IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 November 2023, accepted 29 December 2023, date of publication 3 January 2024,
date of current version 11 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3349409

RESEARCH ARTICLE

FCTformer: Fusing Convolutional Operations and Transformer for 3D Rectal Tumor Segmentation in MR Images

ZHENGUO SANG[1], CHENGKANG LI[1], YE XU[2], YUANYUAN WANG[1], (Senior Member, IEEE), HONGTU ZHENG[2], AND YI GUO[1]

[1]School of Information Science and Technology, Fudan University, Shanghai 200433, China
[2]Shanghai Cancer Center, Fudan University, Shanghai 200032, China

Corresponding authors: Hongtu Zheng (0356153@fudan.edu.cn) and Yi Guo (guoyi@fudan.edu.cn)

This work was supported in part by the Shanghai Science and Technology Development Funds under Grant 20DZ1100101; in part by the School of Information Science and Technology, Fudan University, China; in part by the Shanghai Cancer Center, Fudan University; and in part by Fudan University.

ABSTRACT Accurate and reliable segmentation of rectal cancer in Magnetic Resonance Imaging holds crucial significance in preoperative prediction, tumor staging, and neoadjuvant therapy. Currently, the automated segmentation methods for rectal tumor have predominantly relied on Convolutional Neural Networks (CNNs), which lean heavily on discerning the contrast disparities among locally neighboring MRI voxels. However, these methods tend to exhibit segmentation inaccuracies when confronted with rectal cancer instances characterized by indistinct contrasts and markedly diverse shapes. Here, we propose a FCTformer who Fuses Convolutional operations and Transformer modules for accurate rectal tumor segmentation in 3D MRI. Specifically, first, FCTformer integrates a transformer-based global feature extraction mechanism and a CNN-based local feature extraction approach to obtain a dual-faceted multiscale feature representation. This representation enhances the model's capability to capture both the comprehensive semantic features and intricate details of rectal cancer instances, especially in challenging situations such as low-contrast imaging and substantial shape variations. Second, to capitalize on features captured across different scales, thereby enhancing segmentation accuracy, we have incorporated a Dual-Attention decoder. Third, to enhance the tumor's edges and contours, the Prediction Aggregation Unit is designed to capture sharper tumor boundaries and retain fine details that could be lost during repetitive down-sampling stages. Experimental results involving 362 instances of rectal tumor segmentation demonstrate that our proposed method achieves a Dice Similarity Coefficient of 0.827, surpassing existing methods. The satisfactory results obtained from evaluating our approach on a publicly available prostate dataset validate its generalizability.

INDEX TERMS Rectal tumor segmentation, T2W-MRI, swin transformer, attention, 3D CNN.

I. INTRODUCTION

Colorectal cancer is a highly malignant neoplasm that affects the digestive system, with a worldwide incidence and mortality ranking of 3rd and 2nd, respectively [1]. More than half of colorectal cancer cases occur in the rectal region [2], which is characterized by complex anatomical relationships that make complete surgical excision challenging. There is a rapid increase in the number of rectal cancer cases, with over 30% diagnosed as locally advanced rectal cancer upon confirmation. Research on rectal cancer holds profound significance [3].

Magnetic resonance imaging (MRI) offers several advantages, including good soft tissue resolution and multi-directional examination [4], which helps to display the rectosigmoid mesorectum and anatomical structure, making it

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

an effective method for diagnosing rectal cancer. Radiologists follow the diagnostic criteria for rectal cancer, identifying the location and shape of the tumor in MRI and determining the staging of rectal cancer by analyzing the depth of infiltration into the rectal wall, subsequently guiding treatment such as chemoradiation or surgery [5]. Therefore, accurate segmentation of rectal cancer is essential to guide the staging of rectal cancer and determine an appropriate treatment plan. Currently, identifying and delineating rectal tumors mainly relies on experienced physicians, which is a time-consuming, labor-intensive, and observation-dependent process. A fast and accurate fully automatic segmentation model can avoid segmentation errors caused by individual bias and clinical experience, thus substantially improve the efficiency of rectal tumor segmentation. Furthermore, the rapid progress in translational medicine [6], [7] enhances the effective integration of automatic segmentation models into the clinical practices of clinicians, thereby augmenting the value of the development of such models.

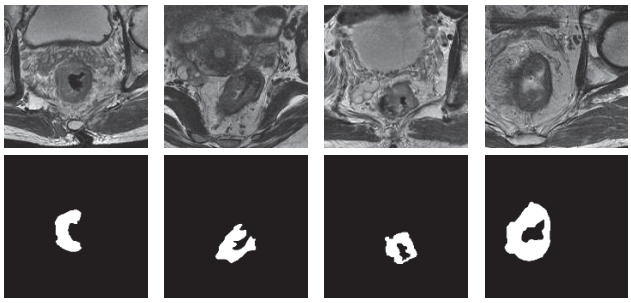

FIGURE 1. Examples of 2D MR rectal cancer images and their corresponding ground truth. (a) demonstrates the tumor's low contrast. (b) illustrates the tumor's irregular shape. (c) and (d) show the complex content and background.

However, as shown in Fig. 1, automated segmentation of rectal MR images poses unique challenges due to several deficiencies, including the following: The rectum, being a hollow visceral organ, undergoes continuous motion due to contraction and expansion. Different scanning planes contribute to the significant unpredictability of shape and position, leading to pronounced morphological changes in rectal tumor imaging. Besides, rectal tumors exhibit varying density distributions, and their low-contrast boundaries pose challenges in distinguishing between cancerous and normal tissues. The presence of rectal contents such as feces and mucus, along with the intricate surrounding background, introduces additional complexities during the rectal tumor segmentation process.

In recent years, the rapid advancement of Convolutional Neural Networks (CNNs) has significantly accelerated the enhancement of image segmentation results. Early benchmarks such as the fully connected network [8] and U-Net [9] have established a foundation for numerous medical image segmentation tasks. Building on this foundation, several improvements have been introduced, including techniques like CRF [10], [11], V-Net [12], ResU-Net series [13], [14], U-Net++ [15] and so on. In the domain of rectal tumor segmentation, initial attempts were made by Trebeschi et al. [16] who employed CNNs for preliminary exploration, directly classifying pixels through fully connected layers and softmax functions. Wang et al. [17] developed a simple 2D CNN model similar to U-Net for rectal cancer, which had limited feature extraction capabilities and struggle to extract deeper features. Li et al. [18] improved U-Net by adding an attention mechanism to the encoder and decoder to extract channel attention features and local spatial attention features. Huang et al. [19] proposed a 3D ROI-Aware U-Net that segmented the ROI region of rectal tumors. After locating the ROI by a module as a preceding operation, three ROI feature maps of various resolutions were sent to the segmentation network. The outputs are averaged to generate the final prediction. The MSBC-Net [20] proposed by Meng et al. integrates classification, regression, and segmentation into one network for rectal tumor and rectal wall segmentation, but additional positioning error it introduced makes false positives more prone in complex contexts.

Existing methods for rectal tumor segmentation primarily rely on CNNs. However, CNNs have limited receptive fields, making it challenging to capture global understanding and leading to the erroneous segmentation of background regions as objects, thereby impeding segmentation performance in complex backgrounds. To address this challenge, researchers have introduced transformer-based networks, which effectively establish long-range dependencies and capture global contextual information. In the context of rectal tumor segmentation, these networks play a crucial role in accurately defining the tumor's shape and delineating its presence within volumetric images. This prevents the misclassification of pixels and ensures precise identification of the tumor's extent.

Dosovitskiy et al. [21] proposed Vision Transformer, which pioneered the application of transformers in computer vision and achieved superior results to CNNs in image classification tasks. Swin Transformer [22] proposed a hierarchical vision transformer that utilized an efficient shifted window approach for local Self-Attention (SA) computation, demonstrating impressive performance across multiple vision tasks. In the field of medical image segmentation, Swin- UNet [23] and DS-TransUNet [24] devised transformer-based architectures with U-shaped structures, drawing inspiration from the Swin Transformer, achieving promising results in 2D image segmentation. Peiris et al. [25] proposed the Volumetric Transformer Encoder-Decoder Structure for multi-modal medical image segmentation, which managed to achieve comparable performance with a more compact model size and reduced computational complexity. However, the self-attention mechanism in transformers lacks translation invariance and is limited in extracting fine-grained features. This limitation can lead to the oversight of small features, thereby potentially resulting in less precise segmentation outcomes. Furthermore, current methods predominantly utilize 2D networks, which fail to preserve rich inter-slice contextual

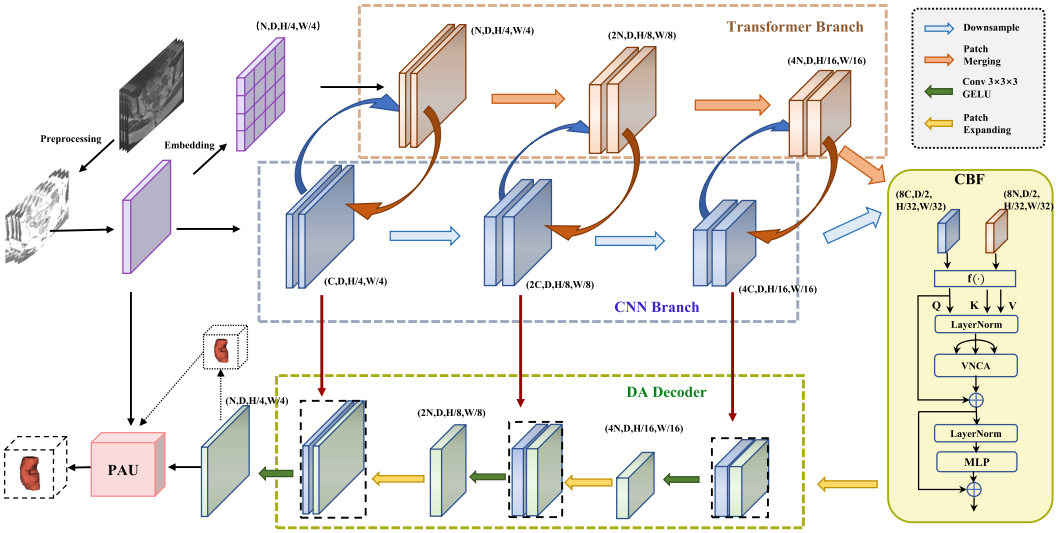

FIGURE 2. An overview of our proposed FCTformer. FCTformer is composed of a dual-branch encoder, a Cross-Branch feature Fusion module in the bottleneck layer, a Dual-Attention decoder, and a Prediction Aggregation Unit.

information and neglect the continuity of tumors between adjacent slices, resulting in suboptimal tumor presence discrimination within 2D slices. Currently, there is a lack of suitable networks that can effectively extract both global and local information from MRI volumetric images, achieving precise rectal tumor segmentation. In an image, local features refer to information such as edges and contours, while global features encompass holistic information like shape and structure.

To address the aforementioned problems, we propose a FCTformer to precisely segment rectal tumor in 3D MRI, which incorporates dual-branch bidirectional fusion encoding and Dual-Attention (DA) decoding techniques. Its objective is to enhance the network's segmentation capability in scenarios characterized by complex backgrounds and significant shape variations. It facilitates a more efficient integration of global and local features, thereby achieving more accurate segmentation results. Firstly, we devised a multiscale feature representation that incorporates bidirectional fusion of global and local features through the integration of transformer modules and convolutional operations. To capitalize on features captured across different scales, thereby enhancing segmentation accuracy, we incorporated a DA decoder. Ultimately, to enhance the tumor's edges and contours, the prediction aggregation unit is designed to capture sharper tumor boundaries and retain fine details that could be lost during repetitive down-sampling stages. The main contributions of this paper include:

- We propose a highly accurate end-to-end segmentation framework called FCTformer that fuses multi-scale Convolutional operations and Transformer. Our framework has achieved state-of-the-art segmentation accuracy in both rectal tumor and prostate segmentation tasks.
- A novel dual-branch encoder and a Cross-Branch feature Fusion (CBF) module are introduced to enhance the extraction capabilities of both the overall tumor shape and the fine-grained details necessary for volumetric segmentation.
- We have designed a DA decoder to maximize the utilization of features captured at different scales. A Prediction Aggregation Unit (PAU) follows, which captures sharper object boundaries and details lost in repeated down-sampling layers.

II. METHOD

An overview of our proposed FCTformer is presented in Fig. 2. The input to our model is a 3D MRI scan $X \in R^{C_0 \times H \times W \times D}$ with C_0 channels (modalities), spatial dimensions H and W, the number of slices D. After preprocessing, the input image is fed into the dual-branch encoder to obtain both global context representation and localized representative features. Subsequently, the CBF module extensively merges deep semantic information before entering the decoder. The decoder extracts fine-grained details and global context from the skip connections of the encoder, as well as the keys and values. Finally, the PAU combines the preliminary segmentation result with tumor features obtained from shallow convolutions to generate the ultimate segmentation mask. The subsequent sections will provide a comprehensive overview of each component.

A. DUAL-BRANCH ENCODER

Due to the irregular variations in size and shape of rectal tumors, as well as the presence of significant background interference, a rectal tumor segmentation network must possess enhanced global feature extraction capability. Given the low contrast between rectal tumors and normal tissues, the network also requires meticulous extraction of local details. Building upon these requirements, we design a multi-scale encoder that combines the CNN and Transformer branches in a mutually fused manner. Considering the complementarity

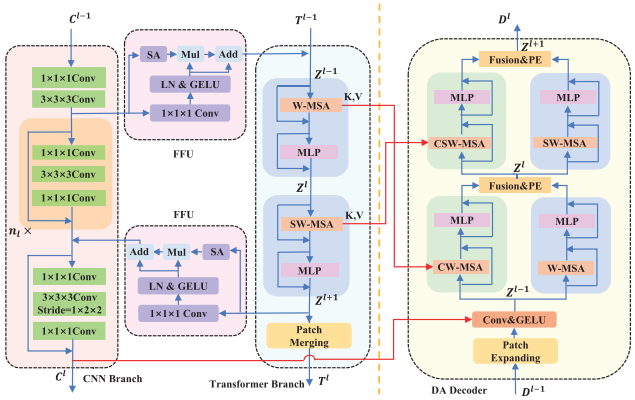

FIGURE 3. Schematic of Encoder-Decoder Structure.

of the two styles of features, in the cascaded encoder module, we continuously feed the local features of the CNN branch back to the patch embedding, enriching the local details of the Transformer branch. Similarly, the object-level relationships from the transformer are fed into the CNN branch to enhance the CNN's perception of a wider context. This interactive dynamic fusion process optimizes the feature representation of both branches to the maximum extent.

Specifically, the initial local features of the shallowest layer are extracted when the input MRI images enter a stem module, and then successively pass through three different levels of dual-branch encoding modules to gain multi-scale information. Notably, due to the anisotropy of data, the resolution between MRI slices is much lower than that between pixels in each slice. In order to protect the integrity of information in each slice, the resolution between slices is kept unchanged in each stage of the network. The output of the last level of both branches is fed into the bottleneck layer.

1) STEM MODULE

The stem module consists of a $1 \times 7 \times 7$ large kernel convolution with a stride of $1 \times 2 \times 2$, followed by a $1 \times 3 \times 3$ max-pooling layer with a stride of $1 \times 2 \times 2$. This configuration is employed to extract image edges, textures, and other local features, thereby reducing the parameter count. The resulting output is a volume of dimensions $C \times D \times H/4 \times W/4$. This output serves as the input for the convolutional branch and the embedding for the Transformer branch.

2) CNN BRANCH

As shown in Fig. 2 and Fig. 3, the CNN branch adopts a feature pyramid structure where the resolution of the feature maps decreases with the increasing depth of the network, with the number of channels increasing. The entire branch is divided into three stages, each consisting of n_l ResNet bottlenecks [26] and a downsampling block. The downsampling blocks perform a stride-2 downsampling on the xy-axis while preserving the z-axis resolution. In our experiments, for simplicity, n_l is set to be {2,2,2}.

CNN and Transformer branches are bidirectionally linked through the Feature Fusion Unit (FFU) module. FFU module initially feeds the input through a $1 \times 1 \times 1$ convolution and LayerNorm. The former converts the features between two branches, and the latter is used to regularize the features. In order to reduce the interference of similar features in the two branches, we incorporate an attention mechanism in the jump connection so that the network pays more attention to the spatial details of the CNN branch and the contextual relationship of the Transformer branch. Formally, we consider a building block from CNN to Transformer branch defined as:

$$F_T = \phi\left(F_C\right) \otimes SA\left(F_C\right) \oplus \phi\left(F_C\right) \tag{1}$$

where $\otimes$ stands for matrix multiplication and $\oplus$ stands for element-level addition. $\phi(\cdot)$ and $SA(\cdot)$ means a $1 \times 1 \times 1$ convolution layer with layer normalization and GELU and a spatial attention layer. To enhance the feature fusion effectiveness of both branches, we ensured the consistency of spatial feature dimensions between CNN and Transformer. This not only facilitates feature transfer but also avoids the frequent alignment of their spatial dimensions.

The Transformer branch projects image patches into a vector, and its self-attention mechanism makes it less sensitive to local details. On the other hand, in CNN, the convolutional kernels slide over overlapping feature maps of high resolution, enabling the preservation of subtle local features. Thus, the CNN branch continuously provides the Transformer branch with local feature details and its inherent biases. The output of each stage's CNN branch is also fed into the decoder branch through skip connections, preserving more details and semantic information from the input data.

3) TRANSFORMER BRANCH

Due to the quadratic complexity of the standard multi-head self-attention (MSA) in vanilla ViT, it is inefficient for segmentation tasks involving larger images. In the case of volumetric images, its complexity exhibits cubic growth. To overcome this limitation, we used the Swin Transformer module and modified it into a 3D form suitable for this task as shown in Fig. 3.

The vanilla MSA block is substituted with two variants: window-based multi-head self-attention (W-MSA) and shifted window-based multi-head self-attention (SW-MSA). Additionally, before each block, layer normalization is applied, and residual connections are incorporated after each block. Notably, SW-MSA introduces a unique windowing configuration that is shifted relative to the input of the W-MSA module, ensuring the presence of cross-window connections. This process is depicted in (2).

$$\begin{aligned}
\hat{z}^{l} &= \mathrm{W-MSA}\left(\mathrm{LN}\left(z^{l-1}\right)\right) + z^{l-1}, \\
z^{l} &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^{l}\right)\right) + \hat{z}^{l}, \\
\hat{z}^{l+1} &= \mathrm{SW-MSA}\left(\mathrm{LN}\left(z^{l}\right)\right) + z^{l}, \\
z^{l+1} &= \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1},
\end{aligned} \tag{2}$$

where $\hat{z}^l$ and z^l denote the output features of the W-MSA module and the MLP module for block l, respectively.

Subsequently, a patch-merging operation is implemented, which involves concatenating adjacent patches in 2×2 groups. This process enlarges the embedding dimensions from C to 2×C while simultaneously reducing the resolution. Moreover, the K and V matrices from the two MSA blocks are introduced into the cross-attention branch in the decoder through a cross connection. This expedites model information transfer and prevents information loss.

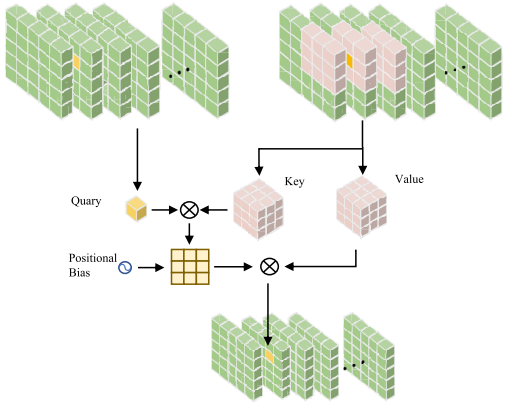

FIGURE 4. Illustration of the volumetric neighborhood cross-attention.

The Transformer branch offers more than just an enlarged receptive field. Its SA mechanism not only enhances the CNN branch's ability to capture image contours and describe shapes but also provides valuable prior information about target types based on long-range dependencies. Furthermore, the inherent biases present in the CNN branch eliminate the necessity of incorporating supplementary positional information into the Transformer branch.

4) BOTTLENECK LAYER

In order to more effectively extract semantic information and correlations from the deep feature space of both branches, thereby further enhancing the ability to identify tumors, we propose a novel CBF module, which takes the resultant deepest levels of two branches as inputs and employs a volumetric neighborhood cross-attention mechanism (VNCA) to fuse information across branches. The primary challenge here lies in the efficient fusion of features at both the CNN and Transformer levels, all while maintaining the integrity and comprehensiveness of the features. A straightforward strategy would involve directly feeding the summation of CNN levels along with their corresponding Swin Transformer levels into the decoders. Such approach, however, fails to ensure feature integrity and combine the more advanced semantic information between them, leading to subpar performance. Furthermore, both simple convolutional layers and vanilla Vision Transformers tend to introduce a large number of parameters. If we continue using the Swin Transformer module, the effectiveness of its window-shift mechanism might be limited in deeper layers of the network, and it is also necessary to use the mechanism in pairs.

We employ the representative sliding mechanism in convolutional operations to compute Cross-Attention (CA), ensuring the preservation of feature continuity and integrity. Given inputs F_N and F_T representing the features from the convolutional and Transformer branches, and F_N and F_T after feature alignment through function $f(\cdot)$, we have linear projections Q, K, V, and relative positional biases $\mathbf{P(i, j)}$. We define the attention weights of the i-th input with neighborhood size k as $\mathbf{W}_i^k$. It is computed as the dot product between the query projection of the i-th input in F_N and the corresponding k nearest key projections in F_T:

$$\mathbf{W}_i^k = \begin{bmatrix} Q_i K_{\rho_1(i)}^T + P_{(i,\rho_1(i))} \\ Q_i K_{\rho_2(i)}^T + P_{(i,\rho_2(i))} \\ \vdots \\ Q_i K_{\rho_k(i)}^T + P_{(i,\rho_k(i))} \end{bmatrix} \tag{3}$$

where $\rho_j(i)$ denotes the j-th nearest neighbor of i. Subsequently, we define the neighboring values $\mathbf{V}_i^k$ as a matrix, with its rows representing the k nearest value projections for the i-th input:

$$\mathbf{V}_i^k = \begin{bmatrix} V_{\rho_1(i)}^T & V_{\rho_2(i)}^T & \cdots & V_{\rho_k(i)}^T \end{bmatrix}^T \tag{4}$$

The volumetric neighborhood attention for the i-th token with neighborhood size k is defined as:

$$\mathrm{VNCA}_k(i) = \mathrm{softmax}\left(\frac{\mathbf{W}_i^k}{\sqrt{d}}\right)\mathbf{V}_i^k \tag{5}$$

where $\sqrt{d}$ is a scaling parameter. This operation is performed for each pixel in the feature map. The graphical representation of this process can be found in Fig. 4, and the channel dimension has been omitted.

B. DECODER

After the bottleneck layer, the decoder initiates with a sequence of successive DA blocks, concluding with the PAU module to generate the ultimate segmentation predictions in 3D rectal MRI images.

In order to maximize the use of local details and global contextual features captured at different scales to improve the perception and learning of tumors, and hence the accuracy of segmentation, we introduce a DA decoder module. As shown in Fig. 3, we employed a parallel decoding approach that combines CA and SA in the decoder. Each decoder block receives tokens generated by the previous decoder block, as well as K and V matrices from the encoder's corresponding transformer branch at the same level. On the left branch of the decoder, the cross-attention mechanism is used in a cross-like connection to fuse the K and V matrices of the same-stage encoder, which can be described as:

$$CA = SA\left(Q_D, K_E, V_E\right) \tag{6}$$

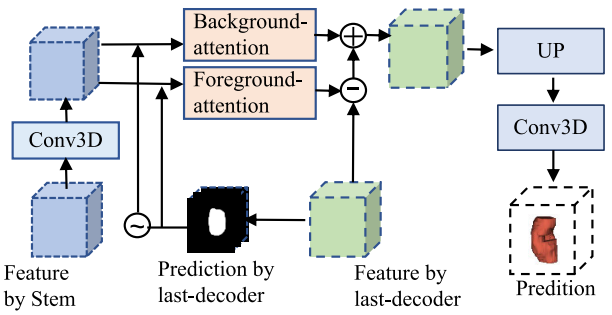

FIGURE 5. Structure of Prediction Aggregation Unit.

The subscripts D and E respectively indicate the information flow from the decoder and encoder.The right branch employs the vanilla SA mechanism. Additional spatial information obtained from the encoder is beneficial. During the backpropagation, the gradient flow of K and V vectors from the decoder to the encoder facilitates faster convergence of the model. Furthermore, the skip connections from the encoder are spatially aligned and combined with deep-level semantic information, thereby complementing the lost edges and textures during the downsampling process to enhance the model's performance. Subsequently, the outputs of the two attention mechanisms are weighted and fused, and an additional three-dimensional fourier positional encoding [27] is incorporated, which helps encapsulate crucial spatial information essential for pixel-level segmentation tasks. This process can be represented as follows:

$$\mathbf{z}^l = \alpha\hat{\mathbf{z}}^l_{CA} + (1-\alpha)\hat{\mathbf{z}}^l_{SA} + \mathcal{F}\left(\hat{\mathbf{z}}^l_{SA}\right) \tag{7}$$

here, $\mathcal{F}(\cdot)$ represents the positional encoding, and α serves as a weight factor controlling the contributions of the CA and SA branches. In our experiments, for simplicity, we choose α = 0.5.

The purpose of patch expansion is to counteract the impact of patch merging, i.e., to upsample and restore the image size.

The PAU takes the stem features, last-decoder block features, and predictions as inputs and produces refined features along with a more precise rectal tumor prediction. As illustrated in Fig. 5, the PAU initially utilizes the predictions from the final decoder and multiplies them with the Stem features to generate foreground attention features and background attention features. Subsequently, it employs two Context Exploration Blocks [28] to identify errors in the predicted foreground (tumor region) attention features and background (non-tumor region) attention features. Following this, it mitigates the blurred background (i.e., false positives of normal tissue) and enhances the missing foreground (i.e., false negatives of tumors) by performing subtraction and addition operations with the features from the final decoder. Finally, the refined features undergo a three-dimensional convolution to produce the ultimate tumor segmentation result. The Context Exploration Block consists of four cascaded dilated convolution modules, and each dilated convolution block is composed of a convolution with a kernel size K_i, a convolution with a kernel size of 1 × 3×3 and a dilation rate of R_i, layer normalization, and ReLU activation function; where K_i and R_i are chosen from {1,3,5,7} and {1,2,4,8} respectively. The outputs of the four dilated convolution modules are combined and passed through a convolution with a kernel size of 3 × 3×3 before the final output.

C. LOSS

In order to enhance the segmentation capability of the model in rectal MRI images with an imbalanced distribution between tumor and background categories, while also expediting the model convergence, we present the following loss functions for training the FCTformer model. The loss function consists of the output $\hat{y}^{dec}$ of the last decoder and the output $\hat{y}^{PAU}$ after PAU. The details about the proposed loss functions are introduced as follows:

$$\mathcal{L} = I\left(\hat{y}^{dec}, T\left(y, \hat{y}^{dec}\right)\right) + I\left(\hat{y}^{PAU}, y\right) \tag{8}$$

$T\left(y, \hat{y}\right)$ is a tri-linear interpolation function that re-scales the ground-truth y to the size of $\hat{y}$. And $I\left(\hat{y}, y\right)$ consists of the following two parts:

$$I\left(\hat{y}, y\right) = \lambda_{focal}\mathcal{L}_{focal}\left(\hat{y}, y\right) + \lambda_{dice}\mathcal{L}_{dice}\left(\hat{y}, y\right) \tag{9}$$

where λ_{dice} and λ_{focal} are the hyper-parameters to balance the Dice and focal loss. we define the Dice loss by (10):

$$\mathcal{L}_{\text{dice}}\left(\hat{y}, y\right) = 1 - \frac{2\|\hat{y}\| \circ \|y\|_1 + \epsilon}{\|\hat{y}\|_1 + \|y\|_1 + \epsilon} \tag{10}$$

where $\circ$ denotes the Hadamard product and ϵ is a smoothing term used to avoid division by zero (we set $\epsilon = 1 \times 10^{-8}$).

In addition to the Dice loss, we utilize the focal loss [29] to mitigate the impact of class imbalance between the lesion and background regions:

$$\begin{aligned}\mathcal{L}_{focal}\left(\hat{y}, y\right) = &-\alpha \cdot \left(1-\hat{y}\right)^{\gamma} \cdot y \cdot \log\left(\hat{y}\right) \\ &- (1-\alpha) \cdot \hat{y}^{\gamma} \cdot (1-y) \cdot \log\left(1-\hat{y}\right)\end{aligned} \tag{11}$$

where α serves as a hyper-parameter to balance the training samples of lesions and background, and γ acts as another hyper-parameter, influencing the degree of emphasis placed on challenging samples within the loss function.

III. EXPERIMENTS

A. DATASET AND PREPROCESSING

1) DATASET

The experiment was performed on clinical rectal cancer dataset obtained from Fudan University Shanghai Cancer Center, from January 2012 to February 2021. This dataset includes 362 cases (243 males, 119 females, age range is 33∼85 years old) with 7732 axial abdominal T2W-MRI 2D slices. Each MR volume involves 16∼27 slices of 512 × 512 pixels, with a voxel ZYX spatial resolution of ([3.9 ∼ 4.1]×[0.39 ∼ 0.41] × [0.39∼ 0.41]) mm^3. The number of cases at different stages of rectal tumor development, namely T1, T2, T3, and T4, are 22 cases (6.1%),

51 cases (14.1%), 184 cases (50.8%), and 90 cases (29.0%), respectively.

Regions of Interest for the segmentation task, are meticulously outlined on each slice of the T2W-MR images by two experienced radiologists using ITK-SNAP, and are redelineated after evaluation by two senior radiologists. In situations where a difference of opinion arises among the radiologists, the final decision regarding the case is reached through a comprehensive discussion involving all radiologists.

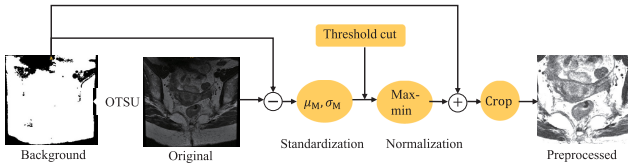

FIGURE 6. The preprocessing approach for the datasets.

Furthermore, we collected a publicly available dataset of prostate images and tested FCTformer on these datasets. The test data comprised four publicly available datasets: NCI-ISBI 2013, I2CVB, PROMISE12, and MSD 2018. From these datasets, we selected a total of 202 3D prostate T2-weighted MRI images along with their corresponding segmentation masks. The number of selections from each dataset was 80, 40, 50, and 32, respectively. By collectively inputting all the data into our model, we aimed to assess the model's generalization performance on multi-center data with complex backgrounds and variations in imaging styles. To ensure consistency across different image views, we center-cropped the images from I2CVB dataset. Subsequently, all samples were resized to 320 × 320 in the axial plane.

2) PREPROCESSING

After applying N4 bias field correction to the images from different patients, we conducted resampling to normalize the varying pixel spacing to the median value of 4.0 × 0.4×0.4 mm^3. To ensure consistency in intensity of input images obtained under different imaging configurations and field of views, and to enhance contrast, we performed preprocessing as illustrated in Fig. 6. After OTSU [30] thresholding, the mean intensity and standard deviation are computed within the foreground according to:

$$Mean\,(X) = \frac{1}{N_{mask}} \sum_{i \in mask} x_i \tag{12}$$

$$std\,(X) = \sqrt{\frac{1}{N_{mask}} \sum_{i \in mask} (x_i - Mean\,(X))^2} \tag{13}$$

where $x_i \in X$ denotes the intensity of a voxel and N_{mask} denotes the count of mask voxels. Threshold cut is employed to mitigate the interference caused by outlier pixels with excessively high values. Then the image is normalized according to standard normalization criterion.

Before feeding the images to the network, we crop the volumes to a fixed size of 16 × 320×320 to remove unnecessary background and reduce the GPU memory footprint. Additionally, in the training stage, data augmentation includes random horizontal flip, random vertical flip, brightness and contrast adjustment, elastic transformation and ROI translation. The dataset was divided into 80% for training and 20% for testing.

B. EXPERIMENTAL SETUP

The framework was implemented using PyTorch and trained on a single Nvidia Tesla V100 GPU, boasting 36 GB of memory. Our training procedure involved employing the Adam optimizer with a learning rate set at $1e^{-4}$ for a total of 600 epochs. We utilized a cosine decay learning rate scheduler and maintained a batch size of 1 during the training process.

C. EVALUATION METRICS

In this study, five widely adopted metrics are used to evaluate the final segmentation results, including Dice Similarity Coefficient(DSC), Precision(Prec.), Specificity(Spec.), recall and Average Symmetric Surface Distance (ASSD) [31], [32]. The DSC index measures the overall overlap rate. The ASSD gauges the average voxel distance between the segmented result and the label. Precision, specificity, and recall metrics assess segmentation from the perspective of voxel classification accuracy. Superior segmentation results are characterized by smaller ASSD values, while larger values are expected for all other metrics.

IV. RESULTS AND DISCUSSION

A. MAIN RESULTS

Our extensive experiments aimed to quantitatively and qualitatively evaluating the performance of the FCTformer. We applied benchmark methods such as U-Net, V-Net, and several other state-of-the-art methods in rectal tumor segmentation or general medical image segmentation to our rectal cancer MRI dataset. The comparison results, as presented in Table 1, demonstrate the superiority of the FCTformer over these aforementioned methods. Specifically, the proposed method achieves outstanding results (mean ± std) with DSC at 82.67±6.14%, precision at 83.48±11.82%, specificity at 99.80±0.11%, recall at 82.02±12.03%, and ASSD at 1.78±1.40mm. Additionally, Fig. 7 shows the prediction effect of FCTformer. In particular, the first three rows depict subtle differences between different methods at the 2D slice level, while the fourth row demonstrates the overall distinctions on the entire 3D image of the same case.

As depicted in Table 1, the traditional convolutional networks, such as 3D UNet [33] and Vnet [12], exhibit poorer performance compared to transformer-based approaches like VT-Unet [25] and nnFormer [34]. This is primarily attributed to the limited receptive field of convolutional networks, where the sliding convolutional kernels can only

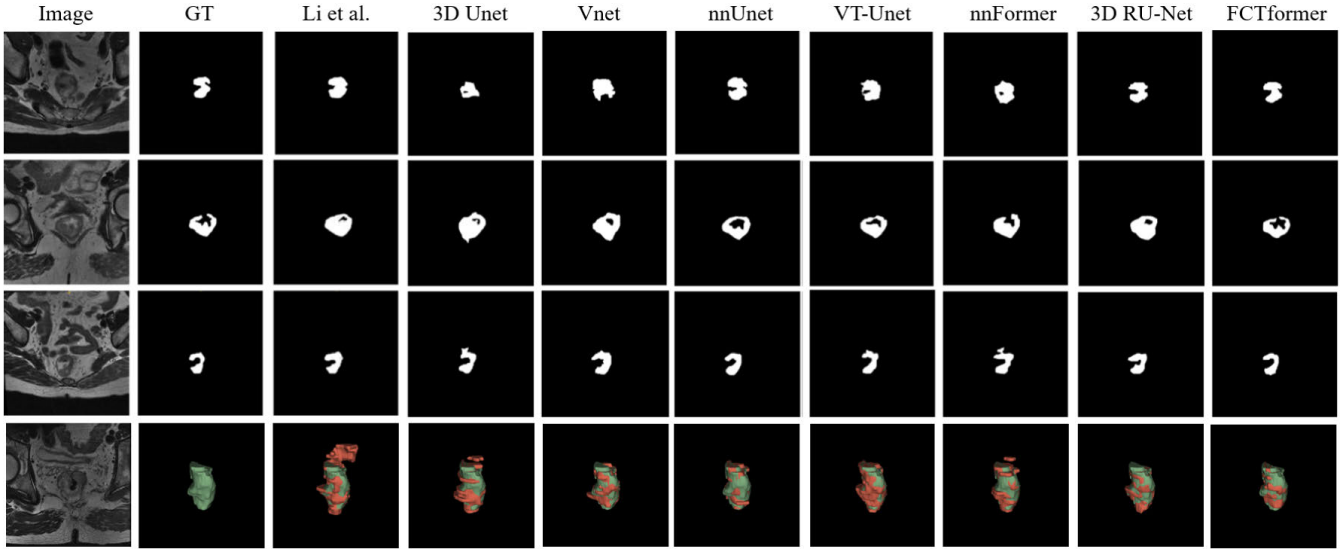

FIGURE 7. Comparative visualization of segmentation results with other state-of-the-art methods. The first three rows showcase the results of 2D slices, while the fourth row displays the visualization of the segmentation on the complete 3D volumetric image. In the fourth row, green represents the ground truth, and red represents the actual segmentation results.

TABLE 1. Comparison to other state-of-the-art architectures (mean±std).

Methods	DSC (%)	Prec. (%)	Spec. (%)	Recall (%)	ASSD (mm)
Li et al.	73.25±11.93	72.32±16.43	99.69±0.16	75.41±14.92	3.30±2.71
3D Unet	71.64±10.31	72.40±16.19	99.63±0.21	73.23±15.28	4.31±3.42
Vnet	73.45±10.12	74.41±13.87	99.68±0.19	73.14±13.95	3.21±2.65
nnUnet	80.75±8.09	82.04±13.13	99.76±0.14	79.26±14.03	2.12±1.45
VT-Unet	77.37±7.75	80.06±12.34	99.74±0.15	76.52±14.51	2.58±2.08
nnFomer	76.58±8.86	78.57±14.63	99.72±0.14	76.10±15.03	2.88±2.06
3D RU-Net	75.48±9.55	78.52±15.17	99.73±0.16	73.87±16.94	3.10±2.52
FCTformer	82.67±6.14	83.48±11.82	99.80±0.11	82.02±12.03	1.78±1.40

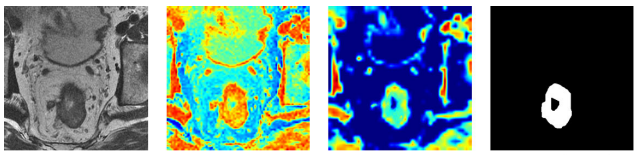

FIGURE 8. Examples of local feature extracted by the CNN branch and global feature extracted by the Transformer branch.

focus on adjacent regions. Consequently, convolutional networks lack the capability to capture global features, making it challenging to accurately delineate the complete shape of tumors. However, segmentation results based on Transformer networks are also unsatisfactory, primarily due to the limited capacity of the self-attention mechanism in extracting fine-grained features. nnUNet [35] achieved the second-best results by adjusting model parameters according to the characteristics of the data. However, the lack of global features has limited the improvement of segmentation performance. Compared to it, our method improves 1.92%, 1.44%, and 2.76% in terms of DSC, precision, and sensitivity, respectively. FCTformer integrates a transformer-based global feature extraction mechanism and a CNN-based local feature extraction approach to obtain a dual-faceted multiscale feature representation. This representation enhances the model's capability to capture both the comprehensive semantic features and intricate details of rectal cancer instances, especially in challenging situations such as low-contrast imaging and substantial shape variations. In models dedicated to rectal tumor segmentation, the 3D RU-Net [19], after employing a module to localize ROI as a preceding step, sends three ROI feature maps of varying resolutions to the segmentation network. The final prediction results are generated by averaging the outputs. However, the localization module introduces additional errors, particularly in complex backgrounds, thereby impacting the ultimate results. Additionally, we deliberately compared this approach with the best-performing model in 2D slice segmentation of rectal tumors. In comparison to the 3D RU-Net, the model proposed by Li et al. [18] experienced a decrease of over 2% in Dice coefficient, primarily attributed to the poorer continuity in segmenting 2D slices. Furthermore, the 2D model tended to erroneously segment tumors in slices without tumors.

As shown in Fig. 7, the segmentation results on tumor images further validate our analysis. In the first and third rows of Fig. 7, convolutional networks such as 3D U-Net and V-Net exhibit diminished capability in resolving adherent

targets and background areas with lower contrast to tumors. The CNN-based method faced a greater challenge in learning explicit global and long-range semantic information interactions compared to the Transformer-based approach. However, the segmentation performance of VT-Unet and nnFormer on edges and contours also falls short of expectations, as depicted in the second and third rows. The primary rationale behind this difficulty lies in the fixed approach to image patch division, which disregards the geometric variations occurring within the same object across different images. Additionally, it's a common occurrence for the local structures of objects in images to become fragmented when using a fixed patch size, making it challenging to encompass the entire local structure associated with the object. While segmentation models based on joint ROI localization effectively reduce model parameter count through a locate-then-segment approach, the localization module is susceptible to significant adverse effects from complex backgrounds, leading to additional localization errors and consequently resulting in numerous false positives. As shown in the fourth row, the 2D Improved U-Net proposed by Li et al. tends to produce significant segmentation errors in slices without tumors. This highlights the crucial role of 3D networks in extracting inter-slice contextual information for comprehensive tumor segmentation. It is evident that in comparison to alternative approaches, our method demonstrates a heightened accuracy in delineating the edges of rectal tumors, and exhibits a more comprehensive forecast of tumor morphology. This proficiency can be attributed to our method's enhanced amalgamation of global context and localized spatial intricacies, consequently augmenting both the global perceptual awareness and the precision of segmentation within our network.

TABLE 2. Comparison between different encoder structures.

Methods	DSC (%)	Prec. (%)	Recall (%)	ASSD (mm)
Only CNN	77.63	76.95	78.91	2.61
Only Transformer	78.72	80.44	77.61	2.38
CNN and Transformer	79.45	80.89	78.64	2.27
1 FFU	80.69	81.43	80.35	2.20
2 FFU	81.81	82.15	81.24	2.11
3 FFU (Ours)	82.67	83.48	82.02	1.78

B. ABLATION STUDY

1) STRUCTURE ANALYSIS

We investigated the impact of various branches and feature fusion units on the segmentation outcomes of the dual-branch network. To conduct this analysis, we performed ablation experiments on a rectal cancer dataset, and the results are summarized in Table 2. In these experiments, we controlled the parameter count of each network by configuring the channel numbers at each layer of the network to be similar. We initially compared different encoding branches. In the model with only the CNN branch, we omitted the transmission of the K and V matrices. In the model with only the Transformer branch, we transferred the output from each stage of the Transformer branch, not the CNN branch, to the decoder through skip connections. Subsequently, to delve deeper into the influence of the FFU on the dual-branch network, we progressively added sets of FFU modules from bottom to top, up to three sets (constrained to three sets when the network depth was four).

Upon examining the first two rows of Table 2, it is evident that the indicators of the single-branch network based on the Transformer with global features surpass those of the single-branch network based on residual modules. The latter network, relying on residual modules, lacks global information, resulting in inferior results. From the second and third rows of Table 2, the dual-branch network demonstrates an improvement of 0.73% in Dice coefficient, 0.45% in precision, and 1.03% in recall compared to the network based on Transformer modules. This enhancement in metrics is attributed to the incorporation of local information into the Transformer branch of the dual-branch network. These outcomes validate the efficacy of the dual-branch network in rectal tumor segmentation. Furthermore, with the increase in the number of FFU modules, their respective metrics consistently improve. In comparison to networks featuring only one FFU module, those with three FFU modules exhibit superior results across all metrics. This advancement is evident in a 1.98% increase in Dice coefficient, a 2.05% increase in precision, a 1.67% increase in recall, and a 0.42 mm decrease in ASSD. This indicates that the introduction of FFU modules enhances the information exchange between the two branch architectures, effectively leveraging global and local information for improved performance.

TABLE 3. Comparison between different modules.

Methods	DSC (%)	Prec. (%)	Recall (%)	ASSD (mm)
Baseline	75.43	76.43	75.39	2.71
Baseline + CBF	76.91	78.23	76.44	2.67
Baseline + PAU	78.12	79.06	77.80	2.38
FCTNet	79.69	81.09	78.96	2.29
FCTNet +CBF	80.89	82.05	80.42	2.14
FCTNet + PAU	81.41	82.25	81.02	2.11
FCTformer	82.67	83.48	82.02	1.78

Furthermore, to further demonstrate the differences in feature extraction between the two branches, we visualized the feature maps of the two branches in the first stage of the encoder, as shown in Fig. 8. From the figure, it can be observed that the features extracted by the CNN branch exhibit a clear perception of the tumor's texture and contrast changes. This can be attributed to the sliding convolutional kernels of CNN, which effectively capture local high-frequency information by sensing changes in adjacent pixels. On the other hand, the Transformer branch can provide a clear depiction of the tumor's shape and structure, reducing misjudgments in the background region. This is primarily because the MSA mechanism in the Transformer can link image patches at longer distances, enabling effective long-range relationship modeling and

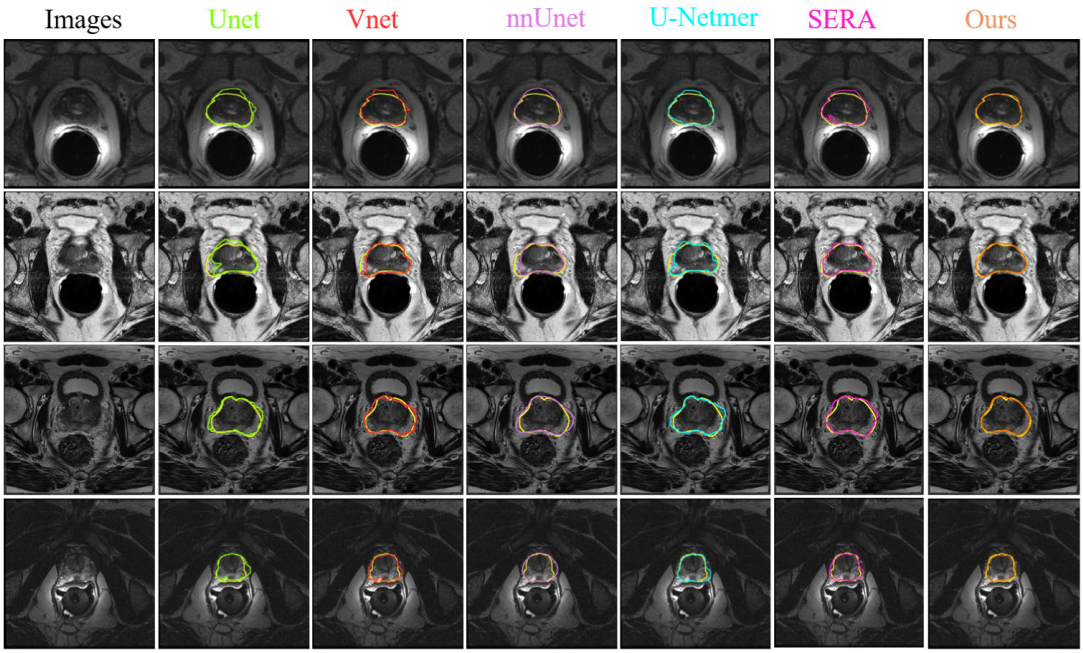

FIGURE 9. Visualization of the prostate segmentation results.

TABLE 4. Comparison between different decoder structures.

Methods	DSC (%)	Prec. (%)	Recall (%)	ASSD (mm)
DC-Only	76.62	75.81	77.74	2.68
DS-Only	77.91	78.82	77.08	2.63
DCDS	79.52	79.41	79.79	2.41
DCtoDS	80.72	82.01	79.66	2.22
DStoDC	81.54	83.12	80.17	2.14
FCT-woPE	82.26	83.05	81.66	2.03
FCTformer	82.67	83.48	82.02	1.78

thereby enhancing the ability to define the complete shape of the tumor.

2) ABLATION STUDY ON EACH MODULE

To assess the distinct impact of various network components on segmentation performance, we carried out ablation experiments across different configurations. Initially, we removed the CBF module, directly summing the features from the two branches, and added a Swin Transformer Block to fuse deep semantic information. We further considered the network removed PAU and CNN branches as the lower baseline. Subsequently, we extended the baseline by (1) expanding baseline by a CNN branch and denoted this network as FCTNet; (2) appending PAU to examine whether adding local details as a supplement enhances the overall segmentation performance; (3) removing the Swin Transformer Block in bottleneck layer and adding CBF. Additionally, Table 3 lists other combinations of the proposed modules. Our FCTformer extended the baseline by incorporating all the proposed components.

The results of the different network components are summarized in Table 3. Compared with the baseline, we observed that using CNN and Transformer modules as the encoder increased DSC more than 4% and decreased ASSD from 2.71 mm to 2.29 mm. This is primarily because the CNN branch and the FFU module provide the Transformer branch with a wealth of detailed edge information and add the relative positional information required by the MSA mechanism. We noticed that adding PAU and CBF to both the baseline and FCTNet resulted in improvements, with an increase of over 1% in the Dice metric. This demonstrates the effectiveness of both modules. The CBF module outperforms the Swin Transformer module, mainly because CBF can more effectively extract semantic information and correlations from the deep feature spaces of both branches, further enhancing the ability to identify tumors. The PAU module, on the other hand, combines features from the Stem layer and the previous decoder layer, focusing on the tumor's edge and using dilated convolutions to produce more accurate segmentation results. Moreover, the PAU module achieves a greater improvement compared to CBF. This is mainly because PAU makes a larger contribution by enhancing the response in the tumor region than CBF, directly benefiting the segmentation task. Lastly, our proposed FCTformer utilized all the proposed modules, resulting in the best performance compared to the baseline, with improvements of 7.24%, 7.05%, 6.63% in DSC, precision, recall and a reduction of 0.93 mm in ASSD, compared with the baseline. The experimental results serve to reinforce the specific decisions made regarding the novel modules introduced in this proposal.

3) ABLATION STUDY ON DECODER

We conducted multiple experiments to validate the effectiveness of the decoder architecture. Based on our definitions,

TABLE 5. Segmentation results on public PROSTATE datasets.

Methods	DSC (%)	Prec. (%)	Spec. (%)	Recall (%)	ASSD (mm)
3D Unet	86.61±3.75	85.67±7.22	99.54±0.24	86.65±6.74	1.34±0.98
Vnet	87.47±3.66	86.89±7.01	99.64±0.22	87.86±6.36	1.25±0.81
nnUnet	89.54±3.20	88.83±6.29	99.80±0.11	90.67±5.18	1.02±0.67
TransUnet	86.23±3.98	85.91±8.10	99.59±0.26	86.51±6.54	1.31±0.96
Swin-Unet	85.81±5.32	83.76±9.43	99.42±0.29	86.72±6.15	1.48±1.03
U-Netmer	88.10±3.48	87.23±6.91	99.53±0.12	88.14±6.48	1.12±0.74
SERA	89.32±3.58	87.94±6.80	99.69±0.15	89.97±6.88	0.98±0.52
FCTformer	90.72±3.15	89.82±5.98	99.81±0.10	90.95±5.18	0.86±0.47

the cross-attention encoder branch is referred to as DC, and the self-attention branch as DS. Keeping the skip connections unchanged, we experimented with the following decoder configurations: (1) using a single branch as the decoder (DC-Only and DS-Only), (2) using both branches without any feature fusion (DCDS), (3) transferring features from one branch to another for unidirectional fusion (DCtoDS and DStoDC), and (4) removing positional encoding (FCT-woPE).

Table 4 presents the results for different decoder configurations. We observed that DC-Only and DS-Only achieved the lowest Dice coefficients, indicating that a single branch alone cannot produce satisfactory results. Using only the DC branch leads to an excessive impact of the K and V matrices on the decoder, disrupting the normal upsampling process. On the other hand, relying solely on the DS branch hinders the effective utilization of Transformer encoder features. Feature reuse is beneficial for information propagation and filtering [36], thereby improving tumor segmentation efficiency and performance. Simultaneously, without fusion or unidirectional fusion of the two branches, there is some improvement in the segmentation results, but it still does not reach the optimum. This is because it is not conducive to multi-directional feature transmission, and the two branches are prone to interference from similar features. In addition, the inclusion of positional encoding, with improvements of 0.41%, 0.43%, 0.36% in DSC, precision, recall, and a reduction of 0.25 mm in ASSD, has also been shown to enhance the model's performance.

C. SEGMENTATION RESULTS ON PUBLIC PROSTATE DATASETS

In order to evaluate the effectiveness and generalizability of our model, we collected a publicly available dataset of prostate images and tested FCTformer on these datasets. The segmentation results obtained by our proposed model were compared against other methods on these datasets, and the performance metrics, presented in the form of mean±std, are listed in Table 5. To showcase the effectiveness and competitive segmentation performance of our proposed model, we present representative visualization results for each of the four datasets in Fig. 9. In the figure, the yellow color represents the ground truth segmentation, while the other colors correspond to the segmentation results obtained by respective model. From the results, it is evident that our method can effectively model the relationships between objects in images with different styles, while paying more attention to edge details, leading to improved localization and better discrimination of the prostate region.

The first three convolution-based networks exhibit difficulty effectively distinguishing the prostate from the background in the first two rows. In contrast, the Transformer-based network is capable of partially suppressing erroneous background distinctions through the incorporation of acquired global information (as demonstrated in the fourth row). Furthermore, from the second row, it becomes apparent that the Transformer-based network possesses a better awareness of prostate regions with significant color disparities. This underscores the Transformer's effectiveness in mitigating challenges posed by object style variations and enhancing the robustness of prostate segmentation. By virtue of the dual-branch encoder's introduction, our approach capitalizes on these benefits while also excelling in distinguishing prostate edges (as shown in the third row). The outcomes presented in Table 5 further underscore the superiority of our method. Notably, 3D networks such as nnU-Net and V-Net outperform 2D methods like TransUNet and Swin-UNet in terms of prediction accuracy. This can be attributed to the crucial contextual information from neighboring slices in images with blurred edges.

It is worth mentioning that SERA also identified the subpar segmentation performance of 2D networks in images lacking the prostate. This method involves localizing slices containing the prostate before segmenting them. However, the simplicity of its convolutional localization network constrains its performance. When compared to SERA, our approach demonstrates improvements of 1.4% in Dice coefficient, 1.88% in precision, 0.98% in recall, and 0.12% in specificity. This substantiates the superiority of our method.

V. CONCLUSION

We proposed FCTformer, an innovative architecture fusing the long-range context of Transformer and the local details of CNN for whole-volume rectal tumor segmentation. Our method effectively distinguishes lesion areas within complex backgrounds by combining inter-slice and long-distance spatial features with local details. Additionally, it refines edge segmentation while obtaining additional spatial information from the encoder, thereby enhancing the learned feature representation. The experiments indicate that on the same

dataset, FCTformer outperforms state-of-the-art CNN and visual Transformer networks, achieving superior segmentation results across various metrics. We also conducted comprehensive ablation experiments to assess the influence of individual network components, thereby reinforcing our design decisions. The aforementioned experimental metrics and visualization results demonstrate that our method can accurately and reliably automate the segmentation of rectal tumors within adjacent normal tissues. In our future research endeavors, we have a strong plan to focus on the development of robust and efficient weakly supervised models for rectal tumor segmentation to further reduce our dependency on data annotation. Furthermore, we will explore neural architectures with more causal relationships to enhance the output results.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.

[2] H. Brenner, M. Kloor, and C. P. Pox, "Colorectal cancer," *Chem. Rev.*, vol. 383, no. 9927, pp. 1490–1502, 2014.

[3] M. Song, S. Li, H. Wang, K. Hu, F. Wang, H. Teng, Z. Wang, J. Liu, A. Y. Jia, Y. Cai, Y. Li, X. Zhu, J. Geng, Y. Zhang, X. Wan, and W. Wang, "MRI radiomics independent of clinical baseline characteristics and neoadjuvant treatment modalities predicts response to neoadjuvant therapy in rectal cancer," *Brit. J. Cancer*, vol. 127, no. 2, pp. 249–257, Jul. 2022.

[4] J. Jian, F. Xiong, W. Xia, R. Zhang, J. Gu, X. Wu, X. Meng, and X. Gao, "Fully convolutional networks (FCNs)-based segmentation method for colorectal tumors on T2-weighted magnetic resonance images," *Australas. Phys. Eng. Sci. Med.*, vol. 41, no. 2, pp. 393–401, Jun. 2018.

[5] U. Tapan, M. Ozbayrak, and S. Tatli, "MRI in local staging of rectal cancer: An update," *Diagnostic Interventional Radiol.*, vol. 20, no. 5, pp. 390–398, Aug. 2014.

[6] M. Mediouni, R. Madiouni, M. Gardner, and N. Vaughan, "Translational medicine: Challenges and new orthopaedic vision (mediouni-model)," *Current Orthopaedic Pract.*, vol. 31, no. 2, pp. 196–200, Mar. 2020.

[7] M. Mediouni, D. R. Schlatterer, H. Madry, M. Cucchiarini, and B. Rai, "A review of translational medicine. the future paradigm: How can we connect the orthopedic dots better?" *Current Med. Res. Opinion*, vol. 34, no. 7, pp. 1217–1229, 2018.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Munich, Germany: Springer, 2015, pp. 234–241.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[12] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[13] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, p. 225.

[14] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[15] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer, 2018, pp. 3–11.

[16] S. Trebeschi, J. J. M. van Griethuysen, D. M. J. Lambregts, M. J. Lahaye, C. Parmar, F. C. H. Bakers, N. H. G. M. Peters, R. G. H. Beets-Tan, and H. J. W. L. Aerts, "Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR," *Sci. Rep.*, vol. 7, no. 1, p. 5301, Jul. 2017.

[17] J. Wang, J. Lu, G. Qin, L. Shen, Y. Sun, H. Ying, Z. Zhang, and W. Hu, "Technical note: A deep learning-based autosegmentation of rectal tumors in MR images," *Med. Phys.*, vol. 45, no. 6, pp. 2560–2564, Jun. 2018.

[18] D. Li, X. Chu, Y. Cui, J. Zhao, K. Zhang, and X. Yang, "Improved U-Net based on contour prediction for efficient segmentation of rectal cancer," *Comput. Methods Programs Biomed.*, vol. 213, Jan. 2022, Art. no. 106493.

[19] Y.-J. Huang, Q. Dou, Z.-X. Wang, L.-Z. Liu, Y. Jin, C.-F. Li, L. Wang, H. Chen, and R.-H. Xu, "3-D RoI-aware U-Net for accurate and efficient colorectal tumor segmentation," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5397–5408, Nov. 2021.

[20] P. Meng, C. Sun, Y. Li, L. Zhou, X. Zhao, Z. Wang, W. Lu, J. Li, and J. Sun, "MSBC-Net: Automatic rectal cancer segmentation from MR scans," TechRxiv, Tech. Rep., 2021.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[23] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 205–218.

[24] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual Swin transformer U-Net for medical image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[25] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A robust volumetric transformer for accurate 3D tumor segmentation," 2021, *arXiv:2111.13300*.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[28] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8768–8777.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[30] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[31] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, pp. 1–28, Dec. 2015.

[32] R. Cárdenes, R. de Luis-García, and M. Bach-Cuadra, "A multidimensional segmentation evaluation for medical image data," *Comput. Methods Programs Biomed.*, vol. 96, no. 2, pp. 108–124, Nov. 2009.

[33] O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Athens, Greece: Springer, 2016, pp. 424–432.

[34] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "nnFormer: Interleaved transformer for volumetric segmentation," 2021, *arXiv:2109.03201*.

[35] F. Isensee, P. F. Jäger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated design of deep learning methods for biomedical image segmentation," 2019, *arXiv:1904.08128*.

[36] L. Wei, W. Cui, Z. Hu, H. Sun, and S. Hou, "A single-shot multi-level feature reused neural network for object detection," *Vis. Comput.*, vol. 37, no. 1, pp. 133–142, Jan. 2021.

ZHENGUO SANG was born in Shandong, China, in 1998. He received the B.S. degree in engineering from Chang'an University, China, in 2020. He is currently pursuing the M.S. degree in biomedical engineering with Fudan University, China. His research interests include medical imaging diagnosis of rectal cancer diseases and medical image processing.

CHENGKANG LI was born in Anhui, China, in 1998. He is currently pursuing the Ph.D. degree in electronic information with Fudan University. He has published several research papers in international journals and conferences. His research interests include medical image analysis, deep causal learning, and automatic segmentation and diagnosis of colorectal cancer. His awards include the Huawei Scholarship and Outstanding Student Title from Fudan University.

YE XU received the Ph.D. degree in surgery from the Harvard Medical School, Brigham and Women's Hospital, USA, in 2022. She is currently a Chief Physician and a Doctoral Supervisor with the Department of Colorectal Surgery, Cancer Hospital, Fudan University. Her research interests include perioperative treatment of locally advanced colorectal cancer and diagnosis and effect evaluation of colorectal cancer with the aid of artificial intelligence.

YUANYUAN WANG (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronic engineering from Fudan University, Shanghai, China, in 1990, 1992, and 1994, respectively. From 1994 to 1996, he was a Postdoctoral Research Fellow with the School of Electronic Engineering and Computer Science, University of Wales, Bangor, U.K. In 1996, he joined the Department of Electronic Engineering, Fudan University, as an Associate Professor, where he was promoted to a Full Professor, in 1998. He has authored or coauthored six books and 500 research articles. His research interests include medical ultrasound techniques and medical image processing.

HONGTU ZHENG received the B.S. and master's degrees from the Shanghai Medical College, Fudan University. Since 2020, he has been an Associate Chief Physician with the Department of Colorectal Surgery, Cancer Hospital, Fudan University. His research interests include perioperative treatment of locally advanced colorectal cancer and diagnosis and effect evaluation of colorectal cancer with the aid of artificial intelligence.

YI GUO received the Ph.D. degree from Fudan University, China, in 2013. She is currently a Full Senior Engineer and a Doctoral Supervisor. She is also the Deputy Director of Fudan University's "Double First Class" Construction Office and the Biomedical Information Committee of Shanghai Biomedical Engineering Society. She is also engaged in interdisciplinary research at the intersection of medical imaging and artificial intelligence.