information

MDPI

Article

Nonlinear Activation-Free Contextual Attention Network for Polyp Segmentation

Weidong Wu [1], Hongbo Fan [2,*], Yu Fan [1] and Jian Wen [1]

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China; wuweidong@stu.kust.edu.cn (W.W.); 20212204259@stu.kust.edu.cn (Y.F.); 20212204261@stu.kust.edu.cn (J.W.)

[2] Faculty of Modern Agricultural Engineering, Kunming University of Science and Technology, Kunming 650300, China

* Correspondence: 20110258@kust.edu.cn

Abstract: The accurate segmentation of colorectal polyps is of great significance for the diagnosis and treatment of colorectal cancer. However, the segmentation of colorectal polyps faces complex problems such as low contrast in the peripheral region of salient images, blurred borders, and diverse shapes. In addition, the number of traditional UNet network parameters is large and the segmentation effect is average. To overcome these problems, an innovative nonlinear activation-free uncertainty contextual attention network is proposed in this paper. Based on the UNet network, an encoder and a decoder are added to predict the saliency map of each module in the bottom-up flow and pass it to the next module. We use Res2Net as the backbone network to extract image features, enhance image features through simple parallel axial channel attention, and obtain high-level features with global semantics and low-level features with edge details. At the same time, a nonlinear n on-activation network is introduced, which can reduce the complexity between blocks, thereby further enhancing image feature extraction. This work conducted experiments on five commonly used polyp segmentation datasets, and the experimental evaluation metrics from the mean intersection over union, mean Dice coefficient, and mean absolute error were all improved, which can show that our method has certain advantages over existing methods in terms of segmentation performance and generalization performance.

Keywords: colorectal polyp segmentation; edge details; image feature extraction; nonlinear activation-free; salient images; uncertainty context attention

check for updates

Citation: Wu, W.; Fan, H.; Fan, Y.; Wen, J. Nonlinear Activation-Free Contextual Attention Network for Polyp Segmentation. *Information* **2023**, *14*, 362. https://doi.org/10.3390/info14070362

Academic Editors: Simone Palazzo and Carmelo Pino

Received: 27 May 2023
Revised: 20 June 2023
Accepted: 23 June 2023
Published: 26 June 2023

1. Introduction

Image segmentation is one of the fundamental and challenging research topics in computer vision, which aims at classifying each pixel in a given image. Image segmentation is applied in different fields, among which the most widely used is medical image segmentation for classifying each organ in a given tomographic image, such as cells in microscopic images [1], pancreatic segmentation [2], or segmentation of pathological regions from normal bodies such as brain tumors [3], as well as polyp segmentation in this paper. Colorectal cancer (CRC) has become one of the most common malignant tumors that endanger human health, mostly evolving from adenomatous polyps, and initially benign polyps are at risk of malignant transformation if timely and effective treatment is not available. Nowadays, colonoscopy is widely used clinically and has become a general method for screening for rectal cancer. Therefore, accurate medical image segmentation plays a crucial role in clinical applications [4]. Early polyp segmentation is usually analyzed in terms of color distribution, textural characteristics, structural changes [5], etc., and manual features are extracted to distinguish the polyps as well as the background. The low contrast between the polyp and the surrounding mucosa, as well as the blurred borders of the abnormal tissue, leads to a high rate of missed diagnoses [6].

The general purpose of the method is to design a segmentation network that generalizes to multiple segmentation tasks. However, designing such a network is extremely challenging due to the large variance between different segmentation tasks. Among the existing solutions, the more common solution is to focus on the commonality between the segmentation tasks but ignore the differences, which simplifies the network to a certain extent. In other words, it is necessary to summarize the common features among multiple tasks, and then deal with them from different aspects. Examples include the extraction of contextual information [7], fusion of local and global information [8], boundary constraints [9], and redesign of skip connections [10]. Although the general method has a good performance in processing multiple segmentation tasks simultaneously, it cannot avoid ignoring the specific features of different segmentation tasks, and the generalization performance is low to a certain extent. Therefore, it is crucial to design a method with high performance in specific object segmentation tasks.

In recent years, several methods have been implemented to solve the segmentation problem of polyps, and the consensus of these methods lies in the specific features of the polyps considered during the design of the network. Methods for polyp segmentation can be broadly divided into traditional methods based on threshold, edge detection, etc., and deep learning methods. Traditional methods are relatively simple to implement, but there are certain limitations in image segmentation [11]. Nowadays, image segmentation using deep learning has become mainstream, and UNet [1] has achieved great success in medical images, following the encoder–decoder architecture, where in the encoder stage, feature extraction units with high-level semantic information and low-level spatial details are used to represent the object. In the decoder stage, features from the encoder are integrated to generate prediction masks. Based on UNet, UNet++ [12] is able to better alleviate the semantic divide between the encoder and decoder feature maps, and its encoder is connected by some nested dense convolutions. For the characteristics of colon polyps, ResUNet++ [13] combines UNet with residual networks [14] and achieves multi-scale feature fusion and information feature extraction by adding a squeeze-and-excitation module (SE) [15] and atrous spatial pyramid pooling (ASPP) [16]. Although the performance of convolution-based methods is satisfactory, they are somewhat limited in learning long distance dependencies between pixels due to the spatial context of the convolution operation [17]. To overcome this limitation, the attention module is incorporated into the architecture of the network, thus enhancing the feature maps and enabling pixel-level classification of medical images. PraNet [18] introduces a parallel reverse attention mechanism network to accurately segment polyps and mitigate the effects of noise from uneven light distribution and randomization of polyp positions. Previous polyp segmentation networks usually adopt the saliency object detection (SOD) method, which can better highlight the object region rather than the background region, and usually, in practice, there are ambiguous boundaries and regions, and the saliency object detection method can generate polyp masks and boundary masks simultaneously. However, this increases the convergence burden of the network and also the cost of obtaining additional edge data is higher. The reverse attention mechanism uses the reverse saliency maps to obtain boundary cues, but since the boundary region is highly correlated with the fuzzy saliency scores, the saliency maps without the reverse operation already have such boundary information.

In this paper, the nonlinear activation-free and uncertainty context attention network (NAF_UCANet) is proposed. The method in this paper calculates regions with fuzzy saliency scores and combines foreground and background regions to better implement the contextual attention module. It can enhance the uncertainty regions on saliency maps that are highly correlated with boundary information, and at the same time enhance image detail feature extraction using a nonlinear activation-free network. The feature map is divided into two parts in the channel dimension by simple gate element-wise multiplication to replace the nonlinear activation function. Based on a modified version of the UNet structured network with additional encoders and decoders, multiple regions are weighted and summed to aggregate the feature maps to obtain a context vector for each region and calculate the similarity of the feature maps. To address the characteristics of previous polyp segmentation methods, the combination of a nonlinear activation-free network and simple parallel axial channel attention proposed in this paper can further improve the feature extraction of polyp images. In this paper, five more well-known polyp segmentation benchmarks, CVC-ClinicDB, Kvasir, CVC-ColonDB, ETIS, and CVC-300, were used to validate the method. In summary, the contributions of this paper are as follows:

1. Instead of simply aggregating feature maps from multiple layers to generate coarse segmentation masks, the complementary information between different layers and the contextual information of each layer are integrated. The nonlinear activation-free uncertainty contextual attention network proposed in this paper is able to enhance uncertainty regions on saliency maps that are highly correlated with boundary information.
2. In order to realize the calculation of areas with fuzzy saliency scores in the case of various polyp locations and the presence of fuzzy areas that are easily confused with polyp areas, a contextual attention module is implemented by combining foreground and background areas. In this paper, a simple parallel axial channel attention is proposed so as to correctly identify polyps from the background region.
3. A nonlinear and non-activating feature detail extraction enhancement technique is introduced, which can fuse feature maps of multiple regions based on an improved U-shaped network with additional encoders and decoders in order to explore the contextual features of each layer more accurately and achieve better segmentation accuracy and computational efficiency.

2. Related Work

2.1. Medical Image Segmentation

Medical imaging has gained immense importance [19] in healthcare throughout history. Nowadays, deep learning is booming, realizing the first pixel-level segmentation fully convolutional network (FCN) [20], combining the skip connection encoder–decoder structure Unet [1], conditional random field, and Deeplab [21] with multiple convolutional layers with different dilation rates and achieving good results in image segmentation. Most of the current medical image segmentation is based on the improvement of UNet, which is used in different segmentation tasks to improve the efficiency of segmentation. UNet++ [12] redesigns the skip connections in UNet by nesting densely, thereby reducing the semantic gap between encoders and decoders, aggregating various semantic features, and capturing objects of different sizes. The pyramid spatial pooling network (PSPNet) processes multi-scale images through grid pooling as well as grids of multiple sizes. For the features of the encoder resolution when corresponding in the decoder, AttentionUNet [22] used attention gating signals to control the different spatial location feature information to further adjust the output features to improve the segmentation efficiency. The dual attention network [23] uses a self-attention mechanism for the segmentation network, including a dual path network in spatial and channel dimensions, which can combine local features with global dependencies. Furthermore, the object context representation (OCR) [24] adds a non-local operation, aggregating pixel representations of each category to consider segmented regions, while calculating the similarity to pixels in the feature maps. Contextual information can represent segmentation objects with different shapes and sizes, and utilizing appropriate context methods is more critical in the field of segmentation.

2.2. Image Feature Details Extraction Enhancement

The purpose of image feature extraction is to obtain useful information and data from images and describe them in a non-image form, so to make it better understood by computers. The introduction of an attention mechanism enables advanced feature extraction while retaining local spatial information. Deep-learning-based methods are able to process unenhanced data [25] by automatically extracting salient features. Some methods improve on UNet, stacking blocks into a U-shaped structure similar to the UNet skip connections. These improvements can better enhance the performance and system complexity of feature extraction, which are mainly divided into inter-block complexity and intra-block complexity.

Inter-block complexity [26] is the result of a multi-level network and each level is U-shaped. The original intention of this design was to divide the difficult image feature extraction task into multiple subtasks. In contrast, [27] is a single-level design and has some competition, but this method introduces more complex connections between feature maps of different sizes. Some keep the simple architecture of the single-level UNet and add intra-block complexity on top of that. The intra-block complexity contains an assortment of intra-block design methods, among which [28] is to reduce the time complexity of the self-attention mechanism by replacing the spatial division with the channel division attention map. Additionally, a gated linear unit [29] (GLU) is added to the feedforward network. The authors of [30] incorporated window-based multi-headed self-attention and introduced locally enhanced feedforward networks in this module, and the use of deep convolution can enhance the ability to capture local information. The gated linear unit [29] is generated element-wise by two linearly varying layers and one is activated by nonlinear activation. From another point of view, GLU is a generalization of the activation function and can replace the nonlinear activation function to some extent. Among them, [30] does not use nonlinear activation functions in GLU, but there is no degradation in performance. The nonlinear activation-free GLU based on nonlinearity is itself characterized by nonlinearity because the product of two linear variations enhances the nonlinearity. In this method, the multiplication of two feature maps is used instead of the nonlinear activation function, which can enhance the feature extraction of image detail edges.

2.3. Polyp Segmentation

The current published literature related to polyps performs better in some specified datasets, using small and validated datasets [31]. Figure 1 shows the associated colorectal polyp maps. Models evaluated on smaller datasets cannot be generalized to a certain extent, and robustness is also difficult to guarantee. Although a full convolutional network [20] can be used to solve the polyp segmentation problem, colonoscopy has a different image domain compared to general images and requires the extraction of semantic features with detailed information. Classical architectures such as the pyramidal spatial pooling network and Deeplab mentioned above employ multi-scale strategies in the backbone network to facilitate the capture of detailed feature information with multiple receptive field sizes, but these methods are usually deployed at the end of the backbone network and are insufficient to recover accurate spatial information. The emergence of DeeplabV3+ [11] connects the low-level feature maps of the backbone network to compensate for the above shortcomings, but it is still not enough to extract relevant detailed features from the images. Both Unet [1] and the feature pyramid network [32] (FPN) employ incremental upsampling of the feature maps and the corresponding scaling of the low-level feature maps. The difference is that FPN uses element-wise addition, while UNet uses channel cascade aggregation features. UNet++ [12] reduces the semantic gap between low and high layers by adding additional layers and dense connections.

As the polyp segmentation research progresses, ResUNet++ [13] constructs a U-shaped network that includes residual blocks from ResNet, the empty space pyramidal pooling module from Deeplab, and the SE attention mechanism from SENet to achieve the task of polyp segmentation. The parallel reverse attention network [18] (PraNet) adopts most of the network techniques in [33] and adds the parallel partial decoder in [34] combined with the low-level feature map of the backbone network, so that the boundary information can be mined better, but the segmentation performance is relatively low for small polyps. The self-attention mechanism [35] better extracts the fine-grained feature maps, making the feature maps with low-level details along with high-level semantic information, but is computationally more intensive. The emergence of axial attention [36] can solve the bottleneck of self-attention by carrying out a series of problems by means of a single axis. In summary, a large number of methods have been used to solve the polyp segmentation problem, including two adjacent layers to extract the contextual information of the polyp representation. Deep convolution contains details and semantic information from shallow to deep layers, and these layers play a larger role in representing polyp segmentation objects with different shapes and sizes. The advantages and disadvantages of some classical segmentation models are shown in Table 1.

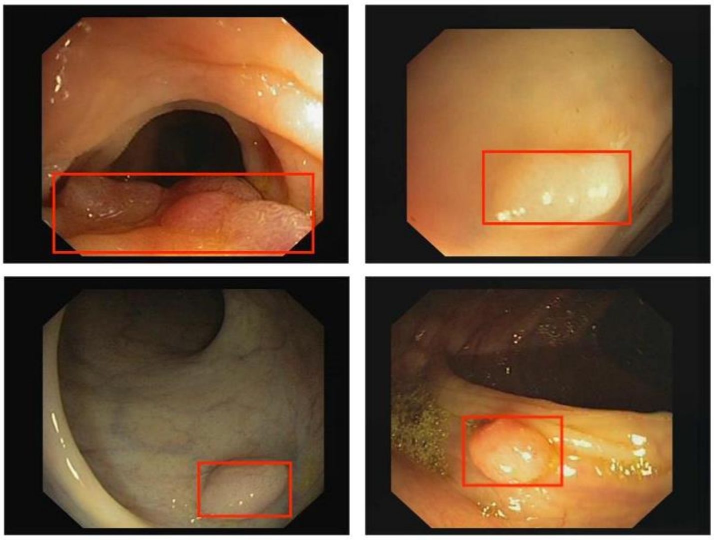


Figure 1. This figure shows the polyp case.

Table 1. Advantages and disadvantages of classical segmentation models.

Segmentation Model	Parameter Amount	Context Information Fusion	Attention Mechanism	Perception of Detail and Edges	Computing Resources
UNet [1]	large	yes	no	normal	high
UNet++ [12]	large	yes	no	well	high
Deeplab [21]	large	yes	no	normal	high
ResUNet++ [13]	small	yes	no	normal	low
AttentionUNet [22]	large	yes	yes	well	high
PraNet [18]	small	yes	yes	well	low

3. Method

3.1. Overall Architecture

The nonlinear activation-free uncertainty contextual attention network (NAF_UCANet) mentioned in this paper is based on the parallel reverse attention network (PraNet) with Res2Net [37] as the backbone network for feature extraction to extract features of polyp segmentation images. Res2Net provides a stronger semantic information modeling capability by introducing multi-scale feature representation and parallel paths with high parameter utilization and scalability and low computational overhead compared to complex network structures. The overall architecture of the network proposed in this paper is shown in Figure 2. The entire network architecture consists of a simple parallel axial channel attention encoder network and a corresponding decoder network. The simple parallel axial channel attention encoder (SPACA -e) is used in the bottom-up flow and skip connection feature fusion path. This module simplifies the parallel axial channel attention by using a parallel nonlinear activation-free network instead of the traditional convolution, incorporating layer normalization as well as a simple gated linear unit, and enhances the nonlinearity by the product of two linear variations, so as to achieve the effect of the architecture mentioned in this paper. Reducing the number of channels in the input feature map achieves each cost of the reduced bottom-up flow and skip connection feature fusion. After extracting the polyp feature maps in the backbone network, the simple parallel axial channel attention encoder is introduced in each layer of the skip connection, and the feature maps obtained by all three SPACA-e modules are used to fuse low-level features with edge information and global semantic high-level features (as shown by the green arrow part in Figure 2). In addition, the resulting feature maps are simultaneously fed into the simple parallel axial channel attention decoder (SPACA-d) and uncertainty context attention (UCA). In the network architecture of this paper, the three feature maps of the SPACA-e module are aggregated for use in the SPACA-d module to initially predict the saliency maps of the input polyps. Immediately afterwards, the feature maps from the SPACA-e and SPACA-d modules are concatenated with the UCA part in skip connection. Among them, the feature maps obtained by the SPACA-d module are used in the upsampling part for context guidance. The feature maps at this point are then element-wise summed with the output salient feature maps from the UCA. From the bottom up, after the first UCA, the feature maps obtained from SPACA-e is channel-connected with the feature maps of the previous UCA to enter the next UCA (blue and black arrows for making the connection in Figure 2). For the network as a whole, the salient feature maps obtained from the previous UCA serves as the contextual guidance for the next UCA (as shown in Figure 2 with purple arrows). Subsequently, the output of the UCA goes into a 1 × 1 convolutional layer and is summed with the previously mentioned contextual guidance. Finally, after completing the three UCA modules, a sigmoid activation function and a bilinear upsampling with a scale factor of 4 are used to obtain the final output feature maps.

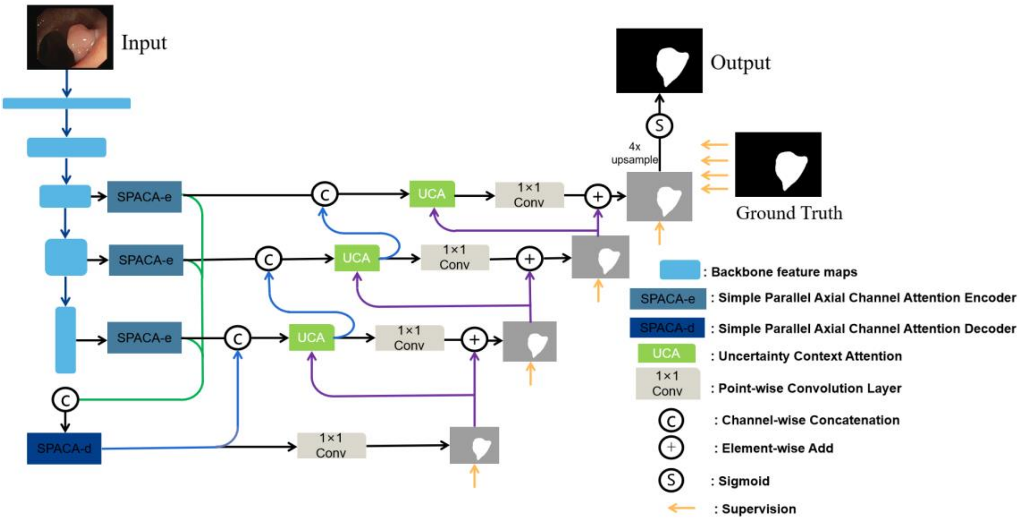

Input
Output
Ground Truth
4x upsample
S
SPACA-e
SPACA-e
SPACA-e
SPACA-d
C
C
C
C
UCA
UCA
UCA
1×1 Conv
1×1 Conv
1×1 Conv
1×1 Conv
+
+
+
: Backbone feature maps
SPACA-e : Simple Parallel Axial Channel Attention Encoder
SPACA-d : Simple Parallel Axial Channel Attention Decoder
UCA : Uncertainty Context Attention
1×1 Conv : Point-wise Convolution Layer
C : Channel-wise Concatenation
+ : Element-wise Add
S : Sigmoid
: Supervision

Figure 2. Overall architecture of NAF_UCANet.

3.2. Simple Parallel Axial Channel Attention

Researchers are currently searching for better techniques to extract fine-grained feature maps so that the feature maps have low-level details along with high-level semantic information. The self-attention mechanism [35] can solve the above problems better, but there is also a bottleneck of a large amount of calculation. The emergence of axial attention [36], which uses a single axis to perform a series of operations and connect them sequentially, solves the above problems at the same time to a certain extent. Based on this, this paper proposes simple parallel axial channel attention, which simplifies axial attention to some extent and introduces a channel attention module. It can extract local information and global dependencies and uses the strategy of axial channel attention to calculate non-local features in the horizontal and vertical directions, and the two are in parallel. Compared to the sequential methods, the parallel configuration of attention in both horizontal and vertical directions has the same degree of contribution to the final output. For the extraction task of enhanced image features, the data needs to be processed at a higher resolution, so it is more critical when choosing the attention mechanism. Local information can be better obtained through deep convolution, and at the same time adding channel attention can avoid large calculation problems and maintain the global information in each feature map. Importantly, the parallel axial channel connections, with the ability to aggregate multiple feature maps between elements through add operations. Compared with the cascade method, due to the parallel connection of the horizontal and vertical axes of the same input, the performance is not degraded. In addition, parallel axial attention uses element-wise summation to avoid the artifacts generated by single axis attention.

The simple parallel axial channel attention proposed in this paper is shown in Figure 3, and the input feature maps are used to calculate the horizontal and vertical axes. Using the encoder and decoder in this module, the channel attention module is introduced to efficiently enhance the feature map of the backbone network output, improving spatial location as well as channel correlation, while allowing better representation of the global refinement function. The simple parallel axial channel attention encoder (SPACA-e) is designed where the encoder first aggregates the low-level feature maps from the top-down stream and uses them in the bottom-up stream. In order to be able to reduce the number of channels while maintaining detailed information, a nonlinear activation-free network of SPACA-e is introduced, as shown in Figure 4a. First, the feature maps from the backbone network are fed into the parallel nonlinear activation-free network. They are then channel-connected and output into successive convolutional layers. As can be seen from the overall architecture of the network, the output of the SPACA-e module is used for the decoder module as well as the upsampling module. Likewise, this paper accordingly introduces the simple parallel axial channel attention decoder (SPACA-d), which additionally adds SPACA to the structure, which is able to better aggregate the feature maps of the encoder outputs from different layers. The simple parallel axial channel attention decoder is shown in Figure 4b, where the yellow arrows in the figure indicate the feature aggregation of different layers of encoders.

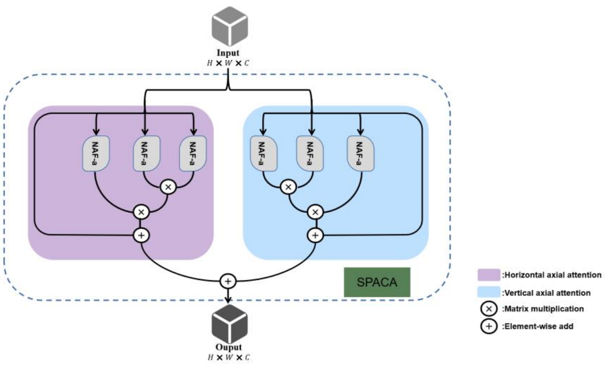
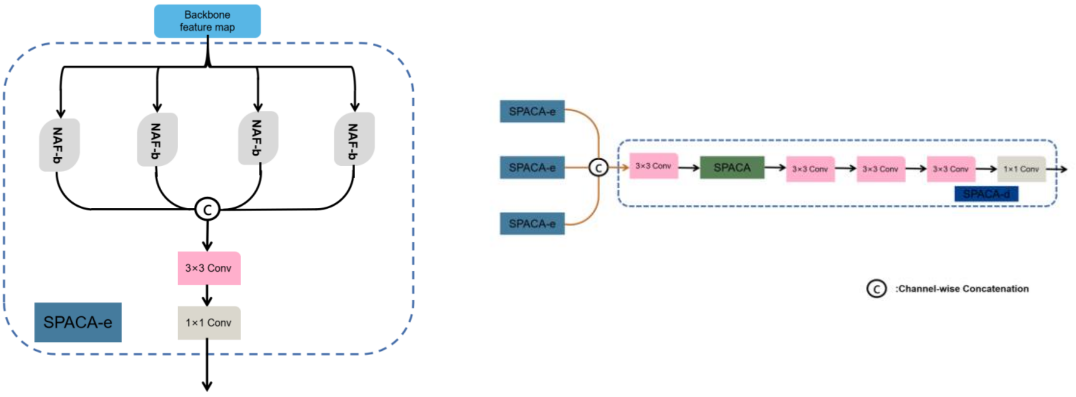


Figure 3. Architecture of simple parallel axial channel attention (SPACA).



(**a**) (**b**)

Figure 4. (**a**) Architecture of a simple parallel axial channel attention encoder (SPACA-e); (**b**) Architecture of a simple parallel axial channel attention decoder (SPACA-d).

3.3. Nonlinear Activation-Free Network

It is understood that the SPACA introduced in the previous subsection can avoid large computational cost while obtaining local information. To perform the task of feature extraction enhancement, the nonlinear activation-free network is introduced in the SPACA, which is unfolded based on the UNet architecture in order to reduce the complexity between blocks. The neural network is stacked in blocks, and how to design the structure within the blocks is the problem. It contains the most common components: convolution, the ReLu activation function, normalization, etc. Normalization is a commonly used module in computer vision tasks. Layer normalization is introduced in the module of this paper, which is extremely important for image feature extraction and can stabilize the training process. Activation functions are also widely used in computer vision and commonly used are ReLu activation functions, while GELU tends to replace ReLu. In the module of this paper, the simple gate is used to instead, which maintains the image deblurring property. The attention mechanism is an unavoidable topic in the design of the intra-block. Since the image feature extraction task processes data with high resolution, it is necessary to maintain the global information of the feature while solving the computational problem. The introduction of channel attention is able to solve these problems simultaneously. The architecture of the nonlinear activation-free function is shown in Figure 5a,b. In this architecture, each component is trivial, such as layer normalization, convolutional layers, simple gates, and simple channel attention, and there are no nonlinear activation functions (ReLu, Sigmoid, etc.) in the entire architecture. However, the combination of these simple modules forms a more functional nonlinear activation-free network.

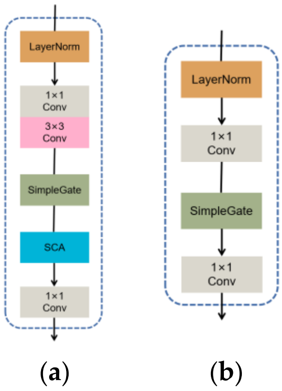


Figure 5. (**a**) Architecture of nonlinear activation-free network A (NAF-a); (**b**) Architecture of nonlinear activation-free network B (NAF-b).

The gated linear unit (GLU) can be expressed by the following formula:

$$Glu(m, f, g, \sigma) = f(m) \odot \sigma(g(m)) \tag{1}$$

where m denotes the feature map, f and g denote linear transformations, and σ denotes a nonlinear activation function, such as sigmoid, etc. $\odot$ denotes the element multiplication. As a special case of GLU, GELU emerges to solve the problem of intra-block complexity. It can be expressed by the following formula:

$$Gelu(m) = m\Phi(m) \tag{2}$$

where Φ denotes the cumulative distribution function of the standard normal distribution. It can be noted that GLU contains nonlinearity itself and does not depend on σ. So, on this basis, the feature map is divided into two parts in the channel dimension for channel-wise and element-wise multiplication. The architecture of the simple gate is shown in Figure 6a. Its formula is expressed as follows:

$$SG(m,n) = m \odot n \tag{3}$$

where m and n denote feature maps of the same size. In addition, a simplified channel attention mechanism is introduced in the nonlinear activation-free network of this paper, since this module can capture global information and also has some advantages in terms of computational efficiency. The spatial information is first compressed into channels, and the SCA architecture is shown in Figure 6b. The channel attention is similar to that of a gated linear unit and can be considered as a special way of GLU. Retaining the role of aggregated global information in channel attention and channel information interaction, the simplified channel attention is expressed as:

$$SCA(m) = m \times Wpool(m) \tag{4}$$

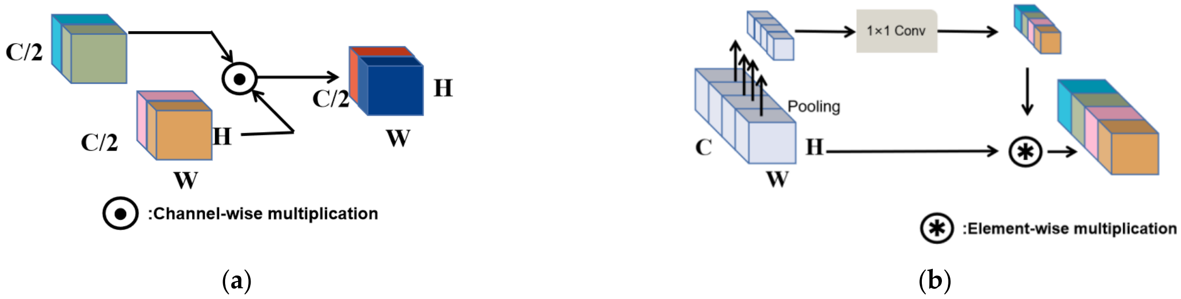


Figure 6. (**a**) Architecture of simple gate; (**b**) Architecture of simplified channel attention (SCA).

Among them, m denotes the feature map, *pool* denotes the global average pooling of spatial information aggregated into channels, and × denotes the channel multiplication operation.

3.4. Uncertainty Contextual Attention

The feature map obtained by the general polyp network after the decoder only exists as the relative position of the polyp in the image, and the details of the edge information are ignored. The method of saliency object detection and the reverse attention mechanism are effective methods in polyp segmentation, paying attention to the reverse saliency map as well as the saliency of reverse attention, and it is found that in both the feature maps obtained by encoder and the feature maps obtained by reverse attention, generally the boundary regions appear where the significance score is ambiguous, and the significance score of the edge regions is usually close to 0.5. This paper designs an uncertainty contextual attention method based on reverse attention to produce reverse saliency features and compute foreground and background features by aggregating pixel features. In the field of image segmentation, the threshold for whether the prediction of the final feature map is the correct pixel is also generally set to 0.5.

Based on this assumption, an uncertainty contextual attention module is designed as a self-attention mechanism that incorporates the semantic feature extraction from uncertainty regions and calculates the foreground feature maps, background feature maps, and uncertainty region maps, respectively, so as to enhance attention to edge features. The detailed structure is shown in Figure 7. The previously calculated input salient map is marked as map, the foreground feature map is $map_f = \max(map - 0.5, 0)$, the background feature map is $map_b = \max(0.5 - map, 0)$, and the uncertainty region map is $map_u = 0.5 - abs(0.5 - m)$. It can be seen that the uncertainty region represents the joint region of the foreground feature map and the background feature map, so there are redundant parts in both, which somehow makes the uncertainty less useful, so the maximum operation needs to be used to calculate the foreground and background features. Then the pixels of the input feature map m are computed and fused with the feature maps of each region, and the feature vectors of the foreground features, background features, and uncertainty region features are calculated separately and expressed as follows:

$$v_f = \sum_{i \in I} map_{fi} m_i \tag{5}$$

$$v_b = \sum_{i \in I} map_{bi} m_i \tag{6}$$

$$v_u = \sum_{i \in I} map_{ui} m_i \tag{7}$$

where $i \in I$ denotes the pixel points belonging to the spatial dimension. v_f denotes the foreground feature vector, v_b denotes the background feature vector, and v_u denotes the uncertainty region feature vector. After that, the similarity between each feature map from the encoder and each feature vector is calculated and expressed as follows:

$$sim_{fi} = \frac{e^{\varphi(mi)^\top \delta(v_f)}}{e^{\varphi(mi)^\top \delta(v_f)} + e^{\varphi(mi)^\top \delta(v_b)} + e^{\varphi(mi)^\top \delta(v_u)}} \tag{8}$$

$$sim_{bi} = \frac{e^{\varphi(mi)^\top \delta(v_b)}}{e^{\varphi(mi)^\top \delta(v_f)} + e^{\varphi(mi)^\top \delta(v_b)} + e^{\varphi(mi)^\top \delta(v_u)}} \tag{9}$$

$$sim_{ui} = \frac{e^{\varphi(mi)^\top \delta(u)}}{e^{\varphi(mi)^\top \delta(v_f)} + e^{\varphi(mi)^\top \delta(v_b)} + e^{\varphi(mi)^\top \delta(v_u)}} \tag{10}$$

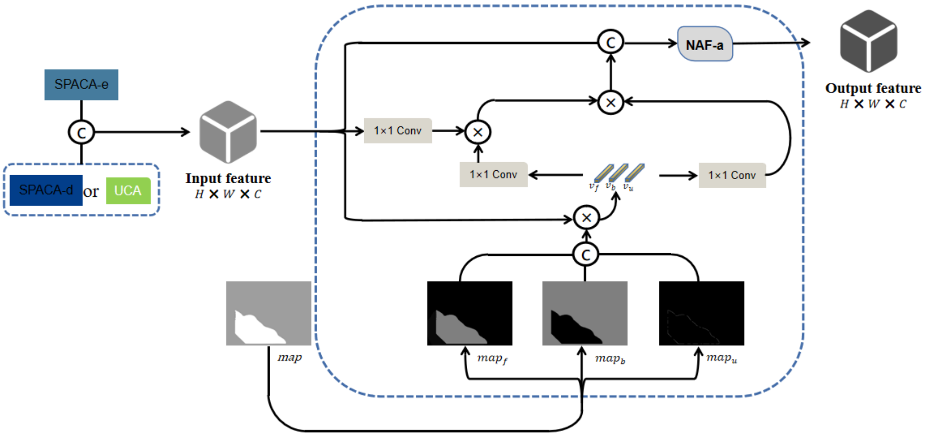


Figure 7. Architecture of uncertainty context attention (UCA).

Among them, the similarity scores sim_f, sim_b, and sim_u are obtained for each feature vector by summing the weights sim_{fi}, sim_{bi}, and sim_{ui}. Then the weighted average is obtained for each pixel of the feature map, which is represented as follows:

$$px_i = \tau\left(sim_{fi}\sigma\left(v_f\right)\right) + \tau(sim_{bi}\sigma(v_b)) + \tau(sim_{ui}\sigma(v_u)) \tag{11}$$

where $\varphi(\bullet)$, $\delta(\bullet)$, $\tau(\bullet)$, and $\sigma(\bullet)$ denote the point-wise convolution operation. px_i represents each pixel in the contextual feature map, which can be represented by the weighted average of three feature representation vectors. The pixel px of the feature map and the input feature map m are channel-connected to obtain the final output feature map.

4. Experimental Analysis

This section discusses the datasets, experimental details, and comparison of results with other experimental benchmark methods. Including the ablation experiments of the module in this paper, five commonly used polyp segmentation benchmarks are used to validate the effectiveness of the polyp segmentation network in this paper. In addition, the interpretability of the network is enhanced through qualitative and quantitative approaches.

4.1. Experimental Datasets and Evaluation Metrics

In this paper, randomly selected images from two datasets, CVC-ClinicDB and Kvasir, were used for training. From these, 900 and 550 pictures were used, respectively; a total of 1450 pictures were used as the training set and other data were used as the test set. Five of the popular datasets in polyp segmentation were used for this paper.

The CVC-ClinicDB contains 612 images from 25 colonoscopies with an image resolution size of 384 × 288 pixels. Of the images in this dataset, 550 images were used as the training set and the remaining 62 images were used as the test set.

The Kvasir dataset contains 1000 polyp images, which are different from other polyp datasets, with image resolution sizes ranging from 332 × 487 pixels to 1920 × 1072 pixels. In addition, the size and shape of polyps in the images varied. Of these, 900 images were selected as the training set and the remaining 100 images were used as the test set.

The CVC-ColonDB dataset is derived from 15 colonoscopy sequences, from which 380 images were selected, all of which were used as a test set.

The ETIS dataset contains 1000 images of endoscopic polyps, and 196 images from 34 colonoscopies with a resolution size of 1225 × 966 pixels were taken from this dataset for this paper. The polyps in these images are all small and have a more similar structure to the margins, so they are somewhat challenging.

The CVC-300 dataset is a test set from EndoScene, which contains 912 images from 44 colonoscopies of 36 patients. It was used as a test set.

In the quantitative analysis of the experiment, the mean intersection over union (mIoU), mean Dice coefficient and mean absolute error (MAE) were used as the evaluation indexes of the experiment. IoU is the ratio of the intersection and union of the predicted result and the ground truth, which lies between 0 and 1. The calculation formula of IoU is as follows:

$$\mathrm{IoU} = \frac{A \cap B}{A \cup B} = \frac{TP}{TP + FP + FN} \tag{12}$$

In addition, the Dice coefficient is the most frequently used metric for medical image segmentation and it is an ensemble similarity metric. The calculation formula is:

$$\mathrm{Dice} = \frac{2|A \cap B|}{|A| + |B|} = \frac{2{*}TP}{2{*}TP + FP + FN} \tag{13}$$

The mean absolute error represents the average of the distance between the predicted value A and the sample true value B to evaluate the pixel-level accuracy. It is expressed as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|B_i - A_i| \tag{14}$$

In the above formula, A denotes the predicted value and B denotes the ground truth. TP is predicted as a positive example and the actual as a positive example; FP is predicted as a positive example and the actual as a negative example; FN is predicted as a negative example and the actual as a positive example; TN is predicted as a negative example and the actual as a negative example.

4.2. Experimental Details

The model was implemented based on Pytorch, and a single NVIDIA TESLA P40 24GB GPU was used to train the model. The batch size for all experiments was set to

32, while using Res2Net as the backbone network. As shown in each sky-blue box in the overall architecture of the network in Figure 2, the intermediate backbone feature maps were extracted from the residual block at the end of each stage. Correspondingly, the step size and expansion rate were modified to increase the spatial size of the feature map. The sizes of the images were uniformly adjusted to 352 × 352 pixels in the training phase, and then the images were adjusted to their initial sizes. During training, additional data enhancement methods were used, including random flipping on the horizontal and vertical axes and random multi-scale scaling of images from 0.75–1.25 times. In addition, because the polyp images may be rotated, the data were randomly rotated within 0–360 degrees for data augmentation and the Adam optimizer was used to optimize the learning rate. The initial learning rate was set to 1×10^{-4}, and the experimental model was iterated 250 times. The loss function $\mathcal{L}$ used in the experiment combines the binary cross entropy (BCE) loss function and the intersection over union ratio (IoU) loss function, which is as follows:

$$\mathcal{L} = \mathcal{L}_{BCE} + \mathcal{L}_{IoU} \tag{15}$$

$$\mathcal{L}_{BCE} = -\sum_{i \in I} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{16}$$

$$\mathcal{L}_{IoU} = 1 - \frac{\sum\limits_{i \in I} y_i \hat{y}_i}{\sum\limits_{i \in I} y_i + \hat{y}_i - y_i \hat{y}_i} \tag{17}$$

where $i \in I$ refers to a pixel in the output value and the ground truth, y is the ground truth, and $\hat{y}$ denotes the output.

5. Results

5.1. Experimental Results under Different Methods

This paper compares the proposed method with previous state-of-the-art methods on five common polyp segmentation benchmarks. In order to effectively validate the advantages of the polyp segmentation network in this paper, UNet, UNet++, ResUNet++, AttentionUNet, and PraNet were selected as the polyp segmentation models for comparison. During the experiment, the network structure was changed while keeping other conditions consistent. Tables 2 and 3 demonstrate the selected polyp segmentation models and the evaluation metrics of the method in this paper on the five polyp segmentation benchmarks. Among them, the higher value of the mean intersection over union(mIoU) and the mean Dice (mDic) indicate the better the performance of this method; the lower value of the mean absolute error (MAE) indicates the better the performance of this method. In addition, the segmentation visualization results of each network model are shown in Figure 8. From the quantitative and qualitative analysis of the experiments, the generalization performance of the proposed nonlinear activation-free uncertainty contextual attention polyp segmentation network as well as the segmentation results have some degree of advantages compared with other methods.

Table 2. Performance comparison of each method on the CVC-ClinicDB and Kvasir datasets.

Method	**CVC-ClinicDB**			**Kvasir**		
	mIoU	**mDic**	**MAE**	**mIoU**	**mDic**	**MAE**
Unet [1]	0.7556	0.8232	0.0193	0.7462	0.8184	0.0550
UNet++ [12]	0.7298	0.7945	0.0226	0.7430	0.8215	0.0482
ResUNet++ [13]	0.7964	0.8092	0.0160	0.7933	0.8132	0.0544
AttentionUNet [22]	0.7745	0.8493	0.0216	0.7675	0.8420	0.0437
PraNet [18]	0.8490	0.8992	0.0092	0.8403	0.8980	0.0302
NAF_UCANet	0.8824	0.9276	0.0063	0.8576	0.9133	0.0250

Table 3. Performance comparison of each method on the CVC-ColonDB, ETIS, and CVC-300 datasets.

Method	CVC-ColonDB			ETIS			CVC-300		
	mIoU	mDic	MAE	mIoU	mDic	MAE	mIoU	mDic	MAE
Unet [1]	0.4442	0.5120	0.0614	0.3352	0.3980	0.0363	0.6274	0.7102	0.0221
UNet++ [12]	0.4104	0.4830	0.0641	0.3446	0.4013	0.0355	0.6240	0.7074	0.0182
ResUNet++ [13]	0.3879	0.4844	0.0783	0.2274	0.2886	0.0552	0.4946	0.5968	0.0253
AttentionUNet [22]	0.5346	0.6224	0.0504	0.3720	0.4228	0.0390	0.7392	0.8254	0.0137
PraNet [18]	0.6403	0.7098	0.0455	0.6273	0.6796	0.0316	0.7971	0.8713	0.0103
NAF_UCANet	0.7120	0.7847	0.0334	0.6782	0.7644	0.0124	0.8496	0.9122	0.0049

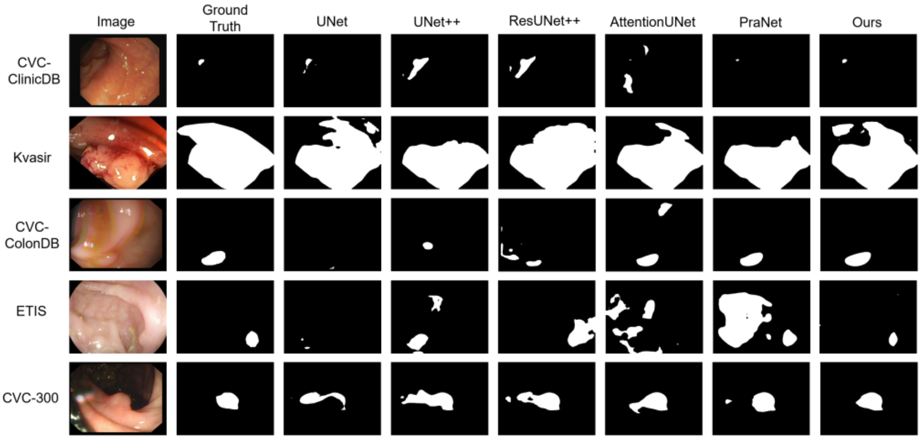


Figure 8. Visualization of different network models on each dataset.

It can be seen from Table 2 that the NAF_UCANet in this paper achieves an mIoU of 0.8824, as well as an mDic coefficient of 0.9276 and an MAE of 0.0063. Compared with the better performing PraNet, the mIoU increased by 3.34%, the mDic coefficient increased by 2.84%, and at the same time the MAE decreased by 0.29%.

From the various evaluation metrics of different polyp segmentation networks on the Kvasir dataset in Table 2, it can be seen that the NAF_UCANet network in this paper obtained an mIoU of 0.8576, an mDic coefficient of 0.9133, and an MAE value of 0.0250. Compared to the current leading performance PraNet network, the mIoU increased by 1.67%, the mDic factor increased by 1.53%, and the MAE decreased by 0.52%.

From Table 3, the method in this paper achieved an mIoU of 0.7120, an mDice coefficient of 0.7847, and an MAE of 0.0334, which is a 7.17% increase in mIoU, a 7.49% increase in mDic coefficient, and a 1.21% decrease in MAE compared to the performance of PraNet on the CVC-ColonDB dataset.

As shown by the data in Table 3, the current PraNet network with better polyp segmentation results had an mIoU value of 0.6273 on the ETIS dataset, an mDic coefficient of 0.6796, and an MAE of 0.0316. It had a better performance compared with some commonly used polyp segmentation networks. The NAF_UCANet in this paper achieved an mIoU of 0.6782, as well as an mDic coefficient of 0.7644 and an MAE of 0.0124, both of which are more significant improvements compared to the better performing PraNet.

From Table 3, the method in this paper achieved an mIoU of 0.8496, an mDic coefficient of 0.9122, and an MAE of 0.0049, which is a 5.25% increase in the mIoU, a 4.09% increase in the mDic coefficient, and a 0.54% decrease in the MAE compared to the performance of PraNet on the CVC-300 dataset.

From the above quantitative and qualitative analysis of different networks in each experimental dataset, the network in this paper has a better effect on the segmentation of polyp images that are smaller and have similar pixels around them. For the five classical polyp segmentation benchmarks used in this paper, all evaluation metrics have been improved to a certain extent. Three datasets, CVC-ColonDB, ETIS, and CVC-300, were not trained with the model, but also showed significant improvements in the evaluation metrics. The mean Dice coefficient, an important evaluation metric for medical image segmentation, increased by 7.49%, 4.09%, and 8.48% on the three datasets that were not involved in model training, respectively, which to some extent indicates the effectiveness and generalization of the method in this paper. In addition, due to the variable size of polyps in the five polyp benchmarks, a longitudinal comparison of the mean Dice coefficients of the different methods in the five datasets showed that our method has improved to a certain extent. Whether for the CVC-ColonDB dataset with a large polyp area or the ETIS and CVC-300 datasets with a small polyp area, the method in this paper can effectively segment the polyp area, which is less affected by the similarity of surrounding pixels as well as blurring. The high agreement with the ground truth better demonstrates the performance of the method in this paper.

5.2. Ablation Experiment of Different Modules

This subsection is designed to validate the effectiveness of the modules in the overall architecture of the network. Ablation experiments with different modules were performed in the five selected polyp datasets. UNet was used as a segmented baseline network for comparison with the other baseline networks in this experiment. Res2Net fuses some decoders as well as multiple features as the backbone network of the experiment and adds each module for comparison experiments to reflect the role of different modules in terms of data. In addition, experiments using ResNet50 as the baseline network, together with the modules in this paper for experimental comparison, further illustrate the effectiveness of the method proposed in this paper and also show that the backbone network used in this paper has a higher performance than the other baseline networks for segmentation. The ablation experiments are shown in Tables 4 and 5. The experimental data reflect the high and low performance judging criteria as in the previous subsection.

Table 4. Ablation experiments for each module on the CVC-ClinicDB and Kvasir datasets.

Method	**CVC-ClinicDB**			**Kvasir**		
	mIoU	**mDic**	**MAE**	**mIoU**	**mDic**	**MAE**
UNet	0.7556	0.8232	0.0193	0.7462	0.8184	0.0550
ResNet50	0.8382	0.8806	0.0119	0.8255	0.8746	0.0291
ResNet50+SPACA	0.8494	0.8920	0.0103	0.8328	0.8823	0.0289
ResNet50+NAF	0.8660	0.9101	0.0080	0.8451	0.8930	0.0276
ResNet50+UCA	0.8545	0.9012	0.0097	0.8436	0.8863	0.0284
Baseline	0.8700	0.9178	0.0079	0.8558	0.9022	0.0248
Baseline+SPACA	0.8632	0.9080	0.0083	0.8470	0.8979	0.0300
Baseline+NAF	0.8735	0.9195	0.0074	0.8529	0.9051	0.0285
Baseline+UCA	0.8703	0.9163	0.0077	0.8523	0.9050	0.0264
NAF_UCANet	0.8824	0.9276	0.0063	0.8576	0.9133	0.0250

Table 5. Ablation experiments for each module on the CVC-ColonDB, ETIS, and CVC-300 datasets.

Method	**CVC-ColonDB**			**ETIS**			**CVC-300**		
	mIoU	**mDic**	**MAE**	**mIoU**	**mDic**	**MAE**	**mIoU**	**mDic**	**MAE**
UNet	0.4442	0.5120	0.0614	0.3352	0.3980	0.0363	0.6274	0.7102	0.0221
ResNet50	0.6545	0.7348	0.0396	0.5525	0.6319	0.0277	0.7659	0.8455	0.0114
ResNet50+SPACA	0.6637	0.7370	0.0370	0.5770	0.6532	0.0247	0.7714	0.8500	0.0112
ResNet50+NAF	0.6722	0.7413	0.0376	0.5748	0.6520	0.0215	0.7958	0.8711	0.0086
ResNet50+UCA	0.6586	0.7393	0.0388	0.5756	0.6584	0.0236	0.7893	0.8672	0.0099
Baseline	0.6741	0.7474	0.0407	0.5898	0.6625	0.0261	0.8196	0.8905	0.0065
Baseline+SPACA	0.6862	0.7628	0.0354	0.6111	0.6921	0.0173	0.8214	0.8929	0.0060
Baseline+NAF	0.6837	0.7542	0.0331	0.5978	0.6696	0.0203	0.8287	0.8992	0.0056
Baseline+UCA	0.6779	0.7507	0.0395	0.6155	0.6938	0.0228	0.8370	0.9025	0.0055
NAF_UCANet	0.7120	0.7847	0.0334	0.6782	0.7644	0.0124	0.8496	0.9122	0.0049

Performing the ablation experiments using each module on the above five polyp segmentation benchmarks is compared with the evaluation metrics of our baseline network on each dataset. In this experiment, UNet was added as the baseline network as well as ResNet50 as the backbone network for the ablation experiments while the overall network architecture remained unchanged. The choice of ResNet50 as the backbone network to replace the backbone network Res2Net in this paper further proves that the selected backbone network has certain advantages in this paper. At the same time, through different baseline networks, the effectiveness of each module can also be better explained. The CVC-Colon300 and ETIS are more challenging in the dataset chosen for this paper, but each module in the method of this paper is fused with the chosen benchmark network, and there is some improvement in performance. In particular, when the modules are fused to form the network proposed in this paper, the mIoU, mDice coefficient, and MAE values are significantly improved. For the simple parallel axial channel attention module, the introduction of this module can make the segmentation performance of the whole network better. Similarly, the introduction of the nonlinear activation-free network results in a better improvement of the average segmentation level of the network. In addition, the inclusion of the uncertainty contextual attention mechanism enables more precise segmentation of the polyp area by quantitative analysis, which is less affected by the mucosa around the polyp and the colon indicating blurred areas. It is better to judge that the uncertainty area is closely related to the polyp boundary, so that the segmentation effect can be better approximated to the ground truth. In addition, the fusion of the backbone network Res2Net selected in this paper and the proposed modules and the segmentation diagram of the NAF_UCANet in each dataset are given, as shown in Figure 9. The visualization of the segmentation by network model visually analyzes the higher performance and accuracy of the method in this paper to segment the polyp image, which is closer to the ground truth. Especially for complex polyp datasets like ETIS, the modules introduced in the method of this paper can better segment the polyp contours. In summary, through quantitative and qualitative experimental analysis, the combination of various modules in the network further proves the effectiveness of the method in this paper.

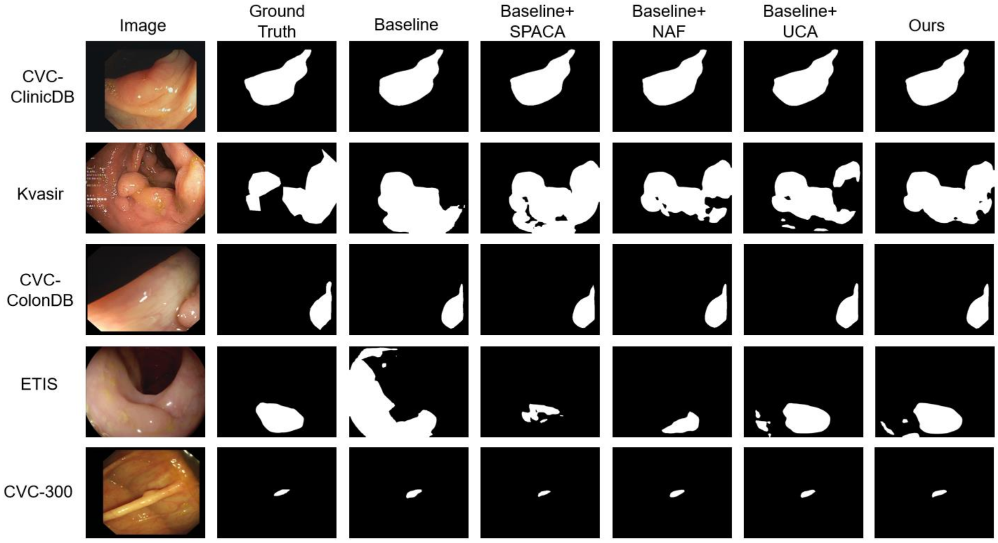


Figure 9. Visualization of ablation experiments for each module of this paper from each dataset.

6. Discussion

The accurate localization of polyps in colonoscopic images facilitates timely treatment. Nowadays, accurate and efficient methods of polyp segmentation have gained much attention and are of great importance to the medical field. Medical image segmentation is also applicable to other parts of the body, and the literature [4] has proposed a hybrid framework to implement brain tumor segmentation that is able to handle complex problems such as noise influence and intensity variations between soft tissues. In addition, the literature [25] has proposed a two-stage 3D UNet based on deep learning to complete the extraction and multi-structure segmentation of brain tumor images. For polyp segmentation, the literature [18] has introduced a parallel reverse attention mechanism network to precisely segment polyps and mitigate the effects of noise caused by uneven light distribution and the randomization of polyp location. In the literature [38], a text-guided attention architecture has been proposed to address the problem of variable size and number of polyps. Table 6 compares the performance of the method proposed in this paper with some related methods for multiple polyp segmentation datasets as well as segmentation of other parts.

Table 6. A comparison of the segmentation performance of the method in this paper and related methods on the same evaluation metrics.

References	Sensitivity%	mDic %	Segmented Parts
[4]	88.7	-	Brain Tumor
[25]	-	87.9	Brain Tumor
[18]	91.7; 90.9; 78.8; 81.2; 95.9	91.8; 89.4; 77.1; 68.0; 89.9	Polyp
[38]	91.3; 94.4; 90.3; 79.2; 82.9	89.8; 94.6; 90.2; 79.8; 81.9	Polyp
NAF_UCANet	94.3; 92.2; 80.1; 81.3; 93.9	92.8; 91.3; 78.5; 76.4; 91.2	Polyp

The method in this paper provides accurate segmentation of polyps to a certain extent, but further improvement is needed. For example, there are obvious areas of fuzzy bubbles in the colonoscopy images and areas with high pixel similarity around the polyps, which may be classified as polyps by the method when it cannot be accurately judged. In addition, the method in this paper is applied to 2D polyp segmentation, and the generalization for some spatial 3D medical image segmentation is insufficient. In future work the judgment of the uncertainty region will be further enhanced, so that the network can meet the segmentation task with higher precision. The spatial context information can also be considered to achieve 3D spatial image segmentation, thus improving the accuracy and generalization of the method in this paper.

7. Conclusions

In this paper, a new polyp segmentation network, NAF_UCANet, is proposed to achieve more effective segmentation for polyps with different shapes and sizes, colors and textures, as well as mucous membrane and blurred areas around polyps. The introduction of a nonlinear activation-free function in the network can enhance the uncertainty regions on the saliency maps that are highly correlated with the boundary information. The calculation of regions with fuzzy saliency scores is achieved using simple parallel axial channel attention, which combines foreground and background regions to implement a contextual attention module, so as to obtain both low-level features with edge details as well as global semantic high-level features. Through experiments and analysis from quantitative and qualitative perspectives, it is demonstrated that the polyp segmentation method proposed in this paper has high segmentation accuracy and generalizability. Furthermore, the effectiveness of the proposed network in performing segmentation of small polyps can be further improved. In future work the judgment of the uncertainty region will be further enhanced, so that the network can meet the segmentation task with higher precision.

Author Contributions: Conceptualization, W.W. and H.F.; methodology, W.W.; software, Y.F. and J.W.; validation, W.W. and Y.F.; formal analysis, W.W. and H.F.; investigation, Y.F. and J.W.; resources, W.W.; data curation, W.W.; writing—original draft preparation, W.W.; writing—review and editing, W.W., H.F. and J.W.; visualization, W.W. and Y.F.; supervision, W.W. and H.F.; project administration, W.W.; funding acquisition, H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data involved in the experiments were downloaded from the respective official websites.

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

References

1. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. In *Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
2. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
3. Haghighi, F.; Hosseinzadeh Taher, M.R.; Zhou, Z.; Gotway, M.B.; Liang, J. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. Medical Image Computing and Computer Assisted Intervention–MICCAI 2020. In *Proceedings of the 23rd International Conference, Lima, Peru, 4–8 October 2020*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 137–147.
4. Bourouis, S.; Alroobaea, R.; Rubaiee, S.; Ahmed, A. Toward effective medical image analysis using hybrid approaches—Review, challenges and applications. *Information* **2020**, *11*, 155. [CrossRef]
5. Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med. Imaging* **2015**, *35*, 630–644. [CrossRef]
6. Puyal, J.G.B.; Bhatia, K.K.; Brandao, P.; Ahmad, O.F.; Toth, D.; Kader, R.; Lovat, L.; Mountney, P.; Stoyanov, D. Endoscopic polyp segmentation using a hybrid 2D/3D CNN. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 295–305.
7. Feng, S.; Zhao, H.; Shi, F.; Cheng, X.; Wang, M.; Ma, Y.; Xiang, D.; Zhu, W.; Chen, X. CPFNet: Context pyramid fusion network for medical image segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 3008–3018. [CrossRef]
8. Song, J.; Chen, X.; Zhu, Q.; Shi, F.; Xiang, D.; Chen, Z.; Fan, Y.; Pan, L.; Zhu, W. Global and local feature reconstruction for medical image segmentation. *IEEE Trans. Med. Imaging* **2022**, *41*, 2273–2284. [CrossRef]
9. Wen, Y.; Chen, L.; Deng, Y.; Zhang, Z.; Zhou, C. Pixel-wise triplet learning for enhancing boundary discrimination in medical image segmentation. *Knowl. Based Syst.* **2022**, *243*, 108424. [CrossRef]
10. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [CrossRef]
11. Yagang, W.; Yiyuan, X.I.; Xiaoying, P.A.N. Method for intestinal polyp segmentation by improving DeepLabv3+ network. *J. Front. Comput. Sci.Technol.* **2020**, *14*, 1243.
12. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Granada, Spain, 20 September 2018*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
13. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; Lange, T.D.; Halvorsen, P. 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; pp. 225–2255.
14. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual U-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
15. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 10–22 June 2018; pp. 7132–7141.
16. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
17. 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.
18. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. Medical Image Computing and Computer Assisted Intervention–MICCAI 2020. In *Proceedings of the 23rd International Conference, Lima, Peru, 4–8 October 2020*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 263–273.
19. Bardhi, O.; Sierra-Sosa, D.; Garcia-Zapirain, B.; Bujanda, L. Deep Learning Models for Colorectal Polyps. *Information* **2021**, *12*, 245. [CrossRef]
20. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
21. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
22. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **2019**, *53*, 197–207. [CrossRef]

23. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA; 2019; pp. 3146–3154.
24. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 173–190.
25. Tomassini, S.; Anbar, H.; Sbrollini, A.; Mortada, M.J.; Burattini, L.; Morettini, M. A Double-Stage 3D U-Net for On-Cloud Brain Extraction and Multi-Structure Segmentation from 7T MR Volumes. *Information* **2023**, *14*, 282. [CrossRef]
26. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H.; Shao, L. Multi-stage progressive image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 14821–14831.
27. Cho, S.J.; Ji, S.W.; Hong, J.P.; Jung, S.W.; Ko, S.J. Rethinking coarse-to-fine approach in single image deblurring. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 4641–4650.
28. Waqas, Z.S.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient Transformer for High-Resolution Image Restoration. *arXiv* **2021**, arXiv:2111.09881.
29. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International Conference on Machine Learning, PMLR, International Convention Centre, Sydney, Australia, 6–11 August 2017; pp. 933–941.
30. Chen, L.; Chu, X.; Zhang, X.; Sun, J. Simple baselines for image restoration. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 17–33.
31. Brandao, P.; Zisimopoulos, O.; Mazomenos, E.; Ciuti, G.; Bernal, J.; Scarzanella, M.V.; Menciassi, A.; Dario, P.; Koulaouzidis, A.; Arezzo, A.; et al. Towards a computedaided diagnosis system in colonoscopy: Automatic polyp segmentation using convolution neural networks. *J. Med. Robot. Res.* **2018**, *3*, 1840002. [CrossRef]
32. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
33. Chen, S.; Tan, X.; Wang, B.; Hu, X. Reverse attention for salient object detection. In Proceedings of the European Conference on Computer Vision (ECCV), München, Germany, 8–14 September 2018; pp. 234–250.
34. Wu, Z.; Su, L.; Huang, Q. Cascaded partial decoder for fast and accurate salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3907–3916.
35. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
36. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
37. 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.* **2019**, *43*, 652–662. [CrossRef]
38. Tomar, N.K.; Jha, D.; Bagci, U.; Ali, S. TGANet: Text-guided attention for improved polyp segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022*; Springer: Cham, Switzerland, 2022; pp. 151–160.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.