Neurocomputing 653 (2025) 131094

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Fine-grained grading network based on sparse transformer and spectral attention for multiparametric MR image segmentation

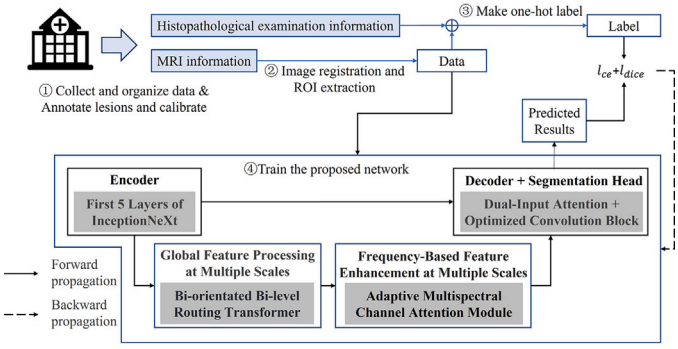
Yatong Liu [a], Wei Wang [b], Yu Zhu [a,*], Hangyu Li [a], Zeyan Zeng [c], Yuhao Zhang [c,d,e]

[a] School of Information Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China
[b] Department of Radiology, Tongji Hospital, School of Medicine, Tongji University, Shanghai, 200065, China
[c] Department of Neurology, Zhongshan Hospital, Fudan University, Shanghai, 200032, China
[d] National Clinical Research Center for Interventional Medicine, Shanghai, 200032, China
[e] Shanghai Clinical Research Center for Interventional Medicine, Shanghai, 200032, China

## HIGHLIGHTS

- Present a novel architecture primarily designed for predicting fine-grained Gleason score (GS) groups, applicable to other lesion segmentation tasks.
- Integrate bi-orientated bi-level routing transformer (BBRT) and adaptive multispectral channel attention (AMSA) modules to boost spatial attention and texture sensitivity.
- Achieve high Dice Similarity Coefficients (DSCs) in prostate cancer grading, prostate region segmentation, and stroke segmentation; ablation studies verify empirically module effectiveness.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

Communicated by W. Wang

*Keywords:*
Magnetic resonance imaging segmentation
Prostate cancer
Sparse transformer
Spectral attention
Brain stroke

## ABSTRACT

Multiparametric magnetic resonance imaging (mpMRI) is crucial in excluding serious diseases by providing clear scans that can quickly locate diseased tissue. Subjective assessment of lesion aggressiveness can vary among clinicians. We propose a network for mpMRI segmentation capable of effectively segmenting fine-grained lesion groups in prostate cancer (PCa), with potential applicability to other magnetic resonance imaging (MRI) segmentation tasks such as brain strokes. Specifically, a bi-oriented bi-level routing transformer (BBRT) with a flexible sparse attention mechanism is designed to enhance the network's target perception. This design optimizes the Transformer's long-range modeling capabilities. An adaptive multispectral channel attention (AMSA) module with an automatically selected frequency components mechanism is designed to enhance target contour recognition accuracy. It improves the network's sensitivity to detect crucial texture information. Experiments are conducted using three MRI datasets. Our method achieved a dice similarity coefficient (DSC) of 76.09 % for grading PCa on PCAMM, demonstrating improvements of 2.77 %, 3.79 %, 3.64 %, and 0.29 % over the suboptimal method in clinically significant Gleason score (GS) groups (GS 3 + 4, GS 4 + 3, GS = 8, and $GS \geq 9$), respectively. The prostate region segmentation obtained DSC of 94.03 % and 84.52 % for the central gland (CG) and peripheral zone (PZ), respectively, on PROSTATEx. The stroke segmentation achieved a DSC of 86.33 % on ISLES2022. The experimental results demonstrate that the proposed model excels in generalization and outperforms other state-of-the-art methods.

* Corresponding author.
*Email addresses:* zhuyu@ecust.edu.cn (Y. Zhu), zhang.yuhao@zs-hospital.sh.cn (Y. Zhang).

https://doi.org/10.1016/j.neucom.2025.131094
Received 31 March 2025; Received in revised form 18 July 2025; Accepted 23 July 2025
Available online 30 July 2025

## 1. Introduction

Multiparametric magnetic resonance imaging (mpMRI) is crucial in cancer detection by providing visual information on lesions such as volume, location, and multifocality [1,2]. The assessment of lesions by mpMRI is a subjective process with large variations among clinicians [3]. Currently, deep learning algorithms have achieved significant success in MRI segmentation tasks [4,5]. However, it is still challenging for complex targets and complex tasks, as represented by the prostate cancer grading task, with the main difficulties including: (1) large differences in target size, with small targets easily missed; (2) unclear target boundaries; and (3) difficulty in quantitatively assessing the extent of cancerous lesions. To address these issues, we selected the prostate and stroke datasets to conduct multiple tasks.

Prostate cancer (PCa) is the second most common cancer in men [6]. MpMRI has been utilized for early detection of PCa, determining biopsy candidates, guiding biopsies, and localizing treatments [7]. To standardize the reporting of prostate mpMRI examinations, expert diagnosticians adhere to the Prostate Imaging Reporting and Data System (PI-RADS) [8]. However, interpreting prostate mpMRI is challenging due to some benign conditions, like benign prostatic hyperplasia (BPH), can interfere with the diagnosis of prostate cancer and the assessment of lesion aggressiveness [9]. Currently, the Gleason score (GS) based on histopathology is the most effective method for assessing the aggressiveness of a lesion [10]. Similarly, in stroke diagnosis, mpMRI has made lesion detection more intuitive, but there still exists significant variability in expert assessments.

Many lesions in medical images, such as prostate cancer and stroke, may be overlooked during feature extraction due to their small size and irregular location. To tackle this issue, researchers commonly opt to incorporate an attention mechanism into the network [11,12]. The attention mechanism utilizes global contextual information from an image, enabling the model to dynamically learn how to prioritize important features of a small target during training. Currently, the Transformer model has shown a global receptive field and multimodal processing in image processing [13]. However, its core module has high computational complexity and incurs heavy memory footprints. To alleviate this issue, researchers introduced sparse attention to Transformer. A significant amount of work [14–16] has reduced key/value tokens using various merging or selection strategies. However, these tokens are shared by all queries, and we would like the network to select key/value pairs that are potentially related to the query and to effectively utilize global information in both spatial and channel dimensions.

For lesions with various shapes and sizes, researchers commonly employ a multi-scale strategy [17–20]. The multi-scale strategy facilitates the transfer and fusion of features at different levels to capture broader and deeper contextual representations [21]. Additionally, numerous lesions exhibit blurred regional boundaries. This necessitates a network capacity to effectively capture and retain crucial details, such as edges, colors, textures, and more. Frequency analysis is a robust tool in signal processing. Qin et al. [22] mathematically proved that global average pooling is a special case of frequency domain feature decomposition. They improved the network's ability to prioritize important details by artificially selecting various frequency components in the frequency representation. However, the selection of frequency components in the model is artificially determined and remains unchanged during training. Suboptimal model optimization can occur when the domain of use is changed. To address this limitation, we reassess the method of selecting frequency components and explore new approaches to improve the network's adaptive attention to detailed information.

We consider two characteristics of prostate cancer grading: (1) lesions in different areas of the gland are graded using different major modalities, and (2) different regions of the gland show varying signal intensities within the same modality of the lesion. Combined with the fact that mpMRI is typically concatenated across the channels, we believe that the model's ability to select important channels is as crucial as its ability to pay spatial attention to regions of interest (ROI). Therefore, we first propose a novel structure for extracting global contextual information by incorporating channel selection and spatially effective sparse attention mechanisms. Secondly, the accurate prediction of segmentation boundaries requires the network's retention of important details, and we propose a new method of assigning higher weights to ROIs based on frequency representation selection. Our contributions are summarized as follows:

(1) This study presents a novel network for mpMRI segmentation tasks. The network is primarily used to predict fine-grained GS groups and further identify clinically significant PCa. Its targeted design and excellent robustness enable it to be applied to other MRI lesion segmentation tasks.
(2) The proposed network has two novel modules bi-orientated bi-level routing transformer (BBRT) and adaptive multispectral channel attention (AMSA). The BBRT is used to enhance the network's spatial attention to lesions and important channel selection capability. The AMSA is used to improve the network's sensitivity to detect and analyze significant texture information.
(3) The proposed method has been evaluated across three datasets representing two medical scenarios. For prostate cancer grading, our method obtained an average dice similarity coefficient (DSC) of 76.09 %. An improvement of 2.77 %, 3.79 %, 3.64 %, and 0.29 % was observed in GS 3+4, GS 4+3, GS = 8, and GS ≥ 9 groups, respectively, compared to the suboptimal method. The prostate region segmentation obtains a DSC of 94.03 % for the central gland (CG) and 84.52 % for the peripheral zone (PZ) on PROSTATEx, while the stroke segmentation achieves a DSC of 86.33 % on ISLES2022. Ablation studies demonstrate the effectiveness of our proposed modules.

## 2. Related work

Our work draws on recent studies regarding the attention mechanism, analytical work on skip connections in U-Net, lesion segmentation methods based on mpMRI, and previous PCa grading methods.

**Attention mechanism:** Attention mechanisms, because they can selectively attend to important information, are widely used in many fields like computer vision (CV) and natural language processing (NLP). Many works such as [23–26] propose their own methods for spatial attention and channel attention, primarily focusing on the spatial domain. In addition, some research focuses on building attention mechanisms in the frequency domain [22], directing the network to concentrate on specific frequency information relevant to the target. Considering the ease of confusion between PCa adjacent grade lesions, we distinguish them using frequency domain signals. In this study, considering the ease of confusion between PCa adjacent grade lesions, we dynamically select key frequency components based on input features and direct the network to focus on significant features corresponding to those frequencies.

**Analytical work on skip connections:** The skip connection, as a fundamental design in the codec structure, fuses low- and high-resolution information to improve the feature representation. However, the direct application of skip connections may create a semantic gap between low- and high-resolution features, resulting in blurred feature maps [27]. To mitigate the above problem, some work has focused on designing additional feature mapping or selection modules. Ibtehaz et al. [28] proposed the MultiResUNet, which incorporates a Residual Path. This allows the encoder features to undergo additional convolution operations before being merged with the corresponding features in the decoder. Seo et al. [29] proposed mU-Net, while Chen et al. [30] proposed FED-Net. Both mU-Net and FED-Net utilize convolutional operations within skip-connection to enhance the performance of segmented medical images.

**Lesion segmentation methods based on mpMRI:** In modern medicine, accurately diagnosing tumors often necessitates the integration of various MRI modalities due to their complexity and diversity. To fully utilize mpMRI information, current research is exploring various approaches and has achieved preliminary results. Yang et al. [31] first applied a single-stage image-level classifier based on CNN for PCa detection. Kohl et al. [32] first introduced an adversarial network for segmenting aggressive PCa, and refined the segmentation results using a discriminator. To address the issue of human cost in practical clinical applications, Alkadi et al. [33] proposed a mono-parametric CAD system that employs a 2.5D segmentation model. Wang et al. [34] used a two-stage segmentation algorithm to segment the prostate and lesions sequentially based on 3D MRI data. For the multimodal brain lesion segmentation task, Zhang et al. [35] extracted modality-invariant representations by explicitly building and aligning global correlations across different modalities. Shi et al. [36] divided the four modalities into two groups based on the characteristics of the perfusion map. They developed a cross-modal cross-attention module based on the exogenous attention mechanism for information interaction.

**Prostate cancer grading methods:** For the PCa grading task, most research efforts have focused on developing efficient classifiers. However, the stratification of clinically significant PCa is crucial for aiding physicians in diagnosis and treatment [37]. Cao et al. [4] first explored the use of mpMRI to predict fine-grained GS groups via CNN. Duran et al. [38] proposed a two-branch model for lesion grading in the peripheral zone (PZ). Mehta et al. [39] proposed a network framework based on patient-level data to address annotation costs. Vente et al. [40] analyzed variations among prostate lesions in different regions and incorporated the regional information into the network framework. Bhattacharya et al. [41] introduced whole-mount histopathology images and designed modules to learn correlation features between radiological and pathological images.

## 3. Method

### 3.1. Data preprocessing

Before inputting the data into the proposed network, it is necessary to register and crop the three mpMRI sequences. Additionally, pixel-level segmentation labels of the GS group need to be performed for processing, as shown in Fig. 1. First, the registration operation is carried out. Using the coordinate information stored in DICOM, the centers of the three imaging sequences corresponding to each slice are aligned. Then, resampling is performed on ADC and DWI. The sampling intervals of images in different modalities obtained from the instrument are inconsistent. We take the T2W image as the reference object for resampling and alignment. Next, a unified cropping process is carried out. During the experiment, it was observed that the cancerous area is small and almost all located within the prostate gland, while the background noise is large, which can cause certain interference to the model. Therefore, we crop out one-fourth of the area of the original image. The cropped image is aligned with the center of the original image, and its side length is half of the original image.

Pixel-level segmentation labels about the GS group included information about location, shape, and degree of aggressiveness. Each patient diagnosed with PCa undergoes a series of diagnostic procedures, including prostate-specific antigen (PSA) testing, transrectal ultrasound (TRUS) biopsy, and MRI. The aggressiveness was evaluated based on the pathology report obtained after the biopsy procedure. Pixel-level annotation of the lesions was performed by a highly experienced radiologist who specializes in diagnosing PCa. The radiologist used the information from the pathology reports. We categorized the annotation pixel-by-pixel into six groups: Non-lesion, GS 3+3, GS 3+4, GS 4+3, GS = 8, and GS ≥9. The categorization was performed using the one-hot encoding method.

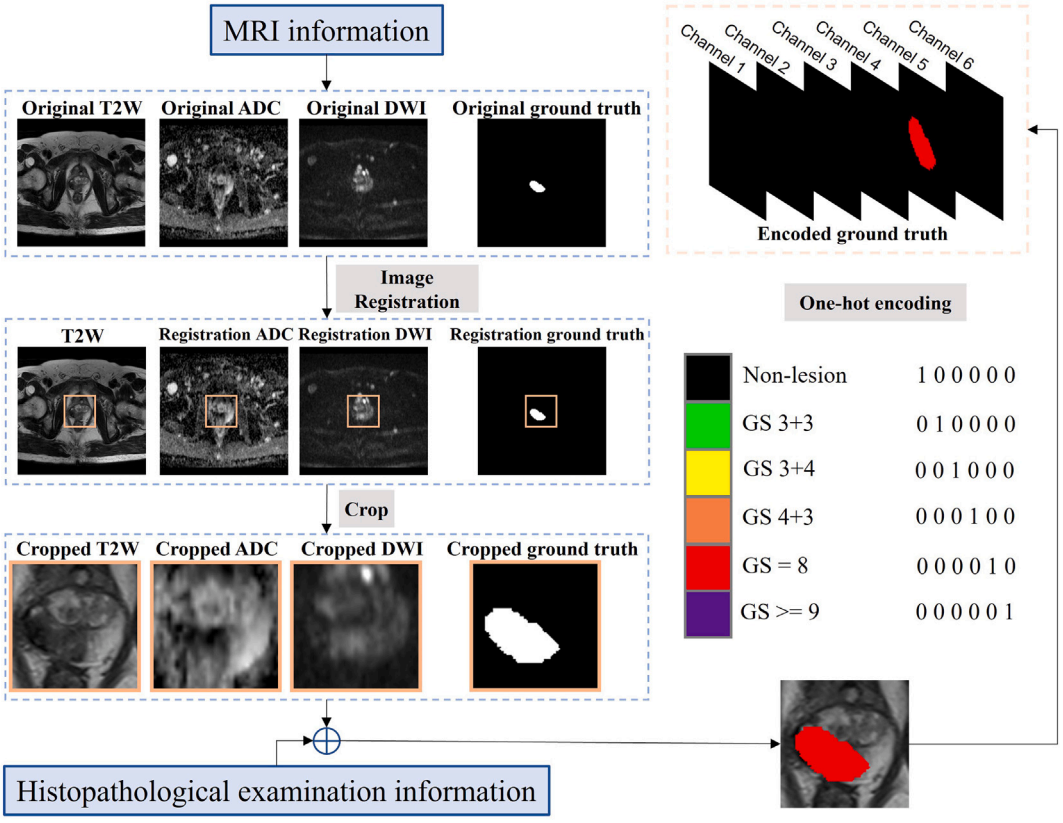



**Fig. 1.** Data preparation pipeline.

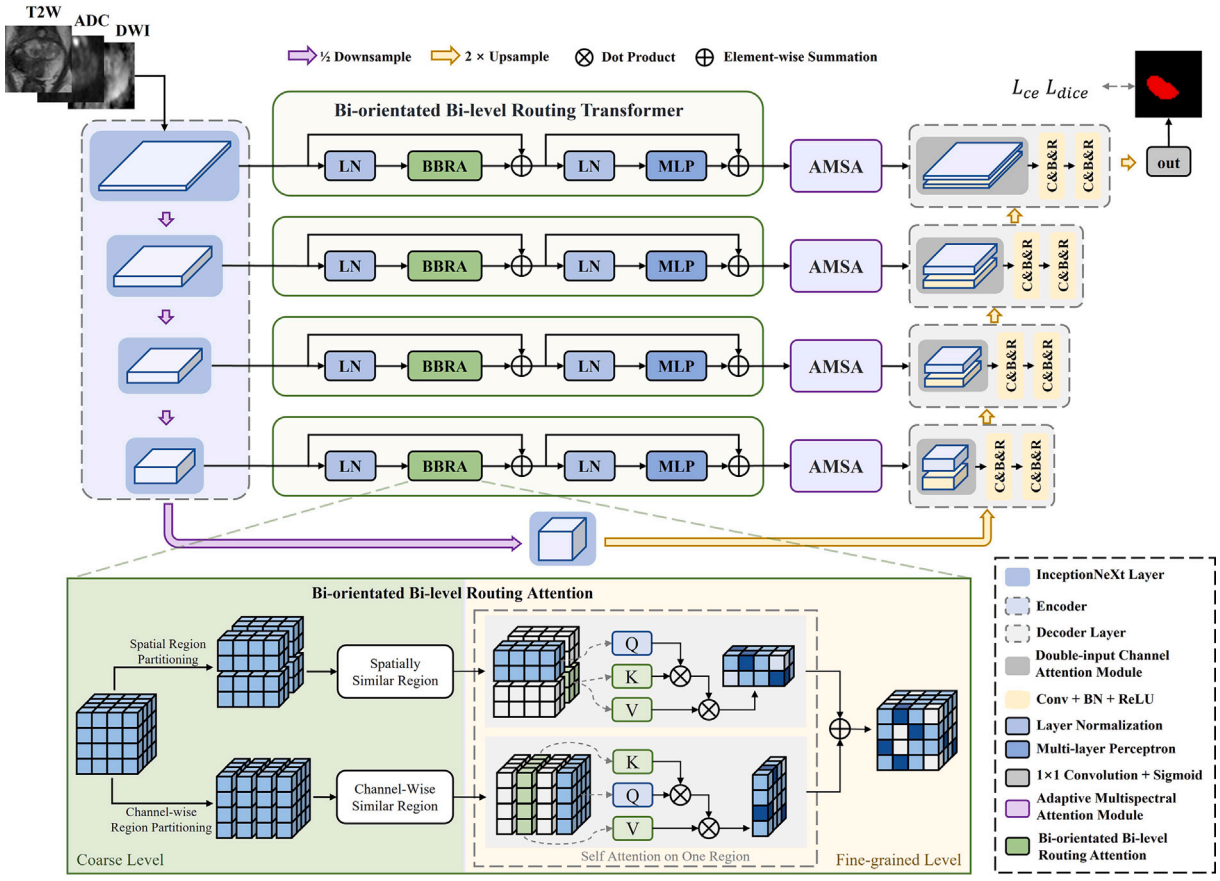



**Fig. 2.** The structure of the proposed network for fine-grained Gleason score (GS) groups prediction.

### 3.2. Network architecture

The network architecture proposed in this study comprises several components, including the InceptionNeXt encoder [42], BBRT, AMSA module, and a double-input channel attention module (DCAM) [43], as shown in Fig. 2. The network receives three imaging sequences as input, which are obtained from Section 3.1. The segmentation results are generated by the network in a six-channel format using an end-to-end approach.

The proposed network utilizes the first five stages of InceptionNeXt as feature encoders. First, the outputs of each layer are fed into a BBRT block, which is based on the Transformer framework and attention mechanism. By prioritizing the calculation of key-value pairs that are closely related to the current semantic region, it can extract important global features related to the target more effectively while reducing computational costs. Then, the features are separately fed into the Adaptive Multi-spectral Attention Module, which enhances the network's ability to acquire important detailed features by adaptively selecting frequency-represented components. It is important to note that the features output by the encoders at different layers are already at different scales. The BBRT block at each layer extracts global information for multi-scale feature extraction, and no additional multi-scale structures are designed within these two modules. Next, the output of the AMSA module is combined with deep semantic features using the DCAM proposed in our previous study. This allows the network to dynamically choose significant components from both the deep and shallow features. The DCAM in each layer is followed by two C&B&R modules. One C&B&R consists of a convolutional layer, a batch normalization layer, and a ReLU function. Finally, the segmentation results are output through a segmentation head.

### 3.3. Bi-oriented bi-level routing attention

Due to the typically small size of the lesion area and the presence of a lot of noise in the background, we hope that the network can efficiently focus on important information from the global context and ignore the interference. Inspired by the work of Zhu et al. [44], we propose a BBRT Block for extracting global information. The structure of the block is shown in Fig. 2. In BBRT, we employ sparse attention to control computational complexity. Unlike manually designed fixed sparse patterns [14,45], our proposed sparse attention mechanism can adaptively distinguish important features from global information and automatically enhance focus on critical regions. As illustrated in Fig. 3, we propose bi-orientated bi-level routing attention (BBRA). This method adopts a bi-level strategy: first performing coarse-grained screening to identify strongly correlated regions, followed by fine-grained attention within these regions. Specifically, BBRA first computes a correlation matrix in coarse-grained regions and then applies fine-grained token-to-token attention in candidate regions based on this matrix. Unlike [23], which was inspired by the intrinsic property of convolutional operations (i.e., extracting effective features by blending channel and spatial information) to employ bi-orientated pooling in attention mechanisms, we innovatively transform this concept into sparse region selection criteria, achieving adaptive region screening along both channel and spatial dimensions.

Given an input feature $X \in \mathbb{R}^{C\times H\times W}$, copy it to obtain two features. Then send each feature into two branches separately. Detailed descriptions are provided for the two branches.

**Spatial branch.** The feature is divided into $S_s \times S_s$ non-overlapping regions. Each region contains $\frac{HW}{S_s^2}$ feature vectors. The reshaped feature

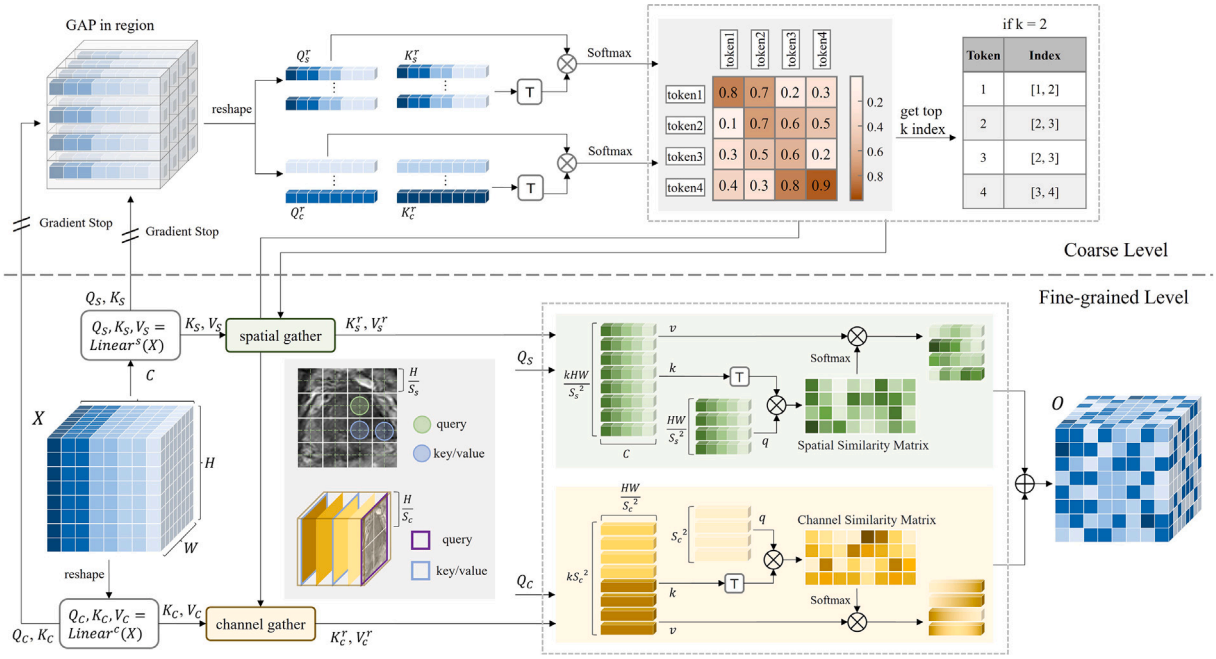



**Fig. 3.** The structure of bi-orientated bi-level routing attention.

$X_s^r \in \mathbb{R}^{S_s^2 \times \frac{HW}{S_s^2} \times C}$ is then obtained. Query, key, and value tensors are obtained by linear mapping, i.e.,

$$Q_s = X_s^r W_s^q, K_s = X_s^r W_s^k, V_s = X_s^r W_s^v \tag{1}$$

Where $W_s^q, W_s^k, W_s^v \in \mathbb{R}^{C \times C}$. Then, the correlation matrix between coarse regions is calculated. Global average pooling is applied to the features in region of $Q_s$ and $K_s$ to obtain region-level features $Q_s^r \in \mathbb{R}^{S_s^2 \times C}$ and $K_s^r \in \mathbb{R}^{S_s^2 \times C}$:

$$Q_s^r = \sum_{i=1}^{S_s^2} GAP(Q_s^i), K_s^r = \sum_{i=1}^{S_s^2} GAP(K_s^i) \tag{2}$$

Where i denotes the number of regions. Subsequently, a similarity matrix of size $S_s^2 \times S_s^2$ is obtained through matrix multiplication of the two features. The magnitude of the values in the similarity matrix reflects the degree of correlation between the regions. The top K regions with high relevance to each region between $Q_s^r$ and $K_s^r$ are retained, and the routing index matrix $I_s^r \in \mathbb{N}^{S_s^2 \times k}$ is obtained to guide fine-grained token-to-token attention:

$$I_s^r = topIndex(Q_s^r \cdot (K_s^r)^\top) \tag{3}$$

In fine-grained attention, first, the top K $K_s^i$ and $V_s^i$ relevant to $Q_s^i$ are summarized based on the routing index matrix $I_s^r$ to obtain continuous blocks $K_s^g, V_s^g \in \mathbb{R}^{S_s^2 \times \frac{kHW}{S_s^2} \times C}$ that are more friendly for GPU memory consolidation:

$$K_s^g = gather(K_s, I_s^r), V_s^g = gather(V_s, I_s^r) \tag{4}$$

Then, token-to-token attention is applied to $Q_s$, $K_s^g$, and $V_s^g$, while the depthwise convolution (DC) is employed, like [44], to preserve richer fine-grained feature information.

$$O = Attention(Q_s, K_s^g, V_s^g) + DC(V_s^g) \tag{5}$$

**Channel branch.** Features are divided into $S_c \times S_c$ non-overlapping patches, so that each patch contains $\frac{HW}{S_c^2}$ feature vectors, where deformed feature $X_c^r \in \mathbb{R}^{S_c^2 \times C \times \frac{HW}{S_c^2}}$. Unlike the spatial branch, C represents the number of partitioned regions, and $S_c^2$ is the length of the vector representing a channel region. After that, linear mapping is applied to obtain query, key, and value tensors, i.e.,

$$Q_c = X_c^r W_c^q, K_c = X_c^r W_c^k, V_c = X_c^r W_c^v \tag{6}$$

Where $W_c^q, W_c^k, W_c^v \in \mathbb{R}^{\frac{HW}{S_c^2} \times \frac{HW}{S_c^2}}$. Then, the correlation matrix between coarse regions in the channel direction is calculated. Global average pooling is applied to the features inside $Q_c$ and $K_c$ for each patch to obtain region-level features $Q_c^r \in \mathbb{R}^{C \times S_c^2}$ and $K_c^r \in \mathbb{R}^{C \times S_c^2}$:

$$Q_c^r = \sum_{j=1}^{C \times S_s^2} GAP(Q_c^j), K_c^r = \sum_{j=1}^{C \times S_c^2} GAP(K_c^j) \tag{7}$$

Where j represents the number of patches. Then, a $C \times C$ similarity matrix is obtained through matrix multiplication between the two features. Similarly, the indices of the top K regions in $K_c^r$ that are highly correlated with each region in $Q_c^r$ are retained, obtaining the routing index matrix $I_c^r \in \mathbb{N}^{C \times k}$, used to guide fine-grained token-to-token attention:

$$I_c^r = topIndex(Q_c^r \cdot (K_c^r)^\top) \tag{8}$$

In fine-grained attention, similarly gather all key-value pairs residing in the union of the K routed regions based on the routing index matrix. Merge them to obtain continuous blocks $K_c^g, V_c^g \in \mathbb{R}^{C \times kS_c^2 \times \frac{HW}{S_c^2}}$ that are more friendly for coalesced memory operations on the GPU:

$$K_c^g = gather(K_c, I_c^r), V_c^g = gather(V_c, I_c^r) \tag{9}$$

Where $K_c^g, V_c^g \in \mathbb{R}^{C \times kS_c^2 \times \frac{HW}{S_c^2}}$ represent the aggregated representation of key and value tokens along the channel direction. Then apply token-to-token attention to $Q_c$, $K_c^g$, and $V_c^g$, and introduce DC:

$$O = Attention(Q_c, K_c^g, V_c^g) + DC(V_c^g) \tag{10}$$

Finally, the features of the two branches are added to obtain the final result, which will be inputted to an ASMA block for further processing.

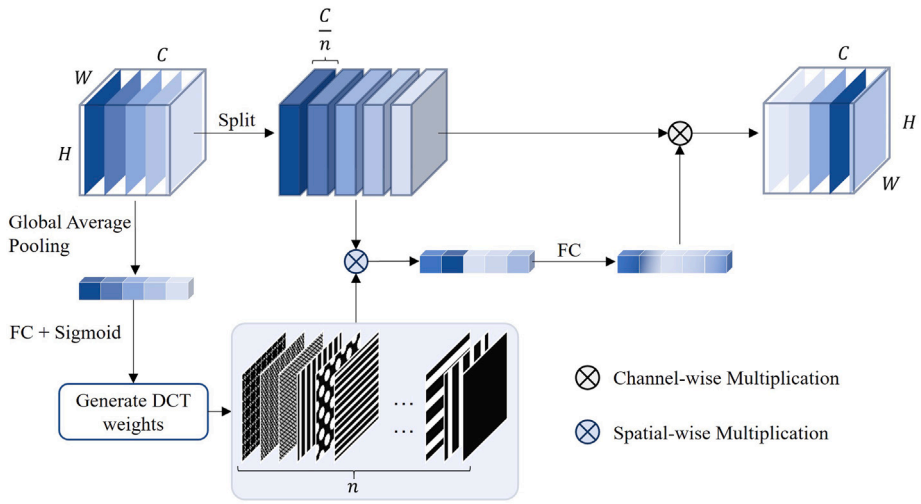



**Fig. 4.** The structure of an adaptive multispectral channel attention module.

### 3.4. Adaptive multispectral channel attention

As discussed above, considering the importance of shallow texture features in detecting lesion boundaries, the proposed ASMA block is used to enhance the network's focus on important detailed information. Its structure is shown in Fig. 4. Qin et al. [22] have proved that discrete cosine transform (DCT) can be viewed as a weighted sum of inputs and GAP corresponds to the lowest frequency component of 2D DCT. They determine the importance of each frequency component in the classification task. We believe that the optimal selection of frequency components is different depending on the task. We hope that the selection of frequency components is guided by input features rather than being manually designed. A detailed description of the proposed AMSA block is as follows.

Two operations are performed on the input feature $X \in \mathbb{R}^{C\times H\times W}$: (1) it is divided into n groups along the channel direction, with each group assigned the same spatial weight. (2) Global average pooling is performed to obtain a vector $P \in \mathbb{R}^{C}$, with global information, which is used to guide the generation of weights.

$$P = GAP(X) \tag{11}$$

For this vector, an FC layer and Sigmoid function are applied, and then it is mapped to the range 0 to 8 and rounded down to obtain the frequency position vector $P^f \in \mathbb{R}^{2\ n}$:

$$P^f = Floor(Map(Sigmoid(W_1\ P))) \tag{12}$$

Where $W_1 \in \mathbb{R}^{2\ n\times C}$ is the learnable linear transformation matrix, $Sigmoid()$ represents the sigmoid activation function, $Map()$ performs linear scaling (mapping feature values from 0–1 to 0–8), and $Floor()$ denotes the rounding down operation (with gradient preservation). Frequency domain coordinates are extracted from $P^f$ to guide the generation of weight maps, and the generation of weight maps still follows the 2D DCT basis function, which is:

$$B^{i,j}_{h,w} = \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi\omega}{W}\left(j+\frac{1}{2}\right)\right)$$
$$s.t.i \in \{0,1,\ldots,H-1\}, j \in \{0,1,\ldots,W-1\} \tag{13}$$

Where $(i, j)$ is spatial domain coordinate, and $(h, w)$ is frequency domain coordinate. Based on the obtained n sets of frequency domain coordinates, the corresponding weight maps are calculated and weighted with the grouped features to obtain a vector $V \in \mathbb{R}^{C}$ with global channel statistics:

$$V = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} X^{i,j} B^{i,j}_{h,w}$$
$$s.t.h, w \in P^f \tag{14}$$

To limit model complexity and aid generalization, referencing Hu et al. [46], we employ a simple gating mechanism with two fully connected (FC) layers and a sigmoid activation, i.e.,

$$V_2 = Sigmoid(W_3\delta(W_2\ V)) \tag{15}$$

where $\delta$ refers to the ReLU function, $W_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_3 \in \mathbb{R}^{C\times\frac{C}{r}}$. The final output of the block is obtained by channel-wise multiplication.

$$O = V_2 \otimes X \tag{16}$$

Where $\otimes$ refers to the channel-wise multiplication.

## 4. Results

In this section, we first present the dataset and implementation details. Second, we investigate the effectiveness of the proposed method on the three segmentation tasks. Next, we show the ablation experiments regarding network structure. Finally, we provide visualization results to demonstrate the advantages of the experimental outcomes.

### 4.1. Dataset and evaluation metrics

Due to the scarcity of publicly available PCa datasets with voxel-level GS annotations, we incorporated segmentation tasks for prostate region and brain stroke. We conducted experiments using three datasets: one private dataset related to the prostate cancer and two publicly available datasets concerning the prostate region and stroke.

**Prostate Cancer Multiparametric MRI (PCAMM):** This dataset, provided by Shanghai Tongji Hospital, is private and dedicated to the PCa grading task. It encompasses mpMRI sequences, specifically including ADC, T2W, and DWI of 171 PCa cases. For each PCa-diagnosed patient, a PSA test, a TRUS biopsy, and an MRI examination were carried out. A radiologist with extensive experience in PCa diagnosis was responsible for annotating the images based on the pathology report. To ensure the accuracy of the annotation, a pathologist provided assistance

**Table 1**
The number of samples for different grades in the PCAMM.

| Grade | GS 3 + 3 | GS 3 + 4 | GS 4 + 3 | GS = 8 | GS ≥ 9 |
|---|---|---|---|---|---|
| Number | 43 | 32 | 32 | 30 | 34 |

during the labeling process, guaranteeing that the biopsy position corresponding to the pathological section was consistent with the labeled lesion position on the MR image. In this study, the index (dominant) lesions, which are the lesions with the highest Gleason grade and most likely to be suspected as tumors, were the primary focus for labeling. Among the 171 PCa cases, there are both clinically significant PCa (Gleason score, GS > 6) and non-clinically significant PCa (GS ≤ 6). The number of samples for different grades is presented in Table 1. The patients' ages range from 47 to 92, with an average age of 73.98 ±9.1 years. Regarding the number of lesions, 132 patients have a single lesion, 31 patients have two lesions, and 8 patients have three lesions. For sampling of the data, the dataset features a diversity in pixel and inter-slice spacing. Specifically, the pixel spacing of the ADC and DWI sequences varies from 1.22 to 1.95 mm, while that of the T2W sequence ranges from 0.28 to 0.39 mm. The inter-slice spacing values are {3.6, 4.2, 4.8, 5.4, 6.0 mm}. Ethical approval for this experimental protocol was obtained from the Ethics Committee of Tongji Hospital Affiliated to Tongji University (Approval Number SBKT-2024–065).

**PROSTATEx:** This dataset is publicly available and used for prostate region segmentation [47]. The dataset includes 98 prostate mpMRI sequences (ADC, T2W, and DWI), which were acquired using two different types of Siemens 3 T MRI scanners (MAGNE-TOM Trio and Skyra) with a body coil. The dataset also provides annotations for the central gland (CG) and PZ regions.

**ISLES2022:** This dataset is publicly available and is used for stroke segmentation. It includes 250 multi-modality MRI scans [48]. Each scan includes diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC), and fluid attenuated inversion recovery (FLAIR) sequences. We chose DWI and ADC as input in this study.

In this study, three metrics are chosen to evaluate the ability of the network to segment the target regions following the common practice: the average Dice similarity coefficient (DSC), average boundary distance (ABD), and relative volume difference (RVD).

### 4.2. Implementation details

Dice loss and cross-entropy loss are used to train the network. We slice all the MRI data and perform five-fold cross-validation on three datasets. The proposed method was implemented in PyTorch. All the cropped prostate region of interest images in this study were resized to 256 × 256. The cropping operation is only performed on the PCAMM. The default size of other images input into the network is also 256×256. To reduce the influence of overfitting caused by limited datasets, several data augmentation operations were employed, including random rotation (0°, 90°, 180°, 270°) and random flipping (up-down or left-right in the x-y planes), and random scaling (0.8–1.2 times). The Adam optimizer was employed for training our model in an end-to-end manner with an initial learning rate of 0.0001 and betas of (0.9, 0.999). The learning rate decayed by a dynamic weight decay of $\left(1-\frac{epoch}{max\_epoch}\right)^{0.9}$. The max _*epoch* settings on the three datasets PCAMM, ISLES2022, and PROSTATEx were 750, 100, and 200 respectively. The batch size was set to 32 by default on an NVIDIA TITAN RTX GPU with 24 GB of memory.The preprocessing phase (including training mask generation) takes 58.1 s in total on the PCAMM dataset, averaging 0.34 s per case, which is negligible compared to the total training time.

### 4.3. Results on the three datasets

We compare the proposed network with eight widely used medical image segmentation networks, four based on CNNs and three based on Transformers. Various networks including UNet, UNet + +, Attention U-Net, nnUNet [49], MedT, TransUNet, Swin-Unet, and UCTransNet [50] are used in this study. We trained these eight networks without utilizing any additional pretraining parameters. The images inputted into the MedT, TransUNet, and Swin-Unet are resized to 224 × 224. Tables 2, 3, and 4 illustrate the experimental results on three datasets.

(1) **Result for PCAMM.** Table 2 shows the average DSC, ABD, and RVD for each model on the PCAMM. The proposed network achieves the highest scores in DSC, ABD, and RVD, with values of 76.09 % (±0.23), 3.63 (±0.02), and 8.64 % (±1.54), respectively. Compared to CNN-based segmentation methods, the three metrics of our method exhibit improvements of 4.69 %, 0.6, and 6.71 % over those of Attention U-Net. Compared to Transformer-based segmentation methods, our method demonstrates metrics that are 2.82 %, 0.42, and 3.12 % higher than those of Swin-Unet. These results demonstrate that the proposed method exhibits significant advantages in segmentation accuracy (evaluated by DSC), boundary delineation precision (assessed by ABD), and over-segmentation control (measured by RVD), enabling more accurate target prediction while reducing boundary errors and volume deviations.

Table 2 also shows the segmentation results on different GS groups. Experimental observations revealed that the prediction accuracy of all models for lesions across five grades (five sample groups) generally exhibited an increasing trend from low-grade to high-grade. This phenomenon aligns well with the pathological characteristics of prostate cancer: high-grade tumors typically exhibit significant morphological differences from normal tissues and larger volumes, rendering their features more distinct. In contrast, low-grade tumors not only share highly similar histological features with normal tissues but are also prone to confusion with benign conditions such as prostatitis and benign prostatic hyperplasia. Such histological similarity, combined with the generally smaller size of low-grade tumors, poses substantial challenges for network models in feature extraction and segmentation tasks. The proposed method demonstrates superior DSC scores compared to other comparative methods across all groups. On samples with GS 3 + 3,

**Table 2**
The segmentation results of different competing methods for various grades of cancer in PCAMM. ↑ indicates that higher values are better and ↓ represents the opposite. The best results are highlighted in bold text and suboptimal results are underlined. (mean ±standard deviation of the Dice similarity coefficient).

| Method | GS 3 + 3 | GS 3 + 4 | GS 4 + 3 | GS = 8 | GS ≥ 9 | Average | | | p value |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | DSC↑ | ABD↓ | RVD↓ | |
| UNet | 31.93±4.81 | 61.85±1.45 | 66.14±0.39 | 70.64±0.87 | 77.00±0.45 | 71.21±0.40 | 3.93±0.36 | 12.30±2.19 | < 0.01 |
| UNet + + | 33.60±0.57 | 60.45±4.44 | 63.85±1.52 | 75.18±2.40 | 78.53±0.20 | 70.72±0.61 | 4.74±0.11 | 12.45±3.25 | < 0.01 |
| Attention UNet | 34.46±4.08 | 57.79±7.04 | 69.93±4.93 | 74.56±2.53 | 74.71±2.83 | 71.40±1.19 | 4.23±0.06 | 15.35±1.82 | < 0.01 |
| MedT | 29.01±5.12 | 49.54±7.17 | 65.78±3.01 | 66.75±1.43 | 75.29±0.89 | 64.11±0.71 | 6.71±0.14 | 18.31±1.99 | < 0.01 |
| TransUNet | 39.34±5.86 | 63.78±1.67 | 68.65±1.39 | 76.19±0.22 | 77.81±1.32 | 70.96±0.48 | 3.90±0.30 | 15.46±1.60 | 0.02 |
| Swin-Unet | 35.04±1.55 | 62.83±1.54 | 67.49±0.48 | 75.58±0.86 | 81.74±2.45 | 73.27±0.35 | 4.05±0.16 | 11.76±2.16 | < 0.01 |
| nnUNet | 36.56±3.09 | 63.40±1.01 | 67.09±2.07 | 70.60±0.84 | 77.19±0.92 | 72.02±0.60 | 3.98±0.22 | 12.31±1.73 | < 0.01 |
| UCTransUNet | 36.04±2.32 | 63.51±1.90 | 68.66±1.84 | 75.02±0.81 | 81.66±1.96 | 73.19±1.03 | 4.11±0.19 | 11.89±1.45 | < 0.01 |
| Ours | **44.13±0.84** | **66.55±2.18** | **73.72±1.32** | **79.83±1.65** | **82.03±0.19** | **76.09±0.23** | **3.63±0.02** | **8.64±1.54** | – |

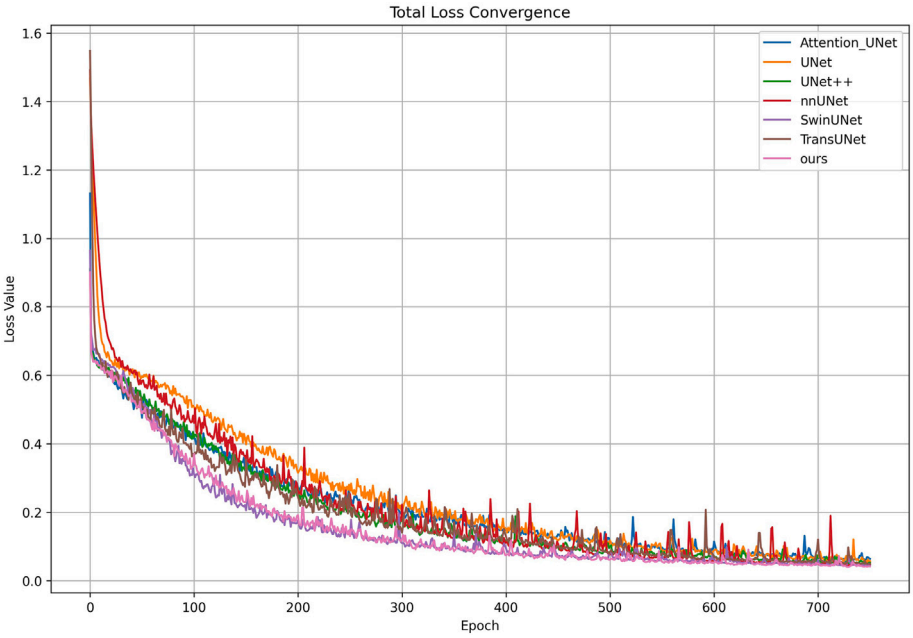



**Fig. 5.** The loss convergence of the model training on PCAMM data.

the score is 4.79 % higher than the suboptimal method (TransUNet). Regarding clinically significant prostate cancer, the proposed method demonstrates enhancements of 2.77 %, 3.79 %, 3.64 %, and 0.29 % for GS 3 + 4, GS 4 + 3, GS = 8, and GS ≥ 9, respectively, compared to suboptimal methods (underlined in Table 2).

The suboptimal segmentation network Swin-Unet performs well on lesions with GS ≥ 9, but its segmentation performance is generally inferior to that of TransUNet on lesions of lower grades. The proposed network achieves optimal segmentation results across all grades. From the perspective of model convergence, the network combining Transformer and CNN structures exhibits slightly faster convergence than the pure CNN architecture, as shown in Fig. 5. The proposed method maintains a relatively rapid convergence rate. Moreover, the standard deviation of the three metrics obtained from the five-fold cross-validation of the proposed network is consistently smaller than that of other methods. This indicates that our network has better stability. Furthermore, Wilcoxon signed-rank tests were used to compare the DSC score differences between our model and baseline models on the test set (all $p < 0.05$), demonstrating statistically significant improvements.

Fig. 6 shows partial segmentation results of the aforementioned methods, encompassing lesions of five different degrees of invasiveness. The lesions in samples (a), (b), and (f) are small and can easily be confused or misidentified among categories. The lesions in samples (e), (g), (h), and (i) are irregular in shape and multifocal, with some areas of the lesions being easily overlooked. The boundaries of the lesions are blurred in all samples. Compared to other methods, the proposed method provides a segmentation that is closer to the actual shapes of the lesions.

(2) **Result for PROSTATEx.** Table 3 displays the experimental results on the prostate region segmentation dataset. The prediction results are evaluated using three metrics for the CG and PZ regions, respectively. For the CG region, the proposed network obtains the best results on DSC and ABD, 94.03 % (±0.24) and 2.29 (±0.11), respectively; for the PZ region, the proposed network achieves the best results on ABD and RVD, 1.81 (±0.57) and 6.04 (±0.89), respectively. Although our method achieves higher DSC than Swin-Unet on the CG segmentation task, it shows slightly lower DSC on the PZ task. We attribute this to the challenge of simultaneously segmenting two closely adjacent target regions. Regarding regional overlap (DSC), Swin-Unet appears more focused on the PZ region whereas our method demonstrates stronger performance on the CG region. Moreover, our model outperforms Swin-Unet in both PZ and CG regions for both boundary matching accuracy (ABD) and volume deviation degree(RVD). Additionally, our network has fewer parameters and delivers superior qualitative results on challenging samples.

Fig. 7 displays the results of the proposed method and the comparison method for the prostate region segmentation. Compared to the CG, the segmentation of the PZ is more challenging, with a higher incidence of PCa. In the relatively simple samples (c)–(d), most methods yield satisfactory qualitative results, especially for the CG region. In the more difficult samples (a)–(b), generally located in the upper 1/3 and lower 1/3 of the gland, the discrepancies in prediction outcomes among different methods become more pronounced, primarily manifested in the capability of boundary delineation. The experimental results demonstrate that the proposed method exhibits superior performance in edge segmentation.

(3) **Result for ISLES2022.** We have identified a challenge in automatically segmenting low-grade PCa due to the small lesion area. To validate the proposed network in segmenting small and complex targets, we opted to conduct further experiments using a stroke dataset. The proposed network achieves the best results on three metrics, namely 86.33 % (±0.30), 1.29 (±0.08), and 0.16 % (±0.03). The results are shown in Table 4. Compared to U-Net + +, the proposed method improved by 0.67 %, 0.25, and 0.02 % on the three metrics, respectively. Compared to TransUnet, it showed improvements of 0.97 %, 0.11, and 0.12 % on the three metrics. The proposed network achieves effective segmentation results for small lesions due to BBRT's improved perception of target information across multiple scales.

Fig. 8 illustrates the segmentation results of the previously mentioned method in stroke cases. Stroke lesions often appear as multiple small and scattered abnormal signal areas on MRI slices, which may be overlooked by the network. As demonstrated in samples (a-d), the proposed network successfully detected the scattered small lesions. In more challenging samples (e-f), the proposed network demonstrated lower false negative rates in its predictions compared to other contrast networks.

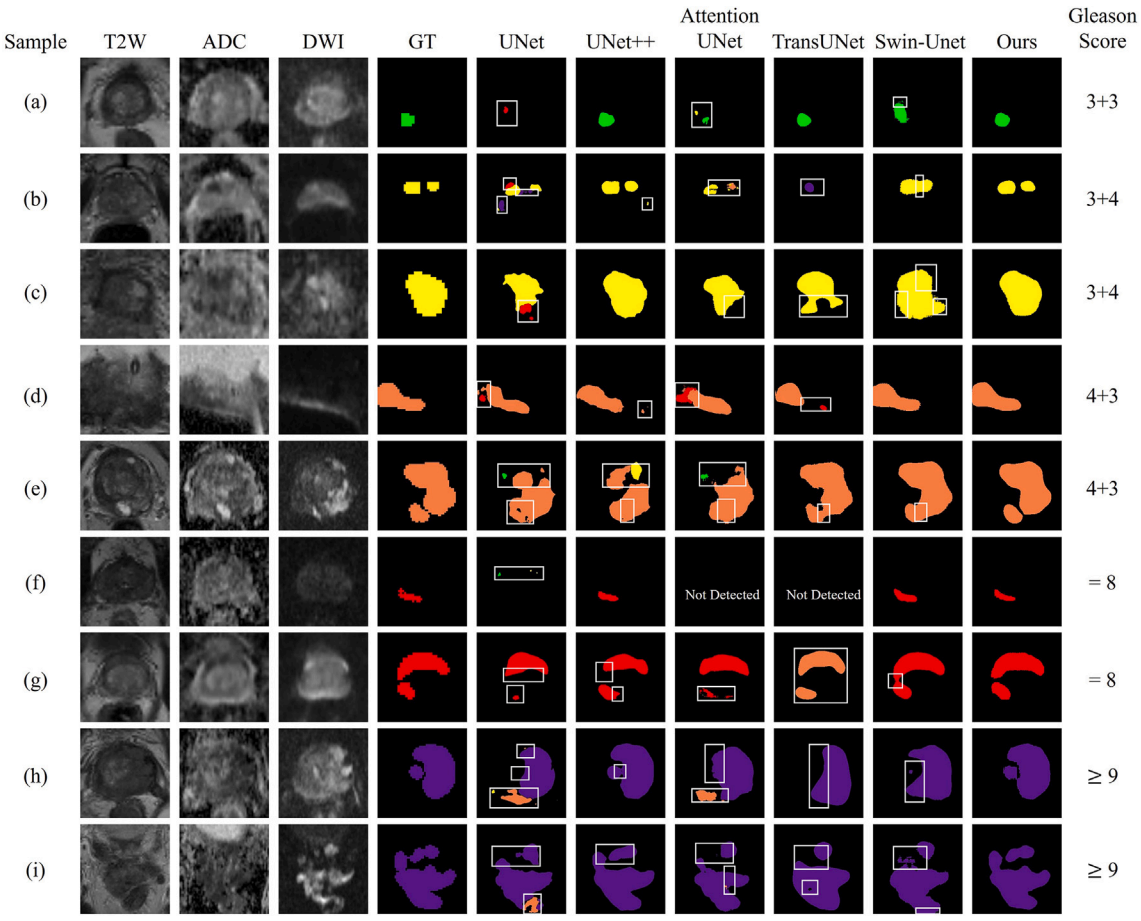



**Fig. 6.** Segmentation results of UNet, UNet++, Attention UNet, TransUNet, Swin-Unet, and the proposed network on PCAMM. The white boxes display the prediction bias between the network's predicted results and actual labels.

**Table 3**
The segmentation results (mean±std) of different competing methods on PROSTATEx. ↑ means the higher the better and ↓ represents the opposite. The best results are marked in bold text and suboptimal results are underlined.

| Method | CG | | | PZ | | |
|---|---|---|---|---|---|---|
| | DSC(%)↑ | ABD(mm)↓ | RVD(%)↓ | DSC(%)↑ | ABD(mm)↓ | RVD(%)↓ |
| UNet | 93.70±0.54 | 2.54±0.17 | 5.81±2.37 | 83.90±0.31 | 2.02±0.16 | 6.42±1.01 |
| UNet++ | 92.96±0.61 | 2.84±0.14 | 6.48±1.93 | 83.59±0.37 | 2.14±0.11 | 7.10±1.23 |
| Attention U-Net | 93.31±0.40 | 2.59±0.25 | 7.12±2.94 | 83.78±0.52 | 1.94±0.27 | 7.78±0.92 |
| MedT | 89.70±0.27 | 3.71±0.20 | 5.13±1.53 | 72.49±0.58 | 3.12±0.12 | 8.13±0.67 |
| TransUNet | 93.41±0.21 | 2.40±0.33 | **4.64±2.15** | 83.64±0.22 | 1.86±0.64 | 6.38±0.84 |
| Swin-Unet | 93.78±0.30 | 2.48±0.28 | 6.84±0.32 | **84.95±0.44** | 2.13±0.29 | 7.15±0.46 |
| Meyer et al. [51] | 87.60±6.60 | - | - | 79.80±5.10 | - | - |
| Ours | **94.03±0.24** | **2.29±0.11** | 5.09±1.60 | 84.52±0.39 | **1.81±0.57** | **6.04±0.89** |

### *4.4. Ablation experiments*

(1) **Ablation for the framework.** To demonstrate the contributions of BBRT and AMSA, we constructed Base + BBRT and Base + AMSA for ablation experiments based on base model. The base model consists of the InceptionNeXt encoder and CAMs. The segmentation results of all networks on three datasets are presented in Table 5. Across the three datasets, the networks incorporating BBRT and AMSA have demonstrated improvements in DSC and ABD compared to the base model, indicating their general effectiveness. The DSC and ABD of Base + BBRT + AMSA network are higher than those of Base + BBRT and Base + AMSA networks, indicating that BBRT and AMSA can enhance each other and facilitate better segmentation performance.

To intuitively illustrate the role of the proposed modules, we visualized several sample feature heat maps. They provide some insights into how the model performs segmentation. The samples are from PCAMM dataset. For each sample, we selected four feature maps and the model's predictions during the inference process. As shown in Fig. 9, map 1 to map 4 correspond to the four features extracted from the first layer. These features represent the output of the encoder's first layer, the output of the BBRT, the output of the AMSA, and the output of the DCAM. It can be seen that the features processed by the BBRT exhibit a higher response to important details within the global context. The features that pass through the AMSA show a more sensitive representation of potential target boundaries. Furthermore, the features processed by the DCAM,

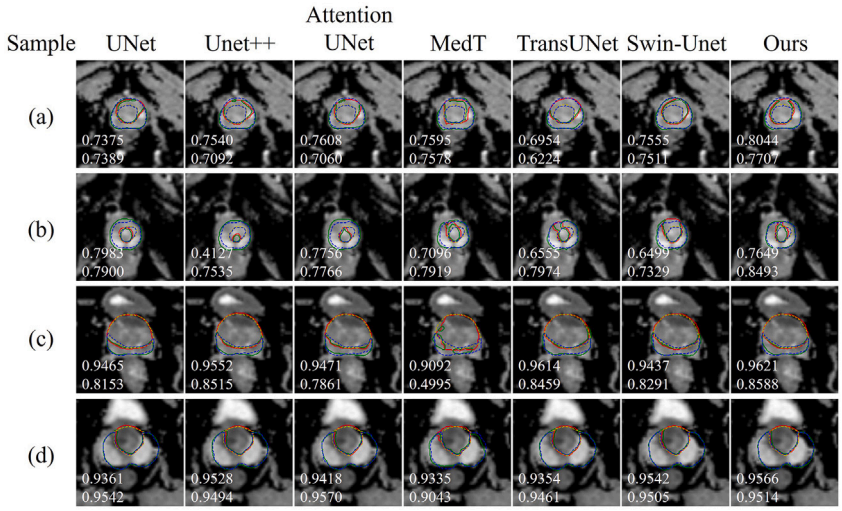



**Fig. 7.** Segmentation results of UNet, UNet++, Attention UNet, MedT, TransUNet, Swin-Unet, and the proposed network on PROSTATEx. The red solid line represents the predicted central gland (CG), the green solid line represents the predicted peripheral zone (PZ), while the yellow dashed line and blue dashed line represent the actual CG and PZ, respectively. The two values represent the dice similarity coefficient for the predicted CG (top) and PZ (bottom).

**Table 4**
The segmentation results (mean±std) of different competing methods on ISLES2022. ↑ means the higher the better and ↓ represents the opposite. The best results are marked in bold text and suboptimal results are underlined.

| Method | DSC(%)↑ | ABD(mm)↓ | RVD(%)↓ |
|---|---|---|---|
| UNet | 85.40±0.28 | 1.72±0.04 | 0.33±0.09 |
| UNet++ | <u>85.66±0.35</u> | <u>1.54±0.07</u> | <u>0.18±0.05</u> |
| Attention U-Net | 85.51±0.55 | 1.58±0.20 | 0.22±0.06 |
| MedT | 80.93±0.11 | 2.23±0.26 | 0.81±0.67 |
| TransUNet | 85.36±0.36 | 1.40±0.05 | 0.28±0.06 |
| Swin-Unet | 85.27±0.40 | 1.57±0.13 | 0.21±0.11 |
| W-Net [52] | 85.6±- | – | – |
| Our | **86.33±0.30** | **1.29±0.08** | **0.16±0.03** |

which incorporate deeper features containing stronger semantics, are able to determine the target location better.

(2) **Ablation for proposed bi-orientated bi-level routing attention.** To demonstrate the superior structural advantages of BBRT and AMSA, we conducted further ablation experiments on the two modules separately. We replaced BBRA in BBRT with self-attention in Vision Transformer (ViT) [53], separate spatial branch, separate channel branch, and two branches in series, respectively. The segmentation results of different core components in BBRT are shown in Table 6. Compared to the experimental results of the Base+AMSA network in Table 5, using separate channel branches or spatial branches alone resulted in only marginal improvement. When the spatial branch and the channel branch are connected in series, the network performance decreases. However, when the two branches are connected in parallel, the BBRT has a positive effect, enhancing the network performance.

(3) **Ablation for proposed AMSA.** To further demonstrate the advantages of AMSA, we compared it with squeeze-and-excitation network (SE) [46] and MSA methods. MSA uses the top 16 frequency components provided in the official code. We conducted five-fold cross-validation on the PCAMM dataset. The segmentation results using different components instead of AMSA in the proposed network architecture are presented in Table 7. AMSA demonstrated a 3.38 % improvement over MSA in DSC, indicating that the adaptive selection of frequency components based on input features can reduce sensitivity to input data, enhance network robustness, and make it more adaptable for transfer to other tasks. In addition, we conducted ablation experiments to determine the optimal number of multispectral components in the adaptive multispectral channel attention. The experiments tested various values of n (1, 2, 4, 8, 16, 32), and as shown in Table 8, the configuration with 16 frequency components achieved the best performance.

(4) **Computational efficiency and resource consumption.** We evaluated the computational complexity and resource consumption of all models, including the number of parameters (Param), floating-point operations (FLOPs) per sample, peak GPU memory usage during training, and inference time per sample (in milliseconds). As shown in Table 9, our network balances computational complexity (Param/FLOPs) and segmentation performance. Notably, MedT required significantly longer inference time (149.44 ms) compared to other models (mean 11.03 ms), while the remaining models showed comparable speeds with variations within ±5.66 ms.

## 5. Discussion

Deep learning algorithms have demonstrated promise in MRI segmentation tasks. For the PCa grading task, most research efforts have focused on building effective classifiers. However, considering the diversity in PCa size and the complexity of its structures, it may be more beneficial for radiologists to identify specific abnormal areas along with grading information. Therefore, our proposed method aims to provide fine-grained grading information. The annotation of each sample's abnormal area is performed by a highly experienced radiologist specializing in PCa diagnosis, in collaboration with pathological reports that provide grading information.

According to the Gleason grading standard, we propose a new segmentation framework for fine-grained grading of PCa, which can also be extended to other MRI segmentation tasks. In the ablation experiment, it is evident from the experimental results in Tables 5–7 that our network architecture settings enable the network to achieve the best performance in grading the lesion areas of PCa. In the comparative experiment section, several conclusions can be drawn. Hybrid CNN-Transformer networks outperform pure CNN architectures, a conclusion that holds true for nnUNet. Although nnUNet exhibits remarkable robustness in medical image segmentation through adaptive parameter selection (including stage number, batch size, and patch size, etc) and model ensembling (2D/3D/cascade 3D UNet), its performance on the PCAMM dataset remains constrained by the underlying 2D UNet structure - even with

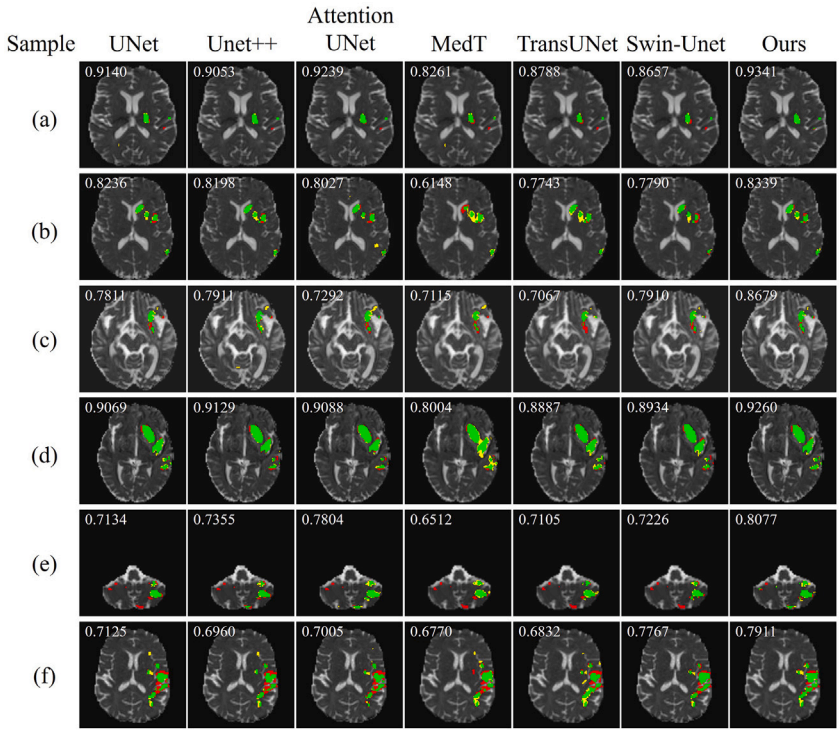



**Fig. 8.** Segmentation results of UNet, UNet++, Attention UNet, MedT, TransUNet, Swin-Unet, and the proposed network on ISLES2022. The green represents true positive areas, the yellow represents false positive areas, and the red represents false negative areas. The values indicate the dice similarity coefficient of the predicted results for this sample.

**Table 5**
The segmentation results (mean±std) of different network components on PCAMM, PROSTATEx, and ISLES2022. ↑ means the higher the better and ↓ represents the opposite. The best results are marked in bold text. The upper part of PROSTATEx is for the central gland (CG), while the lower part is for the peripheral zone (PZ).

| Dataset | model | DSC(%)↑ | ABD(mm)↓ | RVD(%)↓ |
|---|---|---|---|---|
| PCAMM | base | 73.46±0.32 | 4.48±0.02 | 14.99±4.06 |
| | +BBRT | 74.89±0.10 | 3.89±0.13 | 15.21±4.66 |
| | +AMSA | 74.92±0.24 | 3.80±0.08 | 13.62±2.90 |
| | +BBRT+AMSA | **76.09±0.23** | **3.63±0.02** | **8.64±1.54** |
| PROSTATEx | base | 92.79±0.31 | 2.86±0.03 | 8.00±3.64 |
| | +BBRT | 93.68±0.29 | 2.43±0.03 | 7.25±2.32 |
| | +AMSA | 93.36±0.07 | 2.56±0.08 | 6.70±1.92 |
| | +BBRT+AMSA | **94.03±0.24** | **2.29±0.11** | **5.09±1.60** |
| | base | 80.66±0.58 | 2.51±0.24 | 10.59±2.15 |
| | +BBRT | 83.53±0.22 | 2.03±0.13 | 10.91±1.35 |
| | +AMSA | 82.49±0.13 | 2.24±0.15 | 8.63±1.39 |
| | +BBRT+AMSA | **84.52±0.31** | **1.81±0.37** | **6.04±0.89** |
| ISLE2022 | base | 85.12±0.17 | 2.05±0.66 | 0.23±0.10 |
| | +BBRT | 86.01±0.14 | 1.72±0.10 | 0.29±0.12 |
| | +AMSA | 85.84±0.20 | 1.70±0.18 | 0.12±0.8 |
| | +BBRT+AMSA | **86.33±0.30** | **1.29±0.08** | **0.16±0.05** |

its 6-stage configuration (versus 5-stage in baseline UNet). This suggests that the global receptive field of the Transformer is beneficial for extracting target feature representations from MRI and enhancing network segmentation performance. Our proposed method achieves the optimal DSC score on three datasets, demonstrating the network's strong robustness. Furthermore, Tables 2–4 demonstrate that the segmentation performance of comparison methods varies across the three datasets, indicating their sensitivity to different input data and lower robustness.

The primary challenges in the fine-grained grading task of PCa include the difficulty in retaining and identifying subtle lesion features, including (1) the tendency to overlook small lesions and those with low GS. As the grade and size of PCa increase, the features of the lesion area become more obvious, and the network performs better in segmenting high-grade and large lesions. Conversely, the network has lower accuracy in recognizing smaller-sized and lower-grade lesions. (2) The loss of fine edge details. The features of the central region of the lesion are relatively distinct, while the features of the edge region are unclear, making it less distinguishable from normal tissue. (3) It is easy to confuse lesions of different grades. The proposed network addresses the aforementioned issues. As shown in Table 2, the proposed method has improved the detection of lesions of all grades. Compared to suboptimal methods (underlined in Table 2), the proposed network has increased the DSC by 4.79 %, 2.77 %, 3.79 %, 3.64 %, and 0.29 % for lesions at five different grades. Combined with Tables 3 and 4, the proposed network can be well applied to other MRI segmentation tasks. Overall, the quantitative improvement in DSC scores is modest (albeit $< 5\%$ ), we contend that even marginal gains could be clinically significant for prostate cancer Gleason grade segmentation, particularly in: 1) Reducing undergrading risk: Minor improvements in detecting high-grade foci may prevent their misclassification into intermediate-grade groups, which directly influence clinical decisions between active surveillance and radical treatment; 2) Guiding targeted biopsies: More precise localization of grade-specific lesions could enhance the accuracy of MRI-ultrasound fusion biopsy sampling, potentially decreasing the need for repeat biopsies.

Beyond the segmentation task addressed in this study, the proposed network holds potential for other segmentation tasks. The BBRT module in the network enables long-range dependency modeling, while the AMSA mechanism enhances boundary awareness. These capabilities are also crucial for handling other segmentation tasks, such as brain tumour segmentation [54] or abdominal organ segmentation [55]. Although the

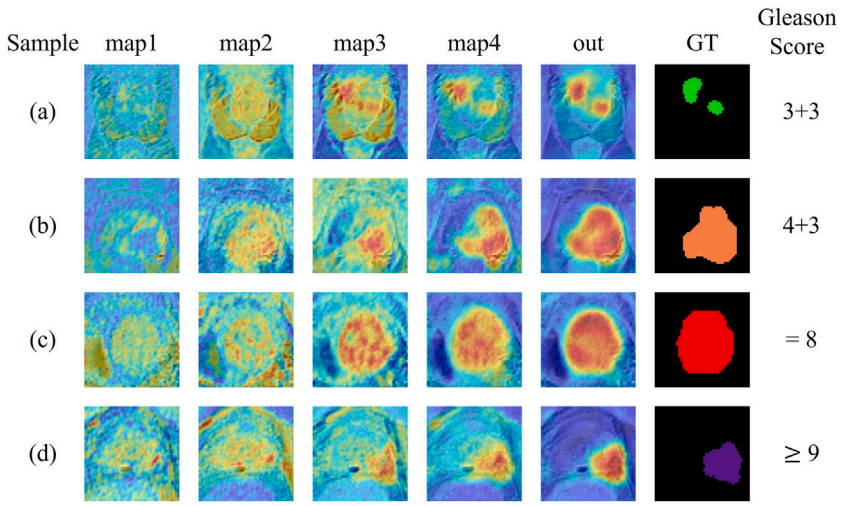



**Fig. 9.** Visualization of feature heat maps and prediction results during inference. From left to right, the first five columns display feature heat maps, all positioned on the first layer of the model. These include the output of the first layer of the encoder, the output of the BBRT, the output of the AMSA, the output of the DCAM, and the output of the segment head (before the activation function). The last column represents the labels for each sample.

**Table 6**
The segmentation results (mean±std) using different core components in bi-oriented bi-level routing transformer (BBRT) on PCAMM. ↑ means the higher the better and ↓ represents the opposite. The best results are marked in bold text.

| Model | DSC(%)↑ | ABD(mm)↓ | RVD(%)↓ |
|---|---|---|---|
| self-attention | 71.34±0.07 | 3.71±0.43 | 12.96±0.80 |
| Spatial branch | 74.93±0.84 | 3.67±0.16 | 15.44±2.34 |
| Channel branch | 75.00±1.05 | 3.91±0.53 | 11.02±2.89 |
| In series | 72.34±0.19 | 3.92±0.59 | 14.64±3.97 |
| In parallel (Ours) | **76.09±0.23** | **3.63±0.02** | **8.64±1.54** |

**Table 7**
The segmentation results (mean±std) using different core components instead of adaptive multi-spectral attention (AMSA) on PCAMM. ↑ means the higher the better and ↓ represents the opposite. The best results are marked in bold text.

| Model | DSC(%)↑ | ABD(mm)↓ | RVD(%)↓ |
|---|---|---|---|
| SE | 71.30±0.75 | 3.76±0.24 | 15.69±4.27 |
| MSA | 72.98±0.66 | 3.81±0.17 | 17.49±4.91 |
| AMSA (Ours) | **76.09±0.23** | **3.63±0.02** | **8.64±1.54** |

**Table 8**
The results of using different numbers of frequency components on PCAMM.

| Number | 1 | 2 | 4 | 8 | 16 | 32 |
|---|---|---|---|---|---|---|
| DSC(%) | 75.37 | 75.81 | 75.77 | 75.90 | **76.09** | 75.94 |

**Table 9**
The computational complexity and resource consumption of all models.

| Method | Param (M) | FLOPs (G) | GPU Memory | Inference Time (ms) |
|---|---|---|---|---|
| UNet | 17.29 | 40.18 | 20.04 | 10.45 |
| UNet++ | 9.07 | 31.18 | 21.27 | 7.20 |
| Attention U-Net | 8.47 | 13.14 | 9.92 | 5.37 |
| MedT | 1.74 | 2.79 | 17.64 | 149.44 |
| TransUNet | 105.32 | 29.33 | 10.34 | 15.76 |
| Swin-Unet | 41.96 | 8.99 | 16.20 | 15.64 |
| nnUNet | 20.65 | 11.55 | 7.62 | 6.89 |
| Ours | 36.92 | 19.69 | 16.18 | 15.93 |

proposed method has obtained satisfactory results, there are still some limitations, as illustrated in Fig. 8 and Table 2. (1) While our method can reduce false negative regions to some extent, we must acknowledge that inherent limitations remain in segmenting extremely small lesions. The identification of small satellite lesions or elimination of their interference remains challenging in stroke imaging data. (2) The detection rate of low-grade PCa is not high. Improving the detection rate of low-grade PCa can decrease the probability of patients missing the optimal treatment opportunity, enabling doctors to promptly manage the tumor's progression. In the future, we will continue to enhance this work by exploring the incorporation of patient-related text information and interactive information between image slices to further improve network performance.

## 6. Conclusion

In the challenging MRI segmentation task, we conducted research focusing on the fine-grained grading of prostate cancer as a representative example. To better address the challenges of PCa grading, we designed two novel modules, BBRT and AMSA. We utilized them to construct a new segmentation framework for fine-grained grading of PCa. This framework can also be extended to other MRI segmentation tasks. In BBRT, BBRA enhances the network's spatial attention to lesions and its capability for important channel selection. AMSA is used to enhance the network's sensitivity to crucial texture information. The framework also includes the InceptionNeXt encoder and DCAM. All modules collectively learn valuable information from multimodal MRI data and provide reliable segmentation results. Extensive experiments with eight widely used medical image segmentation networks show that our method outperforms in prostate lesion segmentation and demonstrates strong robustness, achieving state-of-the-art performance in prostate and stroke segmentation.

## CRediT authorship contribution statement

**Yatong Liu:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wei Wang:** Writing – review & editing, Validation, Funding acquisition, Data curation. **Yu Zhu:** Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Hangyu Li:** Investigation, Data curation. **Zeyan Zeng:** Validation, Funding acquisition. **Yuhao Zhang:** Resources, Funding acquisition.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Yu Zhu reports financial support was provided by National Natural Science Foundation of China. Yu Zhu reports financial support was provided by Exploratory Device R&D Projects of National Clinical Research Center for Interventional Medicine. Yu Zhu reports financial support was provided by Key Clinical Research Projects of National Clinical Research Center for Interventional Medicine. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62476088, the Exploratory Device R&D Projects of National Clinical Research Center for Interventional Medicine (NO. 2021–002), and the Key Clinical Research Projects of National Clinical Research Center for Interventional Medicine (NO. 2019–003).

## Data availability

The data that has been used is confidential.

## References

[1] D.J. Lee, H.U. Ahmed, C.M. Moore, M. Emberton, B. Ehdaie, Multiparametric magnetic resonance imaging in the management and diagnosis of prostate cancer: current applications and strategies, Curr. Urol. Rep. 15 (2014) 1–10.
[2] C. Mingels, L.I. Loebelenz, A.T. Huber, I. Alberts, A. Rominger, A. Afshar-Oromieh, V.C. Obmann, Literature review: imaging in prostate cancer, Curr. Probl. Cancer (2023) 100968.
[3] J.C. Weinreb, J.O. Barentsz, P.L. Choyke, F. Cornud, M.A. Haider, K.J. Macura, D. Margolis, M.D. Schnall, F. Shtern, C.M. Tempany, et al., Pi-rads prostate imaging–reporting and data system: 2015, version 2, Eur. Urol. 69 (1) (2016) 16–40.
[4] R. Cao, A.M. Bajgiran, S.A. Mirak, S. Shakeri, X. Zhong, D. Enzmann, S. Raman, K. Sung, Joint prostate cancer detection and gleason score prediction in MP-MRI via focalnet, IEEE Trans. Med. Imaging 38 (11) (2019) 2496–2506.
[5] C. Li, H. Sun, Z. Liu, M. Wang, H. Zheng, S. Wang, Learning cross-modal deep representations for multi-modal mr image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, October 13–17, 2019, ii ed., vol. 22, Springer, Shenzhen, China, 2019, pp. 57–65.
[6] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer J. Clinic. 71 (3) (2021) 209–249.
[7] A. Stabile, F. Giganti, A.B. Rosenkrantz, S.S. Taneja, G. Villeirs, I.S. Gill, C. Allen, M. Emberton, C.M. Moore, V. Kasivisvanathan, Multiparametric MRI for prostate cancer diagnosis: current status and future directions, Nat. Rev. Urol. 17 (1) (2020) 41–61.
[8] T. Barrett, A. Rajesh, A. Rosenkrantz, P. Choyke, B. Turkbey, Pi-rads version 2.1: one small step for prostate MRI, Clin. Radiol. 74 (11) (2019) 841–852.
[9] V. Panebianco, F. Giganti, Y.X. Kitzing, F. Cornud, R. Campa, G. De Rubeis, A. Ciardi, C. Catalano, G. Villeirs, An update of pitfalls in prostate MP-MRI: a practical approach through the lens of pi-rads v. 2 Guidelines, Insight. Image. 9 (2018) 87–101.
[10] G.J. Van Leenders, T.H.V.D. Kwast, D.J. Grignon, A.J. Evans, G. Kristiansen, C.F. Kweldam, G. Litjens, J.K. McKenney, J. Melamed, N. Mottet, et al., The 2019 international society of urological pathology (ISUP) consensus conference on grading of prostatic carcinoma, Am. J. Surg. Pathol. 44 (8) (2020) e87–e99.
[11] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention u-net: learning where to look for the pancreas, 2018, arXiv preprint arxiv:1804.03999.
[12] Y. Chen, K. Wang, X. Liao, Channel-unet: a spatial channel-wise convolutional neural network for liver and tumors segmentation, Front. Genet. 10 (2019) 492928.
[13] Y. Liu, Y. Zhu, Y. Xin, Y. Zhang, D. Yang, T. Xu, Mestrans: multi-scale embedding spatial transformer for medical image segmentation, Comput. Methods Programs Biomed. 233 (2023) 107493.
[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[15] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, B. Guo, Cswin transformer: a general vision transformer backbone with cross-shaped windows, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12124–12134.
[16] Z. Xia, X. Pan, S. Song, L.E. Li, G. Huang, Vision transformer with deformable attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4794–4803.
[17] I.C. Duta, L. Liu, F. Zhu, L. Shao, Pyramidal convolution: rethinking convolutional neural networks for visual recognition, 2020, arXiv preprint arxiv:2006.11538.
[18] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, October 5–9, 2015, proceedings, part III 18, Springer, Munich, Germany, 2015, pp. 234–241.
[19] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet + +: redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2019) 1856–1867.
[20] Y. Cai, Y. Wang, Ma-unet: an improved version of unet based on multi-scale and attention mechanism for medical image segmentation, in: Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021), vol. 12167, SPIE, 2022, pp. 205–211.
[21] X. Qin, Y. Zhu, W. Wang, S. Gui, B. Zheng, P. Wang, 3D multi-scale discriminative network with multi-directional edge loss for prostate zonal segmentation in bi-parametric MR images, Neurocomputing 418 (2020) 148–161.
[22] Z. Qin, P. Zhang, F. Wu, X. Li, Fcanet: frequency channel attention networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 783–792.
[23] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[24] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.
[25] X. Li, L. Wang, H. Liu, B. Ma, L. Chu, X. Dong, D. Zeng, T. Che, X. Jiang, W. Wang, et al., Syn_segnet: a joint deep neural network for ultrahigh-field 7 t MRI synthesis and hippocampal subfield segmentation in routine 3 t MRI, IEEE J. Biomed. Health Info. (2023).
[26] W. Li, W. Huang, Y. Zheng, Corrdiff: corrective diffusion model for accurate MRI brain tumor segmentation, IEEE J. Biomed. Health Info. (2024).
[27] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, A.K. Nandi, Medical image segmentation using deep learning: a survey, IET Image Proc. 16 (5) (2022) 1243–1267.
[28] N. Ibtehaz, M.S. Rahman, Multiresunet: rethinking the u-net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87.
[29] H. Seo, C. Huang, M. Bassenne, R. Xiao, L. Xing, Modified U-NET (MU-NET) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images, IEEE Trans. Med. Imaging 39 (5) (2019) 1316–1325.
[30] X. Chen, R. Zhang, P. Yan, Feature fusion encoder decoder network for automatic liver lesion segmentation, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 430–433.
[31] X. Yang, C. Liu, Z. Wang, J. Yang, H.L. Min, L. Wang, K.-T.-T. Cheng, Co-trained convolutional neural networks for automated detection of prostate cancer in multiparametric MRI, Med. Image Anal. 42 (2017) 212–227.
[32] S. Kohl, D. Bonekamp, H.-P. Schlemmer, K. Yaqubi, M. Hohenfellner, B. Hadaschik, J.-P. Radtke, K. Maier-Hein, Adversarial networks for the detection of aggressive prostate cancer, 2017, arXiv preprint arxiv:1702.08014.
[33] R. Alkadi, F. Taher, A. El-Baz, N. Werghi, A deep learning-based approach for the detection and localization of prostate cancer in T2 magnetic resonance images, J. Digit. Imaging 32 (2019) 793–807.
[34] Z. Wang, R. Wu, Y. Xu, Y. Liu, R. Chai, H. Ma, A two-stage cnn method for MRI image segmentation of prostate with lesion, Biomed. Signal Process. Control. 82 (2023) 104610.
[35] G. Zhang, W. Wang, D. Yang, J. Luo, P. He, Y. Wang, Y. Luo, B. Zhao, J. Lu, A bi-attention adversarial network for prostate cancer segmentation, IEEE Access 7 (2019) 131448–131458.
[36] T. Shi, H. Jiang, B. Zheng, C 2 ma-net: cross-modal cross-attention network for acute ischemic stroke lesion segmentation based on ct perfusion scans, IEEE Trans. Biomed. Eng. 69 (1) (2021) 108–118.
[37] M. Baboudjian, A. Breda, P. Rajwa, A. Gallioli, B. Gondran-Tellier, F. Sanguedolce, P. Verri, P. Diana, A. Territo, C. Bastide, et al., Active surveillance for intermediate-risk prostate cancer: a systematic review, meta-analysis, and metaregression, Eur. Urol. Oncol. 5 (6) (2022) 617–627.
[38] A. Duran, P.-M. Jodoin, C. Lartizien, Prostate cancer semantic segmentation by gleason score group in bi-parametric MRI with self attention model on the peripheral zone, in: Medical Imaging with Deep Learning, PMLR, 2020, 193–204.
[39] P. Mehta, M. Antonelli, H.U. Ahmed, M. Emberton, S. Punwani, S. Ourselin, Computer-aided diagnosis of prostate cancer using multiparametric MRI and clinical features: a patient-level classification framework, Med. Image Anal. 73 (2021) 102153.
[40] C. De Vente, P. Vos, M. Hosseinzadeh, J. Pluim, M. Veta, Deep learning regression for prostate cancer detection and grading in bi-parametric MRI, IEEE Trans. Biomed. Eng. 68 (2) (2020) 374–383.
[41] I. Bhattacharya, A. Seetharaman, C. Kunder, W. Shao, L.C. Chen, S.J. Soerensen, J.B. Wang, N.C. Teslovich, R.E. Fan, P. Ghanouni, et al., Selective identification and localization of indolent and aggressive prostate cancers via corrsignia: an MRI-pathology correlation and deep learning framework, Med. Image Anal. 75 (2022) 102288.
[42] W. Yu, P. Zhou, S. Yan, X. Wang, Inceptionnext: when inception meets convnext, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5672–5683.

[43] Y. Liu, Y. Zhu, W. Wang, B. Zheng, X. Qin, P. Wang, Multi-scale discriminative network for prostate cancer lesion segmentation in multiparametric mr images, Med. Phys. 49 (11) (2022) 7001–7015.
[44] L. Zhu, X. Wang, Z. Ke, W. Zhang, R.W. Lau, Biformer: vision transformer with bi-level routing attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10323–10333.
[45] W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, X. He, W. Liu, Crossformer++: a versatile vision transformer hinging on cross-scale attention, IEEE Trans. Pattern Anal. Mach. Intell. 46 (5) (2023) 3123–3136.
[46] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[47] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, H. Huisman, Spie-aapm prostatex challenge data (version 2), [dataset] (2017). https://doi.org/10.7937/K9TCIA.2017.MURS5CL.
[48] M.R.H. Petzsche, E. de la Rosa, U. Hanning, R. Wiest, W. Valenzuela, M. Reyes, M. Meyer, S.-L. Liew, F. Kofler, I. Ezhov, et al., A multi-center magnetic resonance imaging stroke lesion segmentation dataset, Sci. Data 9 (1) (2022) 762.
[49] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.
[50] H. Wang, P. Cao, J. Wang, O.R. Zaiane, Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 2441–2449.
[51] A. Meyer, M. Rakr, D. Schindele, S. Blaschke, M. Schostak, A. Fedorov, C. Hansen, Towards patient-individual pi-rads v2 sector map: Cnn for automatic segmentation of prostatic zones from t2-weighted MRI, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 696–700.
[52] Z. Wu, X. Zhang, F. Li, S. Wang, L. Huang, J. Li, W-net: a boundary-enhanced segmentation network for stroke lesions, Expert Syst. Appl. (2023) 120637.
[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, 2020, arXiv preprint arxiv:2010.11929.
[54] Y. Ding, X. Yu, Y. Yang, Rfnet: region-aware fusion network for incomplete multi-modal brain tumor segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3975–3984.
[55] Y. Ding, L. Li, W. Wang, Y. Yang, Clustering propagation for universal medical image segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 3357–3369.

## Author biography

**Yatong Liu** is a doctoral candidate at the School of Information Science and Engineering, East China University of Science and Technology. She is an AI algorithm engineer specializing in medical image processing, deep learning algorithms, intelligent analysis of multi-source medical images, lesion segmentation and detection, and pattern recognition. She has published in journals at the intersection of medical and computer vision, and has been involved in publicly and privately funded projects.

**Wei Wang** received the M.M. degrees from Tongji University School of Medicine, Shanghai, China. He is working in the Department of Radiology, Tongji Hospital of Tongji University School of Medicine. His current research interests include medical image computing, Early diagnosis of urinary tumors, and machine learning.

**Yu Zhu** received the Ph.D degree from Nanjing University of Science and Technology, China, in 1999. She is currently a professor in the department of electronics and communication engineering of East China University of Science and Technology. Her research interests include image processing, computer vision, multimedia communication and deep learning, especially, for the medical auxiliary diagnosis by artificial intelligence technology. She has published more than 90 papers in journals and conferences.

**Hangyu Li** is a Ph.D. Candidate at the School of Information Science and Engineering, East China University of Science and Technology. His research interests primarily revolve around generative models, neural radiance fields, deep learning algorithms and applications, and computer vision. Within these areas, he focuses on topics such as AIGC, medical image reconstruction using generative models, and novel views synthesis of medical images based on neural radiance fields.

**Zeyan Zeng** received the bachelor's degree and master's degree in 2021 and 2024 from Fudan University, Shanghai, China, where he will pursue the Ph.D degree in Neurology. His research focuses on cerebral hemodynamics and numerical simulation of cerebral vessels.

**Yuhao Zhang** received the Ph.D degree in neurology from Fudan University, Shanghai, China, in 2014. Since 2012, he has been an Associate Professor with the Department of Neurology, Zhongshan Hospital, Fudan University. He is a member of the Chinese Stroke Association. His research focuses on diagnostic techniques of Stroke.