scientific reports

Check for updates

OPEN

A mixed Mamba U-net for prostate segmentation in MR images

Qiu Du, Luowu Wang & Hao Chen ✉

The diagnosis of early prostate cancer depends on the accurate segmentation of prostate regions in magnetic resonance imaging (MRI). However, this segmentation task is challenging due to the particularities of prostate MR images themselves and the limitations of existing methods. To address these issues, we propose a U-shaped encoder-decoder network MM-UNet based on Mamba and CNN for prostate segmentation in MR images. Specifically, we first proposed an adaptive feature fusion module based on channel attention guidance to achieve effective fusion between adjacent hierarchical features and suppress the interference of background noise. Secondly, we propose a global context-aware module based on Mamba, which has strong long-range modeling capabilities and linear complexity, to capture global context information in images. Finally, we propose a multi-scale anisotropic convolution module based on the principle of parallel multi-scale anisotropic convolution blocks and 3D convolution decomposition. Experimental results on two public prostate MR image segmentation datasets demonstrate that the proposed method outperforms competing models in terms of prostate segmentation performance and achieves state-of-the-art performance. In future work, we intend to enhance the model's robustness and extend its applicability to additional medical image segmentation tasks.

Keywords Prostate segmentation, Magnetic resonance imaging, Mamba, Feature fusion, Multi-scale

A statistical analysis conducted by the American Cancer Society revealed that in 2023, prostate cancer (PCa) was the most prevalent male cancer in the United States, with an incidence rate of 29% and a mortality rate of 11%[1]. The early detection of prostate cancer depends on the precise localization of the prostate area, which can be facilitated by the Prostate Imaging Reporting and Data System (PI-RADS), which is based on magnetic resonance imaging (MRI)[2]. Radiologists frequently face the need for manual segmentation of prostate regions in MRI scans. However, the accurate identification and segmentation of anatomical structures entail a laborious and highly specialized process[3]. Persistent inter-observer variability poses a challenge to the precise delineation of regions of interest (ROIs) during manual segmentation[4]. Thus, there is an urgent demand for automated and reliable prostate MR image segmentation. The early segmentation algorithms that relied on threshold methods, edge detection filters, and machine learning techniques were cumbersome, time-consuming, and often unreliable[5]. Convolutional neural networks (CNNs), with their efficient hierarchical feature representation, have become widely employed in medical image segmentation with the rapid growth of deep learning. To solve the issue of positive and negative sample imbalance in segmented samples, Milletari et al.[6] proposed V-Net for prostate MRI segmentation by integrating 3D convolution with the U-Net architecture and introducing a dice coefficient loss function. Rundo et al.[7] proposed USE-Net, a convolutional neural network that incorporates Squeeze-and-Excitation blocks[8] into U-Net, to complete the prostate region segmentation task. The intrinsic locality of convolution processes, however, restricts CNN's capacity to learn long-range relationships. The proposal of Transformer[9] solves the limitations of the CNN model in global representation. It can accurately model long-range interdependence by utilizing the multi-head self-attention (MSA) method[10]. Convolutional coupled Transformer U-Net (CCTUnet), a U-shaped architecture based on a convolutional coupled Transformer, was proposed by Yan et al.[11] for prostate MRI peripheral and transition zone segmentation. The MSA method, however, entails quadratic complexity relative to the image size[12], leading to substantial computational overhead in processing downstream dense prediction tasks, such as medical image segmentation[13]. Thus, there is still no clear way to improve long-range reliance on CNNs.

Recently, Mamba[14] has further improved the Structured State Space Sequence model (S4)[15] through a selection mechanism, and it is considered promising for addressing the lack of long-range dependencies in CNNs. In continuous long sequence data analysis, such as natural language processing and genomic research, Mamba with linear complexity provides state-of-the-art performance, even outperforming Transformer. A new universal

Department of Urology, Hunan Provincial People's Hospital, The First Affiliated Hospital of Hunan Normal University, Changsha 410005, People's Republic of China. ✉email: 18874219779@163.com

visual backbone with bidirectional Mamba blocks (Vim), for instance, was established by Zhu et al.[16]. It uses bidirectional compressed vision to express the state space model and labels image sequences with positional embeddings. To handle objects with heterogeneous appearances, Ma et al.[13] proposed a new architecture called U-Mamba for general medical image segmentation. This architecture is thought to be a strong contender to serve as the foundation of future medical image segmentation networks.

Motivated by the aforementioned research, we intend to employ Mamba blocks to optimize CNNs' long-range modeling capabilities and propose MM-UNet, a U-shaped network for prostate MR image segmentation that combines CNN with Mamba. It is composed of three well-designed modules, skip-connections, and the traditional encoder-decoder structure. Specifically, we represent multi-scale features at the granular level using the 3D Res2Net[17] encoder and use the adaptive feature fusion module (AFFM) and multi-scale anisotropic convolution module (MACM) to replace traditional skip connections. In addition, we propose a global context-aware module (GCM) that is integrated into the network's bottleneck layer at the bottom. Using two publicly available prostate MR imaging datasets, we evaluate our MM-UNet and get state-of-the-art performance in comparison to other competing approaches. Our contributions can be summarized as:

- An AFFM based on the channel attention mechanism is proposed to adaptively recognize and select the most discriminating spatial and semantic information, thereby guiding the effective fusion between adjacent hierarchical features;
- A GCM based on Mamba blocks with linear scaling of feature sizes is proposed to capture the global context information in the image with less computational cost, thus achieving the complementarity of local fine-grained features;
- A MACM based on multi-scale anisotropic convolution and 3D convolution decomposition is proposed. It can achieve robust and accurate prostate segmentation by exploiting 3D contextual information and multi-scale features of MR images with anisotropic resolution while reducing complexity.

Methods

Overview

Figure 1 illustrates the proposed MM-UNet design, which uses an encoder-decoder network design to effectively extract local features and global context from prostate MR images. MM-UNet mainly consists of five components: (1) 3D Res2Net encoder for encoding features of different scales; (2) Adaptive feature fusion module (AFFM) for combining low-level encoder adaptively features with high-level decoder features; (3) MR image long-distance

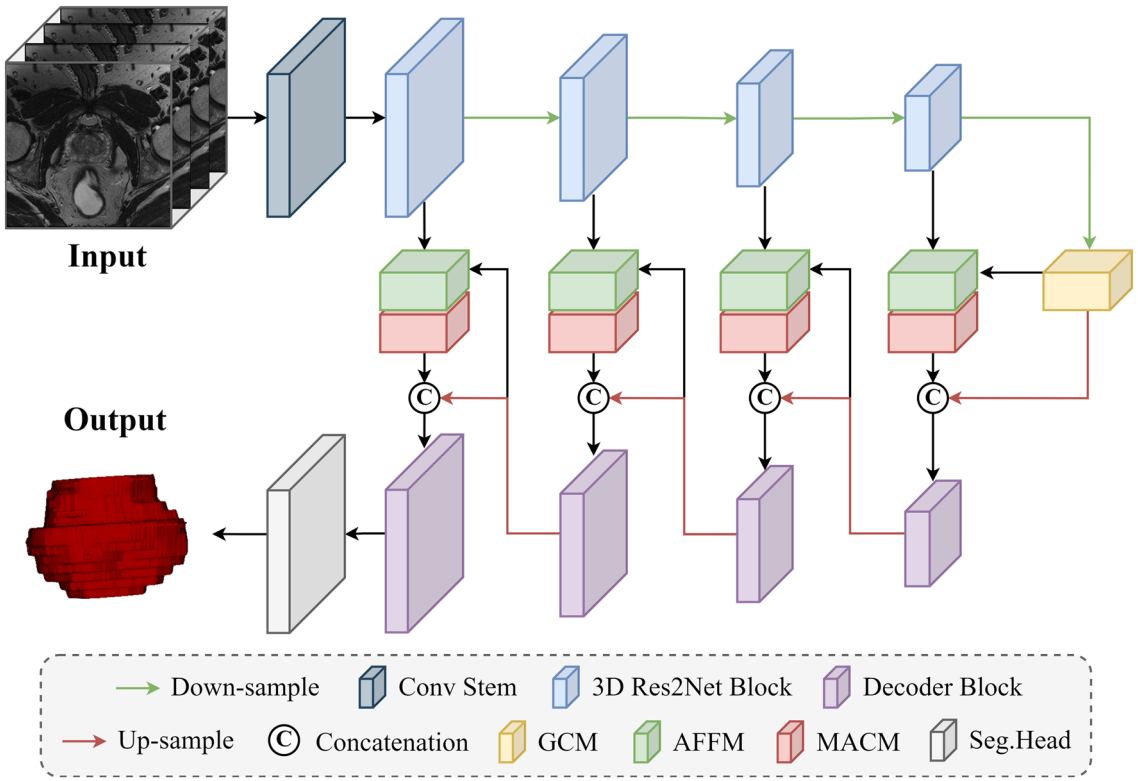

Fig.1. The overall architecture of the proposed MM-UNet. MM-UNet follows the classic encoder-decoder and skip-connections structure and combines three carefully designed novel modules to complete the task of robust prostate MR image segmentation, including adaptive feature fusion module (AFFM), global context-aware module (GCM) and multi-scale anisotropic convolution module (MACM). The encoding module uses the 3D modified Res2Net block to extract multi-scale information, and the decoding module combines MACM with a kernel size of 3 and $1 \times 1 \times 1$ convolution operation to perform layer-by-layer decoding.

modeling is improved by the global context-aware module (GCM), which takes advantage of Mamba's linear scaling properties; (4) Multi-scale anisotropic convolution module (MACM) for capturing multi-scale contextual information and overcoming anisotropic spatial resolution interference; and (5) Anisotropic convolutional layer-based 3D decoder for predicting segmentation results. Next, we'll go into great depth about each component.

3D res2Net encoder

As the encoder backbone, we employ Res2Net[17], which has been pre-trained on the ImageNet dataset[18]. When compared to ResNet[19], Res2Net's hierarchical residual-like connections within each residual block allow it to more effectively use multi-scale information at the granular level, expanding each layer's receptive field. A given image is first passed through a convolution stem to generate rich local information, which consists of parallel $3\times3\times3$ convolutions and $1\times1\times1$ convolutions followed by summation operations[20]. Then enter the four Res2Net residual blocks in sequence to generate features with different spatial resolutions. As Res2Net was first developed for 2D natural image analysis, 3D medical image segmentation requires its extension. To do this, we first replace the 7×7 2D convolution in the first layer with a $7\times7\times3$ 3D convolution, and then the 2D convolution of $x\times x$ is directly transformed into the 3D convolution of $x\times x\times1$ in other layers[21]. It may be specifically said that the $7\times7\times3$ kernel with single-channel 3D input is equivalent to the 7×7 kernel with three-channel 2D input in Res2Net. Thus, our encoder can be initialized using pre-trained Res2Net. When representing prostate MR images, this initialization can help transfer image representation capabilities acquired on large-scale natural image datasets.

Adaptive feature fusion module

Due to the complex anatomical structure and morphology of the prostate, high-level features from the decoder and low-level features from the encoder are crucial for accurate prostate prediction[22]. To this end, we propose an adaptive feature fusion module (AFFM) guided by an attention mechanism. This module can adaptively combine the most effective features between the two to obtain the best representation of the prostate structure. The specific structure is shown in Fig. 2.

First, we upsample the high-level features F_h^{i+1} to the same size as the low-level features $F_l^i(i \in \{1,2,3,4\})$ and fuse the high-level and low-level features through a channel concatenation operation. The specific process is shown in Eq. (1). Secondly, we introduce Squeeze-and-Excitation Block[8] as the channel attention to calculate the weight vector, thereby re-weighting low-level features and suppressing the interference of irrelevant background noise. Specifically, to obtain the global information of each channel F_c^i, we first use the global average pooling operation to stimulate the feature channel. Then the results after learning through two fully connected

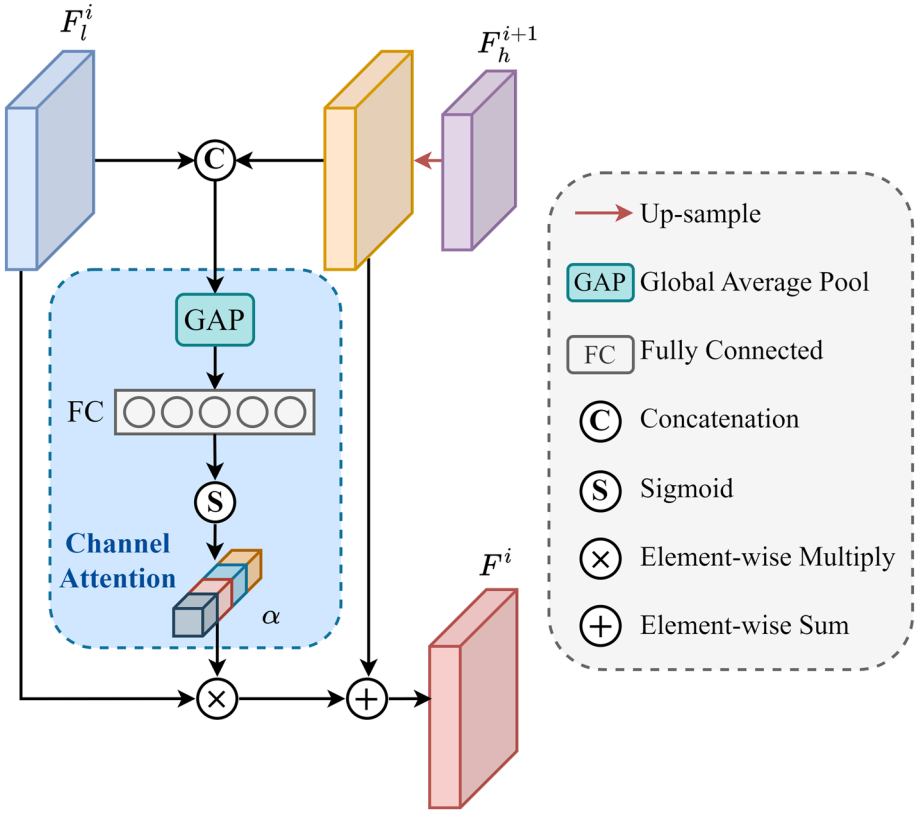

Fig.2. The specific structure of AFFM. It utilizes channel attention to guide adjacent features for adaptive fusion.

layers are input into the Sigmoid function to obtain the channel attention weight α. The above process is shown in Eq. 2. Finally, we sum the low-level features which are weighted by the weight α, and high-level features to get the final result F^i, as shown in Eq. (3). By employing AFFM to gradually guide the fusion between high-level and low-level features, the proposed MM-UNet can suppress irrelevant background noise and retain more semantic information for more precise localization.

$$F_c^i = \mathbb{C}\left(F_l^i, f_u\left(F_h^{i+1}\right)\right) \tag{1}$$

$$\alpha = f_\sigma\left\{f_{FC}^2\left(\max\left(0, f_{FC}^1\left(f_{gap}\left(F_c^i\right)\right)\right)\right)\right\} \tag{2}$$

$$F^i = \alpha \otimes F_l^i + F_h^{i+1} \tag{3}$$

Among them, $\mathbb{C}$ represents the concatenation operation, f_u represents the upsampling operation, f_σ represents the Sigmoid function, $f_{FC}^i (i \in \{1, 2\})$ represents the fully connected layer, f_{gap} represents the global average pool, and $\otimes$ represents element-wise multiplication.

Global context-aware module

Inspired by previous studies[13], we propose a global context-aware module (GCM) to achieve global context understanding in medical image segmentation. It can leverage the linear scaling property of Mamba to enhance the long-range dependencies of CNN.

As shown in Fig. 3, GCM consists of two consecutive Residual blocks[19] and a Mamba block, where the Residual block consists of a normal convolutional layer, instance normalization (IN), and leaky ReLU activation[13]. Subsequently, the image features with size $B \times C \times H \times W \times D$ are flattened and transposed to $B \times L \times C$, where $L = H \times W \times D$, B represents the batch size, C represents the number of channels, H and W represent the height and width of the feature map, and D represents the depth of the feature map.

After layer normalization, the features enter the Mamba block which contains two parallel branches for enhanced long-range dependency modeling. In the first branch, the features are expanded to $B \times 2L \times C$ through a linear layer, followed by a 1D convolutional layer, a HardSwish activation function[23], and an SSM layer. In the second branch, the features are also expanded to $B \times 2L \times C$ through a linear layer, followed by a HardSwish activation function. Then, the features of the two branches are then merged using the Hadamard product, projected back to the original shape, resized, and transposed to a feature size of $B \times C \times H \times W \times D$. GCM can effectively

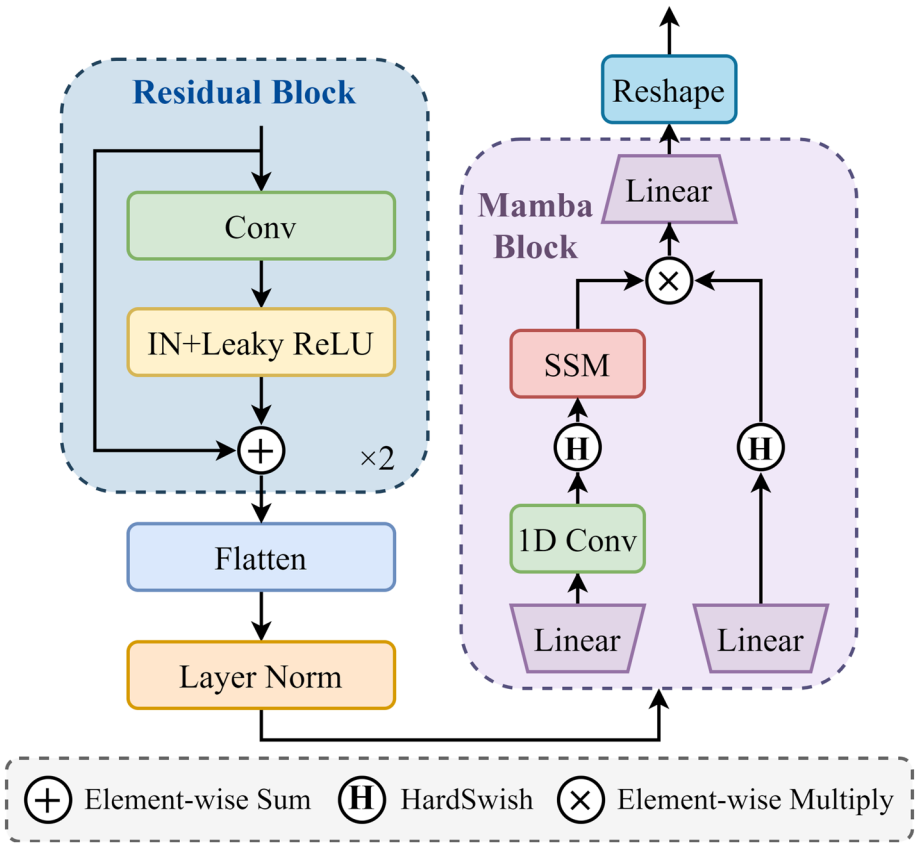

Fig.3. The specific structure of GCM. IN is the abbreviation for instance normalization.

improve the global context capture capability in prostate MR image segmentation while maintaining high efficiency during training and inference.

Multi-scale anisotropic convolution module

We propose a multi-scale anisotropic convolution module (MACM) with large convolution kernels to improve the classification and localization capabilities of the network[24] and then combine it with AFFM as a skip connection at each scale. The specific structure of MACM is shown in Fig. 4a.

The MACM consists of four anisotropic convolutional blocks (ACB) with a kernel size of k ($k \in \{3, 7, 13, 15\}$) constructed in parallel. It can not only enhance the localization performance and voxel classification ability of the network but also fuse the features of different regions and reduce the loss of contextual information at different image scales[25]. Specifically, $1 \times 1 \times 1$ convolution is first applied to the input and output to adjust the number of channels of the features, thereby avoiding an uncontrollable increase in computational complexity. Secondly, the inputs are passed to the ACBs of different branches respectively, and the results of each branch are connected with the inputs as the output.

In the ACB branch, the input is passed to the anisotropic convolutions of $1 \times 1 \times k$ and $k \times k \times 1$ respectively, and then the results are added as the output of this branch. With this design, the $k \times k \times 1$ convolution further exploits the 2D features in the x–y plane, while the $1 \times 1 \times k$ convolution can focus on inter-slice features. At the same time, we apply the 2D convolution decomposition principle[26] to 3D convolution to avoid the possibility that large 3D convolution will increase the computational complexity. As shown in Fig. 4(b), the large kernel 3D convolution $x \times y \times z$ is decomposed into the following combinations: $x \times 1 \times 1 + 1 \times 1 \times z + 1 \times y \times 1$ and $1 \times y \times 1 + 1 \times 1 \times z + x \times 1 \times 1$. After decomposition, the number of parameters can be reduced from o^3 to $3o$.

Results

Dataset

Using two publicly accessible prostate MR image segmentation datasets, we evaluate the efficacy and performance of MM-UNet. The International Medical Image Processing Organizing Committee held the prostate segmentation competition in 2012, using the PROMISE12 dataset[27]. It comprises individuals with benign conditions and prostate cancer, and it originates from four different medical institutes. Of the 80 prostate T2-weighted axial MR images that are accessible, 50 have professional segmentation masks. 60 T2-weighted MR images, all of which are from patients with prostate cancer, make up the ASPS13 dataset[28] that was utilized in the NCI-ISBI 2013 Automated Segmentation of Prostate Structures competition. The doctor marks the actual prostate boundaries in each training case. There are 10 prostate MR images in the test set; the ground truth of these images is not provided.

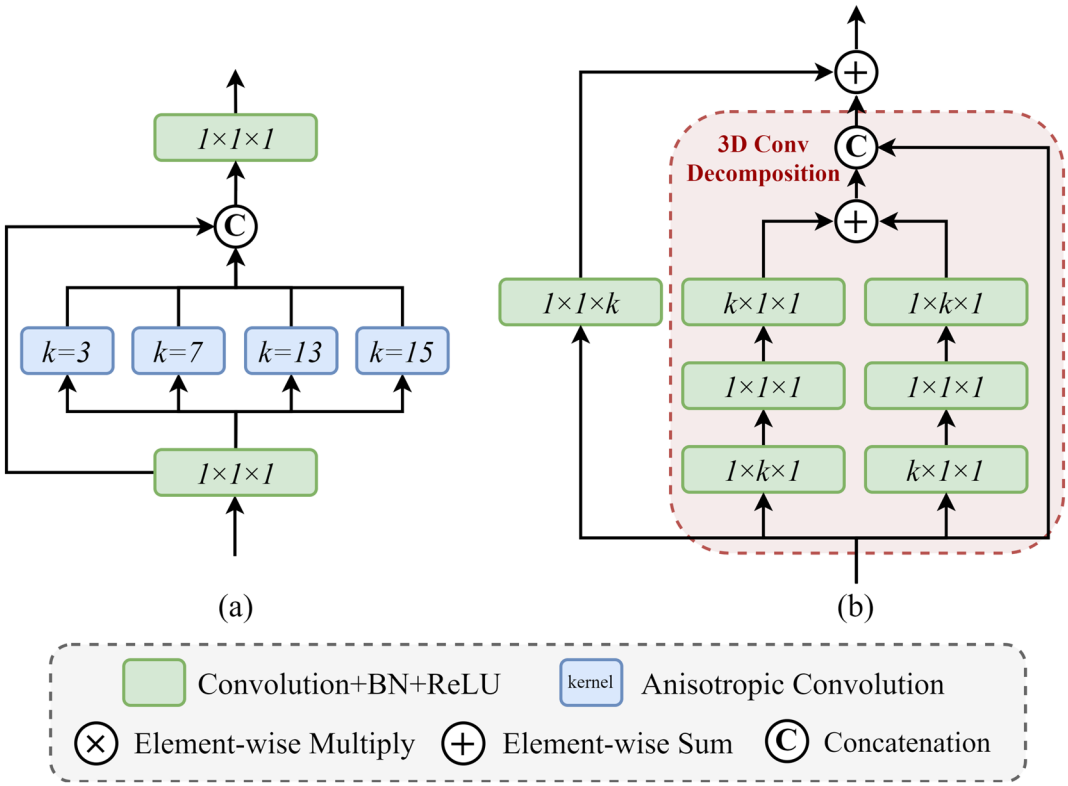

Fig.4. The specific structure of MACM. (**a**) MACM consists of four ACBs with different convolution kernel sizes and constructed in parallel. (**b**) The specific structure of the 3D convolution decomposition principle. k represents the size of the convolution kernel.

Implementation details and evaluation metrics

Python and the PyTorch framework are used in the construction of the proposed model, and experiments are conducted on hardware equipped with four NVIDIA 4090 GPUs. We used a variety of online data augmentation techniques, including random Gaussian noise addition, random flipping, and random rotation, to expand the training dataset after normalizing prostate MR images[21]. During training, a random crop size of $128 \times 128 \times 128$ is used, and the loss function used is called cross-entropy loss. For all experiments, Adam[29] was employed as the optimizer, and a poly learning approach was chosen, with a weight decay of $5 \times 10{-}^{4}$ and an initial learning rate of $1 \times 10{-}^{4}$. We employ a five-fold cross-validation approach for training and testing to acquire a fair and dependable performance of various methods[30], and we apply the same data preprocessing and learning strategy for all experiments to produce a fair comparison[31].

To evaluate prostate segmentation results scientifically and measure model effectiveness, we utilize three performance metrics[32]: Dice Similarity Coefficient (DSC), 95% Hausdorff Distance (95HD), and Average Symmetric Distance (ASD). The better the segmentation accuracy, the higher the DSC value and the lower the 95HD and ASD values. These are determined using the following equations:

$$DSC(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{4}$$

$$95HD = \max_{k95\%} \left(d_{HD}(X, Y), d_{HD}(Y, X)\right) \tag{5}$$

$$d(x, A) = \min_{y \in A} d(x, y) \tag{6}$$

$$ASD = \frac{\sum_{x \in B_X} d(x, B_Y) + \sum_{y \in B_Y} d\left(y, B_X\right)}{|B_X| + |B_Y|} \tag{7}$$

where X represents the prediction result, Y represents the ground truth value, $|\cdot|$ represents the cardinality computation operation, $d_{HD}(X, Y)$ represents the Hausdorff distance between X and Y, $\max_{k95\%}$ represents the maximum value at the 95th percentile, B_X, B_Y represents the boundaries of X and Y, and $d(x, A)$ is the Euclidean distance of the voxels that make up the image's actual spatial resolution.

Comparisons with the state-of-the-art methods

We evaluate our method against nine state-of-the-art methods, which include classic general segmentation networks (Attention U-Net[33], V-Net[6], and 3D U-Net[34]), Transformer-based medical segmentation models (TransUNet[35], SwinUNETR[36] and UNesT[37]), and methods specifically designed for prostate segmentation (MSD-Net[21], CCT-Unet[11], and CAT-Net[38]). When comparing our evaluation metrics to other approaches, we display both the standard deviation and average performance in the form of the mean ± standard deviation.

Table 1 quantitatively demonstrates the segmentation performance of MM-UNet and other comparison methods on the PROMISE12 and ASPS13 datasets. Our method works better than other techniques on two public prostate MR imaging datasets, according to experimental results. In particular, our method achieved an average DSC of 92.39%, 95HD of 3.43 mm, and ASD of 1.42 mm in experiments conducted on the PROMISE12 dataset. Our method gets the best indicator values out of all the methods. The efficacy of the proposed MM-UNet is demonstrated by the improvements in DSC, 95HD, and ASD compared to the traditional V-Net, which are 3.35%, 0.78 mm, and 0.61 mm, respectively. Among Transformer-based medical segmentation models and methods specifically designed for prostate segmentation, UNesT and MSD-Net are the techniques that perform the best, respectively, while our method is 1.42 and 0.51% higher than these two methods in DSC scores, respectively. In

	PROMISE12			ASPS13		
Methods	**DSC (%)**	**95HD (mm)**	**ASD (mm)**	**DSC (%)**	**95HD (mm)**	**ASD (mm)**
Attention U-Net	87.53 ± 5.02	4.94 ± 1.28	2.69 ± 0.73	86.92 ± 5.38	5.16 ± 1.03	2.83 ± 0.64
V-Net	89.04 ± 3.99	4.21 ± 0.63	2.03 ± 0.37	87.10 ± 2.98	4.85 ± 0.73	2.98 ± 0.59
3D U-Net	89.90 ± 4.19	4.56 ± 1.13	1.93 ± 0.58	89.32 ± 3.55	4.71 ± 1.11	2.48 ± 0.69
TransUNet	89.98 ± 5.57	4.07 ± 1.15	1.95 ± 0.74	89.78 ± 2.67	4.58 ± 0.68	2.30 ± 0.29
SwinUNETR	90.42 ± 3.75	3.80 ± 0.84	1.84 ± 0.67	90.22 ± 2.75	4.37 ± 0.78	2.16 ± 0.52
CAT-Net	90.74 ± 3.19	3.87 ± 0.79	1.74 ± 0.35	90.74 ± 2.93	4.03 ± 0.98	1.99 ± 0.47
UNesT	90.97 ± 1.66	3.77 ± 0.90	1.87 ± 0.37	91.05 ± 2.58	3.98 ± 0.64	2.03 ± 0.28
CCT-Unet	91.46 ± 2.11	3.63 ± 1.00	1.82 ± 0.61	91.23 ± 3.56	3.89 ± 0.84	1.81 ± 0.49
MSD-Net	91.88 ± 3.57	3.56 ± 0.64	1.70 ± 0.50	91.54 ± 3.76	3.77 ± 0.59	1.91 ± 0.54
MM-UNet (ours)	**92.39 ± 2.38**	**3.43 ± 0.28**	**1.42 ± 0.16**	**92.17 ± 2.19**	**3.61 ± 0.53**	**1.67 ± 0.32**

Table 1. Quantitative comparison of our method and others on the PROMISE12 and ASPS13 datasets. The best results are in bold.

terms of DSC, 95HD, and ASD, the proposed method generally demonstrated improved reliability for prostate MR image segmentation. Similarly, in experiments on the ASPS13 data set, our method achieved the best performance, with values of 92.17%, 3.61 mm, and 1.67 mm for DSC, 95HD, and ASD, respectively. Compared with CCT-Unet, which ranks second on ASD, DSC, and 95HD, it has improved by 0.94% and 0.28 mm, respectively.

Figure 5 and Fig. 6 display the qualitative evaluation results of our method and those of the competitors in common cases. In the slices of different situations, it can be observed that our method has fewer incorrect segmentations, and the segmentation boundaries are closest to the true situation. The careful model structure design of the proposed MM-UNet allows for considerable and consistent performance benefits on prostate MR images, as demonstrated by quantitative and qualitative analysis. These outcomes show the effectiveness of MM-UNet in solving the varying and complex semantics of prostate regions in this challenging task.

Ablation study

In our proposed network, we conduct an ablation study on the PROMISE12 dataset to evaluate the performance of different methods loaded with various modules, therefore demonstrating the efficacy of diverse components. As a baseline, we employ the 3D U-Net[34] architecture. To lower the computational complexity, we use a 4-layer structure in the encoder as opposed to the 5-layer structure seen in the typical U-Net design. To ensure a fair comparison, all competitors in our ablation experiments performed on the same computing environment with the same data augmentation.

We conduct step-by-step ablation experiments by substituting our presented 3D Res2Net encoder, global context-aware module (GCM), adaptive feature fusion module (AFFM), and multi-scale anisotropic convolution module (MACM) for the corresponding modules in the baseline structure. As can be seen from the quantitative experimental results in Table 2, compared with the baseline, the 95HD of model 1 is reduced by 0.11mm, proving the advantages of the 3D Res2Net encoder. The DSC of Model 2, Model 3, and Model 4 improved by 0.81, 0.98, and 0.57%, respectively, in comparison to Model 1, demonstrating the efficacy of GCM, AFFM, and MACM. The DSC of MM-UNet is enhanced by 0.66, 1.19, and 0.94% in comparison to Models 5, Model 6, and Model 7, respectively. This indicates that utilizing GCM, AFFM, and MACM together can further enhance the model's performance. For DSC and 95HD, MM-UNet gets significant improvements of 2.49% and 1.13mm,

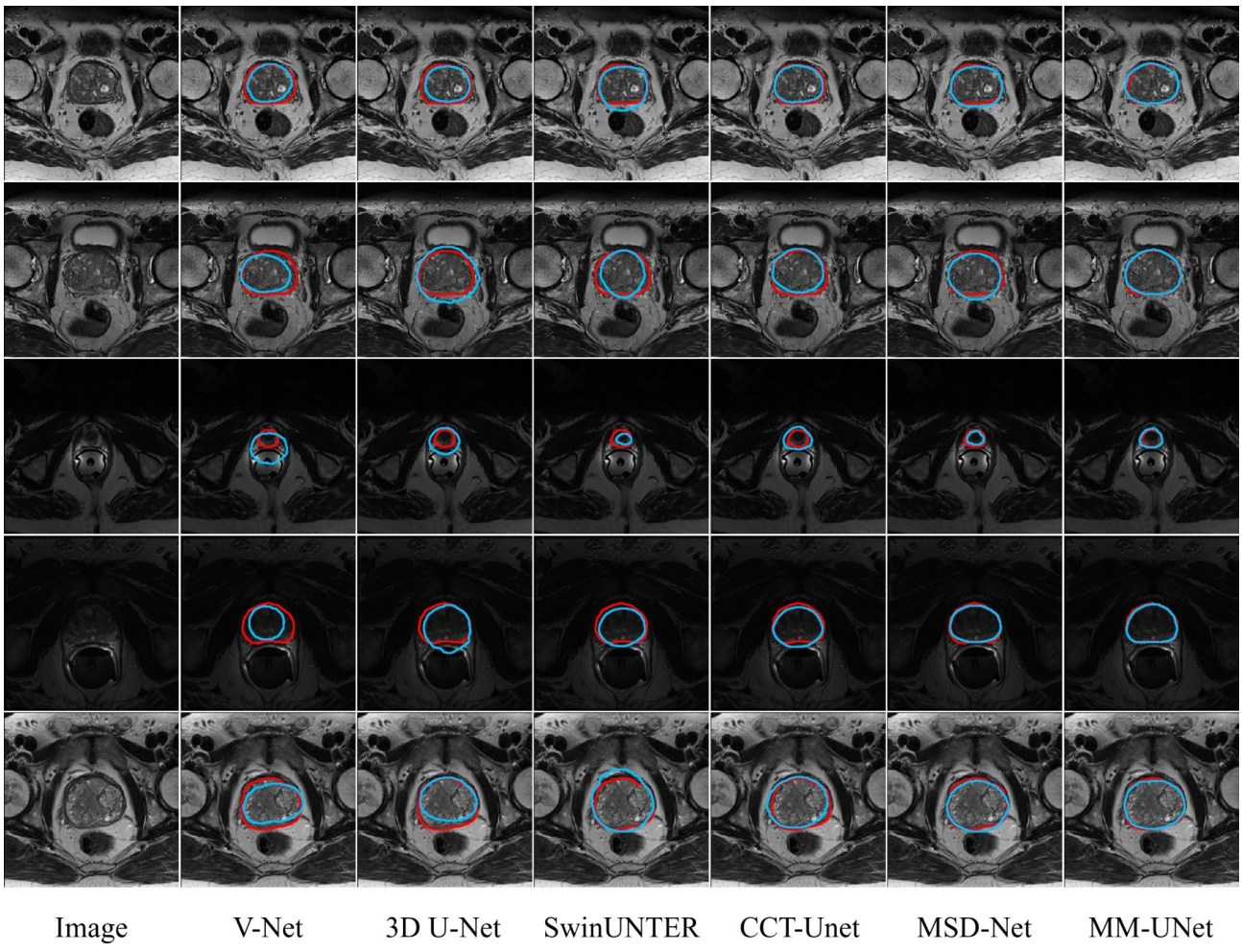

Fig.5. Qualitative comparison of our method and others on the PROMISE12 dataset. The red line represents the real situation, and the blue line represents the segmentation results of various models.

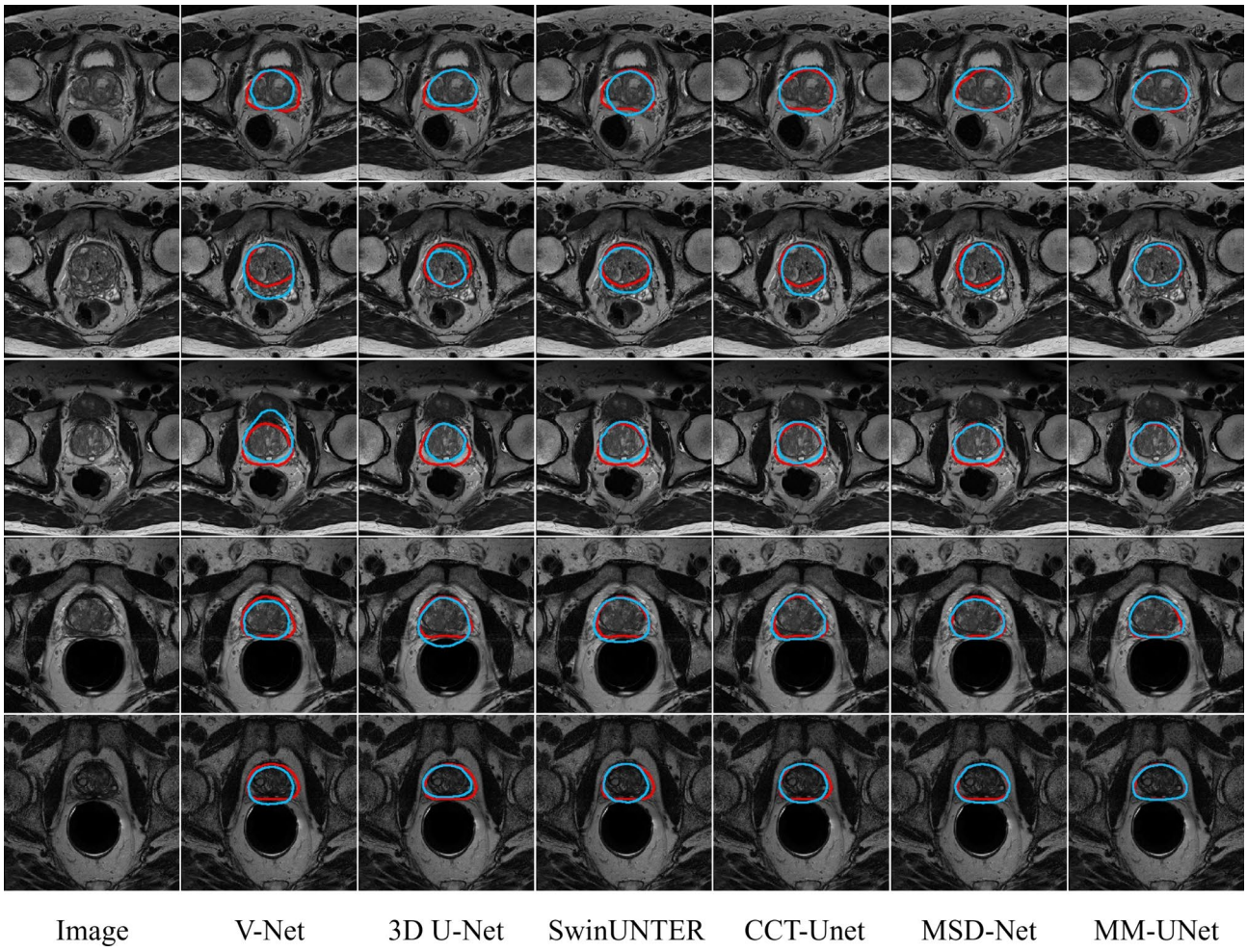

Fig.6. Qualitative comparison of our method and others on the ASPS13 dataset. The red line represents the real situation, and the blue line represents the segmentation results of various models.

Network	3D Res2Net	GCM	AFFM	MACM	DSC (%)	95HD (mm)
Baseline					89.90 ± 4.19	4.56 ± 1.13
Model 1	√				89.98 ± 3.21	4.45 ± 1.09
Model 2	√	√			90.71 ± 3.59	4.07 ± 1.15
Model 3	√		√		90.88 ± 3.74	3.99 ± 0.61
Model 4	√			√	90.47 ± 3.75	4.25 ± 0.93
Model 5	√	√	√		91.73 ± 2.57	3.61 ± 0.72
Model 6	√	√		√	91.20 ± 2.73	3.84 ± 0.92
Model 7	√		√	√	91.45 ± 3.79	3.78 ± 0.52
MM-UNet	√	√	√	√	**92.39 ± 2.38**	**3.43 ± 0.28**

Table 2. Prostate segmentation performances of different models in our system.. The best results are in bold.

respectively, over the baseline, demonstrating the superior segmentation performance of the proposed network. Figure 7 qualitatively demonstrates the advantages of the proposed prostate MR image segmentation module. By comparing the segmentation results of Model 1 and Model 2, we can see that after the introduction of GCM, the network has a stronger ability to identify the prostate area, which proves that GCM can fully capture global context information to optimize segmentation edges. Model 5 shows less-segmentation and under-segmentation as compared to Model 2, indicating that AFFM can adaptively fuse low-level and high-level information to improve the model's segmentation capabilities. MM-UNet integrates GCM, AFFM, and MACM at the same time, and more comprehensive and smooth segmentation results are produced. These results prove that the synergistic effect between modules can well improve the segmentation results of the prostate edge.

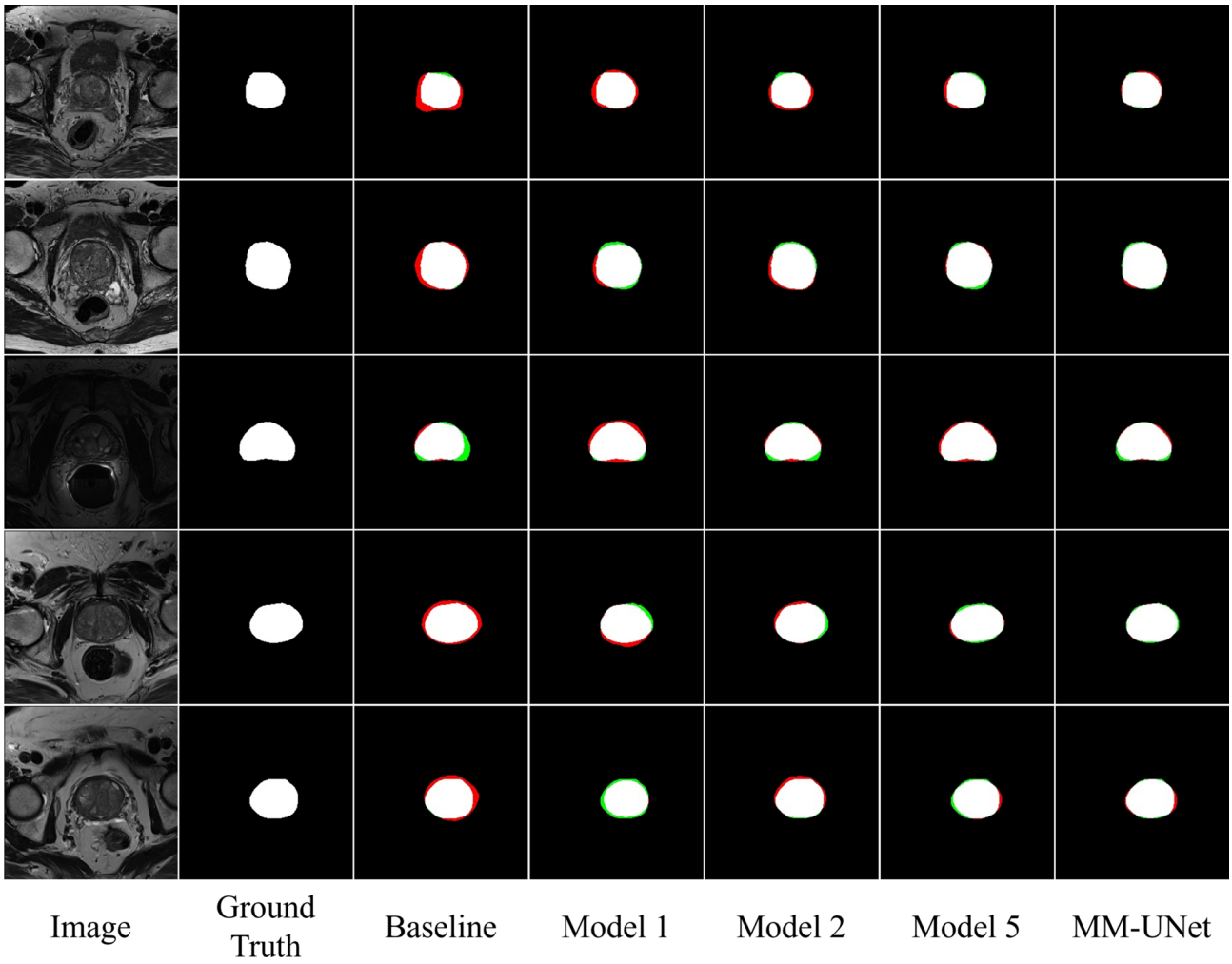

Fig.7. Visual comparison between different models in our system for prostate segmentation. The colors white, green, and red represent the correct segmentation, the under-segmentation, and the over-segmentation, respectively.

Further, we conducted ablation experiments on GCM and MACM to explore the best performance combination. The ablation study results for GCM are displayed in Table 3. It can be seen that the use of Residual Block and HardSwish activation contributed to the final performance. At the same time, we also discussed the selection of multi-scale convolution kernels and the necessity of anisotropic convolution in MACM, and Table 4 presents the results of the study. It can be seen that by combining small kernel convolution and large kernel convolution, multi-scale feature information can be well captured so that scale changes between different image instances can be processed. At the same time, compared with using ordinary 3D convolution, using anisotropic convolution has improved the image quality by 0.12 mm on the 95HD. This indicates that the 3D context information of MR images with anisotropic resolution can be more effectively used by anisotropic convolution.

Residual block	Activation function	DSC (%)	95HD (mm)
	SiLU	92.09 ± 2.61	3.54 ± 0.36
√	SiLU	92.21 ± 2.44	3.47 ± 0.40
√	Swish	92.34 ± 1.12	3.45 ± 0.47
√	HardSwish	**92.39 ± 2.38**	**3.43 ± 0.28**

Table 3. Ablation study on global context-aware module. The best results are in bold.

Kernel Size						
$k=3$	$k=7$	$k=13$	$k=15$	Anisotropic Conv	DSC (%)	95HD (mm)
√				√	91.93 ± 2.58	3.57 ± 0.50
√	√			√	92.04 ± 2.33	3.53 ± 0.67
√	√	√		√	92.20 ± 1.82	3.51 ± 0.57
√	√	√	√	√	**92.39 ± 2.38**	**3.43 ± 0.28**
√	√	√	√		92.25 ± 2.32	3.55 ± 0.47

Table 4. Ablation study on multi-scale anisotropic convolution module. The best results are in bold. Anisotropic Conv represents whether to use anisotropic convolution.

Discussion

Prostate MR image segmentation has been facilitated by the advent of deep learning and other convolutional neural network-based techniques in recent years[6]. The fixed receptive fields increase CNN's computational efficiency, but they also make it less capable of capturing long-range relationships[39]. In contrast to convolutional neural networks, the Transformer uses self-attention to provide each image patch with a global context that is dependent on input or patches. However, 3D medical images often have high resolution, which may cause a significant computational burden for Transformer-based methods[40]. Recently, the structured state space sequence model Mamba has been demonstrated to be an efficient and effective tool for modeling long sequences. This model scales linearly, or nearly linearly, with sequence length.

Motivated by the previous studies, we propose in this study a U-shaped encoder-decoder network for prostate segmentation in MRI called MM-UNet, which is based on Mamba and U-Net. The three main contributions of MM-UNet are the introduction of three novel modules, namely AFFM, GCM, and MACM. The quantitative evaluation results (Table 1) and qualitative evaluation results (Fig. 5 and Fig. 6) on two public prostate MR image segmentation datasets clearly illustrate MM-UNet's efficacy and robustness, which is better than other SOTA methods.

Although our model achieves impressive results, it still has some limitations. First, since manual annotation is a costly and time-consuming undertaking, this results in a limited amount of available MR image training data[25], limiting the peak performance of the model. Secondly, our model is specifically designed for prostate MR image segmentation tasks, and it is uncertain how well it performs on medical image segmentation tasks involving more modalities and organs. To improve model robustness and mitigate overfitting risks, our future research will focus on leveraging large-scale unlabeled medical images through self-supervised learning. Additionally, we'll compare more segmentation backbones and explore additional medical image segmentation tasks from various modalities and targets.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 7 March 2024; Accepted: 23 August 2024
Published online: 28 August 2024

References

1. Siegel, R. L. *et al.* Cancer statistics, 2023. *Ca Cancer J. Clin.* **73**, 17–48 (2023).
2. Fehr, D. *et al.* Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. *Proc. Natl. Acad. Sci.* **112**, E6265–E6273 (2015).
3. Yi, Z. *et al.* Computer-aided diagnosis of prostate cancer based on deep neural networks from multi-parametric magnetic resonance imaging. *Front. Physiol.* **13**, 918381 (2022).
4. Steenbergen, P. *et al.* Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation. *Radiother. Oncol.* **115**, 186–190 (2015).
5. Gillespie, D. *et al.* Deep learning in magnetic resonance prostate segmentation: A review and a new perspective. *ArXiv Prepr. ArXiv201107795* https://doi.org/10.48550/arXiv.2011.07795 (2020).
6. Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. in *2016 fourth international conference on 3D vision (3DV)* 565–571 (Ieee, 2016).
7. Rundo, L. *et al.* USE-Net: Incorporating squeeze-and-excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing* **365**, 31–43 (2019).
8. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 7132–7141 (2018).
9. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv Prepr. ArXiv201011929* https://doi.org/10.48550/arXiv.2010.11929 (2020).
10. Ding, Y. *et al.* FI-Net: Rethinking Feature Interactions for Medical Image Segmentation. *Adv. Intell. Syst.* https://doi.org/10.1002/aisy.202400201 (2024).
11. Yan, Y., Liu, R., Chen, H., Zhang, L. & Zhang, Q. CCT-Unet: A U-shaped network based on convolution coupled transformer for segmentation of peripheral and transition zones in prostate MRI. *IEEE J. Biomed. Health Inform.* **27**, 4341–4351 (2023).
12. Ding, Y. *et al.* CTH-Net: A CNN and transformer hybrid network for skin lesion segmentation. *Iscience* https://doi.org/10.1016/j.isci.2024.109442 (2024).
13. Ma, J., Li, F. & Wang, B. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *ArXiv Prepr. ArXiv240104722* https://doi.org/10.48550/arXiv.2401.04722 (2024).

14. Gu, A. & Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *ArXiv Prepr. ArXiv231200752* https://doi.org/10.48550/arXiv.2312.00752 (2023).
15. Gu, A. *et al.* Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Adv. Neural Inf. Process. Syst.* **34**, 572–585 (2021).
16. Zhu, L. *et al.* Vision mamba: Efficient visual representation learning with bidirectional state space model. *ArXiv Prepr ArXiv240109417* https://doi.org/10.48550/arXiv.2401.09417 (2024).
17. Gao, S.-H. *et al.* Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 652–662 (2019).
18. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. in *2009 IEEE conference on computer vision and pattern recognition* 248–255 (Ieee, 2009).
19. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).
20. Huang, Z. *et al.* STU-Net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *ArXiv Prepr. ArXiv230406716* https://doi.org/10.48550/arXiv.2304.06716 (2023).
21. Jia, H., Cai, W., Huang, H. & Xia, Y. Learning multi-scale synergic discriminative features for prostate image segmentation. *Pattern Recognit.* **126**, 108556 (2022).
22. Dong, C. *et al.* A novel multi-attention, multi-scale 3D deep network for coronary artery segmentation. *Med. Image Anal.* **85**, 102745 (2023).
23. Howard, A. et al. Searching for mobilenetv3. in Proceedings of the IEEE/CVF international conference on computer vision 1314–1324 (2019).
24. Liu, S. et al. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. in Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11 851–858 (Springer, 2018).
25. Jia, H. *et al.* 3D APA-Net: 3D adversarial pyramid anisotropic convolutional network for prostate segmentation in MR images. *IEEE Trans. Med. Imaging* **39**, 447–457 (2019).
26. Peng, C., Zhang, X., Yu, G., Luo, G. & Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 4353–4361 (2017).
27. Litjens, G. *et al.* Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Med. Image Anal.* **18**, 359–373 (2014).
28. Bloch, N. *et al.* NCI-ISBI 2013 challenge: Automated segmentation of prostate structures. *Cancer Imaging Arch.* **370**, 6 (2015).
29. Kinga, D., Adam, J. B., & others. A method for stochastic optimization. in *International conference on learning representations (ICLR)* vol. 5 6 (San Diego, California;, 2015).
30. Whybra, P. *et al.* The image biomarker standardization initiative: Standardized convolutional filters for reproducible radiomics and enhanced clinical insights. *Radiology* **310**, e231319 (2024).
31. Ushinsky, A. *et al.* A 3D–2D hybrid U-net convolutional neural network approach to prostate organ segmentation of multiparametric MRI. *Am. J. Roentgenol.* **216**, 111–116 (2021).
32. Yeghiazaryan, V. & Voiculescu, I. An overview of current evaluation methods used in medical image segmentation. *Dep. Comput. Sci. Univ. Oxf.* 2015 (2015).
33. Oktay, O. *et al.* Attention u-net: Learning where to look for the pancreas. *ArXiv Prepr. ArXiv180403999* https://doi.org/10.48550/arXiv.1804.03999 (2018).
34. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19* 424–432 (Springer, 2016).
35. Chen, J. *et al.* Transunet: Transformers make strong encoders for medical image segmentation. *ArXiv Prepr. ArXiv210204306* https://doi.org/10.48550/arXiv.2102.04306 (2021).
36. Hatamizadeh, A. *et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images International MICCAI Brainlesion Workshop* (Springer, 2021).
37. Yu, X. *et al.* Unest: Local spatial representation learning with hierarchical transformer for efficient medical segmentation. *Med. Image Anal.* **90**, 102939 (2023).
38. Hung, A. L. Y. *et al.* CAT-Net: A cross-slice attention transformer model for prostate zonal segmentation in MRI. *IEEE Trans. Med. Imaging* **42**, 291–303 (2022).
39. Ding, Y. *et al.* HI-MViT: A lightweight model for explainable skin disease classification based on modified MobileViT. *Digit. Health* **9**, 20552076231207196 (2023).
40. Liu, J. *et al.* Swin-UMamba: Mamba-based unet with imagenet-based pretraining. *ArXiv Prepr. ArXiv240203302* https://doi.org/10.4855/arXiv.2402.03302 (2024).

Acknowledgements

We thank Belaydi Othmane for helpful suggestions on the grammar, style and syntax of the manuscript.

Author contributions

QD and HC conceived and supervised the study. LWW contributed to data collection and assembly. QD performed data analysis and interpretation. QD and LWW performed software, visualization and validation. All authors contributed to writing the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.C.