Improved RA-U-Net for Dental Image Segmentation

Du Wenjie[1] and Du Wangchun[2,*]

[1]*Shanghai Zhongqiao Vocational and Technical University, 201514, China*

[2]*Shanghai University of Medcine and Health Science, 201318, China*

Abstract: The diversity of tooth morphology, irregular arrangement, and blurred grayscale characteristics of the alveolar bone present significant challenges in image recognition and segmentation. While deep learning-based U-Net networks have achieved considerable progress in medical image segmentation, this paper proposes the RA-U-Net architecture, an enhanced version of U-Net, achieving high-precision segmentation of multimodal dental images. To address challenges in dental image segmentation, the U-Net architecture is enhanced by incorporating residual modules and attention mechanisms. A group normalization strategy is applied for feature map channel segmentation, improving deep network learning capabilities and model generalization. The network was tested on a dental CBCT image dataset from the Shanghai Ninth People's Hospital. Results demonstrated an average Dice coefficient of 0.839, surpassing other image segmentation networks, with a precision of 0.961 and a recall of 0.719. These findings indicate that the RA-U-Net model delivers superior image segmentation performance, providing reliable support for dental diagnosis and treatment.

Keywords: RA-U-Net, Residual Module, Attention Mechanism, Dental Image Segmentation.

INTRODUCTION

Based on high-definition dental medical images, techniques such as image enhancement, image segmentation, and three-dimensional model reconstruction enable the acquisition of highly precise three-dimensional diagnostic and therapeutic data for teeth. This significantly advances clinical diagnostic capabilities and treatment standards, mitigating human diagnostic errors caused by dentists relying solely on intuition or experience. Consequently, it enhances the efficiency and success rate of dental diagnosis and treatment.

Cone Beam Computed Tomography (CBCT) medical images are computer-reconstructed tomographic representations of target structures obtained through cone-beam projection. CBCT plays a vital role in dental diagnostics and treatment, addressing the limitation of conventional dental CT scans in accurately assessing spatial relationships between teeth.

1. RA-U-NET NETWORK INCORPORATING RESIDUAL MODULES AND ATTENTION MECHANISMS

The U-Net architecture was first proposed in 2015 [1] and was quickly applied to biological image segmentation. In the field of medical image edge segmentation, the U-Net model employs a relatively simple encoder-decoder structure. Its fully convolutional form imposes minimal requirements on sample size. By utilizing skip connections during feature fusion, it requires fewer training samples and does not necessitate model pre-training, as shown in Figure **1**.

When tackling dental image segmentation tasks, the following challenges arise: First, the diverse shapes, varying sizes, irregular arrangements, and uneven grayscale values of teeth in images make feature extraction difficult. Second, slow model training imposes high time requirements. Finally, significant hardware demands are involved. Given these challenges in dental image segmentation and the limitations of the U-Net architecture, this paper proposes the RA-U-Net model, inspired by U-Net, residual modules, and the Attention mechanism. Experimental results demonstrate that RA-U-Net achieves superior performance in dental image segmentation tasks.

1.1. Basic Structure of the RA-U-Net Model

The RA-U-Net model ensures training effectiveness for deep networks and accelerates training speed by incorporating residual modules. Through the Attention mechanism, it focuses on tooth edge regions, enhances weight values in these areas, and suppresses irrelevant regions, thereby increasing the proportion of effective features [2]. To prevent gradient vanishing and accelerate training in deep neural networks, this model incorporates Group Normalization (GN) layers. These divide feature map channels into groups, thereby circumventing constraints imposed by batch size limitations.

The network architecture is shown in Figure **2**. The RA-U-Net model takes a single-channel CBCT image as input and produces a single-channel image with annotated dental segmentation as output. The encoding part of the network downsamples the input

*Address correspondence to this author at the Shanghai University of Medcine and Health Science, 201318, China; E-mail: sddu126@126.com

Received on 23-09-2025 Accepted on 27-10-2025 Published on 29-10-2025

E-ISSN: 2311-8695/25

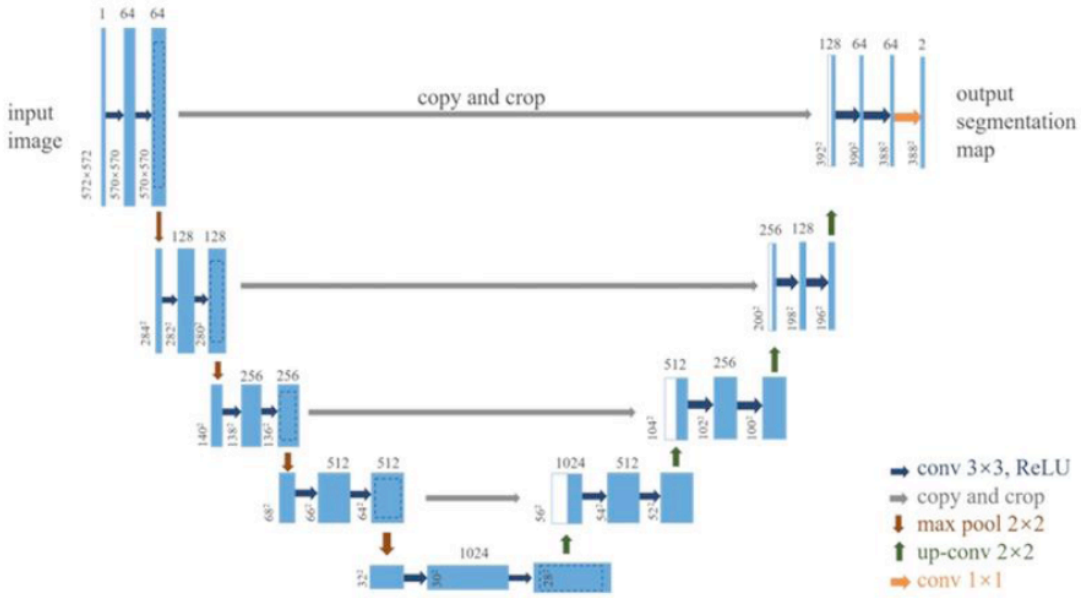

Figure 1: U-Net structure diagram.

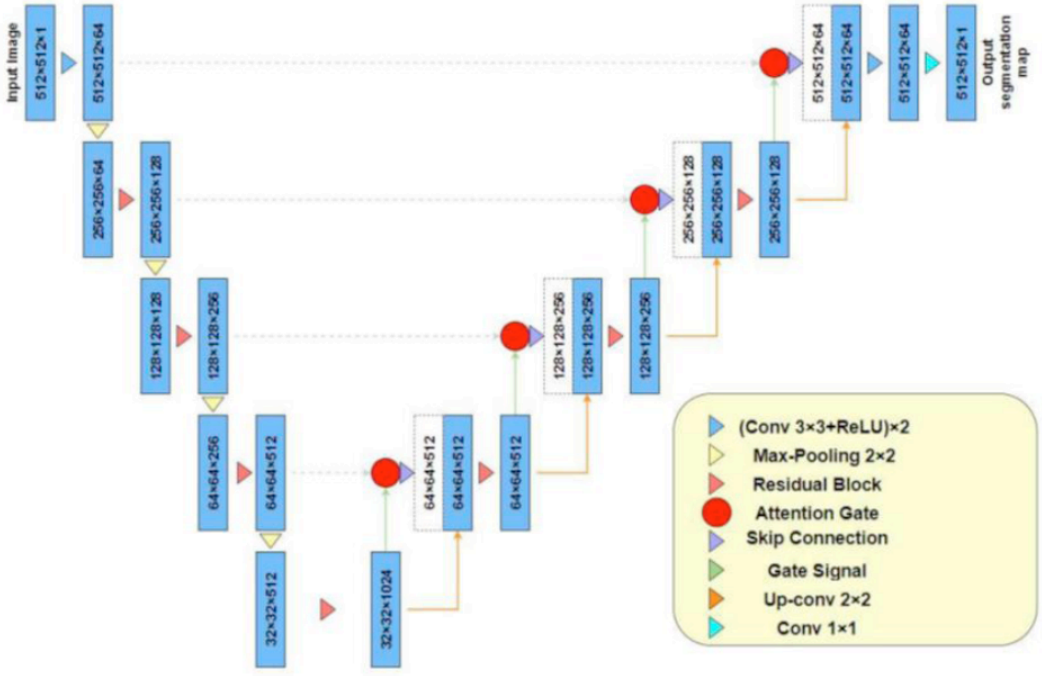

Figure 2: RA-U-Net structure diagram.

using 2×2 max pooling layers, while the decoding part expands the dimensions through deconvolution to restore the output to its original size. This paper replaces the original convolutional layers with residual modules and incorporates attention gates (AGs) into the skip connections.

1.2. Residual Modules

Gradient vanishing and gradient explosion have always accompanied the backpropagation process in deep neural networks. While the introduction of Batch Normalization (BN) layers significantly mitigated these issues, it also led to network degradation. In 2015 Kaiming He *et al*. [3] proposed residual networks(ResNet), introducing the concept of residual connections. Applying residual modules in medical image segmentation addressed the problem of network degradation [4]. This model employs an enhanced bottleneck residual module arranged as follows: 1×1 convolution unit → 3×3 convolution unit → 1×1 convolution unit. The 1×1 convolution unit adjusts the number of data channels. When channel counts are

large, it first reduces the channels, then increases them after the 3×3 convolution operation. This approach reduces computational load while accelerating training speed.

1.3. Attention Mechanism

The attention mechanism draws inspiration from human visual characteristics. When applied to medical image processing, it suppresses irrelevant regions while enhancing weight assignments for target areas, thereby improving both information processing efficiency and accuracy [5]. The attention gate proposed in Attention U-Net [6] enhances the sensitivity and prediction accuracy of neural network models while maintaining low computational complexity. Since lower-level features capture more accurate information, the attention gate structure combines these lower-level features with upper-level features, as illustrated in Figure **3**.

Attention gates represent a soft attention mechanism that progressively amplifies the weight values of locally relevant regions while suppressing activity in irrelevant areas within complex scenes, offering broad application prospects. Although the placement of attention modules imposes few constraints, this model positions them at the skip connections within the original U-Net neural network to optimize convergence speed and facilitate modifications.

1.4. Weighted Loss Function Design

In medical image processing, positive-negative sample imbalance frequently occurs. Furthermore, due to the limited volume of dental image datasets, addressing sample imbalance becomes particularly challenging. When using the commonly employed BCE loss function to handle such imbalanced data, it often occurs that although the loss value decreases significantly, the segmentation performance of the final trained model is unsatisfactory. This is because when the number of positive samples is far smaller than that of negative samples, the model tends to avoid making predictions. Therefore, weighting the loss function is necessary. This model employs a combined loss function of BCE and DSC, ultimately defined as: $L = \alpha L_{BCE} + \beta L_{Dice}$, where α and β are weighting coefficients that adjust the proportion of BCE and Dice loss functions. The objective of this experiment is target region segmentation. The experimental data exhibits an imbalance where target regions occupy a low proportion while background regions dominate. Overreliance on BCE loss may cause the model to favor predicting the dominant background regions, leading to missed detections of target regions. By increasing the weight of the Dice loss, the model focuses on optimizing the intersection-over-union ratio between targets and background, thereby improving segmentation accuracy under class imbalance. To validate the weighting rationale, preliminary weight gradient experiments were conducted, testing five weight combinations (α =1, β =1; α =1, β =0.5; α =0.3, β =1; α =0.5, β =1; α =0.7, β =1). The combination α =0.5, β =1 yielded the highest Dice coefficient and lowest omission rate, simultaneously achieving optimal segmentation accuracy and training stability. After experimentation, this model sets α to 0.5 and β to 1, using this loss function for subsequent experiments.

1.5. Model Evaluation Metrics

Common evaluation metrics in medical image segmentation include: Dice coefficient, precision, and recall.

The Dice coefficient measures the similarity between two samples, as shown in Formula (1).

$$Dice = \frac{2\mathrm{TP} \times \mathrm{recall}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} = \frac{2|X \cap Y|}{|X| + |Y|} \quad (1)$$

Here, X represents the segmentation results predicted by the algorithm, while Y denotes the gold standard annotated by experts. The Dice coefficient ranges from 0 to 1, reflecting the similarity between the neural network model's predictions and the gold

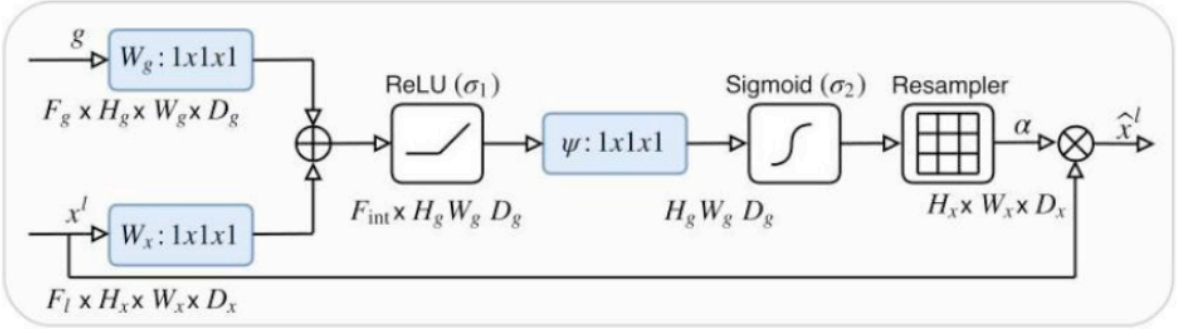

Figure 3: The structure of the attention gate.

standard. A Dice value closer to 1 indicates superior segmentation performance, with minimal discrepancy between the predicted and actual results; conversely, a value closer to 0 signifies poor segmentation quality, indicating significant deviation from the true segmentation.

Precision measures the accuracy of a prediction, representing the proportion of samples correctly predicted as positive among those predicted as positive. It can be calculated using Equation (2).

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\% \quad (2)$$

TP denotes true positives, while FP denotes false positives. The closer the value approaches 1, the fewer negative samples are misclassified as positive samples. In medical image segmentation, pixels matching the predicted results with the gold standard can be classified as positive samples, while mismatched pixels are classified as negative samples. This allows us to determine the proportion of pixels in the predicted results that correspond to the lesion regions outlined by the gold standard.

The recall rate indicates the proportion of actual positive samples that are correctly predicted, as shown in Equation (3).

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \quad (3)$$

FN denotes false negatives. The closer its value is to 1, the more pixels are successfully predicted in the gold standard, indicating better image segmentation performance.

Precision and recall are interdependent and often conflicting metrics. For instance, when a neural network model achieves high precision, it typically exhibits low recall. Therefore, we must strike a balance between the two, striving to maintain both precision and recall at relatively high levels within the neural network model.

2. EXPERIMENTAL DATA AND PROCESSING PROCEDURES

2.1. Experimental Dataset and Gray-Scale Preprocessing

To evaluate the performance of the proposed model, CBCT images from six patients were randomly selected from the dental CBCT image database at Shanghai Ninth People's Hospital. The dataset comprised 95 teeth and 1,150 distinct cross-sectional slices. The number of slices per patient ranged from [184, 198], with each patient having 184, 186, 192, 194, 196, and 198 slices. To ensure data partitioning independence and rationality, this study adopted a “patient-level stratified division” approach. The overall data was divided according to a 17∶4∶4 ratio: the training set contained 780 layers, the validation set 184 layers, and the test set 186 layers. Given the relatively small sample size of this study, to prevent model overfitting, we strictly restricted data distribution from the same patient across different datasets. Cross-validation was introduced, with an additional 4-fold cross-validation conducted on top of the patient-based partitioning. Furthermore, an early stopping mechanism was enabled: training would immediately halt if the validation set showed no decrease after 20 consecutive iterations.

Based on the DICOM data protocol for acquisition and transmission, CBCT image data files in DICOM 3.0 standard format were obtained for these 95 teeth. The CBCT scan parameters were set as follows: voltage 90kV, current 10mA, scan time 300ms, resolution 0.1mm × 0.1mm. DICOM format images represent an industry-standard image format for medical imaging and its associated data, enabling communication across different devices. They are widely adopted in medical imaging. A DICOM standard format file typically consists of one header file and one dataset. The header file contains the original metadata of the image data, while the dataset usually comprises multiple data elements defined by identifiers, types, lengths, and values [7].

Some DICOM files lack preset window width (W) and window level (C) values. To set these parameters, locate the maximum and minimum gray values (g_{max} and g_{min}) across the entire image, referencing formulas (4) and (5). Subsequently, apply windowing to maximize contrast between different gray levels, achieving gray-level histogram equalization.

$$W = g_{max} - g_{min} \quad (4)$$

$$C = \frac{g_{max} - g_{min}}{2} \quad (5)$$

To enhance the clarity of the displayed image, a logarithmic transformation method using a window function is applied on top of the above process. This maps all gray levels to the range of 0 to 255 gray levels, thereby capturing the full range of displayable image information. Refer to Formula (6).

$$G(V) = \frac{255log(V)+1}{log(V_{max})} \quad (6)$$

Here, V denotes the original grayscale value of the windowed image, while $G(V)$ represents the displayed grayscale value of the windowed image.

2.2. Image Enhancement and Noise Reduction

During dental scanning and imaging, varying degrees of geometric and grayscale blurring occur, necessitating image enhancement processing. For issues such as blurred boundaries between teeth and their corresponding root canals and dentin, it is essential to filter out as much noise as possible to enhance image clarity while preserving the integrity of image information, ensuring no blurring or loss of contour details occurs. For addressing blurring in dental images, commonly used adaptive enhancement methods include the CLAHE enhancement algorithm and the Laplace correction algorithm.

CLAHE stands for Contrast-Limited Adaptive Histogram Equalization [8]. When certain regions within an image are significantly brighter or darker than others, conventional histogram equalization algorithms struggle to preserve the fine details in those areas. The enhanced dental images exhibit improved local contrast compared to the original CBCT dental images while preserving edge information, referred to Equation (7).

$$g(x,y) = CLAHE\{f(x-k, y-l), k, l \in W\} \quad (7)$$

$g\ (x,\ y)$ represents the enhanced grayscale value, W denotes a two-dimensional template typically consisting of a 3-pixel or 5-pixel square region; k and l serve as stretching factors, whose values can be adjusted to modulate the intensity of image enhancement.

To address texture blurring caused by variations in lighting conditions during imaging, the Laplace correction algorithm can be applied to correct the issue of grayscale values being concentrated within a narrow range in dental CBCT images.

The Laplace correction algorithm is a method based on the Laplacian operator of second-order derivatives[9]. The Laplacian operator for a two-dimensional function f (x, y) can be referenced in formula (8).

$$\nabla^2 f(x,y) = \frac{\partial y^2 f(x,y)}{\partial x^2} + \frac{\partial y^2 f(x,y)}{\partial y^2} \quad (8)$$

Simplifying the above formula yields the simplified Laplace operator, as shown in equation (9).

$$\nabla^2 f(x,y) = f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) - 4f(x,y) \quad (9)$$

By constructing enhanced filter templates W1 and W2, all pixels in the tooth image are computed using a simplified Laplace operator. Refer to Equations (10) and (11).

$$W1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (10)$$

$$W2 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad (11)$$

First, randomly select 200 representative CBCT slice images exhibiting varying grayscale and blur characteristics. Apply CLAHE enhancement and Laplace correction methods respectively to these images. Based on the enhancement outcomes, determine the optimal enhancement method for the remaining images. For operational convenience, the selected CBCT slice images meet uniform specifications, excluding those with severe artifacts or inconsistent slice thickness. Core information is concentrated in the grayscale channel, and normalization processing is performed simultaneously. Python tools were chosen for CLAHE enhancement with preset parameters: tile grid size 8×8, contrast limit 2.0, and 256 grayscale levels. Laplace correction was performed using a classic 3×3 negative center filter template to enhance edge detection and detail, as referenced in Equation (10). After convolution, the original image was fused with the enhanced edge image at a weight ratio of 1:0.5 to generate the Laplace-enhanced image, requiring no additional parameter adjustments.

This experiment compares the enhanced original dental images with their corresponding gold standard to prevent mismatches between images and labels, as shown in Figure **4**.

Before enhancement, the gray-scale contrast between teeth and surrounding jawbone soft tissues is minimal, resulting in unclear tissue boundaries. After enhancement, the tooth regions are significantly intensified, greatly improving differentiation from other tissues and enabling direct visualization of each tooth's morphology. Before enhancement, fine structures such as tooth cusps and interdental spaces are insufficiently highlighted in gray-scale images. After enhancement, these details are clearly presented, facilitating precise observation of dental anatomy. The background of the unenhanced image contains substantial gray-scale information from non-dental tissues, introducing interference. After enhancement, the background is suppressed, directing visual focus entirely onto the teeth and improving the efficiency of dental analysis.

2.3. Experimental Environment

This experiment utilizes the PyTorch deep learning framework. The batch size parameter is set to 2, with 60 epochs of iteration. The Adam optimization

Figure 4: Original images (**left**) and enhanced images (**right**).

algorithm is employed, with an initial learning rate of 0.0001 and the ReLU activation function. The original data undergoes data augmentation to form the training set, which is fed into the neural network model for training. The validation set does not alter model parameters and serves solely as a basis for hyperparameter tuning. The test set is used to evaluate the model's final training results.

3. EXPERIMENTAL RESULTS AND ANALYSIS EXPERIMENTAL RESULTS AND ANALYSIS

Since the dental images used in the experiment were low-resolution and single-channel grayscale images, CBCT image blocks of varying sizes were first cropped. The input images were uniformly resized to 128×128 for the experiment. When extracting image features, two consecutive 3×3 convolutional kernels were applied for the convolution operation; Following ReLU activation operation, 2×2 max pooling with a stride of 2 was applied. After repeating this process three times, the feature map size was reduced to 16×16 with a dimension of 128.

Then, the feature map undergoes the upsampling to restore it. First, the feature map is upsampled using a 2×2 convolution kernel. The resulting feature map is then merged with the lower-level feature map. Subsequently, two 3×3 convolution kernels are applied to reduce the number of channels. After repeating this operation three times, the feature map size matches the input image dimensions, with the dimension reduced to 16. Finally, a 1×1 convolution operation is performed on the feature map, mapping the 16-dimensional feature to a 2-dimensional segmentation result.

To validate the superiority of this experimental method for dental image segmentation, the RA-U-Net model was compared with U-Net models employing other modules. All models utilized the same dataset, training methodology, and data processing procedures. To evaluate the image segmentation results, three metrics of Dice, Precision, and Recall were statistically analyzed across different models. The test results for each model are presented in Table **1**.

Table 1: Performance Evaluation Results of Different Models

Model	Dice	Precision	Recall
U-Net	0.716	0.825	0.767
U-Net + Att	0.737	0.894	0.582
U-Net + Res	0.795	0.912	0.538
RA-U-Net	0.839	0.961	0.719

This paper compares the image segmentation results of the U-Net model, the U-Net + Att model with attention mechanisms, the U-Net + Res model with residual modules, and the RA-U-Net model. Table **1** shows that the RA-U-Net model achieves a Dice coefficient of 0.839, representing an approximately 17% improvement over the 0.716 achieved by the traditional U-Net model. This higher segmentation accuracy in clinical evaluations helps prevent treatment errors caused by misclassification. The Precision coefficient of the RA-U-Net model is 0.961, representing an approximately 16% improvement over the traditional U-Net network model's 0.825. This higher value indicates fewer non-dental tissues included in segmentation results during clinical evaluation, thereby enhancing the precision of fabricating devices such as dental braces. The Recall coefficient of the RA-U-Net model is 0.719, representing a decrease of approximately 6% compared to the traditional U-Net network model's 0.767. This lower value indicates partial tooth omission in segmentation during clinical evaluation, potentially affecting the quality of three-dimensional tooth reconstruction. This primarily occurs because the model prioritizes ensuring segmented regions are genuine tooth tissue, leading to the omission of some genuinely blurred tooth edges. The RA-U-Net model outperforms the residual-enhanced U-Net + Res model in dental edge region segmentation, which achieving an overall segmentation accuracy of 0.795, precision of 0.912, and recall of 0.538.

Compared to RA-U-Net, the traditional U-Net model lacks residual modules and attention mechanisms, resulting in discontinuity between convolutional layers and lower feature utilization. The U-Net + Res model, which only incorporates residual modules, exhibits insufficient focus on key target regions, overly simplistic skip connections, and inadequate feature fusion between layers.

RA-U-Net features fewer parameters than the traditional U-Net, indicating that RA-U-Net does not rely on increased parameter count to achieve improved dental image segmentation accuracy. The experimental results above demonstrate the superiority of the RA-U-Net model, which combines residual modules and attention mechanisms in extracting features from dental images.

We also observe that the segmentation accuracy of the U-Net + Att model, which only incorporates the attention module, is suboptimal. Analysis reveals the following reasons: After passing through the Attention module, the upper layer and lower layer features of the image are transformed via the Sigmoid function into soft masks with values between [0,1]. When performing dot-product operations with the input features, this process may weaken the influence of deep-layer features and inhibit the segmentation of small objects. As the network deepens, feature propagation distortion and degradation occur, leading to reduced segmentation accuracy. The residual module directly transmits effective features from shallow layers through skip connections, fundamentally preventing this degradation and ensuring critical information is preserved in deeper layers. Fine structures like tooth cusps and interdental spaces are key evaluation points in CBCT segmentation. While attention modules can enhance relevant key features, they cannot recover details lost during forward propagation. The skip connections in the residual module transmit detailed features captured in shallow layers—such as tooth cusps and interdental spaces—to deeper layers, preventing loss of fine details.

4. SUMMARY

This experiment proposes an enhanced U-Net model incorporating residual modules and attention mechanisms, termed RA-U-Net. Building upon the original U-Net architecture, RA-U-Net replaces conventional convolutional layers with residual modules, thereby reducing required model parameters while enhancing the learning capacity of deep networks. Attention modules are incorporated at skip connections to link features across layers and suppress irrelevant regions. Replacing Batch Normalization (BN) layers with Group Normalization (GN) layers significantly accelerates convergence, improves generalization, and enhances segmentation performance during training with small batch sizes. Comparative experiments conclusively demonstrate that the proposed RA-U-Net outperforms other modified U-Net deep learning approaches.

In future dental image segmentation applications, increasing the training dataset size, enhancing data preprocessing effectiveness, and optimizing experimental parameters can further improve the performance of dental CBCT image segmentation methods. Additionally, three-dimensional dental CBCT images can be processed using 3D neural networks.

CONFLICTS OF INTEREST

The authors report no conflict of interest, and this research received no specific grant from any funding agency.

REFERENCES

[1] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//

International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. 2015: 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[2] DU Wenjie, Wang Yuanjun. Improved RA-U-Net network for liver tumor segmentation of CT images [J]. Progress in Biomedical Engineering, 2025, (5)

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, etc. Deep Residual Learning for Image Recognition[C]. CVPR IEEE Xplore, 2015,12,5

[4] Dong X Y, Chen P. Segmentation of liver tumors based on bottleneck residual attention mechanism U-net [J]. CT Theory and Applications, 2021, 30(6): 661-670.

[5] Xu P Q, Liang Y X, LI Y. Medical image segmentation fusing multi-scale semantic and residual bottleneck attention[J]. Computer Engineering, 2023, 49(1 0): 1 62-1 70.

[6] Oktay O, Schlemper J, Folgoc LL, *et al*. Attention U-Net: Learning where to look for the pancreas. arXiv: 1804. 03999, 2018.

[7] HE Bin, JIN Yongjie, LI Yulan. Defining nuclear medical file formal based on DICOM standard [J]. Nuclear Electronics & Detection Technology, 2001, 21(6): 440-443.

[8] Reza A M. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement [J]. Journal of Vlsi Signal Processing, 2004, 38(1): 35-44. https://doi.org/10.1023/B:VLSI.0000028532.53893.82

[9] PAN Zhenkuan, WEI Weibo, ZHANG Haitao. Variational models for image diffusion based on gradient and Laplacian [J]. Journal of Shandong University (Natural Science), 2008, 43(11): 11-16.

https://doi.org/10.12974/2311-8695.2025.13.06