MRAUnet++: A Novel Multi-Scale Residual Attention Network for Enhanced Rectal Cancer Segmentation

Zhengpeng Li, Member, IANEG, Jun Hu, and Zhuang Liang, Jiansheng Wu*

Abstract—**Deep learning (DL) models play a crucial role in medical image analysis, with their performance reliant on the scale and diversity of available training data. However, medical imaging faces challenges like data acquisition difficulty, sample class imbalance, and high annotation costs. To tackle these issues and enable automatic and accurate segmentation of rectal tumor regions in CT images, aiding physicians in diagnosis and treatment, we present the innovative Multi-scale Residual Attention-based Unet++ Network (MRAUnet++). This model replaces UNet++'s convolutional blocks with multi-scale residual blocks, extracting features across different scales for enriched diversity. Residual connections optimize the network structure, enhancing its ability to express complex features. The integrated attention mechanism suppresses irrelevant features, boosting segmentation performance by allowing selective focus on crucial rectal tumor features. Empirical evaluations on a rectal tumor dataset demonstrate MRAUnet++'s outstanding performance, achieving a Dice coefficient of 83.32% and an IoU value of 70.06%, surpassing other comparative models. These results highlight the model's enhanced accuracy in rectal tumor segmentation, providing a reliable diagnostic tool for medical practitioners.**

Index Terms—**Semantic segmentation, Residual Network, Attention mechanism, Multiscale feature**

I. Introduction

In the contemporary healthcare landscape, colorectal cancer ranks as the third most prevalent cancer in the United States, with rectal cancer constituting a significant proportion of diagnoses, posing substantial challenges to the healthcare system[1]. Advances in medical science have demonstrated that accurate tumor screening techniques can effectively reduce both incidence and mortality rates of rectal cancer. Precision in the segmentation of rectal tumor regions, especially in early-stage diagnosis and treatment, is crucial. Traditionally, this segmentation process relies on manual delineation by physicians using CT images, which is a time-consuming task with accuracy dependent on the clinician's experience and expertise.

Manuscript revised Nov. 15, 2023; revised Mar. 4, 2024. The research work was supported by National Natural Science Foundation of China (No.51774179), and Science and Technology Innovation Project of University of Science and Technology Liaoning (LKDYC202109, LKDYC202219).

Zhengpeng Li is a doctoral student of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: lkdlzp0901@163.com).

Jun Hu is a professor of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 320083700074@ustl.edu.cn)

Zhuang Liang is a teaching assistant of the LiaoNing Finance and Trade College, Huludao, 125105, China. (e-mail: 2749936763@qq.com).

Jiansheng Wu is a professor of University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author, e-mail: ssewu@163.com).

Medical image segmentation, an emergent technique in biomedical image processing, has made significant contributions to sustainable healthcare and has become a key research direction in the field of computer vision[2, 3]. Notably, Ronneberger and colleagues[4] introduced the Unet network model, based on an encoder-decoder approach. The encoder extracts contextual information from images through multiple convolutions and downsampling operations, while the decoder precisely localizes targets through multiple convolutions and upsampling operations. A distinctive feature of this network is the incorporation of skip connections, closely linking nodes at the same level in the encoder and decoder, thereby enhancing data information. Zhou et al. [5] developed the Unet++ network model, redesigning a series of nested and dense skip pathways on the Unet model to minimize the semantic gap between feature mappings. Zhang et al.[6] proposed the Deep Residual UNet model, substituting the convolutional units of the Unet model with residual convolutional units to facilitate information propagation, reduce training parameters, and improve the network's training efficiency. Jha and others[7] introduced the ResUnet++ model, building upon the residual UNet framework with the addition of attention mechanisms and spatial pyramid pooling to capture context information at various scales. Shah and colleagues[8] advanced the Unet model by incorporating dilated convolutions and residual connections, resulting in the AtResUnet model, which achieved excellent segmentation results for rectal tumors. However, repeated use of dilated convolutions can lead to information loss, impacting segmentation outcomes.

Despite these advancements, existing methods have not effectively utilized multi-scale information or fully exploited the original features of such data. To address these challenges, we propose the following improvements:

1) Introduction of a novel segmentation network model based on Unet++, namely the Multi-scale Residual Attention Unet++ Network (MRAUnet++), specifically designed for the task of rectal tumor segmentation.

2) Implementation of multi-scale residual blocks, utilizing convolutional sequences with varying kernel sizes to extract multi-scale features from images. This design enables the network to capture a richer and more diverse range of features, and the use of residual connections optimizes the network structure, enhancing its training and feature expression capabilities.

3) The integration of attention blocks to distinguish between foreground feature information and background noise within images, thus suppressing irrelevant feature infor-

mation. The attention gates within these blocks enable the network to focus more on learning key features, increasing the accuracy and robustness of segmentation.

4) Comprehensive evaluation of our model on a public dataset. Experimental results indicate that our improved model demonstrates superior performance, surpassing other advanced methods in rectal tumor segmentation tasks.

II. Related Work

A. Traditional Method

In recent years, the field of medical image segmentation has witnessed significant advancements, with the emergence of numerous innovative methods aimed at enhancing accuracy and adaptability. Shajin et al.[9] introduced a novel Sailfish Optimizer, integrating Levy flight, chaos, and opposition-based multi-level thresholding for precise medical image segmentation. By optimizing the weights of the Otsu method and Kapur's entropy, they achieved lower mean square errors and higher accuracy compared to existing methods. In anomaly detection, Hansen et al. [10] employed a self-supervised approach for segmenting medical images with limited examples. By calculating the anomaly scores of single foreground prototypes, they circumvented the need for explicit modeling of complex backgrounds, thus improving segmentation outcomes. Lu's research[11] proposed a semi-supervised medical image segmentation framework, utilizing uncertainty-aware pseudo-labeling and consistency loss, combined with a student-teacher model, to effectively leverage unlabeled data for enhanced performance. Xie et al. [12] developed a weakly-supervised medical image segmentation method based on detection, growth, and segmentation (DGS), combining YOLOV5 and GradCAM detectors with adaptive region growing for pseudo-label generation. This significantly improved segmentation performance in challenging scenarios like blurred edges, imaging noise, and interferences among multiple organs in medical images, while also reducing the time and requirements for annotations. Guo et al. [13] employed a multi-information fusion convolutional neural network (Mif-CNN) and CNN-based region growing, proposing a coarse-to-fine segmentation framework for accurately and efficiently segmenting the airway tree in chest CT scans, offering feasibility for computer-aided diagnosis systems in pulmonary diseases. In the realm of 3D medical image segmentation, Bennai et al. [14] utilized a multi-agent approach, employing autonomous, interactive agents with an improved region growing algorithm for segmenting 3D medical images. Wang's study[15] introduced a curriculum learning and Fourier domain adaptation-based medical image segmentation method to address domain adaptation challenges in medical data. This method incrementally introduced amplitude information in the frequency space and utilized a training time chain to augment the mixed data distribution, exhibiting robustness and generalizability across multiple test domains, especially suited for retinal and cell nucleus segmentation tasks. He et al. [16] proposed a statistical deformation model-based data augmentation method for medical volumetric image segmentation, significantly enhancing the fully automatic segmentation performance of organs in various body parts through diverse and realistic data augmentation, providing a potent tool for addressing the challenge of limited data access in medical image segmentation. Hu's research[17] introduced a neural memory ordinary differential equation (nmODE) based knowledge distillation method (nmODE-KD), transferring the knowledge of continuous nmODEs to discrete layers, improving the robustness and efficiency of deep neural network models, particularly in scenarios with incomplete labeled training data. Finally, Sun et al. [18] proposed an orthogonal decomposition adversarial domain adaptation (ODADA) framework for medical image segmentation. By designing an orthogonal loss function, they decomposed input features into domain-invariant and domain-specific representations to counteract performance degradation caused by domain shifts. These deep learning methods have brought increased flexibility and precision to medical image segmentation, providing powerful tools for addressing the complexity and diversity of structures and textures in medical images.

B. Deep Learning Method

In the field of medical image segmentation, Chen et al. [19] proposed the Synchronous Medical Image Augmentation (SMIA) framework, addressing difficulties in data acquisition and sample class imbalance through two modules based on random transformation and synthesis. SMIA generates a diversified and annotated medical image training set through random transformations while maintaining medical significance through synthesis-based modules. This positively impacts the performance of deep learning models. Hassanzadeh's work [20] introduced the Evolutionarily Enhanced Unet Framework, EEvoU-Net, the first to utilize a genetic algorithm for designing the structure and parameters of a medical image segmentation network. This innovative method achieved optimization of network architecture and parameters through evolutionary computation, offering a new perspective for medical image segmentation. Xie et al. [21] introduced a 3D end-to-end hybrid multi-atlas segmentation (MAS) and deep convolutional neural network (DCNN) segmentation pipeline with deep label fusion (DLF). This method, by employing weighted voting and fine-tuning of sub-networks, achieves precise segmentation of anatomical structures, introducing modality augmentation strategies for multi-modality imaging and enhancing segmentation accuracy in scenarios with missing data during testing. Wang's Boundary-Aware Context Neural Network (BA-Net) [22] adopted an encoder-decoder structure, capturing rich context and fine spatial details in medical images through pyramid edge extraction, multi-task learning modules, and cross-feature fusion, achieving more accurate segmentation results. He's ProtoSeg approach [23] measured the segmentation ability of deep features in different layers and units by calculating Segmentation Ability (SA) scores, providing interpretability and explanation for the hidden representations of deep learning features. Kumar Singh [24] proposed MISegNet, a robust and lightweight real-time segmentation network for multimodal medical images, combining discrete wavelet transform, global context awareness, residual shuffle attention mechanisms, and a novel boundary-aware loss function FEPE, accurately segmenting ambiguous boundary regions. Shirokikh [25] systematically evaluated deep learning models in 3D medical image segmenta-

tion tasks, providing practical insights for other medical image segmentation tasks by addressing manual segmentation issues, reducing contour variability, and improving contouring process efficiency. Mohagheghi [26] proposed an interpretable deep correction method, refining the outputs of other models through cascaded one-dimensional and two-dimensional models, improving the reliability and accuracy of CNN segmentation model results. Shukla [27] investigated credibility improvement, non-credible prediction amendment, and classifier switching operations through deep learning medical image segmentation model credibility improvement methods, successfully enhancing performance. Finally, Fu's design of deep feature attention blocks, feature refinement blocks, and adaptive feature extraction blocks [28] optimized the accurate processing of lesion edges and fine details in medical images by deep neural networks, demonstrating superior performance on multiple datasets through deep supervision mechanisms. These methods collectively propel the development of the field of medical image segmentation, providing diverse and innovative solutions to enhance accuracy and address challenging scenarios.

III. Research Method

In the MRAUnet++ model, we propose the use of multi-scale residual blocks to replace traditional convolutional layers. These residual blocks, composed of convolutional sequences with varying kernel sizes, are capable of extracting richer and more detailed feature mappings from images. The design employing multi-scale residual blocks enhances the network's capability to express complex image features, thereby increasing the accuracy of segmentation tasks. The introduction of residual connections not only helps to circumvent the "degradation" problem caused by increased network depth but also smoothens the propagation of image features, providing more favorable conditions for network optimization and enhancing its generalization ability. This is crucial for the accurate segmentation of rectal cancer tumors, as the complexity of tumor regions demands a network with robust feature extraction and expression capabilities. Furthermore, we have incorporated attention blocks, designed to accentuate key feature information of the target area through an attention mechanism while suppressing features irrelevant to the task. This focus mechanism aids in improving the network's precision in locating and segmenting rectal tumors, enabling it to concentrate more on learning and emphasizing the important features of the lesion area. Fig. 1 details the basic structure of the network model, illustrating the transformation from convolutional blocks to multi-scale residual blocks, and the integration of attention blocks. This comprehensive depiction showcases the innovative architecture of the MRAUnet++ model, highlighting its advances and functional enhancements.

As depicted in Fig. 1, the network model comprises an encoder and a decoder. In the encoding path, multi-scale residual blocks are employed to extract features from the input image, enabling the acquisition of rich feature information. The extracted features undergo downsampling via pooling layers to reduce feature dimensions. Following each multi-scale residual block, batch normalization layers and ReLU activation functions are integrated, facilitating faster convergence of the model. In the decoding path, transpose convolutions are used for upsampling the input features to precisely localize the segmented area and restore the feature size to its original dimensions. Subsequently, utilizing the inherent restrictive skip connections of Unet++, different levels of the U-shaped network are connected into a cohes ive structure. This aims to extract features at various levels and amalgamate them through stacking, thereby enhancing the precision of the model's segmentation capabilities. Notably, attention mechanisms are employed at the connections of various layers of the model to suppress noise and irrelevant information in the features. This ensures the model's focused learning on key features and concentrated attention on critical areas of the task. Finally, the segmentation map is obtained through the application of 1x1 convolutions and a Sigmoid activation function. The Sigmoid function is computed as follows, where x represents the input feature vector:

$$S(x) = \frac{1}{1+e^{-x}} \tag{1}$$

where x represents the input eigenvector.

A. Backbone

In deep learning, the task of semantic segmentation of medical images aims to achieve precise segmentation of different tissues or structures within medical images, providing detailed interpretations of diseases, abnormalities, or anatomical regions. Mathematically formalizing this task involves constructing a mapping function that translates an input medical image X into a corresponding pixel-level label map Y, where each pixel label denotes the specific tissue or structure it belongs to. The objective of deep learning models is to learn a mapping function F such that for any given input image X_i, the model's output $F(X_i)$ accurately represents the segmentation of different structures in the medical image. Formally, the mapping function F is represented as a deep neural network with parameters θ. Thus, the task of semantic segmentation of medical images can be mathematically formulated as an optimization problem to minimize the loss function:

$$\theta^* = \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(F_\theta(X_i), Y_i\right) \tag{2}$$

where $\mathcal{L}$ represents the loss function, measuring the discrepancy between the model output $F_\theta(X_i)$ and the true label Y_i. N denotes the number of samples, where X_i is the $i-th$ medical image, and Y_i is the corresponding pixel-level label map. By training deep learning models on extensive medical image data, they learn feature representations of various tissues and structures, thereby enabling accurate semantic segmentation on new, unseen data. This approach represents a significant advancement in medical image analysis, offering a potent tool for enhancing diagnosis and treatment planning in healthcare.

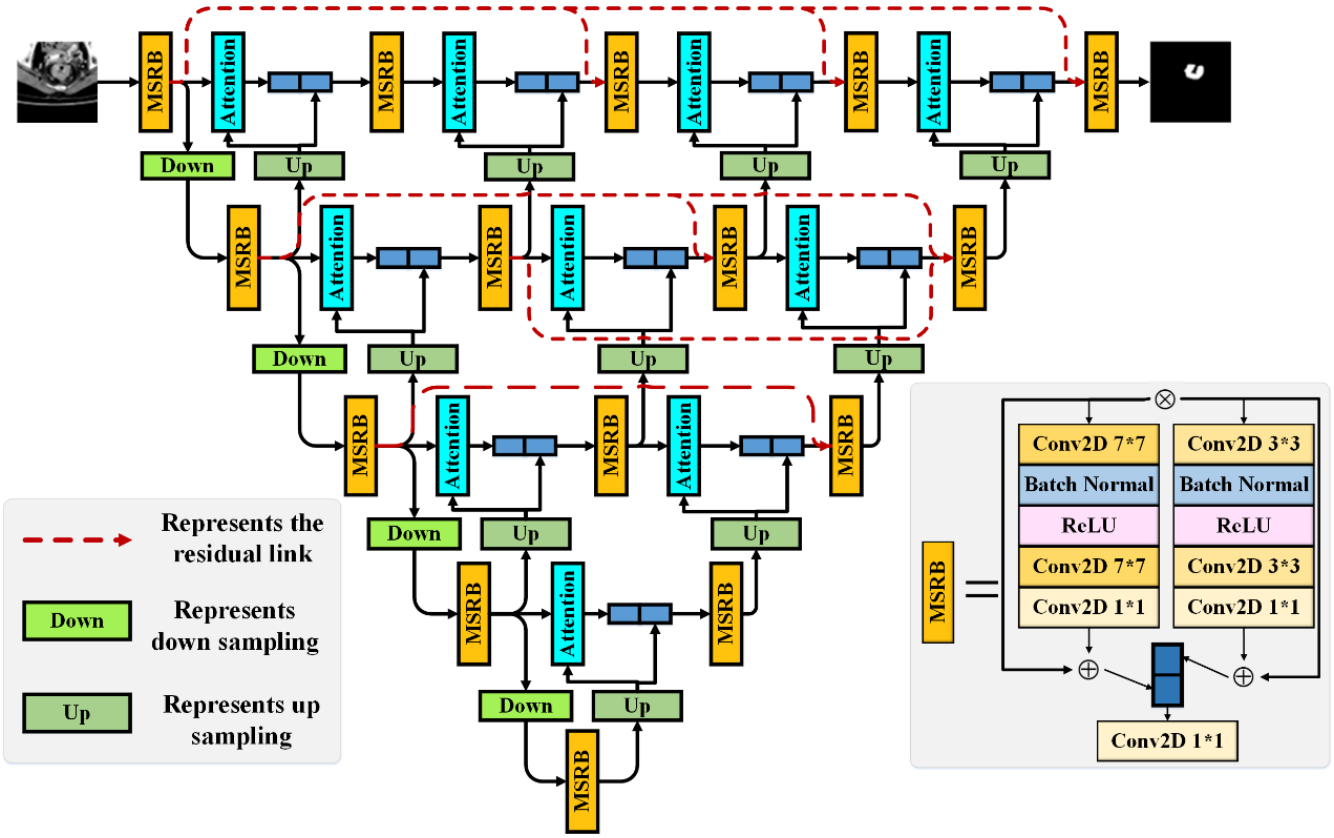


Fig. 1. Overview of the MRAUnet++

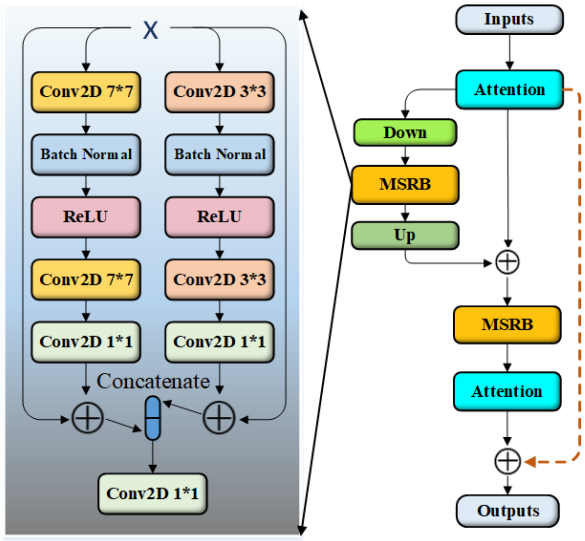


Fig. 2. Overview of the multi-scale residual block.

B. Multi-scale Residual Block

In our designed MRAUnet++ model, we have specifically incorporated an innovative multi-scale residual block (MSRB) module, as illustrated in Fig. 2. The design of this module effectively addresses the limitations encountered with single-scale convolutional kernels in medical image feature extraction. These limitations include the loss of feature dimensions due to the monolithic approach of feature extraction and issues such as gradient explosion, gradient vanishing, and slow convergence rates due to increased network layers, ultimately leading to network "degradation".

The MSRB module employs a multi-scale residual strategy, replacing traditional convolutional blocks in the Unet++ network with multi-scale residual blocks. These blocks comprise convolutional sequences with varying sizes of kernels, extracting a richer array of features from different scales of the image to obtain a more comprehensive representation of image features. These features are then effectively integrated through residual connections.

Initially, an input tensor of size (256, 256, 1) undergoes two convolution operations in our network. These operations employ convolutional kernels of sizes 7x7 and 3x3, respectively, with an output channel count of 32. Taking a grayscale image as an example, the computation for the convolution can be described as per equation (3). Post each convolution operation, the feature matrix undergoes Batch Normalization to maintain its dimensional integrity. Subsequently, the convolution results are activated through the ReLU activation function, culminating in an output feature map of size (256, 256, 32).

$$Z(i) = f(\sum_{x}^{h}\sum_{y}^{d}\alpha_{x,y} \times w_{x,y}^{i} + b^{i}) \tag{3}$$

where $Z(i)$ represents the output value of the i-th local feature map of a grayscale image; f denotes the activation function; h and d signify the length and width of the local feature map, respectively. $\alpha_{x,y}$, $w_{x,y}^{i}$, and b^{i} correspond to the pixel value at the matrix position (x, y) of the local feature map, along with its associated weight and bias values.

Following this, the activated result from the previous step undergoes two additional convolution operations using kernels of sizes 7×7 and 3×3. After each convolution, batch normalization layers and relu activation functions are applied. The resulting feature map, sized (128, 128, 32), is

assigned to the variable X, serving as the output of the residual block.

To implement the residual connection, the adjusted result is processed through a 1×1 convolution layer and normalization operation. Subsequently, the keras add method is used to combine the adjusted result with the output X of the residual block, completing the residual connection. The final residual block function concatenates the output feature maps of the two different residual blocks – one using a 7×7 kernel and the other a 3×3 kernel. This is followed by 1×1 convolution to integrate features and expand channels, resulting in an output feature map of size (128, 128, 64), enriched with more comprehensive feature information.

In our experiments, we extract image features using a sequence of convolution operations with 3×3 and 7×7 kernel sizes. Each convolution sequence includes two convolution layers, a batch normalization layer, and a relu layer. Residual connections are employed to achieve identity mapping. Finally, the outputs of the two different convolution kernels are combined and passed through a 1×1 convolution layer to yield the final result. Let X represent the input features of the image, and Z denote the output of the multi-scale residual block. The computation of Z is as follows:

$$X_1 = Conv2D_{3\times3}(\mathrm{Re}\,LU(BN(Conv2D_{3\times3}(X)))) \quad (4)$$

$$X_2 = Conv2D_{7\times7}(\mathrm{Re}\,LU(BN(Conv2D_{7\times7}(X)))) \quad (5)$$

$$X_3 = X_1 + X \quad (6)$$

$$X_4 = X_2 + X \quad (7)$$

$$Z = Conv2D_{1\times1}([X_3, X_4]) \quad (8)$$

Residual connections facilitate smoother propagation of feature signals in both forward and backward passes within the network. In forward propagation, information is transmitted in a bottom-up manner, aiding in resolving the network's "degradation" issue. During backward propagation, error information can be directly transferred to the subsequent layers without intermediate computations, thereby alleviating problems like gradient explosion and vanishing. This makes the network easier to optimize and simultaneously enhances its generalization capabilities.

Overall, this module's introduction into the MRAUnet++ model marks a significant advancement in medical image segmentation, addressing key challenges and improving the network's ability to handle complex medical image data with enhanced efficacy and accuracy.

C. Attention Block and Attention Gate

The MRAUnet++ model incorporates an attention mechanism block, composed of attention gates, specifically designed to suppress noise and irrelevant feature information while effectively reducing the number of parameters required for network training. This module addresses a critical issue in the Unet++ network model related to skip connections: the transmission of noise and irrelevant features from the encoder to the decoder, leading to semantic discrepancies between them. The introduction of the attention mechanism significantly simplifies the complexity of network training.

The essence of the attention mechanism lies in its ability to selectively filter out vital information from a vast pool of data, concentrating focus on these key details and disregarding the majority of unimportant information. This process is reflected in the computation of weight coefficients, where higher weights signify greater emphasis on focal points. When calculating attention with a task-related query vector Q and a global key vector K, the model can focus more on pertinent information by computing attention weights. To ensure better convergence of the model, the calculated attention weights are subjected to a moderated non-linear mapping and normalized using the SoftMax activation function, resulting in the final attention matrix. This matrix is then multiplied with the corresponding value vector V to produce the final outcome.

$$Attention(Q,K,V) = SoftMax(\frac{QK^T}{\sqrt{d_k}})V \quad (9)$$

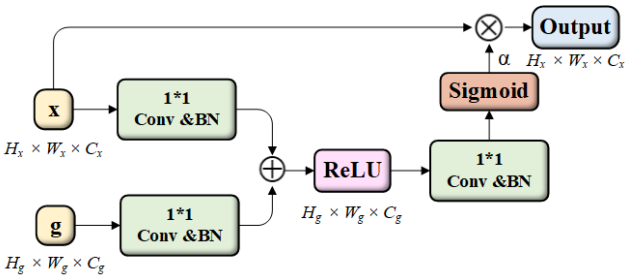
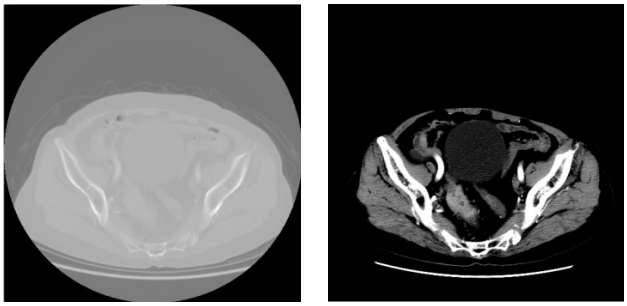


Fig. 3. Overview of the attention gate.

The attention block, as depicted in Fig. 3, consists of attention gates, receiving inputs from features g extracted from the up-sampling output and x, the same level features extracted by the encoder. Feature g acts as a gating control signal, suppressing noise and irrelevant features in x. Initially, 1×1 convolution and normalization operations are performed on features g and x, followed by element-wise addition using keras's add function. The merged information, post-activation by the Relu function, undergoes another 1x1 convolution and normalization. The resultant is then passed through a Sigmoid activation function to obtain the final attention coefficient a. The input feature x is element-wise multiplied with the attention coefficient using the multiply function, yielding the final output value. This structure of the attention gate successfully achieves weighted integration of features, thus enhancing the network's focus on important features and reducing sensitivity to noise and irrelevant information.



(a) (b)

Fig. 4. The effect diagram of window processing. (a) represents the original image data, and (b) displays the image post-window processing.

IV. EXPERIMENT

A. Datasets

In our study, we utilized the open-source rectal cancer

lymph node metastasis (RCLNM) dataset, which comprises two sets of CT images: the arterial phase and the venous phase. Our focus was on the arterial phase CT images, encompassing a total of 3029 images along with their corresponding image masks. Among these, there are 2169 positive samples and 860 negative samples. The image samples are grayscale images of size 512×512 pixels, formatted in DICOM.

Initially, the images were normalized into 48×48 three-channel color images. Subsequently, the pixel values were standardized to fall within the range of [0,255]. The next step involved windowing the images to enhance the contrast of the CT images, with the effects illustrated in Fig. 4. Finally, data class labels were established based on the image masks. This was accomplished by extracting the data matrix of the image masks and performing a summation operation. If the resulting value was zero, the image was labeled as negative (label 0); otherwise, it was labeled as positive (label 1).

B. Experimental Environment

The experiment was conducted on a computer equipped with an NVIDIA GeForce GTX 3080Ti GPU. Regarding the experimental parameters, the number of epochs was set to 80, with a batch size of 6. The chosen optimizer was Adam, featuring a learning rate of 0.0002, and the Mean Squared Error function served as the loss function. To optimize the training process, the ReduceLROnPlateau callback function was employed. It reduces the learning rate to one-tenth of its original value if there is no decrease in the validation loss for 20 consecutive epochs.

To alleviate GPU memory demands, the original images underwent resizing to 256x256 using Python's image slicing techniques. Subsequently, pixel value normalization was applied to ensure they fall within the range of [0, 255]. Following this, windowing processing was applied for further enhancement. As illustrated in Fig. 5, the first row showcases the images post-window processing, while the second row displays the real masks of the post-processed images.

Ultimately, to enhance the diversity of the original dataset, augmentation techniques such as horizontal flipping and rotations of 45°, 90°, and 180° were employed. This augmentation strategy expanded the dataset to five times its original size, totaling 15,145 images. These images were then allocated into training, validation, and test sets in a 6:2:2 ratio, consisting of 9,087 images for training, 3,029 for validation, and 3,029 for testing. This augmented dataset significantly enriched diversity, providing a more robust and comprehensive set of images for training and evaluation. This diversified dataset facilitates the model's learning of a broader range of features and patterns, essential for enhancement.

C. Evaluation Indicator

To assess the performance of the model, two metrics, the Dice coefficient and the Intersection over Union (IoU) coefficient, were employed.

The Dice coefficient measures the similarity between the true image mask (Y) and the predicted output image (Y'). It is commonly used to quantify the overlap between segmented images and actual mask images. The formula for calculating the Dice coefficient is as follows:

$$Dice = \frac{2\times|Y\cap Y'|}{|Y'|+|Y|} = \frac{2\times TP}{FP+FN+2\times TP} \tag{10}$$

The IoU coefficient quantifies the ratio of the intersection to the union of the true image mask and the predicted output image. It also serves to describe the overlap degree between segmented images and actual mask images. The calculation formula for the IoU coefficient is as follows:

$$IoU = \frac{|Y\cap Y'|}{|Y\cup Y'|} = \frac{TP}{FP+FN+TP} \tag{11}$$

In these formulas, TP refers to the number of cases correctly identified as positive; FP denotes the number of cases incorrectly identified as positive; FN represents the number of cases incorrectly identified as negative. These evaluation metrics provide an objective assessment of the model's performance in the segmentation of rectal cancer tumors, allowing for a comprehensive evaluation of its accuracy and effectiveness.

D. Experimental Results and Analysis

As presented in Table I, the MRAUnet++ network model demonstrated superior performance on the rectal tumor dataset. Through extensive experimental evaluation, it achieved a Dice coefficient of 83.32% and an IoU value of 70.06%. In comparison to existing deep learning models such as Unet, Unet++, Attention Unet, ResUnet, and ResUnet++, MRAUnet++ exhibited notable superiority in performance metrics.

TABLE I
THE RESULTS OF DIFFERENT MODELS ON THE RCLNM DATASET

Methods	Dice	IoU
Unet	78.12	61.27
Unet++	80.51	66.28
Attention Unet	81.56	67.86
ResUnet	79.94	65.15
ResUnet++	82.37	68.09
MRAUnet++	83.32	70.06

From the perspective of the Dice coefficient, MRAUnet++ exhibited improvements of 5.2% over Unet, 2.81% over Unet++, 1.76% over Attention Unet, 3.38% over ResUnet, and 0.95% over ResUnet++. The Dice coefficient is a crucial metric for measuring the accuracy of image segmentation. The notable enhancement in this metric indicates that MRAUnet++ has an increased capability to accurately locate and segment lesion areas in rectal tumor segmentation tasks.

In terms of IoU values, MRAUnet++ achieved improvements of 8.79% over Unet, 3.78% over Unet++, 2.2% over Attention Unet, 4.91% over ResUnet, and 1.97% over ResUnet++. The IoU value, or IoU, is another key metric for assessing the performance of segmentation models. The remarkable advantage of MRAUnet++ in IoU values suggests better outcomes in pixel overlap within the lesion areas.

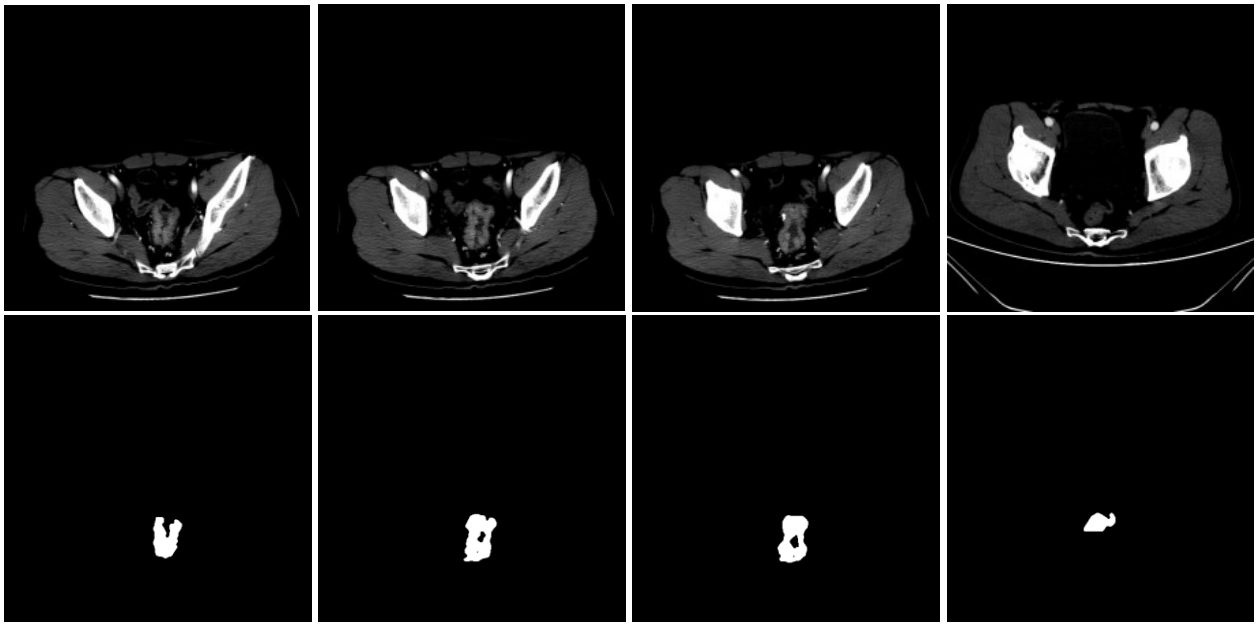


Fig. 5. Sample images of rectal tumors pictures and real masks.

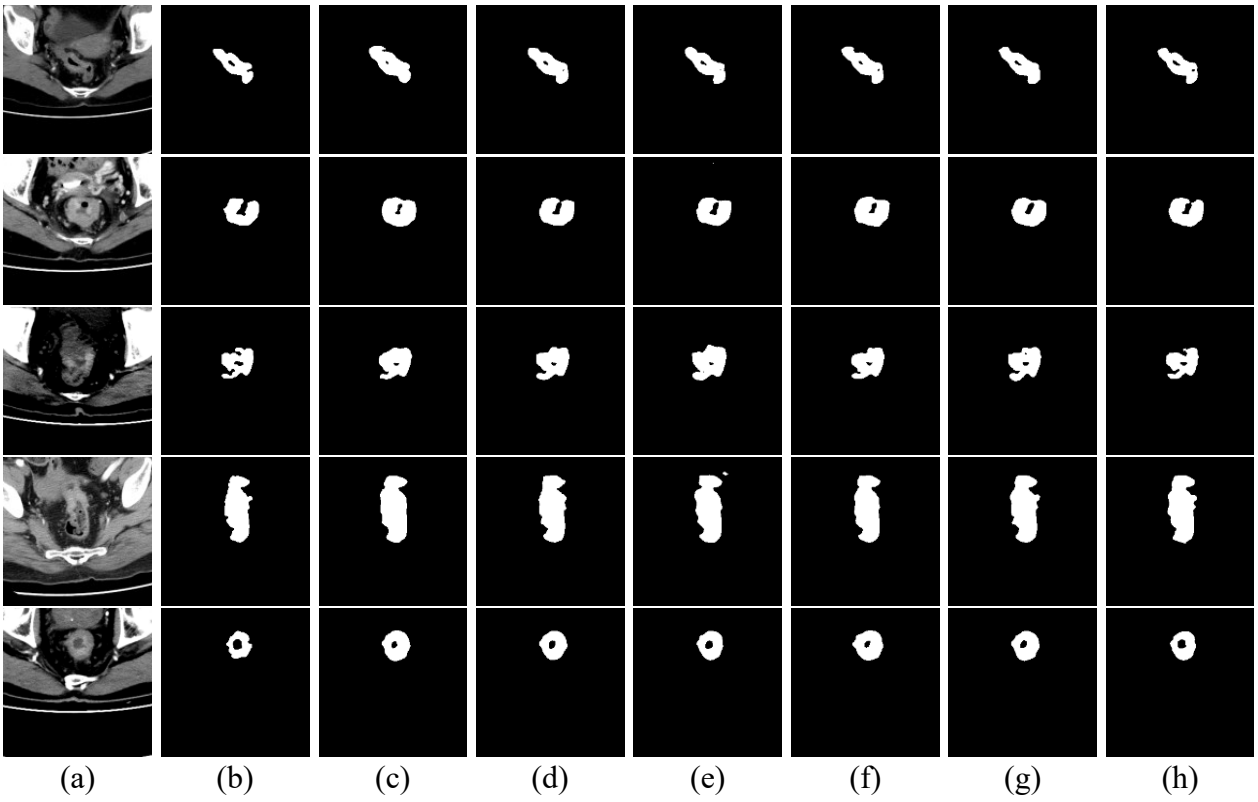


Fig. 6. Qualitative results comparison along on the RCLNM datasets.

For qualitative comparison, we randomly selected five sets of sample data and presented the segmentation results among different deep learning network models in Fig. 6. Column a represents the original images, column b shows the true mask images of the original images, and columns c to h display the segmentation images obtained by the Unet, Unet++, Attention Unet, ResUnet, ResUnet++, and MRAUnet++ models, respectively. As observed in the figure, MRAUnet++ provides more detailed segmentation results for rectal lesions compared to other networks.

This comparative analysis highlights MRAUnet++'s enhanced efficiency and accuracy in medical image segmentation, particularly in complex tasks such as rectal tumor segmentation, making it a promising tool in medical imaging analysis.

V. CONCLUSION

In this study, we introduced the MRAUnet++ architecture—an innovative network structure based on deep supervision and multi-scale fusion strategies, featuring the attention mechanism as a core component. This architecture is specifically designed for the precise segmentation of rectal cancer tumors. Experimentally validated on the meticulously constructed RCLNM dataset, our architecture successfully achieved effective segmentation of rectal cancer lesion areas, demonstrating impressive performance. In terms of evaluation metrics, our algorithm yielded satisfactory results, achieving a Dice coefficient of 83.32% and an IoU index of 70.06%. These outcomes distinctly validate the superior performance and accuracy of our proposed method in rectal cancer tumor segmentation tasks. The successful application of this technology offers positive prospects and support for the diagnosis and treatment of rectal cancer in medical clinical settings. Moving forward, we plan to refine the proposed network structure further, considering the integration of advanced technologies such as additional transformer blocks, to enhance its performance and adaptability. This ongoing exploration and improvement will contribute to addressing the critical issue of rectal cancer tumor segmentation in the field of medical image segmentation more effectively, providing more reliable solutions for clinical medical applications.

REFERENCES

[1] W. Xue *et al.*, "Use of patient-derived tumor organoid platform to predict the benefit of postoperative adjuvant chemotherapy for poor responders to neoadjuvant chemoradiotherapy in locally advanced rectal cancer," *Bioengineering and Translational Medicine,* vol. 8, no. 6, 2023. DOI:10.1002/btm2.10586.

[2] Y. Li, B. Wang, K. Zhang, Z. Jiang, W. Shi, and W. Liu, "A Novel Automatic Method Based on U-Net for Lung Fields Segmentation," *Engineering Letters,* vol. 30, no. 2, pp. 636-643, 2022.

[3] A. J. Santoso and Pranowo, "Medical Image Segmentation Using Phase-Field Method based on GPU Parallel Programming," *Engineering Letters,* vol. 30, no. 1, pp. 214-220, 2022.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *18th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015, October 5, 2015 - October 9, 2015*, Munich, Germany, 2015, vol. 9351: Springer Verlag, pp. 234-241.

[5] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *4th International Workshop on Deep Learning in Medical Image Analysis, DLMIA 2018 and 8th International Workshop on Multimodal Learning for Clinical Decision Support, ML-CDS 2018 Held in Conjunction with MICCAI 2018, September 20, 2018 - September 20, 2018*, Granada, Spain, 2018, vol. 11045 LNCS: Springer Verlag, pp. 3-11.

[6] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geoscience and Remote Sensing Letters,* vol. 15, no. 5, pp. 749-753, 2018.

[7] D. Jha *et al.*, "ResUNet++: An Advanced Architecture for Medical Image Segmentation," in *21st IEEE International Symposium on Multimedia, ISM 2019, December 9, 2019 - December 11, 2019*, San Diego, CA, United states, 2019: Institute of Electrical and Electronics Engineers Inc., pp. 225-230.

[8] N. A. Shah, D. Gupta, R. Lodaya, U. Baid, and S. Talbar, "Colorectal Cancer Segmentation Using Atrous Convolution and Residual Enhanced UNet," in *5th International Conference on Computer Vision and Image Processing, CVIP 2020, December 4, 2020 - December 6, 2020*, Prayagraj, India, 2021, vol. 1376 CCIS: Springer Science and Business Media Deutschland GmbH, pp. 451-462.

[9] F. H. Shajin, B. Aruna Devi, N. B. Prakash, G. R. Sreekanth, and P. Rajesh, "Sailfish optimizer with Levy flight, chaotic and opposition-based multi-level thresholding for medical image segmentation," *Soft Computing,* vol. 27, no. 17, pp. 12457-12482, 2023.

[10] S. Hansen, S. Gautam, R. Jenssen, and M. Kampffmeyer, "Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels," *Medical Image Analysis,* vol. 78, 2022. DOI:10.1016/j.media.2022.102385.

[11] L. Lu, M. Yin, L. Fu, and F. Yang, "Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation," *Biomedical Signal Processing and Control,* vol. 79, 2023. DOI: 10.1016/j.bspc.2022.104203.

[12] Y. Xie, Z. Zhang, S. Chen, and C. Qiu, "Detect, Grow, Seg: A weakly supervision method for medical image segmentation based on bounding box," *Biomedical Signal Processing and Control,* vol. 86, 2023. DOI: 10.1016/j.bspc.2023.105158.

[13] J. Guo *et al.*, "Coarse-to-fine airway segmentation using multi information fusion network and CNN-based region growing," *Computer Methods and Programs in Biomedicine,* vol. 215, 2022. DOI: 10.1016/j.cmpb.2021.106610.

[14] M. T. Bennai, Z. Guessoum, S. Mazouzi, S. Cormier, and M. Mezghiche, "A stochastic multi-agent approach for medical-image segmentation: Application to tumor segmentation in brain MR images," *Artificial Intelligence in Medicine,* vol. 110, 2020. DOI: 10.1016/j.artmed.2020.101980.

[15] A. Wang, M. Islam, M. Xu, and H. Ren, "Curriculum-Based Augmented Fourier Domain Adaptation for Robust Medical Image Segmentation," pp. 1-13, 2023. DOI: 10.1109/TASE.2023.3295600.

[16] W. He *et al.*, "A statistical deformation model-based data augmentation method for volumetric medical image segmentation," *Medical Image Analysis,* vol. 91, 2024. DOI: 10.1016/j.media.2023.102984.

[17] J. Hu, C. Yu, Y. Zhang, and H. Zhang, "Enhancing Robustness of Medical Image Segmentation Model with Neural Memory Ordinary Di®erential Equation," *International Journal of Neural Systems,* vol. 33, no. 12, 2023. DOI: 10.1142/S0129065723500600.

[18] Y. Sun, D. Dai, and S. Xu, "Rethinking adversarial domain adaptation: Orthogonal decomposition for unsupervised domain adaptation in medical image segmentation," *Medical Image Analysis,* vol. 82, 2022. DOI: 10.1016/j.media.2022.102623.

[19] J. Chen, N. Yang, Y. Pan, H. Liu, and Z. Zhang, "Synchronous Medical Image Augmentation framework for deep learning-based image segmentation," *Computerized Medical Imaging and Graphics,* vol. 104, 2023. DOI: 10.1016/j.compmedimag.2022.102161.

[20] T. Hassanzadeh, D. Essam, and R. Sarker, "EEvoU-Net: An ensemble of evolutionary deep fully convolutional neural networks for medical image segmentation," *Applied Soft Computing,* vol. 143, 2023. DOI:10.1016/j.asoc.2023.110405.

[21] L. Xie *et al.*, "Deep label fusion: A generalizable hybrid multi-atlas and deep convolutional neural network for medical image segmentation," *Medical Image Analysis,* vol. 83, 2023. DOI: 10.1016/j.media.2022.102683.

[22] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, "Boundary-aware context neural network for medical image segmentation," *Medical Image Analysis,* vol. 78, 2022. DOI: 10.1016/j.media.2022.102395.

[23] S. He, Y. Feng, P. E. Grant, and Y. Ou, "Segmentation ability map: Interpret deep features for medical image segmentation," *Medical Image Analysis,* vol. 84, 2023. DOI: 10.1016/j.media.2022.102726.

[24] V. Kumar Singh *et al.*, "Prior wavelet knowledge for multi-modal medical image segmentation using a lightweight neural network with attention guided features," *Expert Systems with Applications,* vol. 209, 2022. DOI: 10.1016/j.eswa.2022.118166.

[25] B. Shirokikh *et al.*, "Systematic Clinical Evaluation of a Deep Learning Method for Medical Image Segmentation: Radiosurgery Application," *IEEE Journal of Biomedical and Health Informatics,* vol. 26, no. 7, pp. 3037-3046, 2022.

[26] S. Mohagheghi and A. H. Foruzan, "Developing an explainable deep learning boundary correction method by incorporating cascaded x-Dim models to improve segmentation defects in liver CT images," *Computers in Biology and Medicine,* vol. 140, 2022. DOI: 10.1016/j.compbiomed.2021.105106.

[27] S. Shukla, L. Birla, A. K. Gupta, and P. Gupta, "Trustworthy Medical Image Segmentation with improved performance for in-distribution samples," *Neural Networks,* vol. 166, pp. 127-136, 2023.

[28] Z. Fu, J. Li, Z. Hua, and L. Fan, "Deep supervision feature refinement attention network for medical image segmentation," *Engineering Applications of Artificial Intelligence,* vol. 125, 2023. DOI: 10.1016/j.engappai.2023.106666.

Zhengpeng Li received his Bachelor's degree in engineering from University of Science and Technology Liaoning in 2016 and his Master's degree from University of Science and Technology Liaoning in 2020. He is now a PhD student at Liaoning University of Science and Technology, focusing on remote sensing image processing and deep learning.

In 2020, he studied at Liaoning University of Science and Technology with a research direction of the deep learning, medical image segmentation, medical image analysis, remote sensing scene understanding, remote sensing semantic image segmentation. Now, he is now a reviewer for journals such as Neurocomputing, Engineering Letters, and Scientific Reports.

Jun Hu received his Bachelor's degree in Hydraulic Engineering from North China University of Water Conservancy and Electric Power (now North China University of Water Resources and Electric Power) in 2001, and his Doctorate in Disaster Prevention and Mitigation Engineering and Protective Engineering from Dalian University of Technology in 2008, through a direct Bachelors-to-PhD program. In November 2014, he was exceptionally promoted to Professor.

He currently serves as a member of the Standing Committee of the Party Committee and Vice President at Liaoning University of Science and Technology. His research interests are focused on remote sensing imagery, geotechnical engineering numerical simulation, and the automation of safety monitoring and early warning systems for tailings dams. He has led over 30 key projects sponsored by the Education Department of Liaoning Province, general projects, and corporate commissioned projects, with total funding exceeding 30 million yuan. As the first author or corresponding author, he has published more than 60 papers in international and domestic journals. He has been awarded the title of Outstanding Teacher of Liaoning Province, among other honors.

Zhuang Liang received her Bachelor's degree in engineering from University of Science and Technology Liaoning in 2015 and his Master's degree from University of Science and Technology Liaoning in 2019.

Now he is a teaching assistant at the College of Big Data College at LiaoNing Finance and Trade College with a research direction of the deep learning, medical image segmentation, medical image analysis.

Jiansheng Wu received his Master's degree in Computer Application Technology from Anshan University of Science and Technology (now University of Science and Technology Liaoning) in 2005. He was promoted to Professor in 2010.

He is currently a faculty member in the School of Computer and Software Engineering at University of Science and Technology Liaoning. His research focuses on computer vision, image processing and analysis, multimodal remote sensing image analysis, and medical image analysis. He has led or participated in projects funded by the National Natural Science Foundation of China, key projects of the Education Department of Liaoning Province, general projects, and corporate commissioned projects. He has published over 30 papers as the first author or corresponding author.