposium on Medical Information Processing and Analysis. 2021; 12088: 95–104.

[18] AlQarni S, Chandrashekar G, Bumann EE, Lee Y. Incremental learning for panoramic radiograph segmentation. 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE. 2022; 557–561.

[19] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems. 2015; 28.

[20] Lee J, Han S, Kim YH, Lee C, Kim I. Application of a fully deep convolutional neural network to the automation of tooth segmentation on panoramic radiographs. Oral Surgery, Oral Medicine, Oral Pathology, and Oral Radiology. 2020; 129: 635–642.

[21] Zhao Y, Li P, Gao C, Liu Y, Chen Q, Yang F, *et al*. TSASNet: tooth segmentation on dental panoramic X-ray images by two-stage attention segmentation network. Knowledge-Based Systems. 2020; 206: 106338.

[22] Chung M, Lee M, Hong J, Park S, Lee J, Lee J, *et al*. Pose-aware instance segmentation framework from cone beam CT images for tooth segmentation. Computers in Biology and Medicine. 2020; 120: 103720.

[23] Silva B, Pinheiro L, Oliveira L, Pithon M. A study on tooth segmentation and numbering using end-to-end deep neural networks. 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE. 2020; 164–171.

[24] Wang K, Liew JH, Zou Y, Zhou D, Feng J. PANet: few-shot image semantic segmentation with prototype alignment. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE. 2019; 9197–9206.

[25] Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, *et al*. ResNeSt: split-attention networks. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE. 2022; 2736–2746.

[26] Chen K, Pang J, Wang J, Xiong Y, Li X, Sun S, *et al*. Hybrid task cascade for instance segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2019; 4974–4983.

[27] Leite AF, Gerven AV, Willems H, Beznik T, Lahoud P, Gaêta-Araujo H, *et al*. Artificial intelligence-driven novel tool for tooth detection and segmentation on panoramic radiographs. Clinical Oral Investigations. 2021; 25: 2257–2267.

[28] Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, *et al*. Dual attention network for scene segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2019; 3146–3154.

[29] Zhou Y, He X, Huang L, Liu L, Zhu F, Cui S, *et al*. Collaborative learning of semi-supervised segmentation and classification for medical

images. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2019; 2079–2088.

[30] Kumar A, Fulham M, Feng D, Kim J. Co-learning feature fusion maps from pet-CT images of lung cancer. IEEE Transactions on Medical Imaging. 2020; 39: 204–217.

[31] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2016; 770–778.

[32] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, *et al*. SSD: single shot multibox detector. Computer Vision—ECCV 2016. 2016; 104: 21–37.

[33] Redmon J, Farhadi A. Yolov3: an incremental improvement. arXiv. 2018; 1804.02767.

[34] Wang C, Bochkovskiy A, Liao HM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2023; 7464–7475.

How to cite this article: Chen Wang, Jingyu Yang, Hongzhi Liu, Peng Yu, Xijun Jiang, Ruijun Liu. Co-Mask R-CNN: collaborative learning-based method for tooth instance segmentation. Journal of Clinical Pediatric Dentistry. 2024; 48(6): 161-172. doi: 10.22514/jocpd.2024.136.