mathematics

MDPI

Article

LCANet: A Lightweight Context-Aware Network for Bladder Tumor Segmentation in MRI Images

Yixing Wang [1], Xiang Li [2] and Xiufen Ye [1,*]

[1] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China
[2] Center for Medical Artificial Intelligence, Shandong University of Traditional Chinese Medicine, Qingdao 266112, China
* Correspondence: yexiufen@hrbeu.edu.cn

Abstract: Accurate segmentation of the lesion area from MRI images is essential for diagnosing bladder cancer. However, the precise segmentation of bladder tumors remains a massive challenge due to their similar intensity distributions, various tumor morphologies, and blurred boundaries. While some seminal studies, such as those using CNNs combined with transformer segmentation methods, have made significant progress, (1) how to reduce the computational complexity of the self-attention mechanism in the transformer while maintaining performance and (2) how to build a better global feature fusion process to improve segmentation performance still require further exploration. Considering the complexity of bladder MRI images, we developed a lightweight context-aware network (LCANet) to automatically segment bladder lesions from MRI images. Specifically, the local detail encoder generates local-level details of the lesion, the lightweight transformer encoder models the global-level features with different resolutions, the pyramid scene parsing module extracts high-level and multiscale semantic features, and the decoder provides high-resolution segmentation results by fusing local-level details with global-level cues at the channel level. A series of empirical studies on T2-weighted MRI images from 86 patients show that LCANet achieves an overall Jaccard index of 89.39%, a Dice similarity coefficient of 94.08%, and a Class pixel accuracy of 94.10%. These advantages show that our method is an efficient tool that can assist in reducing the heavy workload of radiologists.

Keywords: bladder tumor segmentation; MRI images; transformer; multiscale; intelligent modeling

MSC: 68T02; 68U02

Citation: Wang, Y.; Li, X.; Ye, X. LCANet: A Lightweight Context-Aware Network for Bladder Tumor Segmentation in MRI Images. *Mathematics* **2023**, *11*, 2357. https://doi.org/10.3390/math11102357

Academic Editors: Fang Liu and Qianyi Liu

Received: 13 April 2023
Revised: 12 May 2023
Accepted: 15 May 2023
Published: 18 May 2023

1. Introduction

Bladder cancer originates in the cells of the bladder [1]. According to statistics, bladder cancer is the tenth most prevalent cancer globally, with approximately 573,000 new cases and 213,000 deaths in 2020, placing a heavy burden on patients, families, and society [2]. Achieving early diagnosis and treatment is essential for eliminating the incidence and mortality of bladder cancer. In current clinical practice, the gold standard for diagnosing bladder cancer is the transurethral resection of tumors and optical cystoscopy [3]. However, these methods are invasive and insensitive to minute tumors, which may delay diagnosis. As a noninvasive method with superior contrast, MRI plays an increasingly important role in the clinical diagnosis of bladder cancer [4]. In particular, T2-weighted imaging can provide structural images, which is advantageous in tumor segmentation [5]. The accurate segmentation of tumor areas from MRI images is essential for grading and staging bladder cancer. However, as shown in Figure 1, due to their similar intensity distribution, complex and variable tumor morphology, and blurred borders, it is difficult for radiologists to segment bladder cancer based only on this. Moreover, the manual segmentation process

is susceptible to the subjective experience of the physician, resulting in inconsistent segmentation results. Thus, developing an accurate image analysis method to assist physicians in accurately segmenting bladder cancer has become increasingly critical [6].

The use of bladder tumor segmentation algorithm has made some progress. Early algorithms resorted to traditional computer vision methods, such as Markov Random Fields [7], level-sets-based methods [8], or region growing [9]. For example, Duan et al. [10] proposed an adaptive window-setting scheme for segmenting bladder tumors in T1-weighted MRI images. However, the direct application of these techniques to segment bladder cancer is hardly satisfactory due to the complex distribution of tissues around the bladder.

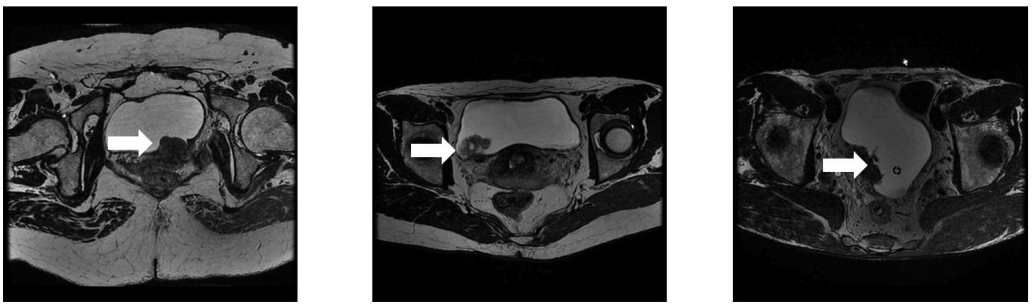

Figure 1. MRI images of various bladder tumors: The white arrow points to the tumor lesion. As seen in these images, the blurred borders, morphology of the tumor, and similar surrounding tissues severely affect the segmentation accuracy, especially for minute tumors.

Powerful nonlinear learning capability allows for full convolutional networks, such as UNet [11], to achieve extremely successful results in the field of medical image segmentation [12,13]. Inspired by this finding, many deep learning methods have been reported to segment bladder tumors from MRI images in recent years [14–17]. To continuously improve the segmentation results, researchers have proposed various optimization strategies, such as leveraging prior knowledge [18], enlarging the receptive field [19], multiscale feature utilization [20], and fusing global features [21]. For example, Dong et al. [22] proposed a deep network based on an attention UNet consisting of content and shape attention modules, which introduces a shape prior to ensuring closed segmentation results. The model was evaluated on T2-weighted images with a Dice similarity coefficient of 84%. In addition, prior information, such as tumor size and location, may help to improve the segmentation accuracy of bladder tumors. Huang et al. [18] proposed a semantic cancer segmentation network, fusing tumor size and location as the prior information, and the experimental results verify the effectiveness of combining prior knowledge. Li et al. [23] proposed a position-guided deformable UNet (PGD-UNet), which used deformable convolution to guide tumor segmentation. The above methods mostly use prior information as a constraint to optimize the network parameters. However, the segmentation results of such methods still differ significantly from those of manually annotated images due to the network's single receptive field.

Theoretically, dilated convolution provides a larger field of perception that can utilize more semantic features. The increased dilation rate allows for the use of multiscale features to better meet the segmentation requirements of small and large objects [24]. Dolz et al. [25] added dilated convolution to the UNet model, and the expansion rate of the dilated convolution gradually increased. The experimental results on T2-weighted MRI images show the effectiveness of introducing dilated convolution and achieving segmentation results that outperform traditional methods. Liu et al. [26] proposed the pyramid-in-pyramid network (PiPNet), which is composed of a pyramid backbone and atrous spatial pyramid pooling of four parallel atrous convolutions with increasing dilation rates. The model was evaluated on MRI images with a Dice similarity coefficient of 95%. Xu et al. [27] proposed a multiscale network based on dilated convolution (DMC-UNet), using dilated convolutions with different dilation rates to obtain different receptive fields. Pan et al. [28] used a stacked dilated UNet

with focal loss to accurately segment the bladder wall and tumor regions, with a mean Dice similarity coefficient of 66% for tumor region segmentation. However, the use of dilated convolution to obtain a larger field of perception does not fully deal with the perturbations caused by the surrounding tissue and blurred boundaries [29].

In terms of multiscale feature utilization, a more classical network is the PylNet network proposed by Zhang et al. [20]. The PylNet designs a novel pyramidal convolution block, which can extract images' multiscale feature information, resolve the loss of detail information in tumor regions of different shapes and sizes, and is more beneficial to the extraction of minute tumor regions. Simultaneously, the decoding structure of the algorithm is optimized, the complex upsampling convolution operation is not introduced in the decoding process, and the designed multiscale feature fusion module improves the decoding efficiency and effectively reduces the number of parameters. Unlike the above work, Yu et al. [5] proposed a cascade path-augmentation UNet (CPA-UNet) to mine multiscale features for segmenting tumor areas in T2-weighted images. The CPA-UNet achieves superior segmentation results regarding the Dice similarity coefficient of 87.40%. In addition, Wang et al. [30] proposed a dual-attention Vnet triple plus (DAVnet3+) network to accurately segment tumor regions. They used full-scale skip connections to fully utilize the information of multiscale feature maps. Then, they introduced spatial and channel attention to adaptively suppress irrelevant background regions in high-level semantic features. These methods further mitigate the interference of different factors and improve the accuracy of tumor lesion segmentation. However, as they are limited by the perceptual field of the convolutional kernel, the above works fail to establish explicit global dependency. In recent years, transformers and self-attention mechanisms have reformed the fields of computer vision and natural language processing [31]. SETR [32] is one of the earlier semantic segmentation networks that uses pure transformer encoders. However, pure transformer segmentation networks cannot satisfactorily extract detailed feature information [33]. Hence, to obtain segmentation results that are closer to the real situation, some segmentation networks that integrate CNNs and transformers have been proposed [34,35]. For example, He et al. [36] proposed a hybrid CNN-transformer network (HCTNet) to boost breast lesion segmentation in ultrasound images. He et al. [37] proposed an efficient hierarchical hybrid vision transformer (H2Former) for medical image segmentation. We proposed an auxiliary segmentation algorithm named MSEDTNet that integrates a multiscale encoder and decoder with a transformer [21]. Specifically, the multiscale encoder generates feature maps containing detailed local features. The transformer models the long-range dependency among high-level tumor semantics from a global space. A decoder with a spatial context fusion module fuses the contextual information and progressively produces high-resolution segmentation results. Experimental results from T2-weighted MRI scans show an overall Jaccard index of 83.46% and a Dice similarity coefficient of 92.35% for MSEDTNet. Although the segmentation performance of the network can be optimized via integrating CNNs and transformers, there are still some issues that need to be overcome. These are (1) how to reduce the computational complexity of the self-attention mechanism in the transformer while maintaining performance and (2) how to build a better means of global feature fusion to improve segmentation performance.

To alleviate the above challenge, we propose a lightweight context-aware (LCANet) framework for bladder tumor segmentation. The LCANet mainly consists of four parts: a local detail encoder, a lightweight transformer encoder, pyramid scene parsing, and a decoder. First, we use a local detail encoder to capture the local details of the tumor areas. Second, the lightweight transformer encoder is adapted to global model features with different resolutions, which can reduce the computational complexity of the self-attention mechanism while maintaining performance (question 1). Then, a special fusion manner was designed to integrate local detail features with global features (question 2). We also adopted the pyramid scene parsing to extract high-level and multiscale semantic features. Last, we used the decoder to obtain high-resolution segmentation results. Context-aware means that multiscale information is extracted, including coarse and fine-grained semantic

and long-range feature information, using a contextual fusion of both features. Extensive experiments demonstrate that LCANet leads to significant and consistent improvements in bladder tumor segmentation.

The main contributions are presented as follows:

- Differing from previous studies, we design a novel lightweight transformer encoder that models interpixel correlations at a fine-grained scale while reducing the computational complexity of the self-attention mechanism.
- A lightweight context-aware U-shaped network was developed to segment bladder tumors from MRI images. The network designs use special fusion to integrate local detail features with global features, which improves the segmentation accuracy of bladder lesions by learning multiscale global representations from bladder MRI images.
- Extensive experiments have shown that our method consistently improves the segmentation accuracy of bladder cancer lesions, which surpasses the robust baseline and state-of-the-art medical image segmentation methods.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed method in detail. In Section 3, we present the datasets and experimental settings. The abundant experimental results are shown in Section 4. Finally, a discussion and conclusions are given in Sections 5 and 6, respectively.

2. Methods

Figure 2 illustrates our proposed lightweight context-aware (LCANet) framework for bladder tumor segmentation. Our LCANet shares the same basic design as UNet, mainly including downsampling, upsampling, and skip connections. The difference is that we use ResUNet as the backbone to better capture the local-level details of the tumor (in Section 2.1). Then, we propose a lightweight transformer encoder to model global-level features with different resolutions (in Section 2.2). The pyramid scene parsing module is also utilized to extract high-level and multiscale semantic features (in Section 2.3). We provide high-resolution segmentation results by fusing local-level details with global-level cues at the channel level and using layer-by-layer upsampling convolution (in Section 2.4). The above modules are described in detail below.

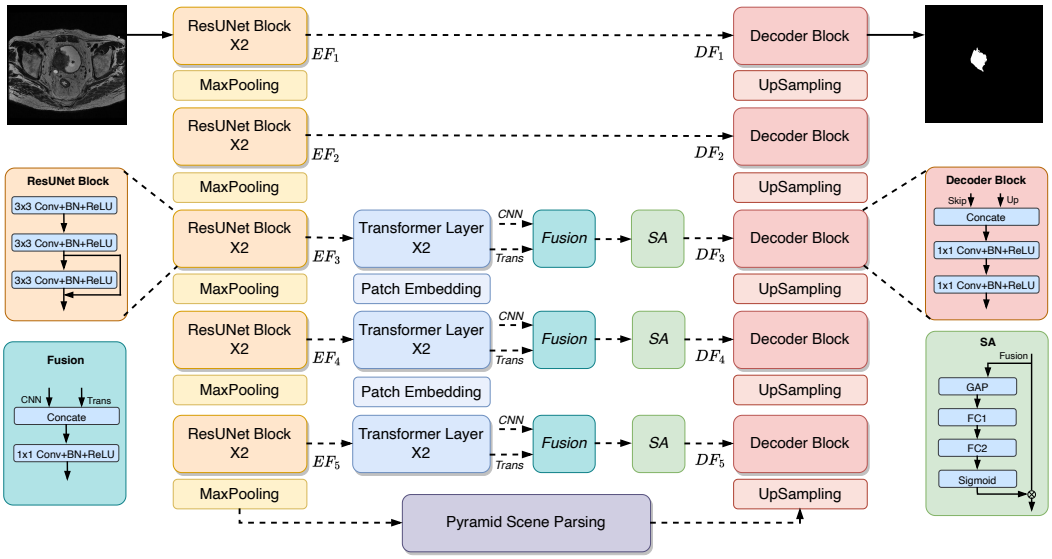

Figure 2. The description of the lightweight context-aware (LCANet) framework. The network is a U-shaped network that includes downsampling, skip connections, and upsampling operations.

2.1. Local Detail Encoder

Medical image segmentation is the division of medical images into separate parts based on the various attributes of pixels in computer vision tasks. As a low-level or pixel-level

vision task, in contrast to classification and object detection, the spatial information of the medical image is crucial when semantically segmenting various regions. Here, a local detail encoder is designed to segment MRI images of bladder cancer to better extract detailed semantic features in the tumor region. As shown in Figure 2, the local detail encoder consists of five ResUNet convolution blocks that extract fine-grained image feature information at various stages. The resolution of the feature map gradually decreases as the network depth increases, as does the number of channels in the feature map. Figure 2 depicts the network structure of the ResUNet convolution block, which includes the convolution layer, batch normalization layer, nonlinear activation function, and residual structure.

Specifically, given the layer $l-1$ output feature of the local detail encoder $x_{cnn}^{l-1} \in \mathbb{R}^{h \times w \times c}$, where c is the number of channels in the feature map and h and w are its dimensions, the layer l special feature map x_{cnn}^{l-1} is expressed as follows:

$$CFM_1 = ReLU(BN(Conv_{3\times3}(x_{cnn}^{l-1})) + b^l) \tag{1}$$

$$CFM_2 = ReLU(BN(Conv_{3\times3}(CFM_1)) + b^l) \tag{2}$$

$$CFM_3 = BN(Conv_{1\times1}(CFM_2)) + b^l) \tag{3}$$

$$x_{cnn}^{l} = ReLU(CFM_1 + CFM_3) \tag{4}$$

where the convolution kernel size is 3×3 for Formulas (1) and (2) and 1×1 for Formula (3). b^l is the bias term of the l th layer. $CFM_i \in \mathbb{R}^{h \times w \times c^{'}}$ is the feature map generated after convolution. $ReLU(\cdot)$ and $BN(\cdot)$ denote the nonlinear activation function and batch regularization operation, respectively.

2.2. Lightweight Transformer Encoder

Although the local detail encoder can extract detailed feature information from an image, its field of perception is still limited, and it cannot extract information from the entire image at once. Even though different convolutions can be cascaded to expand the field, the effective field occupies only a small fraction of the theoretical field, making the direct extraction of long-distance-dependent information impossible [38]. The feature vectors extracted by transformers usually contain more global contextual information [39]. Thus, the lightweight transformer encoder is designed to model the different scale spatial information of the local detail encoder output, as well as long-range correlation. Figure 2 depicts the structure of the lightweight transformer encoder, which is composed of three layers of transformer blocks. The transformer block of the first layer receives only the local detail encoder branch inputs from the same stage, whereas the transformer blocks of other layers not only receive the local detail encoder branch inputs from the current layer but also fuse the global information from the previous transformer blocks to better aggregate the coarse and fine-grained semantic information at different scales.

Here, let i indices represent the lightweight transformer encoder branch, and N_t represents the total number of layers for a lightweight transformer encoder. The stack of feature map x_{trans}^{i} is expressed as

$$x_{trans}^{i} = \begin{cases} Trans(x_{cnn}^{i}), i = 1 \\ Trans([x_{cnn}^{i}, PE(x_{trans}^{i-1})]), i = 2, \ldots, N_t \end{cases} \tag{5}$$

where $PE(\cdot)$ denotes the slice reassembly operation, which is implemented by a convolution operation with a step of 2, followed by batch regularization and the ReLU activation function. $[\cdot]$ and $Trans(\cdot)$ represent the cascading operation and transformer block, respectively. Each standard transformer block consists primarily of a dual-axis multihead self-attention (dual-axis MHSA) and a feed-forward network (FFN). Its structure is depicted in Figure 3a. The term multihead in dual-axis MHSA refers to the multiple independent self-attention

mechanism calculations, in which the input sequence is calculated in segments to obtain the segmented subspace representation, and then the segmented subspace representation is connected. Dual-axis MHSA improves calculation efficiency, serves an integrated function, keeps the model from overfitting, and prevents the model from paying too much attention to its position when encoding current position information [40].

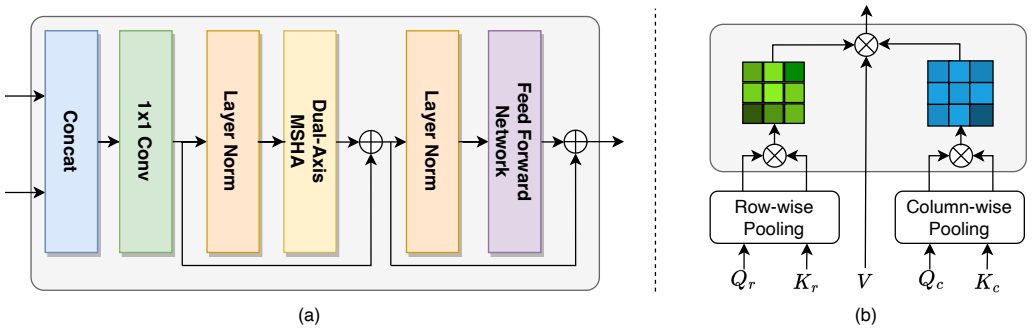

Figure 3. (**a**) An illustration of the transformer structure and (**b**) the dual-axis multihead attention mechanism.

As shown in Figure 3b, the input feature map $X \in \mathbb{R}^{H\times W\times C}$ is converted to Q, K, and V matrices to calculate the self-attention. The computational complexity of the self-attention in the original MHSA is $O(H^2W^2C)$. To reduce the computational complexity, we calculate relationships at the row and column level, rather than between pixels over the entire image size $(HW \times HW)$ [41]. First, to adaptively learn the representation of row-level and column-level relations, the Q and K matrices are divided into row-level matrices Q_r and K_r, and column-level matrices Q_c and K_c. Second, the feature map is transferred to the average pool operation in both directions. The relationship between the two directions is calculated so that the self-attention mechanism can be calculated for the whole feature map. The above process is expressed as follows:

$$A(Q,K,V) = (Soft_r(\frac{Q_{rp}K_{rp}^T}{\sqrt{d_k}}))V(Soft_c(\frac{Q_{cp}K_{cp}^T}{\sqrt{d_k}})) \tag{6}$$

where $Soft_r$ and $Soft_c$ are row-level self-attention and column-level self-attention, respectively. When calculating the row-level self-attention, the matrix V is reshaped into $H \times WC$. While using column-level self-attention, the matrix V is reshaped into $HC \times W$. In this way, the complexity of the algorithm is reduced to $O(H^2WC + HW^2C)$, which is less than the original MHSA computation complexity.

Formulaically, in the l layer transformer layer, given the input Z_{cnn}^{l-1} and Z_{trans}^{l-1}, the output Z_l is expressed as

$$Z_h = Conv_{1\times1}([Z_{cnn}^{l-1}, Z_{trans}^{l-1}]) \tag{7}$$

$$Z_h^{'} = MHSA(Norm(Z_h)) + Z_h \tag{8}$$

$$Z_l = FFN(Norm(Z_h^{'})) + Z_h^{'} \tag{9}$$

where $[\cdot]$ is the cascade operation. $Conv_{1\times1}$ is the 1×1 convolution operation. $Norm(\cdot)$, $MHSA(\cdot)$, and $FFN(\cdot)$ represent the normalization of the layer, dual-axis multihead attention mechanism, and feed-forward network, respectively.

2.3. Pyramid Scene Parsing

A large receptive field is necessary for the segmentation task, as a small receptive field is incapable of performing multiclass segmentation, and the segmented mask boundary

may be very rough [42]. After integrating the semantic information output by the local detail encoder and the lightweight transformer encoder, we also add the pyramid scene parsing module, which is composed of four different scale pooling functions. Please refer to [43] for the architecture of the pyramid scene parsing module. The application of pyramid scene parsing has the following two advantages: (1) It avoids the problem that the size of the CNN input image must be fixed, so the aspect ratio and size of the input image can be arbitrary. (2) Feature extraction and the reaggregation of features from different angles increase the accuracy of semantic segmentation tasks. Based on the original coarse-grained semantic feature information, four different scales of spatial feature information are fused to further strengthen the coding ability of multiscale features.

2.4. Decoder

As shown in Figure 2, to obtain high-quality segmentation results for bladder cancer MRI images in the original $(H \times W \times C)$ space, a decoder with a multilayer upsampling structure was designed to gradually recover the multiscale feature map output from the pyramid scene parsing module to the original input dimensions via a series of successive convolution and upsampling structures.

Lower-level features in the semantic segmentation task have more spatial information (segmentation localization features) and less semantic information (category judgment features), while higher-level features have less spatial information and more semantic information. As a result, when the decoder recovers bladder cancer MRI segmentation images, adding as much high-resolution information and low-resolution information as possible to the recovered features, i.e., combining low-resolution information in downsampling (to provide the basis for the tumor foreground and background category identification) and high-resolution information in upsampling (to provide the basis for accurate segmentation localization), and then complementing the undistributed information high-quality tumor segmentation results is critical. Thus, to increase the spatial resolution of the semantic information, we add several additional skip-connection structures to the encoder. Note that we add a spatial attention mechanism to the skip-connection structures corresponding to the local detail encoder and the lightweight transformer encoder, which can weigh the important feature maps to further highlight tumor feature information. The above process is expressed as follows:

$$x^l_{combind} = [x^l_{cnn}, z^l_{trans}] \tag{10}$$

$$value = FC2(FC1(GAP(x^l_{combind}))) \tag{11}$$

$$x^l_{combind} = x^l_{combind} \times Sigmoid(value) \tag{12}$$

where x^l_{cnn} and z^l_{trans} are the convolution result and the transformer operation result of the l layer, respectively. $GAP(\cdot)$ is the global average pooling operation function. $FC1(\cdot)$ and $FC2(\cdot)$ are two full connection layers.

The skip-connection structure is only used to aggregate the features extracted by the local detail encoder in the branch that does not use the lightweight transformer encoder. After the five-layer upsampling structure, the image is restored to the same size as the input image, the pixel values are compressed to the range [0, 1] using the sigmoid function, and the segmentation result of the bladder cancer lesion region is obtained using 0.5 as the threshold.

2.5. Loss Function

The LCANet is end-to-end-optimized using the binary cross-entropy function. In two-class segmentation tasks, the cross-entropy function is frequently utilized because

of its capacity to accurately reflect the distinction between the predicted mask and the ground-truth labels. The loss function is calculated as follows:

$$Loss = -\sum_{(a,b)} Y(a,b) log\hat{Y}(a,b) + (1 - Y(a,b)) log(1 - \hat{Y}(a,b)) \quad (13)$$

where $Y(a,b) \in [0,1]$ denotes the ground-truth label, and $\hat{Y}(a,b) \in [0,1]$ denotes the predicted mask.

3. Datasets and Experimental Settings

In this section, we first describe the dataset used for the experiments (in Section 3.1). Then, to verify the effectiveness of the modules and methods, we describe the experimental setup and implementation in detail (in Section 3.2). Finally, we introduce the experimental evaluation metrics (in Section 3.3).

3.1. Dataset

The data for this study were derived from the Affiliated Hospital of Shandong University of Traditional Chinese Medicine and include 86 patients and 1320 MRI images of bladder cancer. A GE Discovery MR 750 3.0T MRI was used to scan all patients, with slices having a 1 mm thickness and slice interval and a repetition time and echo time of 2500 ms and 135 ms, respectively. The image size was uniform at 224 × 224, and three experienced clinicians marked the tumor areas on each image using the labelme toolkit.

3.2. Experimental Settings

To verify the effectiveness of different components, we first conducted ablation experiments in three aspects: network architecture, position, and parameter. For architecture, we investigated the efficacy of different modules, such as lightweight transformer encoders, pyramid scene parsing, and fusion skip connections in the semantic segmentation task of bladder tumors with the specific experimental configurations shown in Table 1. The position aspect investigated the impact of different light transformer encoder positions on the segmentation performance, i.e., the lightweight transformer encoder was placed at EF_1, EF_2, and EF_3 or EF_2, EF_3, and EF_4, and discussed the effect of various configurations on the results. Finally, the parameter aspect investigated the impact of different numbers of transformer layers in the lightweight transformer encoder on the segmentation results.

Table 1. Ablation models.

Model	Description
BaseNet	Vanilla ResUNet baseline
BaseTNet	baseline + LTE
BaseTPSPNet	baseline + LTE + PSP
LCANet	baseline + LTE + PSP + FSK

LTE: lightweight transformer encoder; PSP: pyramid scene parsing; FSK: fusion skip connection.

To demonstrate the effectiveness of LCANet, we first compare it with five state-of-the-art bladder tumor segmentation methods, including PylNet [20], Res-UNet [44], MD-UNet [19], Dolz et al. [25], and MSEDTNet [21]. MSEDTNet is our preproposed bladder tumor segmentation algorithm. Second, LCANet is compared with the current general semantic segmentation algorithms, including UNet [11], Swin-UNet [33], DeepLabv3+ [45], and TransUNet [46]. UNet includes a batch normalization layer as well as a complementary zero operation at convolution to ensure that the output is the same size as the input. Swin-UNet is a U-shaped semantic segmentation algorithm with a Swin transformer structure for both the encoder and decoder, which can better learn an image's global and long-term semantic information and has a strong segmentation performance on COVID-19 lesions in CT Images [47]. DeepLabv3+ introduces atrous spatial pyramid pooling (ASPP) to integrate the bottom features with the top features to improve the segmentation boundary accuracy.

Here, we employed ResNet101 as the backbone network to extract features. TransUNet, with the advantages of both transformers and UNet, is a powerful alternative to medical image segmentation. TransUNet performs better than various competing methods in both multiorgan and heart segmentation. Here, we segmented the bladder tumor region with the help of a pretrained model.

All models were trained and tested on an Intel 6246 CPU and NVIDIA Tesla V100 32 GB GPU using the MMSegmentation [48] deep learning framework. The optimizer was set to stochastic gradient descent (SGD), and the initial learning rate was 0.01. The momentum and weight decay were 0.9 and 0.0005, respectively. Moreover, we utilized poly-police to update the learning rate in each round of training. The mini-batch size was 8, and the training epochs were 200 for convergence. We also used data enhancement strategies, such as random flipping, random rotation, and random crop, to optimize the overfitting phenomenon. Note that the proposed method did not require pretraining on any large datasets. In addition, for a fair comparison, all models were trained using our dataset according to the same setup. Furthermore, we partitioned the dataset using a fivefold cross-validation method to test the generalization performance of the proposed network. Specifically, in each fold of the training set, we randomly further divided the training set into four parts, leaving one part for the validation set. In this way, the training, validation, and test set ratios in each data fold were 3:1:1, respectively.

3.3. Evaluation Metrics

To quantitatively evaluate the segmentation performance of different methods on bladder tumors, we used three widely used segmentation metrics, the Jaccard index (*JI*), Dice similarity coefficient (*DSC*), and Class pixel accuracy (*CPA*), respectively. Larger values of these three metrics represent a more robust segmentation performance of the model. The calculation formulas are presented as follows:

$$JI = \frac{TP}{FP + TP + FN} \tag{14}$$

$$DSC = \frac{2TP}{FP + 2TP + FN} \tag{15}$$

$$CPA = \frac{TP}{TP + FP} \tag{16}$$

where TP refers to the number of positive samples correctly classified as positive, FP represents the number of negative samples incorrectly classified as positive, and FN denotes the number of positive samples incorrectly classified as negative. In addition, a paired Student's *t*-test was also performed to ensure that the experimental results were statistically significant compared with the state-of-the-art segmentation methods.

4. Results

In this section, we perform a series of ablation experiments on the proposed network (in Section 4.1). Then, we compare our method with the start-of-the-art deep learning methods (in Section 4.2). Finally, we report the experimental results.

4.1. Ablation Study

4.1.1. Ablation of the Different Module Architectures

To evaluate the performance of various network components, we conducted ablation experiments. Table 2 shows the experimental results for the different components. All results are expressed as means and standard deviations, and the best results are highlighted in bold. As shown in Table 2, the results indicate that these network components are vital to improving the segmentation performance. The pyramid scene parsing module generates the most significant performance improvement of 0.61% for the network in the Jaccard index. The role of the pyramid scene parsing module is to extract multiscale semantic

information about bladder tumors. This finding suggests that integrating multiple scales can ensure a better adaption to the shape and location changes in tumors and strengthen the algorithm's ability to extract multiscale high-level semantic feature information. For the Dice similarity coefficient metric, we found that the lightweight transformer encoder improves the segmentation results by as much as 1.20%. The results mean that the transformer effectively compensates for the traditional convolutional difficulties of modeling global relevance, further enriching the extracted coarse-grained high-level semantic features and ultimately improving the segmentation performance of the algorithm.

Table 2. Segmentation results of different network components. Red arrows indicate increases. The best results are marked with bold text.

Model	JI(%) ↑	DSC(%) ↑	CPA(%) ↑
BaseNet	88.03 ± 0.69	92.39 ± 0.44	93.19 ± 1.14
BaseTNet	88.51 ± 0.79	93.59 ± 0.59	93.77 ± 0.39
BaseTPSPNet	89.12 ± 0.54	93.68 ± 0.46	93.19 ± 0.61
LCANet	**89.39 ± 0.60**	**94.08 ± 0.37**	**94.10 ± 0.73**

JI: Jaccard Index; DSC: Dice Similarity Coefficient; CPA: Class pixel accuracy.

4.1.2. Ablation of the Lightweight Transformer Encoders on Various Stages

We investigated how the lightweight transformer encoder impacts the segmentation performance at different scales, ranging from EF_5 (the deepest layer with a size of 14 × 14) to EF_2 (the shallowest layer with a size of 112 × 112). EF_1 was not evaluated due to GPU capacity limitations. As shown in Figure 4a, the accuracy was improved by stacking lightweight transformer encoders, and the highest results were achieved when extracting local–global features in four consecutive stages (EF_5, EF_4, and EF_3). However, applying the lightweight transformer encoder to large-scale features, such as EF_2, increases the parameters and computational consumption of the model, which may cause overfitting problems. Therefore, according to the experimental results, we introduced a three-layer lightweight transformer encoder, which provides the best segmentation results and appropriate computational consumption. However, the location of this addition still needs to be further discussed.

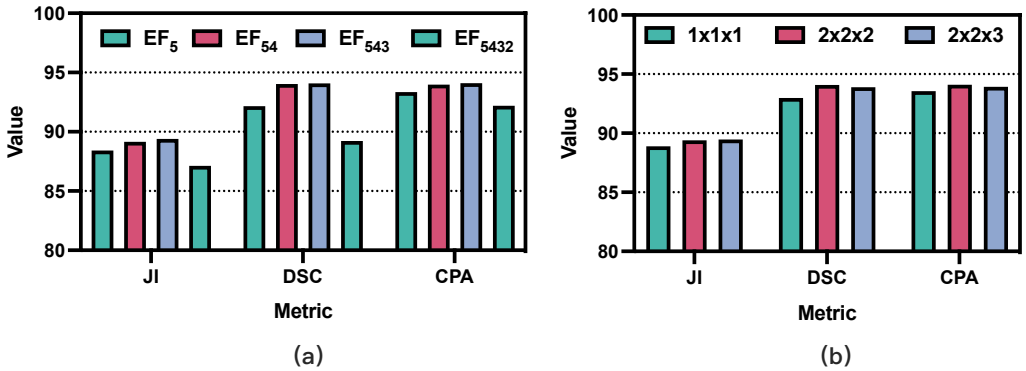

Figure 4. Segmentation results of different ablation experiments: (**a**) ablation results of the lightweight transformer encoders on various stages; (**b**) ablation results of the block numbers on different transformer encoders. JI: Jaccard Index; DSC: Dice Similarity Coefficient; CPA: Class pixel accuracy.

4.1.3. Ablation of the Three Lightweight Transformer Encoders at Various Stages

To further evaluate the effect of the lightweight transformer encoder's placement on segmentation performance, we analyzed the impact of different insertion locations on the segmentation results. Specifically, we placed the lightweight transformer encoders at EF_1, EF_2, and EF_3 and at EF_2, EF_3, and EF_4, respectively. The experimental results are shown in Table 3. From the experimental results, it can be seen that the stacked lightweight

transformer encoder improves segmentation performance, and its segmentation results increase with the stacked positions, i.e., the highest segmentation results are obtained when modeling global correlation in the EF_3, EF_4, and EF_5 stages. However, since the feature maps extracted by the shallow layers of the algorithm (e.g., EF_1 and EF_2) are only downsampled by 1–2 layers, their feature maps are still large in scale, and applying the transformer to EF_1, EF_2, and EF_3 will affect the performance and generate additional computational cost. Therefore, this paper achieves a relative balance between performance and computational complexity by applying the transformer to EF_3, EF_4, and EF_5.

Table 3. Segmentation results of different lightweight transformer positions. Red arrows indicate increases. The best results are marked with bold text.

Model	JI(%) ↑	DSC(%) ↑	CPA(%) ↑
EF_1, EF_2 and EF_3	87.92 ± 1.15	92.45 ± 0.54	92.19 ± 1.23
EF_2, EF_3 and EF_4	88.31 ± 0.84	93.62 ± 0.59	93.77 ± 0.93
EF_3, EF_4 and EF_5	**89.39 ± 0.60**	**94.08 ± 0.37**	**94.10 ± 0.73**

JI: Jaccard Index; DSC: Dice Similarity Coefficient; CPA: Class pixel accuracy.

4.1.4. Ablation of the Block Numbers on Different Transformer Encoders

Convolution is a local operation that usually only models the relationships between neighboring pixels. In contrast, a transformer is a global operation with a powerful ability to model the relationship between all pixels. The use of convolution combined with transformers can better represent the semantic features of an image [49]. Table 2 also verifies the effectiveness of the transformer structure. However, the trade-off between the number of layers and performance in the transformer remains disputed. We analyzed the impact of the number of layers in the transformer, and the experimental results are shown in Figure 4b. Note that K, M, and N represent the number of transformer layers in the first, second, and third lightweight transformer encoders, respectively. For example, $1 \times 1 \times 1$ means that each lightweight encoder uses only one transformer layer. The results show that the best segmentation performance was achieved by choosing two transformer layers in each lightweight transformer encoder. Thus, as the number of layers increases, the number of network parameters also gradually increases, but the performance gain decreases.

4.2. Comparison with State-of-the-Art Methods

To verify the segmentation performance of the proposed algorithm, we compared it with different algorithmic models. The experimental results are shown in Table 4. All results are presented as the mean and standard deviation, and the best results are highlighted in bold. Table 4 shows that our method achieves the best results on three evaluation metrics. The three evaluation index values of our method are 89.39%, 94.08%, and 94.10%, respectively. Compared with the worst results, these metrics are improved by 10.96%, 7.75%, and 11.08%. Compared with our preproposed MSEDTNet network, these metrics are improved by 0.57%, 0.74%, and 1.65%, respectively. To further validate the merits of our method, paired Student's t-tests with secondary results were performed to ensure that the experimental results were statistically significant. The results are shown in Table 4, and the p-value ($p < 0.05$) denotes a significant difference between our method and the comparison method. From the above analysis, it can be concluded that our method achieves superior performance in bladder tumor segmentation.

Table 4. Segmentation results of different competing methods. Red arrows indicate increases. The best results are marked with bold text. The green asterisk indicates that the difference between our method and the competitor is significant after the paired Student's *t*-test. (*: $p < 0.05$).

Model	JI(%) ↑	DSC(%) ↑	CPA(%) ↑
PylNet [20]	87.99 ± 0.61	93.20 ± 0.39	92.77 ± 0.83 *
Res-UNet [44]	87.42 ± 0.63	92.83 ± 0.41	92.08 ± 0.73
MD-UNet [19]	88.81 ± 0.59 *	92.92 ± 0.85	92.51 ± 1.11
Dolz et al. [25]	88.56 ± 0.37	93.56 ± 0.23 *	92.73 ± 0.51
MSEDTNet [21]	88.22 ± 0.59	93.34 ± 0.38	92.45 ± 1.40
UNet [11]	86.48 ± 1.25	92.20 ± 0.83	91.10 ± 2.08
DeepLabv3+ [45]	83.11 ± 1.16	87.38 ± 0.74	85.32 ± 2.13
TransUNet [46]	81.02 ± 1.36	90.87 ± 1.01	92.54 ± 1.32
Swin-UNet [33]	78.43 ± 1.17	86.33 ± 0.95	83.02 ± 1.63
LCANet	**89.39 ± 0.60**	**94.08 ± 0.37**	**94.10 ± 0.73**

JI: Jaccard Index; DSC: Dice Similarity Coefficient; CPA: Class pixel accuracy.

We also illustrate the floating point operations (FLOPs) and frames per second (FPS) of different segmentation methods in Figure 5. The FLOPs refer to the amount of computation and are used to measure the complexity of the model. The larger the value of each FLOPs, the higher the computational complexity required by the model. The FPS is defined in the image domain as the number of frames per second transmitted by the screen, which reflects the inference speed of the model. As shown in Figure 5, compared with MSEDTNet, our method achieved the best results in both FLOPs and FPS. Although convolution-based segmentation models, such as PylNet, require less computational effort than our method, the semantic segmentation performance of their network is shown to be unsatisfactory in Table 4. This finding illustrates the important role of transformers in driving performance improvements in segmentation networks and reflects our introduction of a new attention computation that can effectively reduce the complexity of the network and further improve its inference speed.

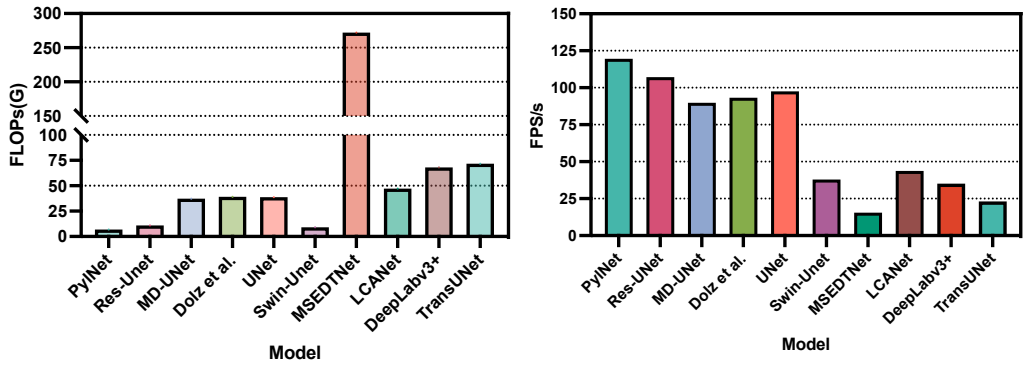

Figure 5. The FLOPs and FPS histogram of different segmentation methods: (**a**) the FLOPs histogram of different segmentation methods; (**b**) the FPS histogram of different segmentation methods.

Figure 6 displays the visual segmentation results of different methods. Compared with the segmentation results of other methods, LCANet not only effectively mitigates the perturbation of tumor size and surrounding tissues but also achieves segmentation results that are closer to ground-truth masks. In addition, LCANet can alleviate the influence of heterogeneous structures on the segmentation results and optimize the unsmooth edge regions as much as possible. Comprehensive evaluation results shows that LCANet achieves the best segmentation results in bladder tumor segmentation.

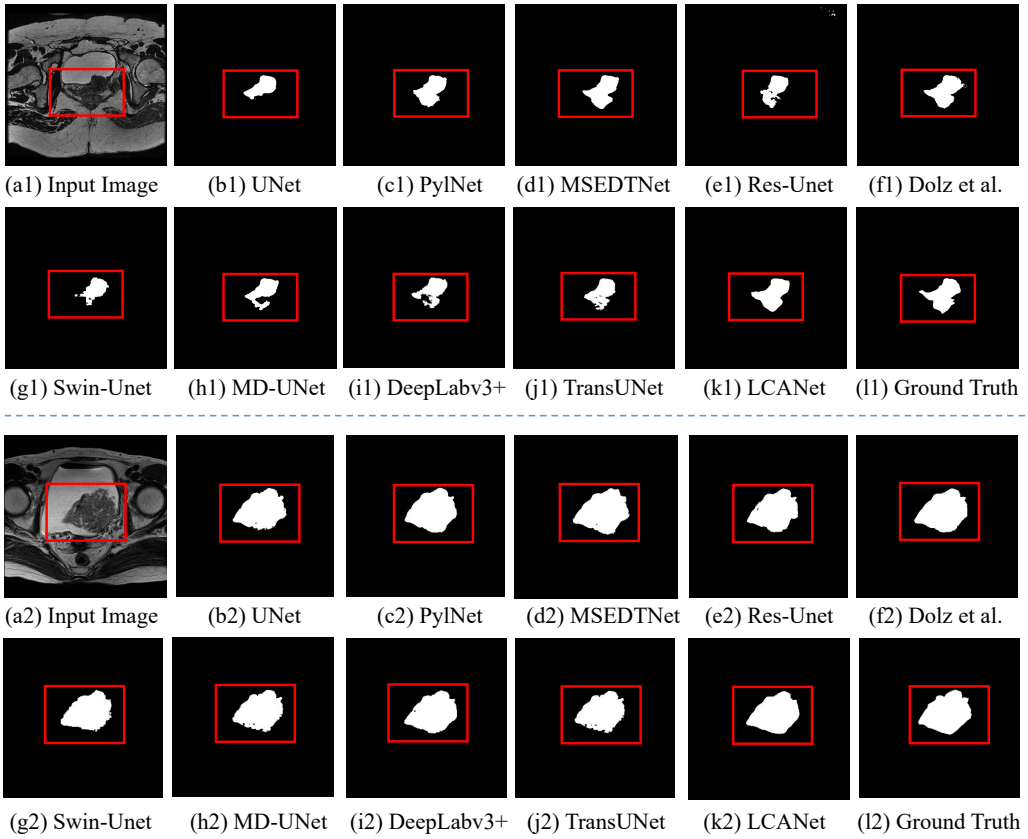

Figure 6. Segmentation results of different competing methods: (**a1**,**a2**) The original image of the input; (**b1**–**k1**) and (**b2**–**k2**) are images of UNet, PylNet, MSEDTNet, Res-UNet, Dolz et al. [25], Swin-UNet, MD-UNet, DeepLabv3+, TransUNet, and ours; (**l1**,**l2**) the ground-truth image. The red rectangle is the tumor area.

5. Discussion

In this paper, we propose the use of a lightweight context-aware network (LCANet) for the semantic segmentation task of bladder cancer. The network not only has a solid ability to extract local detail information but also captures the global semantic features of the context using a transformer encoder. In fusing local and global features, we added the channel attention module to the skip-connection structures to enhance the valuable feature for bladder tumor segmentation. In addition, the pyramid scene parsing module is introduced, aiming to achieve the feature extraction of tumor information at different scales. We conducted the ablation experiments to evaluate the effectiveness of network modules. According to the experimental results in Tables 2 and 3, we can observe that the setup of our network components resulted in optimal network performance in bladder tumor segmentation.

We also compared the performance of LCANet with current state-of-the-art methods. Based on the experimental results, we can draw several conclusions. In general, the segmentation performance of the networks based on UNet variants (such as Res-UNet) is better than that of the original UNet network, which suggests that using residual structures to fuse low-level features during the encoding phase facilitates the segmentation of bladder lesions. We observed that DeepLabv3+ achieved a suboptimal performance, possibly because the network has a robust feature representation and is prone to overfitting on simple bladder cancer segmentation tasks. We also observed that some improved networks (e.g., PylNet, MD-UNet, and Dolz et al. [25]) also achieved relatively good segmentation results, because these networks are modified as necessary for the properties of bladder tumor images, such as by adopting dilated convolution or atrous pyramid pooling to enhance the multiscale feature extraction ability of the network. From the segmentation

results of U-shaped networks with transformers (such as MSEDTNet), it can be concluded that transformers can also improve segmentation performance. However, Swin-UNet uses a transformer to replace the convolution operation, yielding poor segmentation results. This finding suggests that combining transformers with CNNs may improve the segmentation performance of bladder tumor images [50,51]. However, this process requires necessary modifications to fit the complex bladder tumor images according to the properties of the images [52]. For example, TransUNet achieved an unsatisfactory performance, which may be because the network is not explicitly designed for bladder cancer segmentation.

Our proposed method combines all the merits mentioned above, such as using CNNs to parse detailed features, using transformers to model global feature relationships, and using pyramid scene parsing to expand multiscale features. Using these modules, we achieved satisfactory results. Based on the different methods of FLOPs and the FPS histogram in Figure 5, we observed some interesting phenomena. Generally, the computational complexity required by the convolution-based approach is much lower than that of the transformer-based method. Thus, the FPS of the convolution-based method is much higher than that of the transformer-based method. Compared with MSEDTNet, LCANet improves the self-attention calculation and significantly reduces computational complexity.

According to the visual segmentation results shown in Figure 6, various methods can segment the tumor region. Specifically, the original UNet fails to model multiscale semantic features and global semantic relationships; thus, the boundary segmentation is coarse, especially for regions with muscle coverage. In contrast, PylNet and Res-UNet also obtain unsatisfactory results for image boundary segmentation, although multiscale low-level and high-level semantic feature fusion is considered. In addition, whereas MSEDTNet models global semantic relationships in bladder cancer images, it models the higher-level semantic information extracted by convolution and does not model interpixel correlations at the fine-grained scale of MRI images. Our method models interpixel and long-range pixel correlations from the fine-grained scale of MRI images and fully considers the role of skip connections fusing local and global information in bladder cancer segmentation, which can better describe the tumor region and generate more accurate tumor segmentation results.

Although the LCANet achieved the best segmentation performance on the bladder tumor task, Figures 5 and 6 show that LCANet still has some shortcomings: (1) maintaining segmentation performance while further reducing the complexity of the self-attention mechanism remains a challenging task; (2) obtaining accurate tumor boundaries remains a challenging task, especially in severe situations with similar intensity distributions of surrounding tissues. To address the above challenges, we will design more lightweight self-attention computation schemes to further reduce the complexity of self-attention while maintaining segmentation performance. Furthermore, we will introduce boundary loss functions to optimize the segmentation boundaries, which improves segmentation performance. In addition, this paper focuses on the task of tumor segmentation, and the considerations of future work will include the classification and staging of bladder cancer.

6. Conclusions

To better address the problem of bladder tumor segmentation, we designed a novel lightweight transformer encoder and used it to construct a lightweight context-aware network (LCANet) for bladder lesion segmentation. LCANet includes a local detail encoder to generate local-level details of the lesion, and a lightweight transformer encoder to model the global-level features with different resolutions, which can reduce the computational complexity of the self-attention mechanism while maintaining performance. Then, the pyramid scene parsing module extracts high-level and multiscale semantic features. The decoder provides high-resolution segmentation results by fusing local-level details with global-level cues at the channel level. Comprehensive experimental results demonstrate that LCANet performs better in terms of bladder tumor segmentation and can address more complex bladder lesion segmentation, showing its promise as a clinical tool for segmenting bladder tumors and reducing the heavy workload of radiologists. In addition, the LCANet

framework is scalable and sustainable, meaning that it can easily be extended to other tumor disease segmentation tasks.

Author Contributions: Investigation, X.Y.; methodology, Y.W. and X.L.; writing—original draft, Y.W. and X.L.; writing—review and editing, X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 42276187) and the Fundamental Research Funds for the Central Universities, China (Grant No. 3072022FSC0401).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Ethics Committee of the Affiliated Hospital of Shandong University of Traditional Chinese Medicine.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

JI	Jaccard Index
DSC	Dice Similarity Coefficient
CPA	Class pixel accuracy
FLOPs	Floating point operations
FPS	Frames per second
LTE	Lightweight transformer encoder
PSP	Pyramid scene parsing
FSK	Fusion skip connection
MHSA	Multihead self-attention
FFN	Feed-forward network

References

1. Tran, L.; Xiao, J.F.; Agarwal, N.; Duex, J.E.; Theodorescu, D. Advances in bladder cancer biology and therapy. *Nat. Rev. Cancer* **2021**, *21*, 104–121. [CrossRef] [PubMed]
2. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
3. Li, M.; Jiang, Z.; Shen, W.; Liu, H. Deep learning in bladder cancer imaging: A review. *Front. Oncol.* **2022**, *12*, 930917. [CrossRef] [PubMed]
4. Cha, K.H.; Hadjiiski, L.; Chan, H.P.; Weizer, A.Z.; Alva, A.; Cohan, R.H.; Caoili, E.M.; Paramagul, C.; Samala, R.K. Bladder cancer treatment response assessment in CT using radiomics with deep-learning. *Sci. Rep.* **2017**, *7*, 8738. [CrossRef]
5. Yu, J.; Cai, L.; Chen, C.; Fu, X.; Wang, L.; Yuan, B.; Yang, X.; Lu, Q. Cascade Path Augmentation Unet for bladder cancer segmentation in MRI. *Med. Phys.* **2022**, *49*, 4622–4631. [CrossRef]
6. Gandi, C.; Vaccarella, L.; Bientinesi, R.; Racioppi, M.; Pierconti, F.; Sacco, E. Bladder cancer in the time of machine learning: Intelligent tools for diagnosis and management. *Urol. J.* **2021**, *88*, 94–102. [CrossRef]
7. Li, L.; Liang, Z.; Wang, S.; Lu, H.; Wei, X.; Wagshul, M.; Zawin, M.; Posniak, E.J.; Lee, C.S. Segmentation of multispectral bladder MR images with inhomogeneity correction for virtual cystoscopy. In Proceedings of the Medical Imaging 2008: Physiology, Function, and Structure from Medical Images, San Diego, CA, USA, 17–19 February 2008; Volume 6916, pp. 279–283.
8. Duan, C.; Liang, Z.; Bao, S.; Zhu, H.; Wang, S.; Zhang, G.; Chen, J.J.; Lu, H. A coupled level set framework for bladder wall segmentation with application to MR cystography. *IEEE Trans. Med. Imaging* **2010**, *29*, 903–915. [CrossRef]
9. Garnier, C.; Ke, W.; Dillenseger, J.L. Bladder segmentation in MRI images using active region growing model. In Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, MA, USA, 30 August–3 September 2011; pp. 5702–5705.
10. Duan, C.; Yuan, K.; Liu, F.; Xiao, P.; Lv, G.; Liang, Z. An adaptive window-setting scheme for segmentation of bladder tumor surface via MR cystography. *IEEE Trans. Inf. Technol. Biomed.* **2012**, *16*, 720–729. [CrossRef]

11. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
12. Liu, Y.; Li, X.; Li, T.; Li, B.; Wang, Z.; Gan, J.; Wei, B. A deep semantic segmentation correction network for multi-model tiny lesion areas detection. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 1–9. [CrossRef]
13. Sohail, A.; Nawaz, N.A.; Shah, A.A.; Rasheed, S.; Ilyas, S.; Ehsan, M.K. A Systematic Literature Review on Machine Learning and Deep Learning Methods for Semantic Segmentation. *IEEE Access* **2022**, *10*, 134557–134570. [CrossRef]
14. Ali, N.; Bolenz, C.; Todenhöfer, T.; Stenzel, A.; Deetmar, P.; Kriegmair, M.; Knoll, T.; Porubsky, S.; Hartmann, A.; Popp, J.; et al. Deep learning-based classification of blue light cystoscopy imaging during transurethral resection of bladder tumors. *Sci. Rep.* **2021**, *11*, 11629. [CrossRef]
15. Wu, E.; Hadjiiski, L.M.; Samala, R.K.; Chan, H.P.; Cha, K.H.; Richter, C.; Cohan, R.H.; Caoili, E.M.; Paramagul, C.; Alva, A.; et al. Deep learning approach for assessment of bladder cancer treatment response. *Tomography* **2019**, *5*, 201–208. [CrossRef]
16. Yang, Y.; Zou, X.; Wang, Y.; Ma, X. Application of deep learning as a noninvasive tool to differentiate muscle-invasive bladder cancer and non–muscle-invasive bladder cancer with CT. *Eur. J. Radiol.* **2021**, *139*, 109666. [CrossRef]
17. Moribata, Y.; Kurata, Y.; Nishio, M.; Kido, A.; Otani, S.; Himoto, Y.; Nishio, N.; Furuta, A.; Onishi, H.; Masui, K.; et al. Automatic segmentation of bladder cancer on MRI using a convolutional neural network and reproducibility of radiomics features: A two-center study. *Sci. Rep.* **2023**, *13*, 628. [CrossRef]
18. Huang, X.; Yue, X.; Xu, Z.; Chen, Y. Integrating general and specific priors into deep convolutional neural networks for bladder tumor segmentation. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
19. Ge, R.; Cai, H.; Yuan, X.; Qin, F.; Huang, Y.; Wang, P.; Lyu, L. MD-UNET: Multi-input dilated U-shape neural network for segmentation of bladder cancer. *Comput. Biol. Chem.* **2021**, *93*, 107510. [CrossRef]
20. Zhang, N.; Zhang, Y.; Li, X.; Cong, J.; Li, X.; Wei, B. Segmentation algorithm of lightweight bladder cancer MRI images based on multi-scale feature fusion. *J. Shanxi Norm. Univ. (Nat. Sci. Ed.)* **2022**, *50*, 89–95.
21. Wang, Y.; Ye, X. MSEDTNet: Multi-Scale Encoder and Decoder with Transformer for Bladder Tumor Segmentation. *Electronics* **2022**, *11*, 3347. [CrossRef]
22. Dong, Q.; Huang, D.; Xu, X.; Li, Z.; Liu, Y.; Lu, H.; Liu, Y. Content and shape attention network for bladder wall and cancer segmentation in MRIs. *Comput. Biol. Med.* **2022**, *148*, 105809. [CrossRef]
23. Li, Z.; Pan, H.; Zhu, Y.; Qin, A.K. PGD-UNet: A position-guided deformable network for simultaneous segmentation of organs and tumors. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July; pp. 1–8.
24. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
25. Dolz, J.; Xu, X.; Rony, J.; Yuan, J.; Liu, Y.; Granger, E.; Desrosiers, C.; Zhang, X.; Ben Ayed, I.; Lu, H. Multiregion segmentation of bladder cancer structures in MRI with progressive dilated convolutional networks. *Med. Phys.* **2018**, *45*, 5482–5493. [CrossRef]
26. Liu, J.; Liu, L.; Xu, B.; Hou, X.; Liu, B.; Chen, X.; Shen, L.; Qiu, G. Bladder cancer multi-class segmentation in MRI with Pyramid-In-Pyramid network. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 28–31.
27. Xu, J.; Kang, L.; Han, W.; Jiang, J.; Zhou, Z.; Huang, J.; Zhang, T. Multi-scale network based on dilated convolution for bladder tumor segmentation of two-dimensional MRI images. In Proceedings of the 2020 15th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 6–9 December 2020; Volume 1, pp. 533–536.
28. Pan, H.; Li, Z.; Cai, R.; Zhu, Y. Accurate segmentation of bladder wall and tumor regions in MRI using stacked dilated U-Net with focal loss. In Proceedings of the MIPPR 2019: Parallel Processing of Images and Optimization Techniques; and Medical Imaging, Wuhan, China, 2–3 November 2019; Volume 11431, pp. 69–76.
29. Chen, G.; Li, L.; Dai, Y.; Zhang, J.; Yap, M.H. AAU-net: An Adaptive Attention U-net for Breast Lesions Segmentation in Ultrasound Images. *IEEE Trans. Med. Imaging* **2022**, *42*, 1289–1300. [CrossRef] [PubMed]
30. Wang, L.; Cai, L.; Chen, C.; Fu, X.; Yu, J.; Ge, R.; Yuan, B.; Yang, X.; Shao, Q.; Lv, Q. A novel DAVnet3+ method for precise segmentation of bladder cancer in MRI. *Vis. Comput.* **2022**, 1–13. [CrossRef]
31. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
32. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
33. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv* **2021**, arXiv:2105.05537.
34. Luthra, A.; Sulakhe, H.; Mittal, T.; Iyer, A.; Yadav, S. Eformer: Edge enhancement based transformer for medical image denoising. *arXiv* **2021**, arXiv:2109.08044.
35. Shen, J.; Lu, S.; Qu, R.; Zhao, H.; Zhang, L.; Chang, A.; Zhang, Y.; Fu, W.; Zhang, Z. A boundary-guided transformer for measuring distance from rectal tumor to anal verge on magnetic resonance images. *Patterns* **2023**, *4*. [CrossRef]
36. He, Q.; Yang, Q.; Xie, M. HCTNet: A hybrid CNN-transformer network for breast ultrasound image segmentation. *Comput. Biol. Med.* **2023**, *155*, 106629. [CrossRef]

37. He, A.; Wang, K.; Li, T.; Du, C.; Xia, S.; Fu, H. H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2023**. [CrossRef]
38. Song, Q.; Li, J.; Guo, H.; Huang, R. Denoised Non-Local Neural Network for Semantic Segmentation. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–13. [CrossRef]
39. Zhang, H.; Hu, W.; Wang, X. Parc-net: Position aware circular convolution with merits from convnets and transformer. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XXVI; Springer: Berlin/Heidelberg, Germany, 2022; pp. 613–630.
40. Azad, R.; Al-Antary, M.T.; Heidari, M.; Merhof, D. Transnorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model. *IEEE Access* **2022**, *10*, 108205–108215. [CrossRef]
41. Huang, H.; Xie, S.; Lin, L.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Tong, R. ScaleFormer: Revisiting the Transformer-Based Backbones from a Scale-Wise Perspective for Medical Image Segmentation. *arXiv* **2022**, arXiv:2207.14552.
42. Chalavadi, V.; Jeripothula, P.; Datla, R.; Ch, S.B. mSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions. *Pattern Recognit.* **2022**, *126*, 108548. [CrossRef]
43. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
44. Liang, Y.; Zhang, Q.; Liu, Y. Automated Bladder Lesion Segmentation Based on Res-Unet. In Proceedings of the 2021 Chinese Intelligent Systems Conference, Fuzhou, China, 16–17 October 2021; pp. 606–613.
45. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
46. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
47. Gao, Z.J.; He, Y.; Li, Y. A Novel Lightweight Swin-Unet Network for Semantic Segmentation of COVID-19 Lesion in CT Images. *IEEE Access* **2022**, *11*, 950–962. [CrossRef]
48. Contributors, M. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. 2020. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 10 April 2023).
49. Li, X.; Wei, B.; Li, T.; Zhang, N. MwoA auxiliary diagnosis via RSN-based 3D deep multiple instance learning with spatial attention mechanism. In Proceedings of the 2020 11th International Conference on Awareness Science and Technology (iCAST), Qingdao, China, 7–9 December 2020; pp. 1–6.
50. Fang, J.; Lin, H.; Chen, X.; Zeng, K. A hybrid network of cnn and transformer for lightweight image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1103–1112.
51. Alahmadi, M.D.; Alghamdi, W. Semi-Supervised Skin Lesion Segmentation with Coupling CNN and Transformer Features. *IEEE Access* **2022**, *10*, 122560–122569. [CrossRef]
52. Luo, X.; Hu, M.; Song, T.; Wang, G.; Zhang, S. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In Proceedings of the International Conference on Medical Imaging with Deep Learning, Zurich, Switzerland, 6–8 July 2022; pp. 820–833.