mathematics

MDPI

Article

MDER-Net: A Multi-Scale Detail-Enhanced Reverse Attention Network for Semantic Segmentation of Bladder Tumors in Cystoscopy Images

Chao Nie [1,2], Chao Xu [1,2,*] and Zhengping Li [1,2]

[1] School of Integrated Circuits, Anhui University, Hefei 230601, China; wb22101003@stu.ahu.edu.cn (C.N.); 04173@ahu.edu.cn (Z.L.)
[2] Anhui Engineering Laboratory of Agro-Ecological Big Data, Hefei 230601, China
* Correspondence: 04166@ahu.edu.cn; Tel.: +86-133-3919-9368

Abstract: White light cystoscopy is the gold standard for the diagnosis of bladder cancer. Automatic and accurate tumor detection is essential to improve the surgical resection of bladder cancer and reduce tumor recurrence. At present, Transformer-based medical image segmentation algorithms face challenges in restoring fine-grained detail information and local boundary information of features and have limited adaptability to multi-scale features of lesions. To address these issues, we propose a new multi-scale detail-enhanced reverse attention network, MDER-Net, for accurate and robust bladder tumor segmentation. Firstly, we propose a new multi-scale efficient channel attention module (MECA) to process four different levels of features extracted by the PVT v2 encoder to adapt to the multi-scale changes in bladder tumors; secondly, we use the dense aggregation module (DA) to aggregate multi-scale advanced semantic feature information; then, the similarity aggregation module (SAM) is used to fuse multi-scale high-level and low-level features, complementing each other in position and detail information; finally, we propose a new detail-enhanced reverse attention module (DERA) to capture non-salient boundary features and gradually explore supplementing tumor boundary feature information and fine-grained detail information; in addition, we propose a new efficient channel space attention module (ECSA) that enhances local context and improves segmentation performance by suppressing redundant information in low-level features. Extensive experiments on the bladder tumor dataset BtAMU, established in this article, and five publicly available polyp datasets show that MDER-Net outperforms eight state-of-the-art (SOTA) methods in terms of effectiveness, robustness, and generalization ability.

Keywords: bladder tumor segmentation; cystoscopy images; transformer; multi-scale; attention mechanism; computer-aided diagnosis

MSC: 68T07

Citation: Nie, C.; Xu, C.; Li, Z. MDER-Net: A Multi-Scale Detail-Enhanced Reverse Attention Network for Semantic Segmentation of Bladder Tumors in Cystoscopy Images. *Mathematics* **2024**, *12*, 1281. https://doi.org/10.3390/math12091281

Academic Editors: Radu Tudor Ionescu and Samaneh Mazaheri

Received: 24 March 2024
Revised: 12 April 2024
Accepted: 21 April 2024
Published: 24 April 2024

1. Introduction

Bladder cancer is one of the most common malignant tumors of the urinary system [1]. The early screening of bladder cancer is of great significance for the treatment of patients. Cystoscopy is currently the gold standard for screening bladder tumors. Under cystoscopy, urologists can detect bladder tumors and completely remove all visible tumors in the bladder through transurethral resection of bladder tumor (TURBT) [2]. However, the detection of bladder tumors heavily relies on visual examination by doctors, which is cumbersome and subjective, and it is estimated that up to 20% of bladder tumors are overlooked in cystoscopy [3]. Bladder tumor segmentation based on deep learning algorithms is an important technology in medical-assisted diagnosis, which can automatically and accurately locate tumors, help doctors improve detection accuracy, and save time and cost. However,

as shown in Figure 1, bladder tumors vary in shape and size, with uneven brightness; automatically and accurately segmenting bladder tumors is a challenging task.

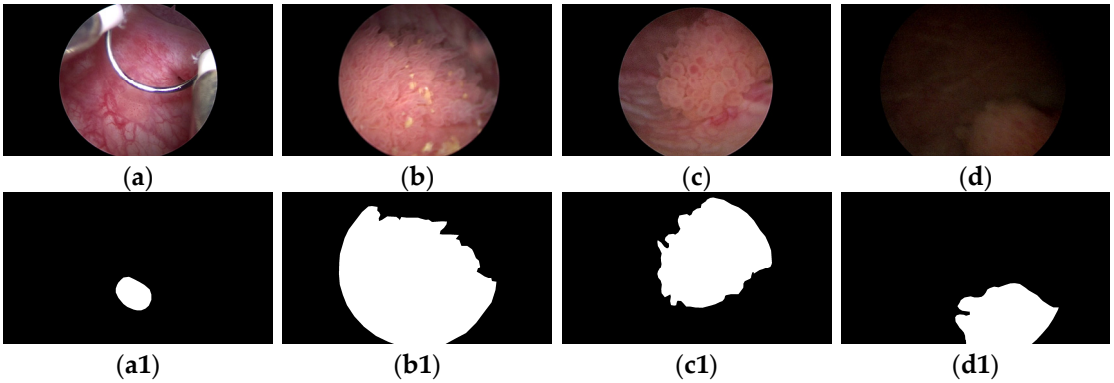

Figure 1. Cystoscopy images. The first row (**a**–**d**) is the original bladder tumor image, and the second row (**a1**–**d1**) is the corresponding ground truth (GT) image. (**a**–**d**) Different in shape and size; (**d**) uneven brightness.

With the rapid development of artificial intelligence technology, deep learning algorithms have been widely applied in medical image segmentation. The existing deep-learning-based medical image segmentation methods can be roughly divided into three types: CNN-based methods, Transformer-based methods, and CNN–Transformer hybrid methods.

Compared to traditional image segmentation methods, convolutional neural networks (CNNs) perform better. Shelhamer et al. [4] proposed fully convolutional networks (FCNs), which achieved excellent performance in image semantic segmentation tasks. Ronneberger et al. modified FCNs and proposed UNet [5], with a U-shaped architecture, which was the first medical image segmentation model based on an encoder–decoder structure. Subsequently, many UNet variant models (UNet++ [6], ResUNet [7], ResUNet++ [8], and DoubleUNet [9]) emerged to improve segmentation accuracy. However, as the feature scale of these models gradually decreases during the encoding stage, some detailed information is lost. Although attempts are made to supplement the lost detailed information during the decoding stage, the semantic gap between the encoder and decoder, as well as the issue of background noise, still exists. Recently, PraNet [10], ACSNet [11], HarDNet-MSEG [12], CaraNet [13], DCRNet [14], FTMFNet [15], and FRBNet [16] have gradually emerged, further improving the accuracy of medical image segmentation. These methods all use CNNs as the backbone to extract features and combine some fine modules for feature enhancement, performing well in extracting local detail information. However, because the convolution operation is essentially a local operation, these CNN-based methods have a relatively weak ability to capture global information and may result in incomplete segmentation results. Especially for lesions with significant differences in size, shape, and texture, CNN-based methods make it difficult to extract the appearance features of lesions and are prone to overfitting.

The emergence of Vision Transformer [17] overcame the limitations of CNNs in capturing remote dependencies. Unlike CNNs, Transformers use multi-head self-attention (MHSA) to capture remote dependencies in images, and then, generate global contextual information based on these remote dependencies. ViT [17] is the first transformer-based image recognition model. ViT divides each image into fixed-sized patches and models the remote dependencies of each patch to generate global information. Subsequently, PVT [18], Twins [19], Swin Transformer [20], and PVT v2 [21] gradually emerged, and these Transformer-based backbone networks further improved the segmentation performance. However, these backbone networks can only perform well when trained on large datasets, and their segmentation performance is limited when trained on small medical datasets. Recently, Poly-PVT [22], SSFormer [23], HSNet [24], MSRAformer [25], PVT-

CASCADE [26], and CAFE-Net [27] overcame this limitation by using transformer weights pre-trained on other non-medical big datasets, such as ImageNet [28]. Due to their expertise in capturing global contextual information, these methods perform well in medical image segmentation tasks. However, these Transformer-based methods have limitations in restoring fine-grained detail information and local boundary information of features and have limited adaptability to multi-scale features of lesions.

To combine the advantages of CNNs and Transformer, some CNN–Transformer hybrid methods [29–32] directly combine CNN and Transformer backbone networks. However, these methods only structurally combine CNNs and Transformer, ignoring the interaction between the two semantics. The more obvious disadvantage is the large computational load and high computational complexity.

To automatically and accurately segment bladder tumors in cystoscopy images, this paper proposes a new multi-scale detail-enhanced reverse attention network MDER-Net based on Transformer, aiming to capture the multi-scale global and local detail features of bladder tumors. We choose PVT v2 [21] as the encoder to capture global contextual information. Subsequently, the proposed multi-scale efficient channel attention module (MECA) is utilized to process the four different levels of features extracted by the encoder. Multi-scale feature information is extracted through convolutions of different kernel sizes, and trainable weight parameters and an ECA module [33] are used to suppress information gaps caused by different receptive fields, fuse feature information, enhance valuable information, and obtain multi-scale features with channel weight information to adapt to changes in the size and morphology of bladder tumors. Next, using a DA module [12], the second, third, and fourth layers of multi-scale advanced features are aggregated through step-by-step fusion to locate the approximate location of the bladder tumor and generate a rough initial segmentation map. Due to the rich texture, color, and edge detail information contained in low-level features, we use the SAM module [22] to fuse the feature map, which can roughly locate the bladder tumor with the lowest-level features, complementing each other in position and detail information, and generating a global prediction map containing detail information. Next, the proposed new detail-enhanced reverse attention module (DERA) is utilized to supplement fine-grained detail information from low-level features to high-level features, and then, extract local feature details from each level of fused features, gradually supplementing them to the global prediction map from the previous stage. In addition, before extracting local detail features from the lowest-level features, we propose a new and efficient channel space attention module (ECSA) to suppress irrelevant information in the bottom-level features, capture bladder tumor details from both the channel and spatial dimensions, enhance local context, and further extract edge detail information by inputting it into an RA module [34]. MDER-Net achieves accurate localization of bladder tumors by mixing six deep supervisions, ranging from locating the approximate location of tumors to supplementing detailed information to repairing tumor boundary information. On the bladder tumor dataset BtAMU, established in this article, our method outperforms other state-of-the-art (SOTA) segmentation methods in the mDice, mIoU, MAE, accuracy, F_β^ω, and HD metrics, with mDice and mIoU reaching 0.9108 and 0.8543, respectively; meanwhile, our method's visual segmentation results are also superior to other SOTA methods. On five publicly available polyp datasets (Kvasir-SEG [35], CVC- ClinicDB [36], CVC-ColonDB [37], ETIS [38], and CVC-300), our method consistently achieved SOTA segmentation performance in the mDice, mIoU, and HD metrics, especially on the CVC-300 dataset, where MDER-Net's mDice and mIoU results improved by 3.03% and 3.08%, respectively, on those of MSRAformer [25].

The main contributions of this article are as follows:

- A new Transformer-based network architecture MDER-Net is proposed, which can capture multi-scale global features of bladder tumors and enhance local feature representation.
- A new multi-scale efficient channel attention module (MECA) is proposed, which improves the network's multi-scale adaptability to lesions, enabling it to adapt to

various changes in the size and morphology of bladder tumors, and improving its generalization ability.
- A new detail-enhanced reverse attention module (DERA) is proposed, which restores fine-grained detail information and local boundary information of features, and can help the network generate prediction masks containing clear tumor boundaries, solving the problem of the Transformer's insufficient ability to recover local detail features.
- A new efficient channel space attention module (ECSA) is proposed, which can reduce the impact of noise and irrelevant information in low-level features, more effectively preserve bladder tumor details in different dimensions of low-level features, and improve segmentation performance.
- A new bladder tumor dataset, BtAMU, is established, which contributes to the development of state-of-the-art (SOTA) semantic segmentation algorithms on images captured by cystoscopys.

2. Related Works

2.1. CNN-Based Methods

In recent years, CNNs [10–16] have been widely used in medical image segmentation tasks. PraNet [10] abandoned the use of shallow features and utilized parallel partial decoders (PPDs) to aggregate advanced semantic features, using reverse attention (RA) modules [34] to mine boundary clues in advanced semantic features; however, PraNet [10] almost ignored global information. HarDNet-MSEG [12] uses HarDNet [39] as the backbone network, expands the receptive field through multi-branch extended convolutional layers, and combines it with the use of PPDs to achieve efficient inference speed in polyp semantic segmentation tasks. However, PraNet [10] and HarDNet-MSEG [12] only use high-level features and ignore the rich detailed information in shallow features. CaraNet [13] combines the CFP module and A-RA module to improve the segmentation performance related to small medical objects, enhancing edge information but ignoring contextual information at different scales. DCRNet [14] utilizes two parallel modules (ICR and ECR) to obtain contextual information within and between images, respectively. FTMF-Net [15] performs well on small medical objects by extracting more accurate boundary information through Fourier transform (FT) modules; however, FTMF-Net has limitations in feature aggregation. FRBNet [16] uses a boundary detection module (BD) to detect tumor boundaries in breast ultrasound images, and then, fuses the boundary information into the coarse prediction map through a feedback refinement module (FRM); however, since the FRM module uses Laplacian convolution operators to directly detect boundaries from low-level features extracted by the encoder, some low-level feature information unrelated to tumor boundaries is excessively extracted.

2.2. Transformer-Based Methods

Transformers [17–21] perform excellently in capturing remote dependencies. ViT [17] utilizes a multi-head self-attention mechanism to capture global contextual information between pixels. To reduce the computational cost of ViT [17], PVT [18] proposed a spatial reduction attention mechanism, and Swin Transformer [20] used sliding window operations to extract visual features at different levels. However, the self-attention used in PVT [18] and Swin Transformer [20] can lead to an insufficient ability to learn local contextual relationships. To overcome this limitation, PVT v2 [21] embeds convolutional layers between the fully connected layers of the feedforward network, but the ability to capture local context is still limited. In recent medical image segmentation research, Polyp-PVT [22] uses a cascaded fusion module (CFM) for advanced feature fusion, and then, uses a similarity aggregation module (SAM) to explore the relationship between advanced and low-level features for feature fusion. However, due to insufficient exploration of the local detail information of each stage feature output by the Transformer encoder, Polyp-PVT [22] has limitations in restoring the boundary detail information of lesions. SSFormer [23] proposed a progressive local decoder (PLD) to emphasize local features and limit attention dispersion.

However, PLD cannot fully recover fine-grained detail information. MSRAformer [25] uses Swin Transformer [20] as the backbone network and supplements boundary detail information through the spatial reverse attention module (SRA). However, due to the direct aggregation of high-level and low-level features in the feature aggregation stage, MSRAformer [25] ignores the semantic gap between high-level and low-level features, which may introduce noise and cannot effectively recover detailed information. HSNet [24], PVT-CASCADE [26], and CAFE-Net [27] all use PVT v2 [21] as the backbone network to extract four layers of features. Among them, CAFE-Net [27] uses FSEM modules to explore potential information in the features extracted by the encoder, while the CADM module effectively preserves lower-level features. Although these models perform well in medical image segmentation tasks, they all overlook the importance of restoring boundary details.

2.3. CNN–Transformer Hybrid Methods

Recently, some CNN–Transformer hybrid methods [29–32] have utilized Transformers to capture remote dependencies and a CNN to capture local contextual relationships between pixels. TransUNet [29] improves the encoder part based on the original UNet [5] network, transforming the original CNN encoder into a two-stage encoder structure, where the upper CNN is used to extract local information and the lower Transformer is used to extract global information. TGDAUNet [32] utilized Res2Net [40] and Swin Transformer [20] as dual-branch encoders to jointly extract features and designed an RGF module to capture nonimportant boundary features. However, these methods require a large amount of computation and are not suitable for medical image segmentation tasks that require real-time performance in clinical practice.

2.4. Segmentation Methods for Bladder Tumors

At present, research on using deep learning algorithms for bladder tumor segmentation [41–43] mainly focuses on CT [44,45] or MRI images [46–50], and there is relatively little research on tumor segmentation for cystoscopy images [3,51–54]. Shkolyar et al. [3] constructed a CNN-based image analysis platform, CystoNet, for automatic bladder tumor detection. Varnyu et al. [51] studied and analyzed the tumor segmentation performance of eight existing deep learning algorithms in cystoscopic images, proving that deep learning technology may be very useful in the real-time diagnosis and treatment of bladder cancer. Yoo et al. [52] used a Mask RCNN model with ResNeXt-101-32 $\times$ 8d-FPN as the backbone to segment tumors in white light and narrowband cystoscopy images. Zhang et al. [53] improved the UNet model by using a hybrid attention module to mine global information in tumor regions, and a guidance and fusion attention module to fuse low-level features of the encoder with high-level features of the decoder; the Dice of this model reached 82.7%. However, these CNN-based models cannot capture global information and find it difficult to extract the appearance features of tumors in bladder mirror images with significant differences in size, shape, and texture, resulting in poor segmentation performance. CystoNet-T [54] improved the tumor detection performance under cystoscopy by adding a Transformer encoder module to the pyramid layer of the feature pyramid network (FPN) and introducing a self-attention mechanism; however, the average precision of this model on the test set was only 91.4%. There is still a lot of room for improvement in the semantic segmentation algorithm for detecting bladder tumors from cystoscopy images.

3. Proposed Method

In this section, we first describe the overall architecture of the proposed MDER-Net, then provide a detailed introduction to the Transformer encoder PVT v2 [21] and the proposed new multi-scale efficient channel attention module (MECA), detail enhanced reverse attention module (DERA), and efficient channel spatial attention module (ECSA). Finally, we provide a loss function for training the network.

3.1. Overall Architecture

The overall architecture of our proposed MDER-Net is shown in Figure 2, which includes six modules: multi-scale effective channel attention module (MECA), dense aggregation module (DA) [12], similarity aggregation module (SAM) [22], detail-enhanced reverse attention module (DERA), efficient channel space attention module (ECSA), and reverse attention module (RA) [34].

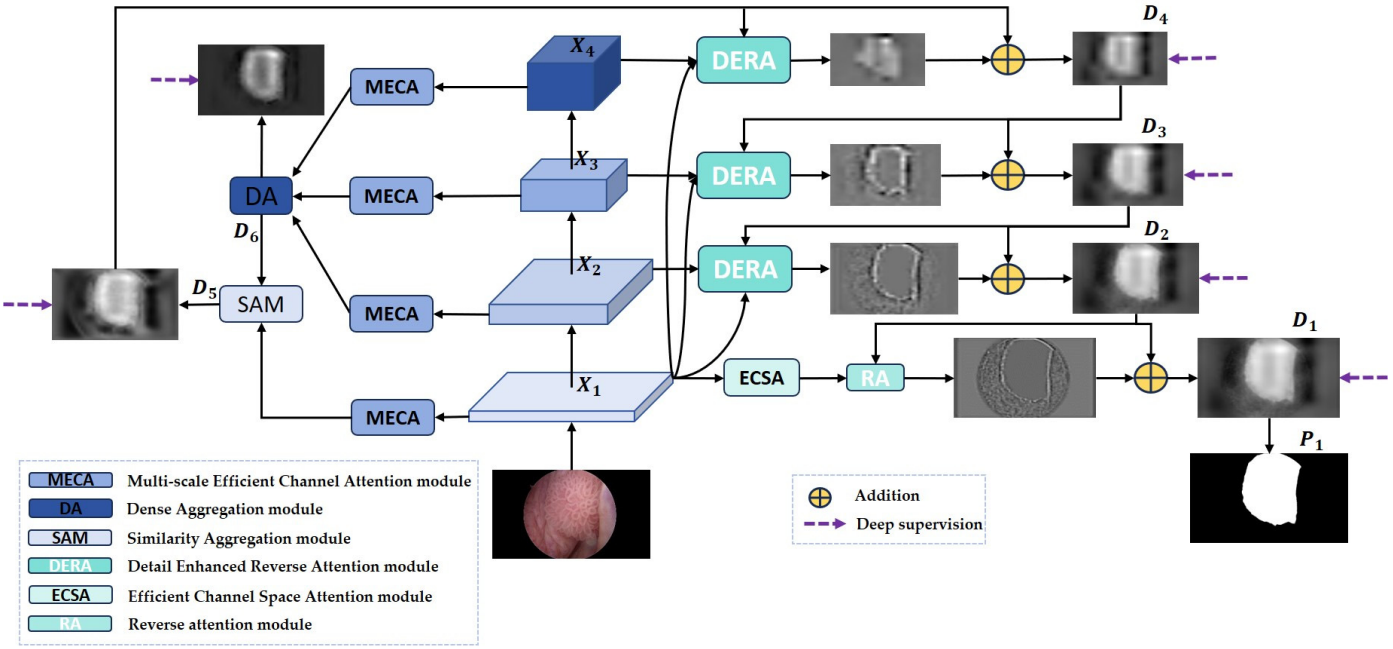

Figure 2. The architecture of the proposed MDER-Net, which consists of the backbone PVT v2, the multi-scale effective channel attention module (MECA), the dense aggregation module (DA), the similarity aggregation module (SAM), the detail-enhanced reverse attention module (DERA), the efficient channel space attention module (ECSA), and the reverse attention module (RA).

Specifically, given an input image $X \in R^{H\times W\times 3}$, we extract four pyramid features $X_i \in R^{\frac{H}{M_i}\times\frac{W}{M_i}\times C_i}$ from the PVT v2 [21] backbone, where $i \in \{1,2,3,4\}$, $M^i \in \{4,8,16,32\}$, $C_i \in \{64,128,320,512\}$, and C_i is the channel dimension of the ith layer. PVT v2, as an encoder, can capture global contextual information and establish remote dependency relationships. Then, we input four pyramid features X_i into the MECA module to obtain multi-scale features X_{Mi} with channel weight information; compared with X_i, X_{Mi} is more adaptable to changes in the size and morphology of bladder tumors. The DA module [12] receives three multi-scale advanced features, X_{M2}, X_{M3}, and X_{M4}, and aggregates them to generate an initial global prediction map D6. D6 can only capture the relatively rough position of the bladder tumor without structural details. Next, we send the coarse segmentation result D6 and the multi-scale low-level feature X_{M1}, containing rich texture, color, and edge detail information, to the SAM module [22], generating a global feature map D5 containing detailed information. Next, we input X_1, X_i, and $D_{i+1}(i = 2,3,4)$ into three DERA modules in sequence; X_1 is used to provide more low-level detail information to each high-level feature $X_i(i = 2,3,4)$. The DERA module uses the global feature map D_{i+1} from the previous layer to sequentially delete the current predicted bladder tumor area, capture fine-grained detail information and local boundary information, and then, integrate this information into D_{i+1} to obtain a predicted image D_2 containing the details of the bladder tumor edge structure. Next, we input the low-level feature X_1 into the ECSA module, and then, input the low-level features that suppress noise and irrelevant information and D_2 into the RA module [34]. We further refine the edge detail information and integrate it into D_2 to obtain the final predicted image D_1. During the training process,

we perform mixed supervision on the prediction map $D_i(i = 1, 2, 3, 4, 5, 6)$ generated in six stages. The overall network structure of MDER-Net is defined as follows:

$$X_i = PVT\ v2(X), (i = 1, 2, 3, 4) \tag{1}$$

$$X_{Mi} = MECA(X_i)(i = 1, 2, 3, 4) \tag{2}$$

$$D_6 = DA(X_{M2}, X_{M3}, X_{M4}) \tag{3}$$

$$D_5 = SAM(X_{M1}, D_6) \tag{4}$$

$$D_i = D_{i+1} + DERA(X_1, X_i, D_{i+1})(i = 2, 3, 4) \tag{5}$$

$$D_1 = D_2 + RA(ECSA(X_1), D_2) \tag{6}$$

3.2. Transformer Encoder PVT v2

Recent studies [55] have shown that Visual Transformers have a stronger ability to capture remote dependency relationships than CNNs. Inspired by this, we use PVT v2 [21] as an encoder to extract global information, obtaining four different levels of pyramid features from PVT v2. Among them, X_1 is considered a low-level feature, which contains rich texture, color, and edge details, as well as more noise and irrelevant information; X_2, X_3, and X_4 are considered high-level features that contain more feature information that can locate bladder tumors.

3.3. Multi-Scale Effective Channel Attention Module

Due to the variable morphology and size of bladder tumors in cystoscopy images, existing Transformer-based methods have limited adaptability to multi-scale features of bladder tumors. To enhance the multi-scale adaptive ability of the network for bladder tumors and accurately locate bladder tumors of different shapes and sizes, we propose a multi-scale effective channel attention module (MECA), as shown in Figure 3. Specifically, first, we extract multi-scale feature information from the multi-level feature $X_i(i = 1, 2, 3, 4)$ output by the encoder through convolutions with three different kernel sizes. Since convolutions of different kernel sizes determine the receptive field range on the cystoscopy image, compared to using convolutions of a single kernel, convolutions with three different kernel sizes can better capture global and local features. Next, when concatenating the feature information extracted by convolutions of different kernel sizes on the channel dimension, to suppress the information gap caused by receptive fields of different sizes we design two trainable weight coefficients w_1 and w_2, which are, respectively, multiplied by the features extracted by the convolutions of the two larger kernels. The ECA (efficient channel attention) [33] module is a local cross-channel interaction strategy that achieves dimensionality reduction through 1D convolution. It can strengthen valuable features and suppress irrelevant ones. Therefore, we use the ECA module to reassign weights to feature maps, so that the feature information that is conducive to segmenting bladder tumors receives attention from the network. At the same time, we use one 1×1 convolution and batch normalization operation on the input feature map $X_i(i = 1, 2, 3, 4)$. The 1×1 convolution reduces the number of feature channels to 32 to reduce computational resources. Finally, it is residually connected with multi-scale feature information with channel weight information to enhance the original features. In this way, the output feature X_{Mi} of the MECA module can adapt to changes in the size and shape of bladder tumors. This process can be described using Equation (7):

$$X_{Mi} = R(BN(Conv1(X_i)) + ECA(Conv1(Cat(M_{conv1}(X_i), w_1 \times M_{conv2}(X_i)\ , w_2 \times M_{conv3}(X_i))))) \tag{7}$$

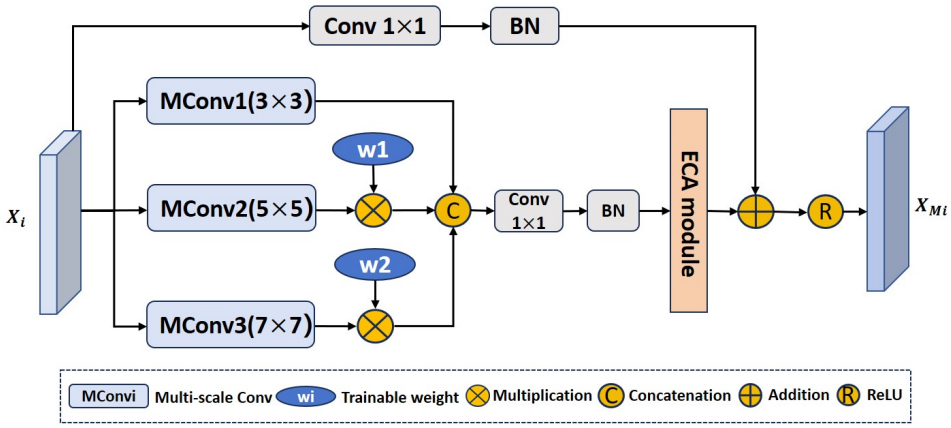

Figure 3. Structure of the proposed multi-scale efficient channel attention module (MECA).

In Equation (7), $M_{convi}(i = 1, 2, 3)$ represent convolutional layers with kernel sizes of 3×3, 5×5, and 7×7, respectively. w_1 and w_2 represent two trainable weight parameters, $Cat(\cdot)$ represents a connection operation in the channel dimension, $Conv1(\cdot)$ represents a convolutional layer with a kernel size of 1×1, $BN(\cdot)$ represents batch normalization, and $R(\cdot)$ represents the ReLU activation function.

3.4. Detail-Enhanced Reverse Attention Module

Due to the limited ability of Transformers to process local contextual information between pixels, existing Transformer-based methods have unclear tumor segmentation boundaries when predicting tumor areas in cystoscopy images. To obtain clear bladder tumor boundaries, we propose a new detail-enhanced reverse attention module (DERA) to capture inconspicuous boundary features and gradually explore supplementing tumor boundary feature information and fine-grained detail information. The DERA module consists of two parts: feature detail enhancement and local detail extraction. The structure of this module is shown in Figure 4.

- Feature detail enhancement: To supplement the fine-grained detail information contained in the low-level feature X_1 to high-level features $X_i(i = 2, 3, 4)$, we first use the predicted image D_{i+1} obtained in the previous stage to reduce the influence of background information; secondly, to preserve as much fine-grained detail information as possible in the low-level feature X_1, we use morphological dilation to expand the prediction area, and then, multiply the inflated prediction mask by the low-level feature X_1 to obtain the feature map X_1'. Next, we fuse X_1' with the high-level feature $X_i(i = 2, 3, 4)$ through downsampling and concatenation operations. Then, we extract and denoise the fused features through a 3×3 convolutional layer, and finally, use a 1×1 convolutional layer to recover the number of feature channels to obtain the feature map $X_i^e(i = 2, 3, 4)$ with enhanced details. This process can be summarized as Equation (8):

$$X_i^e = Conv1(BN(Conv3(Cat(X_i, X_1 \times Dilate(\sigma(D_{i+1})))))) \quad (8)$$

In Equation (8), $\sigma(\cdot)$ represents the sigmoid activation function used to generate prediction mask, $Dilate(\cdot)$ represents morphological dilation operations, $Cat(\cdot)$ represents connection operations in the channel dimension, $Conv3(\cdot)$ represents a convolutional layer with a kernel size of 3×3, $BN(\cdot)$ represents batch normalization, and $Conv1(\cdot)$ represents a convolutional layer with a kernel size of 1×1.

- Local detail extraction: The predicted image D_5 generated by the SAM module [22] lacks boundary details. We use the RA module [35] to extract local feature details from the detail-enhanced feature map $X_i^e(i = 2, 3, 4)$ and explore local boundary clues. The

RA module [34] first uses the prediction map D_{i+1} obtained in the previous stage to generate reverse attention weights R_i:

$$R_i = 1 - \sigma(D_{i+1}) \tag{9}$$

Then, the feature map $X_i^e(i = 2, 3, 4)$ with enhanced details is multiplied into the reverse attention weight $R_i(i = 2, 3, 4)$, the previously predicted tumor regions are deleted, the detailed information of the tumor boundaries is explored, and then, the subsequent feature exploration proceeds through three convolutional units to extract edge detail information and generate a feature map $DR_i(i = 2, 3, 4)$:

$$DR_i = Conv(X_i^e \times R_i) \tag{10}$$

We sequentially integrate $DR_i(i = 2, 3, 4)$ into the prediction map $D_{i+1}(i = 2, 3, 4)$ of the previous stage to refine the details of the edge structure of the bladder tumor. The DERA module can supplement fine-grained detail information and help the network generate prediction masks containing clear tumor boundaries.

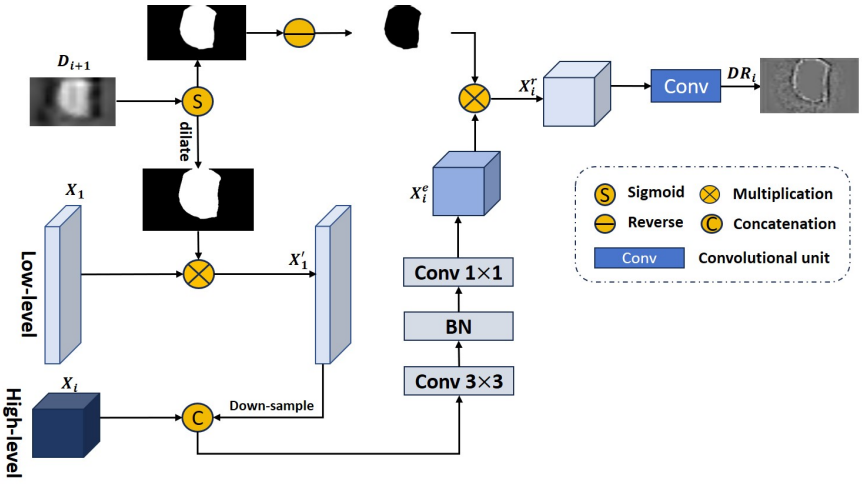

Figure 4. Structure of the proposed detail-enhanced reverse attention module (DERA).

3.5. Efficient Channel Space Attention Module

Due to the rich texture, color, and edge details contained in low-level features, as well as more noise and irrelevant information, to more effectively extract important information from low-level features and capture details of bladder tumors in different dimensions, we propose an efficient channel space attention module (ECSA). The module structure is shown in Figure 5; it refines the feature map by concatenating linear effective channel attention [24] $LECA(\cdot)$ and space attention [56] $SA(\cdot)$:

$$ECSA(X_1) = SA(ECA(X_1)) \tag{11}$$

Linear efficient channel attention $LECA(\cdot)$ identifies which feature maps to focus on, and then, assigns greater weights to these feature maps to enhance these features. The process of $LECA(\cdot)$ can be summarized as Equation (12):

$$LECA(X_1) = \sigma(R^t(Conv1d(R(AVG(X_1))))) * X_1 \tag{12}$$

In Equation (12), $\mathrm{AVG}(\cdot)$ represents the global average pooling layer, used to perform global information statistics by channel; $\mathrm{R}(\cdot)$ refers to the feature reshape operation, which converts the 2D tensor to 1D; $\mathrm{Conv1d}(\cdot)$ is a 1D convolutional layer with a kernel size of $\mathrm{k} = 5$ and a stride of $\mathrm{s} = 1$; $\mathrm{R^t}(\cdot)$ represents the inverse operation of $\mathrm{R}(\cdot)$; and $\sigma(\cdot)$ represents the sigmoid activation function, used to generate channel attention maps.

Space attention $SA(\cdot)$ identifies where to focus in a feature map, and then, gives these regions greater weights in the spatial direction to enhance those features. The process of $SA(\cdot)$ can be summarized as Equation (13):

$$\mathrm{SA}(x) = \sigma(Conv7(Cat(C_m(x), C_a(x)))) * x \tag{13}$$

In Equation (13), $\sigma(\cdot)$ represents the sigmoid activation function, used to generate spatial attention maps; $Conv7(\cdot)$ is a 7×7 convolutional layer with padding 3 to enhance spatial contextual information; $Cat(\cdot)$ represents the connection operation on the channel dimension; $C_m(\cdot)$ and $C_a(\cdot)$ represent the maximum and average values obtained along the channel dimension, respectively; and x represents the input tensor $LECA(X_1)$.

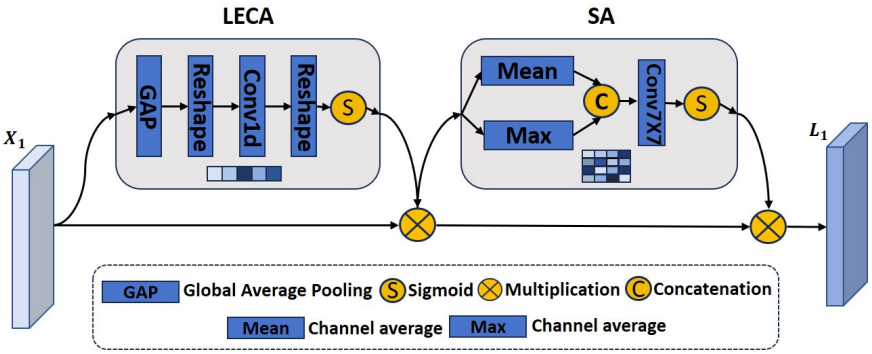

Figure 5. Structure of the proposed efficient channel space attention module (ECSA).

3.6. Loss Function

We used a prediction head for each of the six stages in the proposed MDER-Net, including $D_i(i = 1, 2, 3, 4, 5, 6)$, using addition aggregation to obtain the final prediction mask:

$$output = \sum_{i=1}^{6} P_i \tag{14}$$

In Equation (14), P_i is the prediction mask obtained by applying the sigmoid activation function to D_i, and the $output$ is the final prediction mask.

To monitor the prediction quality of the six stages of the network, we designed a multi-stage joint loss function, which is defined as

$$L_{overall} = \sum_{i=1}^{6} L_i \tag{15}$$

$$L_i = L_{wiou}(D_i, G) + L_{wbce}(D_i, G) \tag{16}$$

In the above equation, $L_i(i = 1, 2, 3, 4, 5, 6)$ represents the loss function for each stage, which is composed of weighted intersection over union (wiou) [57] and weighted binary cross-entropy (wbce) [57]; wiou and wbce limit the prediction mask from global and local perspectives, respectively; G represents the ground truth.

Finally, we provide the training process of the proposed MDER-Net algorithm, as shown in Algorithm 1.

Algorithm 1: The training process of the proposed MDER-Net algorithm.

Input: Image set I_1, I_2,.....I_n, Label(ground truth) set G_1, G_2,.....G_n
Output: Prediction maps D_i, Model parameters $\emptyset$

While not converging **do**
  Sample I_i, G_i from I_1, I_2,.....I_n, G_1, G_2,.....G_n
  Acquire feature maps of four different levels
  X_1,X_2,X_3, $X_4 = Transformer\ Backbone - PVT\ v2(I_i)$
  for feature maps X_i **do**
    for $i = 1$ to 4 **do**
      Use Equation (7) to obtain multi-scale features X_{Mi}
    end for
  end for
  Multi-scale high-level features aggregation $D_6 = DA(X_{M2}, X_{M3}, X_{M4})$
  Multi-scale fusion of high-level and low-level features $D_5 = SAM(X_{M1},\ D_6)$
  for feature maps X_i **do**
    for $i = 2$ to 4 **do**
      Use Equation (8) to obtain feature maps X_i^e with enhanced details
      Calculate reverse attention weights R_i by Equation (9)
      Use Equation (10) to generate edge detail feature map DR_i
      Generate prediction map $D_i = D_{i+1} + \mathrm{D}R_i$
    end for
  end for
  for feature map X_1 **do**
    Generate linear effective channel attention maps LX_1 by Equation (12)
    Generate space attention maps SX_1 by Equation (13)
    Edge detail information extraction $\mathrm{D}R_1 = RA(SX_1, D_2)$
    Generate the final prediction map $D_1 = D_2 + \mathrm{D}R_1$
  end for
  Use Equations (15) and (16) to calculate the total loss $L_{overall}$
  The Adam optimizer and the loss $L_{overall}$ to update the model parameters $\emptyset$
end while

4. Experimental Results and Discussion

In this section, we first introduce the dataset, evaluation metrics, and implementation details. Then, we compare the results of the proposed MDER-Net with state-of-the-art (SOTA) methods to demonstrate its superiority. We also conduct ablation experiments to demonstrate the effectiveness of our proposed three new modules.

4.1. Datasets

4.1.1. Bladder Tumor Dataset: BtAMU

In the field of bladder tumor segmentation, there is a lack of a cystoscopy image dataset for comparative evaluation. Therefore, we have established a bladder tumor dataset, BtAMU. Specifically, BtAMU consists of 1948 bladder tumor images and their corresponding ground truth (GT) labels extracted from 110 cystoscopy examinations and surgical videos provided by the Department of Urology at the First Affiliated Hospital of Anhui Medical University; the GT is manually annotated by professional urology experts. The image resolution is 1920 × 1080.

4.1.2. Polyp Dataset

To verify the robustness and generalization of the proposed MDER-Net, we selected five publicly challenging polyp datasets, including Kvasir-SEG [35], CVC-ClinicDB [36], CVC-ColonDB [37], ETIS [38], and CVC-300.

Kvasir-SEG: This dataset consists of 1000 polyp images extracted from colonoscopy videos, with a resolution distribution range of 332 × 487 to 1920 × 1072.

CVC-ClinicDB: This dataset consists of 612 polyp images extracted from 25 colonoscopy videos, with an image resolution of 384 × 288.

CVC-ColonDB: This dataset consists of 380 polyp images extracted from 15 colonoscopy videos, with an image resolution of 574 × 500.

ETIS: This dataset consists of 196 polyp images extracted from 34 colonoscopy videos, with an image resolution of 1225 × 996.

CVC-300: This dataset is a subset of the polyp dataset EndoScene [58], containing 60 images of colorectal polyps with a resolution of 574 × 500.

As shown in Table 1, when conducting bladder tumor segmentation experiments using the BtAMU dataset we randomly selected 80% of the dataset for training and 20% for testing. For the polyp segmentation task, we use the same data distribution settings as PraNet [10]. Specifically, 90% of the Kvasir-SEG dataset and 90% of the CVC-ClinicDB dataset are used for training, and 10% of the Kvasir-SEG dataset and 10% of the CVC-ClinicDB dataset are used for testing. In addition, to evaluate the generalization performance of the model on unseen CVC-ColonDB, ETIS, and CVC-300 datasets, we trained using 0% of the CVC-ColonDB, ETIS, and CVC-300 datasets and tested 100% of the CVC-ColonDB, ETIS, and CVC-300 datasets. In other words, these three datasets were not used for training the model, but only for testing the model.

Table 1. The image size, quantity, and division of training and testing datasets for the bladder tumor dataset BtAMU and five polyp datasets.

Dataset	Image Size	Image Number	Train Number	Test Number
BtAMU	1920 × 1080	1948	1562	386
Kvasir-SEG [35]	Variable	1000	900	100
CVC-ClinicDB [36]	384 × 288	612	550	62
CVC-ColonDB [37]	574 × 500	380	0	380
ETIS [38]	1225 × 996	196	0	196
CVC-300	574 × 500	60	0	60

4.2. Evaluation Metrics

In the bladder tumor segmentation experiment, we selected mean Dice (mDice), mean intersection over union (mIoU), mean absolute error (MAE), Accuracy, weighted Fmeasure (F_β^ω), and Hausdorff distance (HD) as evaluation metrics for quantitative analysis. In the polyp segmentation task, we used mDice, mIoU, and HD as evaluation metrics. mDice and mIoU are similarity measures at the regional level, representing the degree of agreement between the predicted and actual results of the model. MAE is a pixel-level evaluation metric that represents the average absolute error between the algorithm's predicted value and the true value. Accuracy represents the proportion of pixels correctly predicted by the model to actual pixels. F_β^ω is a metric that comprehensively considers precision and recall, and it trades them off through a parameter β. HD is used to measure the accuracy of boundary segmentation. Among these metrics, the higher the value of the mDice, mIoU, accuracy, and F_β^ω metrics, the better the algorithm's prediction performance, while the lower the values of the MAE and HD metrics, the better the model's segmentation performance. The metrics' definitions are as follows:

$$mDice = \frac{1}{N+1}\sum_{i=0}^{N}\frac{2TP}{FP+2TP+FN} \tag{17}$$

$$mIoU = \frac{1}{N+1}\sum_{i=0}^{N}\frac{TP}{FP+TP+FN} \tag{18}$$

$$MAE = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}|P(x,y) - Y(x,y)| \tag{19}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{20}$$

$$\begin{cases} Precision = \frac{TP}{FP + TP} \\ Recall = \frac{TP}{TP + FN} \\ F_\beta^\omega = \frac{(\beta^2 + 1)Precision^\omega \cdot Recall^\omega}{\beta^2 \cdot Precision^\omega + Recall^\omega} \end{cases} \tag{21}$$

$$\begin{cases} h(A, B) = max_{a \in A}\, min_{b \in B} \|a - \mathrm{b}\| \\ HD(A, B) = \frac{1}{N+1} \sum_{i=0}^{N} max(h(A, B), h(B, A)) \end{cases} \tag{22}$$

In the above equations, TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false positive; N represents the number of test images. In Equation (19), P represents the predicted map, Y represents the ground truth, and W and H are the width and height of the images, respectively. In Equation (20), β is a parameter, $Precision^\omega$ and $Recall^\omega$ denote weighted precision and weighted recall, respectively. In Equation (22), A and B represent the true image and predicted map, respectively, and $\|a - \mathrm{b}\|$ represents a distance function, such as the Euclidean distance.

4.3. Implementation Details

We implement MDER-Net on an NVIDIA GeForce RTX 3090 GPU card using PyTorch. When training, the optimizer is the Adam optimizer, the learning rate is set to 10^{-4}, the decay rate is set to 0.1, and the batch size is set to 16. The proposed network is trained for a total of 100 epochs. In addition, we adjust the size of the input image to 352×352 and use a multi-scale {0.75, 1.0, 1.25} training strategy with a gradient clip limit of 0.5.

4.4. Experimental Results and Discussion on the BtAMU Dataset

4.4.1. Quantitative Results

To verify the effectiveness of the proposed MDER-Net for bladder tumor segmentation, we compared the quantitative results of the proposed MDER-Net with the results of UNet [5], PraNet [10], CaraNet [13], HarDNet-MSEG [12], DCRNet [14], MSRAformer [25], HSNet [24], and TGDAUNet [32], and quantitatively evaluated the segmentation performance of all models on the bladder tumor dataset BtAMU using six medical image segmentation evaluation metrics: mDice, mIoU, MAE, accuracy, F_β^ω, and HD. Table 2 presents a comparison of the quantitative results of the different algorithms on the BtAMU dataset. From Table 2, it can be seen that our proposed MDER-Net achieves state-of-the-art segmentation performance on the BtAMU dataset compared to other models. In terms of the mDice and mIoU metrics, our MDER-Net improved by 1.22% and 1.28%, respectively, compared to the second-best-performing MSRAformer [25], indicating that MDER-Net can better distinguish bladder tumors from normal tissue backgrounds. The accuracy and F_β^ω of MDER-Net were also improved by 0.18% and 0.50%, respectively, compared to the second-best-performing MSRAformer [25]. In addition, MDER-Net showed a decrease of 0.24% and 0.3454 in the MAE and HD indicators, respectively, compared to the second-best-performing HSNet [24]. The HD metric quantitative results showed that MDER-Net improved the accuracy of bladder tumor boundary segmentation and reduced boundary segmentation errors. To improve the clarity of the quantitative results comparison between different models on the BtAMU dataset, we designed bar charts for the mDice and mIoU metrics, as shown in Figure 6. These results validate the effectiveness and superiority of the proposed MDER-Net for bladder tumor segmentation in cystoscopy images.

Table 2. Comparison of quantitative results of our model MDER-Net with other state-of-the-art (SOTA) methods on the BtAMU dataset. ↑ indicates higher is better, ↓ indicates lower is better, and bold indicates the best-performing result.

Methods	mDice ↑	mIoU ↑	MAE ↓	Accuracy ↑	F_β^ω ↑	HD ↓
UNet [5]	0.8241	0.7572	0.0277	0.9729	0.9043	11.8904
PraNet [10]	0.8670	0.8093	0.0241	0.9774	0.9305	11.1110
CaraNet [13]	0.8253	0.7641	0.0385	0.9664	0.9147	12.2071
HarDNet-MSEG [12]	0.8594	0.8003	0.0254	0.9763	0.9213	11.2059
DCRNet [14]	0.8592	0.8008	0.0309	0.9721	0.9281	11.3585
MSRAformer [25]	0.8986	0.8415	0.0205	0.9809	0.9460	10.7913
HSNet [24]	0.8982	0.8411	0.0197	0.9808	0.9455	10.7512
TGDAUNet [32]	0.8951	0.8330	0.0202	0.9798	0.9451	10.7861
Ours	**0.9108**	**0.8543**	**0.0173**	**0.9827**	**0.9510**	**10.4058**

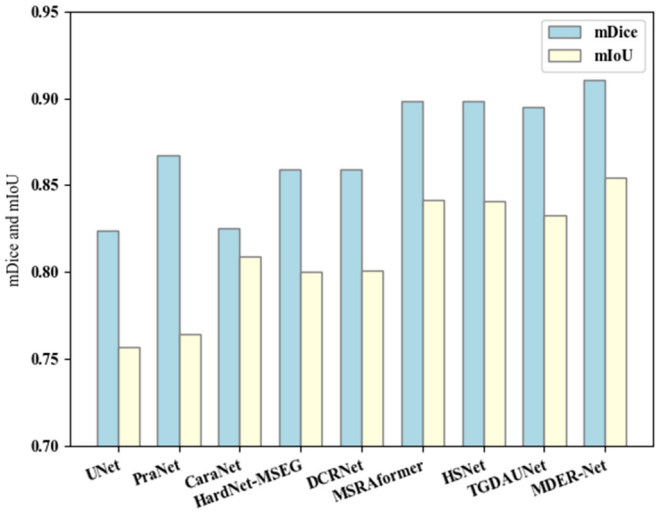

Figure 6. mDice and mIoU metrics bar charts of the different models on the BtAMU dataset.

4.4.2. Qualitative Results

To demonstrate the superiority of the proposed MDER-Net more clearly and intuitively, we compared the qualitative results of the proposed MDER-Net with the results of UNet [5], PraNet [10], CaraNet [13], HarDNet-MSEG [12], DCRNet [14], MSRAformer [25], HSNet [24], and TGDAUNet [32]. Figure 7 shows a visual comparison of the results of the different models on the BtAMU dataset. From Figure 7, it can be seen that compared with other models, our proposed MDER-Net has more accurate prediction results. From Figure 7a,b,f, our proposed MDER-Net is more sensitive to the boundary features of bladder tumors, i.e., it can better outline the boundaries of bladder tumors and remove noisy areas. This is because the DERA module can effectively recover local boundary information, while the ECSA module also reduces the impact of noise and irrelevant information. For Figure 7c–e, the predicted masks of MDER-Net are closer to the ground truth (GT) images, especially for multiple tumor regions (Figure 7c,e) and smaller tumor regions (Figure 7e). Compared with other comparative algorithms, it is more robust. This is because the proposed MECA module improves the network's multi-scale adaptability to lesions, enabling it to adapt to various changes in the size and morphology of bladder tumors. In addition, MDER-Net can remove noise in images with lower brightness (Figure 7f), achieving more accurate bladder tumor segmentation. The comparison of qualitative visual results proves that MDER-Net can better solve the challenges brought by the different shapes and sizes of bladder tumors and the uneven brightness. At the same time, it once again verifies the effectiveness and robust stability of the proposed MDER-Net for bladder tumor segmentation.

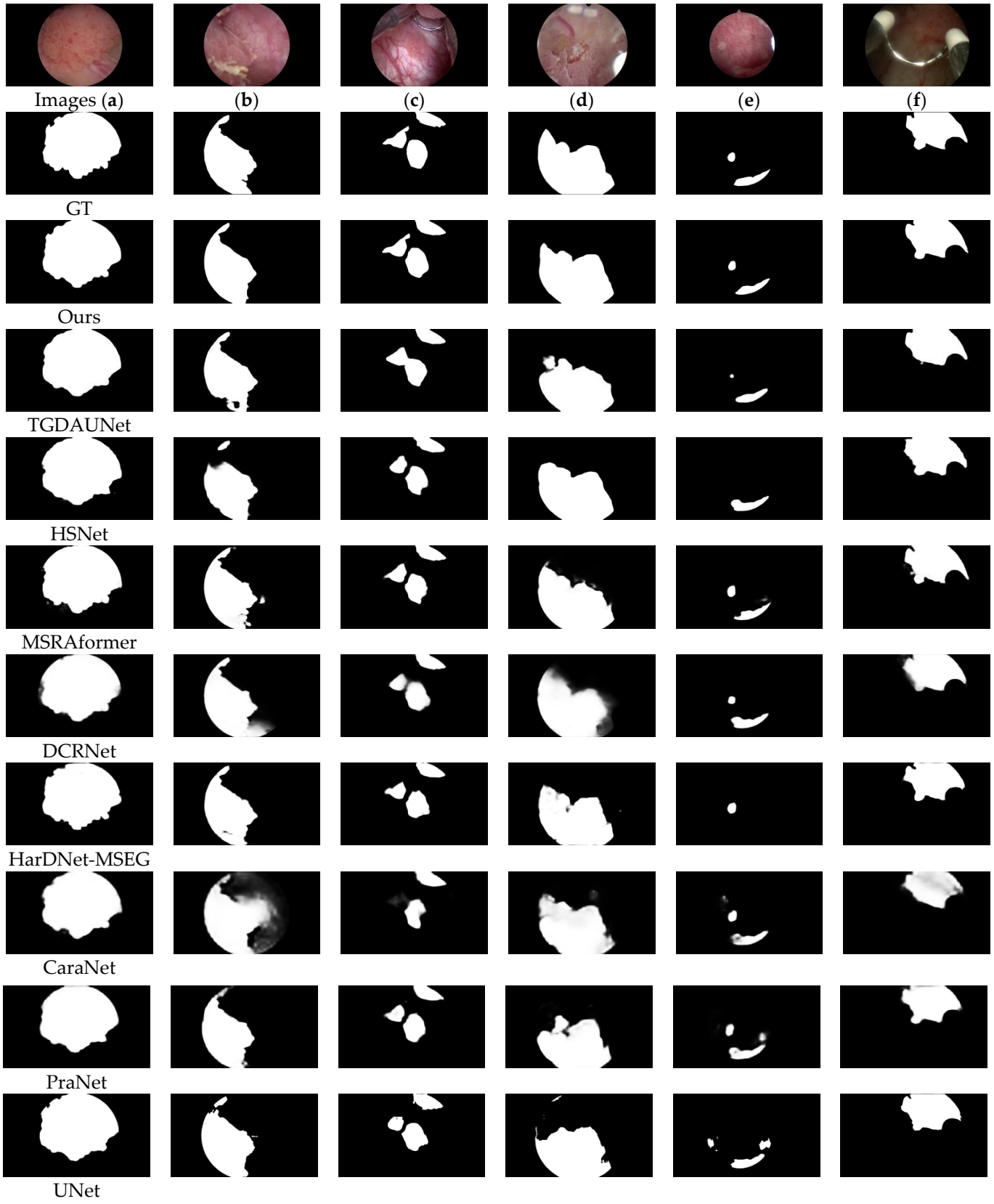

Figure 7. Comparison of qualitative results of other state-of-the-art (SOTA) methods and our MDER-Net on the BtAMU dataset. From top to bottom are the original image (**a**–**f**), ground truth (GT), and the segmentation results of ours, TGDAUNet [32], HSNet [24], MSRAformer [25], DCRNet [14], HarDNet-MSEG [12], CaraNet [13], PraNet [10], and UNet [5].

The quantitative and qualitative results in Table 2 and Figure 7 demonstrate the superiority of MDER-Net in the bladder tumor segmentation task. Compared with other SOTA models, MDER-Net can improve the segmentation accuracy of bladder tumors in cystoscopy images, reduce boundary segmentation errors, and achieve more accurate tumor localization and boundary delineation.

4.5. Experimental Results and Discussion on Polyp Datasets

To verify the robustness and generalization of the proposed MDER-Net, we compared the polyp segmentation results of the proposed MDER-Net with the results of UNet [5], UNet++ [6], PraNet [10], HarDNet-MSEG [12], DCRNet [14], SSFormer [23], MSRAformer [25], and TGDAUNet [32], and quantitatively evaluated the segmentation performance of all models on the polyp tumor datasets Kvasir-SEG [35], CVC-ClinicDB [36], CVC-ColonDB [37], ETIS [38], and CVC-300, using three evaluation metrics: mDice, mIoU, and HD. Tables 3–7, respectively, present the quantitative results comparison of different algorithms on the Kvasir-SEG dataset, CVC-ClinicDB dataset, CVC-ColonDB dataset, ETIS dataset, and CVC-300 dataset. From Tables 3–7, it can be seen that the results of MDER-Net on the five polyp datasets are superior to the other compared methods, consistently achieving state-of-the-art segmentation performance. From Tables 3 and 4, it can be seen that MDER-Net has a stronger feature learning ability than other comparative models. Among them, the mDice metric results of MDER-Net on the Kvasir-SEG dataset and the CVC-ClinicDB dataset reached 92.65% and 92.19%, respectively. From Tables 5–7, it can be seen that MDER-Net has stronger generalization ability compared to the other models on the three unseen polyp datasets. On the CVC-ColonDB dataset, the mDice metric results of MDER-Net were 3.01% and 5.95% higher than the second-ranked MSRAformer [25] and third-ranked TGDAUNet [32], respectively. On the ETIS dataset, our mDice results were 0.27% and 1.88% higher than the second-ranked MSRAformer [25] and third-ranked TGDAUNet [32], respectively. On the CVC-300 dataset, the mDice and mIoU results of MDER-Net increased by 3.03% and 3.08%, respectively, compared to the second-ranked MSRAformer [25], while HD decreased by 0.1862. These experimental results validate the robustness and generalization ability of MDER-Net.

Table 3. Comparison results on the Kvasir-SEG dataset. ↑ indicates higher is better, ↓ indicates lower is better, and bold indicates the best-performing result.

Methods	mDice ↑	mIoU ↑	HD ↓
UNet [5]	0.8209	0.7559	7.9805
UNet++ [6]	0.8237	0.7532	7.6587
PraNet [10]	0.8997	0.8475	6.7757
HarDNet-MSEG [12]	0.8974	0.8434	6.5981
DCRNet [14]	0.8875	0.8385	6.6422
SSFormer [23]	0.8463	0.7566	8.9949
MSRAformer [25]	0.9075	0.8604	6.1880
TGDAUNet [32]	0.9207	0.8699	6.1904
Ours	**0.9265**	**0.8786**	**6.1075**

Table 4. Comparison results on the CVC-ClinicDB dataset. ↑ indicates higher is better, ↓ indicates lower is better, and bold indicates the best-performing result.

Methods	mDice ↑	mIoU ↑	HD ↓
UNet [5]	0.8243	0.7668	4.7971
UNet++ [6]	0.7970	0.7412	4.8488
PraNet [10]	0.9005	0.8576	4.0658
HarDNet-MSEG [12]	0.9104	0.8673	3.8894
DCRNet [14]	0.8979	0.8565	4.1236

Table 4. *Cont.*

Methods	mDice ↑	mIoU ↑	HD ↓
SSFormer [23]	0.8291	0.7496	5.6388
MSRAformer [25]	0.9158	0.8717	4.0575
TGDAUNet [32]	0.9213	0.8689	4.0352
Ours	**0.9219**	**0.8760**	**3.7691**

Table 5. Comparison results on the CVC-ColonDB dataset. ↑ indicates higher is better, ↓ indicates lower is better, and bold indicates the best-performing result.

Methods	mDice ↑	mIoU ↑	HD ↓
UNet [5]	0.5113	0.4402	8.6059
UNet++ [6]	0.4894	0.4110	8.9454
PraNet [10]	0.7156	0.6450	6.8284
HarDNet-MSEG [12]	0.7371	0.6686	6.7526
DCRNet [14]	0.7065	0.6419	7.2245
SSFormer [23]	0.6966	0.6966	9.1187
MSRAformer [25]	0.8037	0.7291	6.3099
TGDAUNet [32]	0.7743	0.7034	6.3336
Ours	**0.8338**	**0.7492**	**6.2414**

Table 6. Comparison results on the ETIS dataset. ↑ indicates higher is better, ↓ indicates lower is better, and bold indicates the best-performing result.

Methods	mDice ↑	mIoU ↑	HD ↓
UNet [5]	0.4059	0.3430	11.5915
UNet++ [6]	0.4134	0.3420	10.7764
PraNet [10]	0.6294	0.5761	9.1532
HarDNet-MSEG [12]	0.7004	0.6345	8.1429
DCRNet [14]	0.5489	0.5065	12.7186
SSFormer [23]	0.5567	0.5567	15.3920
MSRAformer [25]	0.7821	0.7165	7.9340
TGDAUNet [32]	0.7660	0.6789	8.6597
Ours	**0.7848**	**0.7174**	**7.8749**

Table 7. Comparison results on the CVC-300 dataset. ↑ indicates higher is better, ↓ indicates lower is better, and bold indicates the best-performing result.

Methods	mDice ↑	mIoU ↑	HD ↓
UNet [5]	0.7166	0.6390	6.5059
UNet++ [6]	0.7144	0.6362	5.8897
PraNet [10]	0.8717	0.8038	5.0731
HarDNet-MSEG [12]	0.8749	0.8081	5.1673
DCRNet [14]	0.8573	0.7975	5.0207
SSFormer [23]	0.7908	0.7005	6.4926
MSRAformer [25]	0.8749	0.8068	4.9494
TGDAUNet [32]	0.8730	0.8084	5.1858
Ours	**0.9052**	**0.8392**	**4.7632**

4.6. Ablation Experiments

In our proposed MDER-Net, we propose three new modules (MECA, DERA, and ECSA) to improve the segmentation performance of bladder tumors. To verify the effectiveness of each module, we conducted ablation experiments on the BtAMU dataset to explore the impact of each module on the segmentation performance of bladder tumors.

Specifically, our baseline is PVT v2. For the validation of the effectiveness of the MECA module, we replace it with a 1×1 convolutional layer for comparison; for the DERA

module, we compare and verify by replacing all DERA modules with RA modules; and for the ECSA module, we directly remove it to verify its effectiveness. We label the ablation experiments of the MECA, DERA, and ECSA modules as "w/o MECA", "w/o DERA", and "w/o ECSA", respectively. The experimental results are shown in Table 8. Specifically, in terms of the mDice and mIoU metrics, these decreased by 0.76% and 1.00%, respectively, in "w/o MECA" compared to MDER-Net, indicating that compared to multi-scale features processed by the MECA module, the original features extracted by the encoder have limited adaptability to changes in tumor shape and size. The mDice and mIoU metrics of "w/o DERA" were reduced by 1.1% and 1.17%, respectively, compared to MDER-Net, and the HD was 0.3486 higher. This is due to the failure to supplement fine-grained detail information from low-level features to high-level features, resulting in a decrease in segmentation performance. The mDice and mIoU metrics of "w/o ECSA" decreased by 1.16% and 1.38%, respectively, compared to MDER-Net, indicating that the lack of the ECSA module leads to excessive extraction of noise and irrelevant information in low-level features. We designed bar charts of the mDice and mIoU metrics on the BtAMU dataset to visually represent the results of the ablation experiments, as shown in Figure 8. From Table 8 and Figure 8, it can be seen that replacing or removing any of the proposed modules leads to a significant decrease in the segmentation performance of bladder tumors, proving the effectiveness of the proposed modules. The segmentation performance of "w/o MECA", "w/o DERA", and "w/o ECSA" is higher than the baseline, reflecting the effectiveness of cooperation between the various proposed modules.

Table 8. Comparison table of segmentation results on the BtAMU dataset for ablation experiments of our proposed MDER-Net. ↑ indicates higher is better, ↓ indicates lower is better, and bold indicates the best-performing result.

Methods	mDice ↑	mIoU ↑	MAE ↓	Accuracy ↑	F_β^ω ↑	HD ↓
Baseline	0.8768	0.8106	0.0232	0.9780	0.9406	11.7165
w/o MECA	0.9032	0.8443	0.0207	0.9793	0.9495	10.6668
w/o DERA	0.8998	0.8426	0.0207	0.9798	0.9492	10.7544
w/o ECSA	0.8992	0.8402	0.0212	0.9788	0.9457	10.7059
MDER-Net	0.9108	0.8543	0.0173	0.9827	0.9510	10.4058

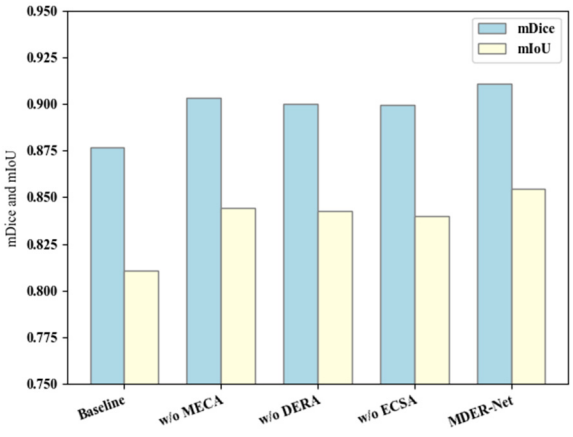

Figure 8. The ablation experiment results of MDER-Net on the BtAMU dataset.

In addition, we analyzed the impact of each module on computational efficiency, including floating-point operations (GFLOPs) and inference time. From Table 9, it can be seen that the inference time of all models meets the real-time requirements. MDER-Net has similar inference time and GFLOPs to "w/o MECA", "w/o DERA", and "w/o ECSA", and slightly higher than baseline. However, compared to them, MDER-Net has a significant advantage in segmentation performance. We believe that in medical image

segmentation tasks, it is more important to achieve higher segmentation accuracy while ensuring real-time performance.

Table 9. Comparison table of computational efficiency results on the BtAMU dataset for ablation experiments of our proposed MDER-Net.

Methods	GFLOPs (G)	Inference Time (s)
Baseline	9.65	0.0205
w/o MECA	12.62	0.0256
w/o DERA	14.05	0.0269
w/o ECSA	16.63	0.0275
MDER-Net	16.63	0.0279

5. Conclusions

This paper proposes a new multi-scale detail-enhanced reverse attention network MDER-Net for bladder tumor segmentation in cystoscopy images. It uses a PVT v2 encoder and six modules (MECA, DA, SAM, DERA, ECSA, RA) to capture multi-scale global features of bladder tumors and enhance local feature representation. By mixing six deep supervisions, it effectively and accurately locates bladder tumors in cystoscopy images. We propose three new modules: MECA, DERA, and ECSA. The MECA module is used to obtain multi-scale features with channel weight information to adapt to changes in the size and morphology of bladder tumors; the DERA module gradually integrates the fine-grained detail information and local boundary information contained in the fused features at all levels into the global feature map of the previous layer, refining the edge structure details of bladder tumors; and the ECSA module is used to suppress irrelevant information in the underlying features, capture details of bladder tumors in different dimensions, and enhance local context. We also established a new bladder tumor dataset, BtAMU, for comparative evaluation. The quantitative and qualitative results of MDER-Net on the bladder tumor dataset BtAMU are superior to eight state-of-the-art (SOTA) methods, demonstrating the superiority of MDER-Net in bladder tumor segmentation tasks. Meanwhile, the visualized qualitative results also prove that MDER-Net can better solve the challenges caused by the varying shapes and sizes of bladder tumors and uneven brightness. The experimental results of MDER-Net on five publicly available polyp datasets are superior to the compared SOTA models, verifying the robustness and generalization ability of the proposed MDER-Net. In the future, we will collect more cystoscopy images and further improve the tumor segmentation performance of the network. At the same time, we hope that our proposed MDER-Net can help clinical decision-making in bladder tumors and provide new ideas for the diagnosis and treatment of bladder tumors. In addition, we will explore the application of MDER-Net in other types of medical image segmentation tasks, such as skin injuries, blood vessels, retina, and 3D medical images.

Author Contributions: Methodology, C.N.; software, C.X.; validation, C.X. and Z.L.; resources, C.N. and C.X.; data curation, Z.L.; writing—original draft preparation, C.N.; writing—review and editing, C.X.; visualization, C.N.; supervision, C.X.; project administration, C.X.; funding acquisition, C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Key Research and Development Program of China (No.2019YFC0117800).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author (accurately indicate status).

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

References

1. Antoni, S.; Ferlay, J.; Soerjomataram, I.; Znaor, A.; Jemal, A.; Bray, F. Bladder Cancer Incidence and Mortality: A Global Overview and Recent Trends. *Eur. Urol.* **2017**, *71*, 96–108. [CrossRef] [PubMed]
2. Kumarasegaram, V.; Drejer, D.; Jensen, J.B. Detection Rate of Carcinoma In Situ during TURBT Following Shift from Photodynamic Diagnosis to Narrow Band Imaging in a Single University Hospital. *Urology* **2022**, *161*, 83–86. [CrossRef]
3. Shkolyar, E.; Jia, X.; Chang, T.C.; Trivedi, D.; Mach, K.E.; Meng, M.Q.H.; Xing, L.; Liao, J.C. Augmented Bladder Tumor Detection Using Deep Learning. *Eur. Urol.* **2019**, *76*, 714–718. [CrossRef]
4. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]
5. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015.
6. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Proceedings of the 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Granada, Spain, 20 September 2018*; Springer: Cham, Switzerland, 2018; Volume 11045, pp. 3–11.
7. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
8. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; De Lange, T.; Halvorsen, P.; Johansen, H.D. Resunet++: An advanced architecture for medical image segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019.
9. Jha, D.; Riegler, M.A.; Johansen, D.; Halvorsen, P.; Johansen, H.D. DoubleU-Net: A deep convolutional neural network for medical image segmentation. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020; pp. 558–564.
10. Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. PraNet: Parallel reverse attention network for polyp segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; pp. 263–273.
11. Zhang, R.; Li, G.; Li, Z.; Cui, S.; Qian, D.; Yu, Y. Adaptive Context Selection for Polyp Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2020, Lima, Peru, 4–8 October 2020; pp. 253–262.
12. Huang, C.H.; Wu, H.Y.; Lin, Y.L. Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. *arXiv* **2021**, arXiv:2101.07172.
13. Lou, A.; Guan, S.; Ko, H.; Loew, M.H. CaraNet: Context axial reverse attention network for segmentation of small medical objects. In Proceedings of the SPIE Medical Imaging 2022: Image Processing, San Diego, CA, USA, 20 February–28 March 2022.
14. Yin, Z.; Liang, K.; Ma, Z.; Guo, J. Duplex Contextual Relation Network For Polyp Segmentation. In Proceedings of the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India, 28–31 March 2022; pp. 1–5.
15. Liu, G.; Chen, Z.; Liu, D.; Chang, B.; Dou, Z. FTMF-Net: A Fourier Transform-Multiscale Feature Fusion Network for Segmentation of Small Polyp Objects. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 5020815. [CrossRef]
16. Li, W.; Zeng, G.; Li, F.; Zhao, Y.; Zhang, H. FRBNet: Feedback refinement boundary network for semantic segmentation in breast ultrasound images. *Biomed. Signal Process. Control.* **2023**, *86*, 105194. [CrossRef]
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Online, 3–7 May 2021.
18. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 11–17 October 2021; pp. 548–558.
19. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. In Proceedings of the 35th Conference on Neural Information Processing Systems, NeurIPS 2021, Online, 6–14 December 2021; pp. 9355–9366.
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 11–17 October 2021; pp. 9992–10002.
21. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. PVT v2: Improved baselines with Pyramid Vision Transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [CrossRef]
22. Dong, B.; Wang, W.; Fan, D.P.; Li, J.; Fu, H.; Shao, L. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. *arXiv* **2021**, arXiv:2108.06932. [CrossRef]
23. Wang, J.; Huang, Q.; Tang, F.; Meng, J.; Su, J.; Song, S. Stepwise Feature Fusion: Local Guides Global. In Proceedings of the 25th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2022, Singapore, 18–22 September 2022; pp. 110–120.
24. Zhang, W.; Fu, C.; Zheng, Y.; Zhang, F.; Zhao, Y.; Sham, C.-W. HSNet: A hybrid semantic network for polyp segmentation. *Comput. Biol. Med.* **2022**, *150*, 106173. [CrossRef] [PubMed]

25. Wu, C.; Long, C.; Li, S.; Yang, J.; Jiang, F.; Zhou, R. MSRAformer: Multiscale spatial reverse attention network for polyp segmentation. *Comput. Biol. Med.* **2022**, *151*, 106274. [CrossRef] [PubMed]
26. Rahman, M.M.; Marculescu, R. Medical Image Segmentation via Cascaded Attention Decoding. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2023; pp. 6211–6220.
27. Liu, G.; Yao, S.; Liu, D.; Chang, B.; Chen, Z.; Wang, J.; Wei, J. CAFE-Net: Cross-Attention and Feature Exploration Network for polyp segmentation. *Expert Syst. Appl.* **2024**, *238*, 121754. [CrossRef]
28. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2012**, *60*, 84–90. [CrossRef]
29. 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.
30. Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2021, Online, 27 September–1 October 2021.
31. Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.R.; Xu, D. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In Proceedings of the 7th International Brain Lesion Workshop, BrainLes 2021, Held in Conjunction with the Medical Image Computing and Computer Assisted Intervention, MICCAI 2021, Online, 27 September 2021.
32. Song, P.; Li, J.; Fan, H.; Fan, L. TGDAUNet: Transformer and GCNN based dual-branch attention UNet for medical image segmentation. *Comput. Biol. Med.* **2023**, *167*, 107583. [CrossRef] [PubMed]
33. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 14–19 June 2020.
34. Chen, S.; Tan, X.; Wang, B.; Hu, X. Reverse attention for salient object detection. In Proceedings of the 15th European Conference on Computer Vision, ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 236–252.
35. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; De Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-seg: A Segmented Polyp Dataset. *Int. Conf. Multimed. Model.* **2020**, *26*, 451–462.
36. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. [CrossRef]
37. Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE Trans. Med. Imaging* **2016**, *35*, 630–644. [CrossRef]
38. Silva, J.; Histace, A.; Romain, O.; Dray, X.; Granado, B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.* **2014**, *9*, 283–293. [CrossRef]
39. Chao, P.; Kao, C.-Y.; Ruan, Y.; Huang, C.-H.; Lin, Y.-L. HarDNet: A low memory traffic network. In Proceedings of the 17th IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Republic of Korea, 27 October–2 November 2019.
40. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef] [PubMed]
41. Bandyk, M.G.; Gopireddy, D.R.; Lall, C.; Balaji, K.C.; Dolz, J. MRI and CT bladder segmentation from classical to deep learning based approaches: Current limitations and lessons. *Comput. Biol. Med.* **2021**, *134*, 104472. [CrossRef] [PubMed]
42. Borhani, S.; Borhani, R.; Kajdacsy-Balla, A. Artificial intelligence: A promising frontier in bladder cancer diagnosis and outcome prediction. *Crit. Rev. Oncol. Hematol.* **2022**, *171*, 103601. [CrossRef] [PubMed]
43. Li, M.; Jiang, Z.; Shen, W.; Liu, H. Deep learning in bladder cancer imaging: A review. *Front. Oncol.* **2022**, *12*, 930917. [CrossRef] [PubMed]
44. Gordon, M.N.; Hadjiiski, L.M.; Cha, K.H.; Samela, R.K.; Chan, H.-P.; Cohan, R.H.; Caoili, E.M. Deep-learning convolutional neural network: Inner and outer bladder wall segmentation in CT urography. *Med. Phys.* **2019**, *46*, 634–648. [CrossRef] [PubMed]
45. Ma, X.; Hadjiiski, L.M.; Wei, J.; Chan, H.-P.; Cha, K.H.; Cohan, R.H.; Caoili, E.M.; Samala, R.; Zhou, C.; Lu, Y. U-Net based deep learning bladder segmentation in CT urography. *Med. Phys.* **2019**, *46*, 1752–1765. [CrossRef] [PubMed]
46. 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]
47. 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 16th IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice, Italy, 8–11 April 2019.
48. Wang, Y.; Li, X.; Ye, X. LCANet: A Lightweight Context-Aware Network for Bladder Tumor Segmentation in MRI Images. *Mathematics* **2023**, *11*, 2357. [CrossRef]
49. Wang, Y.; Ye, X. MSEDTNet: Multi-Scale Encoder and Decoder with Transformer for Bladder Tumor Segmentation. *Electronics* **2022**, *11*, 3347. [CrossRef]
50. 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 15th IEEE International Conference on Signal Processing, ICSP 2020, Beijing, China, 6–9 December 2020.

51. Varnyu, D.; Szirmay-Kalos, L. A Comparative Study of Deep Neural Networks for Real-Time Semantic Segmentation during the Transurethral Resection of Bladder Tumors. *Diagnostics* **2022**, *12*, 2849. [CrossRef]
52. Yoo, J.W.; Koo, K.C.; Chung, B.H.; Lee, K.S. Deep learning diagnostics for bladder tumor identification and grade prediction using RGB method. *Eur. Urol.* **2023**, *83*, S846. [CrossRef]
53. Zhang, Q.; Liang, Y.; Zhang, Y.; Tao, Z.; Li, R.; Bi, H. A comparative study of attention mechanism based deep learning methods for bladder tumor segmentation. *Int. J. Med. Inform.* **2023**, *171*, 104984. [CrossRef] [PubMed]
54. Jia, X.; Shkolyar, E.; Laurie, M.A.; Eminaga, O.; Liao, J.C.; Xing, L. Tumor detection under cystoscopy with transformer-augmented deep learning algorithm. *Phys. Med. Biol.* **2023**, *68*, 165013. [CrossRef]
55. Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; Veit, A. Understanding Robustness of Transformers for Image Classification. In Proceedings of the 18th IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, 11–17 October 2021.
56. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.-S. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017.
57. Wei, J.; Wang, S.; Huang, Q. F^3Net: Fusion, feedback and focus for salient object detection. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, 7–12 February 2020.
58. Vazquez, D.; Bernal, J.; Sanchez, F.J.; Fernandez-Esparrach, G.; Lopez, A.M.; Romero, A.; Drozdzal, M.; Courville, A. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *J. Healthc. Eng.* **2017**, *2017*, 037190. [CrossRef] [PubMed]