IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 April 2023, accepted 16 May 2023, date of publication 19 May 2023, date of current version 19 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3278109

RESEARCH ARTICLE

AGCNet: A Precise Adaptive Global Context Network for Real-Time Colonoscopy

LIANTAO SHI[1], ZHENGGUO LI[1], JIANYANG LI[2], YUFENG WANG[2], HONGYU WANG[3], AND YUBAO GUO[2]

[1]Institute for Carbon-Neutral Technology, Shenzhen Polytechnic, Shenzhen 518055, China
[2]School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan 114000, China
[3]School of Automation, Wuhan University of Technology, Wuhan 430000, China

Corresponding author: Zhengguo Li (Lizhengguo@szpt.edu.cn)

ABSTRACT Colonic endoscopy is the gold standard for detecting rectal polyps and rectal cancer. In which polyps are a major predisposing factor for colorectal cancer, the precise diagnosis of polyps within colorectal endoscopy is highly dependent on a physician of professional level. With the development of deep learning, some semantic segmentation methods have recently been applied to polyp detection, but there are problems with insufficient accuracy and segmentation speed. To this end, we propose a precision adaptive global context network (AGCNet) based on real-time colon endoscopy. Firstly, in order to adapt to the problem of large-scale variation of polyps, we designed a multi-scale semantic fusion module (MSFM), which enhances the representation capability by varieties of filters to collect contextual information at different scales, thus adapting to the problem of large variation of polyp size, especially smaller polyps. In addition, modelling long-range dependence by simply using complex spatial pixels tends to introduce more background noise and increase the computational effort. To this end, a context-aware pyramid aggregation module (CPAM) was designed, which internally includes a novel dual attention mechanism whereby the CPAM aggregates feature information across different regions to boost the network's ability to utilize global context and model long-range dependency through dual attention to further reinforce the features information of important regions and efficiently suppress features in non-important regions. Additionally, the CPAM performs multi-level pooling on the input features to extract multi-scale context information from the image and uses an attention mechanism to selectively highlight informative regions of the image that are most relevant to the segmentation task. The module fuses the multi-level pooled features with the attention map to produce enhanced feature representations that capture both global and local information. Thereby achieving precise polyp segmentation and taking real-time into account. Our proposed AGCNet performed extensive experimental studies on datasets Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB and ETIS-LaribPolypDB. Specifically, AGCNet achieved an IoU of 87.40% and a Dice score of 92.63% on the Kvasir dataset, achieving accurate segmentation results faster than many current state-of-the-art models.

INDEX TERMS Colonoscopy, polyp segmentation, multi-scale semantic feature, context-guided pyramid aggregation module, feature aggregation.

I. INTRODUCTION

Early diagnosis of colorectal cancer (CRC) improves the patients' survival rate. Most CRCs start as adenomatous polyps: surface protrusions on the colon and rectum lining. Over time, they grow into malignant tumours and spread to surrounding organs. The survival rate drops from 95% in the first stage to 35% in the fourth and fifth stages [1]. Early screening and removal of polyps can increase the survival rate. Colonoscopy is the standard for screening CRC, but even so, accurate identification of polyps remains challenging due to (1) the large variation in scale between polyps (Fig.1). (2) the blurring of polyp border information (Fig.1(c)-(d)). (3) low contrast between polyps and gastrointestinal tract.

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

(4) endoscopist's skill [2], [3]. Adenomatous polyps detection rate (ADR) measures physician quality, which indicates the percentage of polyps diagnosed in patients after a complete colonoscopy. ADR varies from 7% to 53%, yet statistically, there is a 25% probability that polyp will be missed in each patient's diagnosis [4]. It is worth mentioning that there is a causal relationship between ADR and reduced CRC mortality. According to the data, a 1% increase in ADR is associated with a 3% decrease in interval cancer [5]. The main factors that can affect ADR are mescal intubation rates, withdrawal times, and quality of bowel preparation, which depend on human intervention by endoscopists. However, the human approach brings great uncertainty and reliability to the diagnostic results, so there is an urgent need for an automated polyp segmentation method to reduce the misdiagnosis brought by human factors.

Polyp segmentation during colonoscopy screening can prevent colorectal cancer (CRC), which is mainly caused by polyps. Polyps are surface protrusions on the colon and rectum lining that can grow into malignant tumours and spread to surrounding organs. Segmentation is challenging because some polyps are flat, with low contrast to the mucosal border [6]. This requires the expertise of endoscopists to reduce the rate of missed examinations during colonoscopy. Most CRC patients (91% to 94%) do not have endoscopy before the disease; others (6% to 9%) have endoscopy but misdiagnose flat polyps [7]. Despite timely colon endoscopy, the leading cause of cancer is the high rate of misdiagnosis of flat polyps, which are also the focus of segmentation.

The subsequent efforts have been made to develop appropriate protocols to address the many challenges in polyp segmentation in colon endoscopy. Previous studies used hand-crafted methods [8], [9] to train a classifier model based on shape, color, texture and appearance. This method could not segment heterogeneous polyps well, resulting in low accuracy and performance [10]. Later studies used deep learning methods to extract features automatically [11]. Although these approaches improve traditional manual segmentation, using bounding boxes alone can only yield image-wise results that cannot distinguish the boundary line between the polyp and the mucosa [12]. FCN was proposed to segment polyps using pre-trained weights for pixel-wise results. However, this method lost spatial information at low dimensions and did not calibrate deep semantic information with shallow information. Inspired by FCN, UNet with an encoder-decoder U-shaped architecture was proposed [13], [14]. The network has been widely used in medical image segmentation since its introduction. However, the UNet model had limitations, such as generating redundant information and increasing computation with traditional convolution. This also provided an opportunity to improve performance further.

Since the encoding process of UNet directly uses a pooling operation to compress the resolution, it is easy to cause the loss of some spatial information. In addition, in skip architecture, the feature map of the encoder-decoder is directly concatenated as feature input easily increases the amount of redundant information. MRUNet provides a multi-scale and residual scheme, using multi-parallel and multi-scale convolution instead of the traditional convolution of the UNet encoder and decoder can effectively reduce the semantic information gap. Simultaneously, the skip connection part is replaced by the residual model. However, MRUNet's operation-only models skip architecture by specifying the feature maps of the encoding layers at a certain level to reduce the semantic information gap of the corresponding decoders. UNet++ takes into account the semantic differences between the encoding and decoding layers and designs a series of nested and dense jump paths at the skip architecture, allowing the decoding layer to take advantage of more rich contextual information in the encoding layer [15]. However, this approach introduces more complex computation and leads to more difficulty of optimizer and back-propagation [16]. To further enhance the ability of context information extraction and computing optimization, CPFNet discards a single stage to model the context information, using and modelling a double pyramidal module to extract the global context information [17]. Besides, PraNet applies region and boundary cues to design a parallel reverse attention mechanism that corrects some misaligned predictions. Considering the uncertainty region of the salient features of polyp segmentation, UACANet proposed an uncertainty-enhanced contextual attention model [18]. However, these methods cannot bridge the semantic information gap between different levels and utilize the global information to achieve an excellent segmentation result while maintaining real-time performance.

In this paper, we propose an adaptive global context architecture named AGCNet, equipped with two new multi-scale semantic extraction and dual attention approaches that can meet the current challenges in segmenting polyps under colonoscopy video. AGCNet compares with the currently existing methods. It models a long-range dependency by using contextual information at different scales without introducing extra computational costs to build confidence in the network to identify the large variation scale and shape of polyps. In addition, a novel method of dual attention mechanism is proposed to effectively suppress background noise without employing sophisticated non-local modelling techniques. The contribution of this work can be summarized as follows:

- We propose a novel MSFM module to enhance the network's multi-scale representation at a more granular level and aggregate multi-scale contextual information to model long-range global dependency. Thus, it can be self-adaptive to scale-variant polyps.
- In order to solve the interference of background noise in the gastrointestinal channel, we further designed the CPAM module to extract more discriminative features by suppressing the interference of irrelevant information through a dual-channel attention approach.

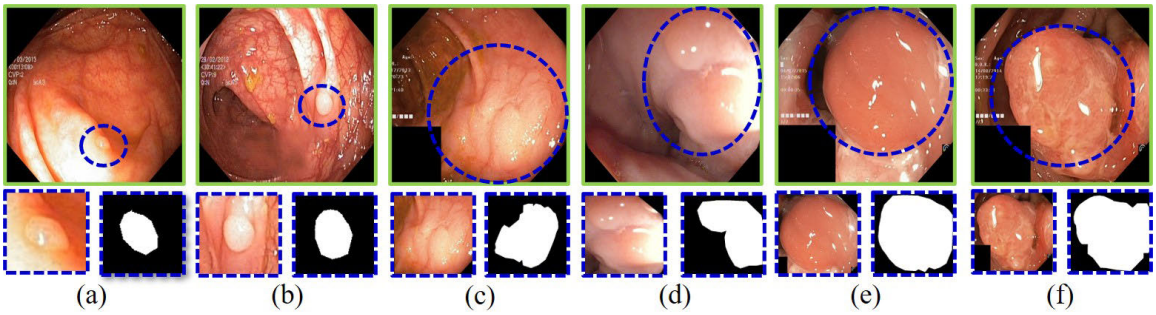


FIGURE 1. Typical challenging polyp image segmentation case: (a)-(b) show images of smaller polyps, while (c)-(f) show some images of polyps with large scale and blurred borders.

- Extensive experimental studies on five publicly available datasets have confirmed that AGCNet can produce more competitive results when compared with other state-of-the-art network models.

II. METHOD

Colorectal endoscopic images from four publicly available datasets, Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB and ETIS-LaribPolypDB, are preprocessed and passed through our AGCNet for feature extraction and segmentation of the polyp region. The whole network can be divided into a classical symmetric encoding and decoding region. The network framework is shown in Fig.2, which employs two novels and effectively validated new components Multi-scale Semantic Fusion Module (MSFM) and Context-aware Pyramid Aggregation Module (CPAM). To capture multi-scale contextual information, we use our proposed MSFM module in the encoding part of each layer, where the features are processed by multi-scale convolution and then passed back by the residual unit to obtain more discriminative representative features to fit the polyp size at different scales. At the bottom of the network, in order to mitigate the interference of the background noise of the polyps on the high-level semantic information, we further employ the CPAM module, which is used to enhance the target region and weaken the background region employing pyramid aggregation and a two-channel attention mechanism to exploit the global contextual information. The operation of the different modules in the AGCNet is discussed in detail in the following subsections.

A. MULTI-SCALE SEMANTIC FUSION MODULE

Due to the diversity of polyps and their different scales, the network tends to lose the boundary information in the downsampling process, resulting in the inability to identify variable polyps accurately. In this case, if the contextual information of the shallow represents information can be reasonably used, it will help the network to identify polyps of different shapes and scales. Inspired by Res2Net [19], we propose a Multi-scale Semantic Fusion Module (MSFM). Most existing methods use input feature maps with different resolutions to improve the multi-scale representation ability. However, it is easy to cause the loss of boundary information by reducing the fine-grained. We use convolution kernels with different scales to extract features at the more granularity level to increase the network's perception field and maintain the model's multi-scale representation ability. Then concatenate to output the final feature map.

As shown in Fig.3 We extract feature from input feature $X \in R^{C\times H\times W}$ by, including 1×1 convolution, Batch Normalization (BN) and activation function ReLU. Keeping the original input scale constant, we get a new feature map $X' \in R^{C\times H\times W}$, where C, H and W represent the number of channels, length and width of the feature map, respectively. Dividing features into four feature maps with an equal number of channels in channel dimension $X'' = [X_0, X_1, X_2, X_3] \in R^{\frac{C}{4}\times H\times W}$, where X_1, X_2, X_3 are transformed via $W_2(\cdot)$. Note that $W_2(\cdot)$ includes 3×3 convolution and BN operations. We concatenate the feature maps $W_2(X_1), W_2(X_2), W_2(X_3)$ and X_0 transformed by $W_2(\cdot)$ in the channel dimension in turn, as shown in the following equation.

$$X_{cat} = \mathrm{CONCAT}\left(W_2(X_1), W_2(X_2), W_2(X_3), X_0\right) \quad (1)$$

From the above equation, we obtain contextual information at different scales based on the extraction of different convolutional kernels, thus increasing the receptive field of the layer network. Finally, we fuse the resulting feature outputs with the original input features via a residual network operation, as defined by the following equation.

$$X_{Out} = W_3(X_{Cat}) \oplus X \quad (2)$$

where $\oplus$ represents the pixel-level additive summation operation, $W_3(\cdot)$ represents the 1×1 convolution, BN and ReLU nonlinear activation functions. To summarize the advantages of MSFM, firstly, it is different from using the resolution feature map to enhance the representational power of the network. MSFM enhances the network's representational power at a finer granularity level. Designing different convolutional kernels for feature extraction and then concatenating enables the network to obtain multi-scale contextual information increasing the perceptual field of the network. Finally, the model models the long-range dependence between the model

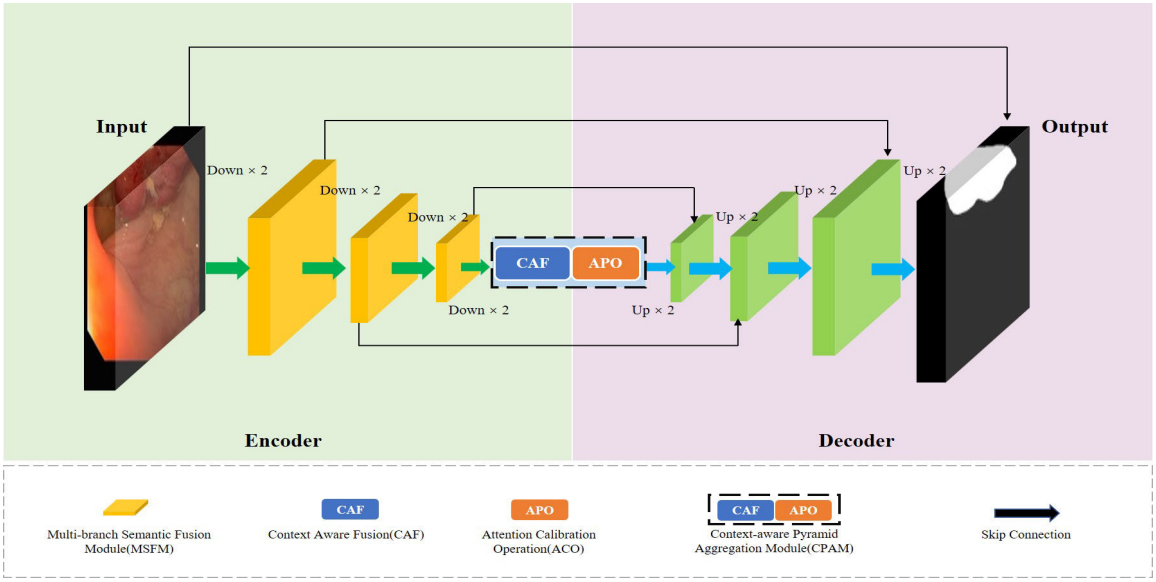


FIGURE 2. Schematic of the AGCNet architecture, which internally contains two main modules, MSFM and CPAM.

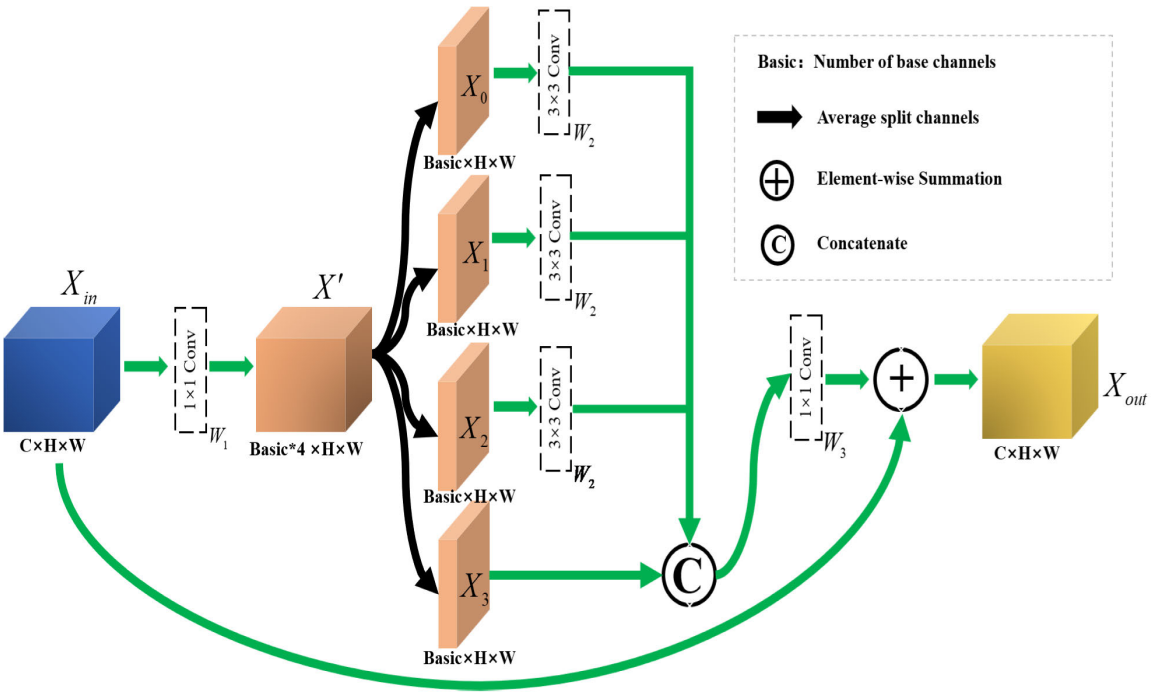


FIGURE 3. Schematic diagram of MSFM.

and the original input image, allowing the network to retain sufficient spatial detail information.

B. CONTEXT-AWARE PYRAMID AGGREGATION MODULE

In order to reduce the misjudgment of foreground and background information, we need to exploit a more extensive range of contextual information, which requires not only modelling the global context to capture the long-range dependencies more efficiently but also deepening the network to guide it to focus on the region of interest. Current state-of-the-art approaches model more complex long-range dependencies mainly by correlating pixels or channels, which not only increases the computational effort but also introduces some unavoidable background noise, thus reducing the segmentation accuracy of the network.

Inspired by ECA-Net [20] and PSPNet [21], we propose a Context-aware Pyramid Aggregation Module (CPAM), which adopts a more efficient context modelling to establish long-range dependencies and effectively enhances the information of cross-channel interactions. The procedure is divided into Context-Aware Fusion and Attention Calibration Operation.

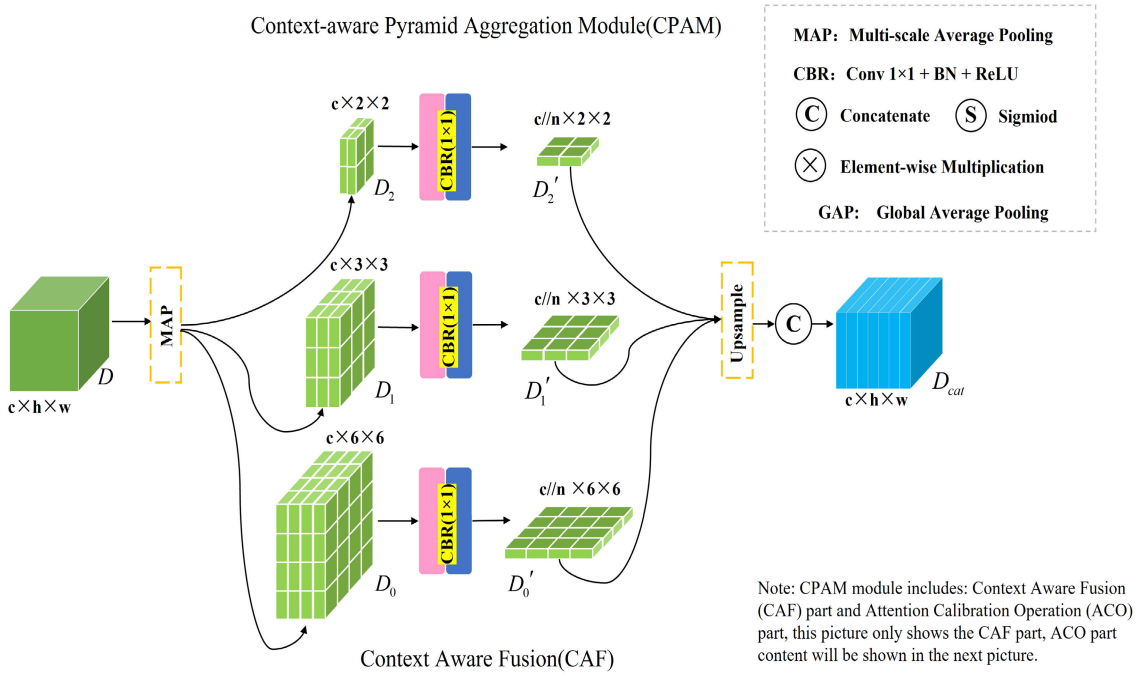


FIGURE 4. Schematic diagram of CPAM.

Operation of the Context-Aware Fusion, as shown in Fig.3. Given the input feature map as $D \in R^{C\times H\times W}$, first extract four feature map $D_0 \in R^{C\times 6\times 6}, D_1 \in R^{C\times 3\times 3}, D_2 \in R^{C\times 2\times 2}$ and $D_3 \in R^{C\times 1\times 1}$ with different resolutions of constant number of channels using multiple pooling operations (MPA) at different scales, and then reduce the dimensionality of the above four feature maps by CBR.

$$D_i'' = \left(Up\left(D_i', \beta_i\right)\right) \tag{3}$$

where $D_i' \in R^{\frac{C}{4}\times H\times W}$, $i = 0, 1, 2, 3$, and Up(·) denote bilinear interpolation up-sampling and β_i is the correlation coefficient. Through the operation of the context-aware fusion part, we obtain rich contextual information that is sufficient for subsequent long-range dependency modelling, which enhances feature differentiation. We connect the obtained feature mappings into the channel dimension as follows.

$$D_{\text{Cat}} = \text{CONCAT}\left(D_0'', D_1'', D_2'', D_3''\right) \tag{4}$$

The Attention Calibration Operations section is designed to enhance the network's representation capabilities further. In this section, Dual attention mechanisms are designed to model long-range dependency as shown in Fig.5. In the spatial attention mechanism, we first use convolution to reduce the dimension of the channel, get the attention weights through the Sigmoid activation function, then perform the attention matrix multiplication operation to reshape the weights of the original input feature maps, as follows.

$$D_{\text{Spatial}} = D_{\text{Cat}} \otimes \left(\sigma\left(S_0\left(D_{Cat}, \alpha\right)\right)\right) \tag{5}$$

where $\otimes$ is the multiplication of the attention matrix, $\sigma(\cdot)$ is the Sigmoid activation function to get the attention weight map, $S_0(\cdot)$ is the 1×1 convolution operation, and α is the S_0 correlation coefficient. In addition, to explicitly model the relationship between context and channel, we used an efficient channel attention mechanism in the other part, which is expressed as follows.

$$D_{\text{Chamnel}} = D_{\text{Cat}} \otimes \left(\sigma\left(F_{\text{Adaptive}}\left(G\left(D_{\text{Cat}}, \theta\right)\right)\right)\right) \tag{6}$$

where $F_{\text{Adaptive}}(\cdot)$ enables local cross-channel information interaction, i.e., how many neighbours are involved in the prediction of a channel's attention. Also, $F_{\text{Adaptive}}(\cdot)$ can adapt to the size of the kernel. $G(\cdot)$ denotes the global average pooling. $G(D) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} D(i,j)$ generates channel-wise statistics, θ is the correlation coefficient of $G(\cdot)$, and finally pixel-level multiplication is used to recalibrate the channel weights of D_{Cat}. We aggregate the feature maps processed by the channel and spatial attention mechanisms.

$$D_{\text{out}} = D_{\text{Spatial}} \oplus D_{\text{Channel}} \tag{7}$$

where $\oplus$ denotes the pixel-level addition operation that achieves feature fusion, our proposed multiscale context module based on a dual-channel attention mechanism, which does not need to build complex pixel-level and channel-level long-distance dependency, is equally capable of capturing different contextual information and can obtain the same ability to suppress background noise. In addition, enhanced model discriminative power for different features and precise segmentation can be achieved between polyps' background and foreground information.

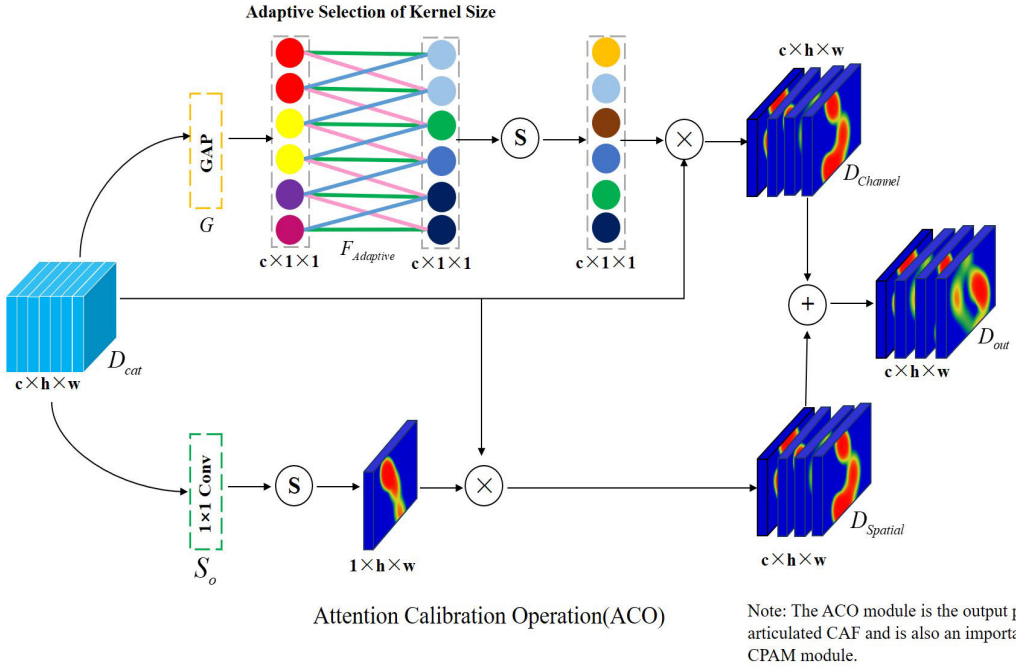


FIGURE 5. Schematic diagram of CPAM.

III. EXPERIMENT

A. LOSS FUNCTION

The loss function is essential for polyp image segmentation in colorectal endoscopic scenarios. Since there is a severe imbalance between the foreground and background information in the polyp image, choosing the most appropriate loss function can help suppress the background noise and accelerate the better convergence of the network. The current mainstream and advanced two loss functions are binary cross-entropy loss, which is formulated as follows [22], [23].

$$L_{BCE} = -\sum \left(p_i \ln\left(\hat{p}_i\right) + (1-p_i)\ln\left(1-\hat{p}_i\right)\right) \quad (8)$$

$$L_{\text{Dice}} = \frac{\left\|p_{(h,w)}\right\|_i + \hat{p}_{(h,w)}\|_i - 2\cdot\left\langle p_{(h,w)}, \hat{p}_{(h,w)}\right\rangle}{\left\|p_{(h,w)}\right\|_i + \left\|\hat{p}_{(h,w)}\right\|_i + \alpha} \quad (9)$$

where p_i and $\hat{p}_i$ denote the label value of the polyp and the predicted value of the polyp region, respectively, (h, w) denotes the image pixel coordinates, α represents the LaPlace smoothing factor used to accelerate the aggregation of the network. We set it to 1e-8 in our network and set up ablation experiments to find the best loss function for training our network model. Table 1 shows the segmentation results under three different loss functions, as shown in the table. The optimal segmentation results can be achieved when the combined mode's loss function is used, mainly due to the different scales of polyps and the interference of a large amount of background noise under the colorectal endoscope.

When using a single loss function for optimization, the gradient of the polyp region will be affected by the gradient of other background regions, increasing the difficulty of network training and affecting the accuracy of the final training results. However, the combination of two loss functions can be used for targeted learning and optimization of polyp regions in the process of backpropagation to alleviate the severe imbalance between foreground and background information. Finally, we obtain the final loss function.

TABLE 1. The ablation study of loss function on Kvasir-SEG dataset.

Loss function	IoU(%)	Dice(%)	Precision(%)
Dice	80.03	79.04	92.71
BCE	83.84	88.62	96.77
Dice+BCE	**87.40**	**92.63**	**95.83**

$$L_{\text{total}} = \lambda_1 \cdot L_{BCE} + \lambda_2 \cdot L_{\text{Dice}} \quad (10)$$

where λ_1 and λ_2 denote the loss function's relevant weight coefficient, we set them to 0.6 and 0.4, respectively.

B. DATASET AND EVALUATION

Four publicly available datasets were chosen for training and evaluating the performance of AGCNet. The CVC-ClinicDB dataset [24] consists of 612 images taken from 31 videos of different types of polyps, where the ground truth masks are hand-labelled by industry professionals, and all images have a resolution size of 384 × 288. The Kvasir-SEG dataset [25] contains 1000 images of polyps, and endoscopic experts annotated the corresponding ground truth mask maps at Oslo University Hospital. The ETIS-Larib dataset [29] contains 196 images of polyps with 1225×966 resolution. It was used as one of the dedicated test sets for automatic polyp segmentation in MICCAI 2015, which was a challenge to evaluate different polyp detection methods. The CVC-ColonDB dataset [1] contains 300 polyp images and their corresponding pixel-level annotated polyp mask maps with a resolution

of 574 × 500, and it was extracted from 15 video sequences, each containing one polyp. The image resolution sizes ranged from 332 × 482 to 1920 × 1072. The experimental part of AGCNet unified the resolution of all images to 512 × 512 size. We followed PraNet's [6] setup of the dataset and used 900 and 550 images for training, respectively, from the datasets Kvasir-SEG and CVC-ClinicDB. To effectively validate the model, we used a mixture of datasets developed by different medical centers to test the generalization ability of the model, including the CVC-ColonDB and ETIS-Larib datasets. We kept 100 and 62 images for testing, respectively.

In order to validate the performance of AGCNet from multiple perspectives, we used four main evaluation metrics to measure the effectiveness of different models for polyp image segmentation, mainly Recall, Precision, the Dice score and the Jaccard similarity coefficient, and also these metrics are widely used in the field of medical image segmentation. The specific equation is as follows.

TP and TN represent true positives and negatives, indicating the network's ability to segment the polyp's foreground and background pixels correctly. Similarly, FP and FN are false positives and false negatives, respectively, representing the network metrics misclassifying the foreground and background pixels of polyps. A represents the set of polyp segmentation result pixels and B refers to the set of actual polyp data label pixels. In addition, Recall and Precision metrics only focus on the distribution of independent pixels, which may affect the final evaluation results, so we introduced the area under the curve (AUC) value in addition to the four main metrics mentioned above to enable a complete evaluation of the results. The GFLOPS stands for Giga Floating-point Operations Per Second, which is a unit of measure for the computational speed of a computer or a processor. The Parameter means the number of parameters the model learns during the training process, and we use millions as the unit of measurement. Finally, considering the need for inherent real-time performance in the colorectal endoscopy scenario, we introduce FPS (Frames Per Second) and execution time to evaluate the real-time performance to meet the clinical needs.

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Dice = \frac{2 \cdot TP}{FP + FN + 2 \cdot TP} \tag{13}$$

$$IoU(A, B) = \frac{A \cap B}{A \cup B} \tag{14}$$

1) IMPLEMENTATION DETAILS

The choice of different hyperparameters is of great importance for improving network performance. For the two main current optimizers, SGD optimization [26] and Adam optimization [27]. We conducted comparative experiments as shown in Table 2, seeing that the SGD optimizer is more suitable for AGCNet and set the initial learning rate to 1e-3. We set the batch size to 4 and used an NVIDIA RTX3090 24GB graphics card for the experiments. We also used several data enhancements: RandomRotate, HorizontalFlip and RandomBrightnessContrast. The network tends to be stable when the number of epochs is 80.

TABLE 2. The ablation study of optimizer on Kvasir-SEG dataset.

Optimizer	IoU(%)	Dice(%)	Precision(%)
Adam	86.31	91.79	**96.55**
SGD	**87.40**	**92.63**	95.83

2) ABLATION STUDIES

To demonstrate the performance of AGCNet, this paper leads ablation experiments to evaluate the effectiveness of both MSFM and CPAM modules on the dataset Kvasir-SEG.

As shown in Fig.6, our ablation experiments compared some representative cases of visual challenges. As shown in the third column of Fig.6, the baseline model shows a mismatch of semantic information. After MSFM extracted the contextual information of multi-scale polyps and calibrated the semantic information of the encoding layer, as shown in the fourth column of Fig.6, we obtained more explicit polyp segmentation images. In addition, we also performed a comparison on the Kvasir-SEG dataset as shown in Table 3, and the Baseline+MSFM approach achieved 77.69%, 82.06% and 94.88% for IoU, Dice and Precision, respectively, which were ahead of Baseline by 7.45%, 3.99% and 5.37%. The result also means that MSFM can enhance the network to adapt to multi-scale polyps at a more granular level using rich semantic information. CPAM suppressed background noise in deep semantic information through a hybrid attention mechanism, which effectively enhanced the model's discriminatory ability between the target region and background tissue. Compared with the Baseline, our Baseline+CPAM can better suppress the background noise and obtain better polyp segmentation results, as shown in the fifth column of Fig.6. In addition, the advantages of CPAM for processing polyp images with low contrast between foreground and background can be seen in Table3, where CPAM achieves 71.53%, 79.25% and 89.66% for IoU, Dice and precision, respectively. Finally, we have seamlessly integrated MSFM and CPAM onto AGCNet. In this way, we not only obtain a more granular level of feature information in the encoder part and a wider range of contextual information through multi-level feature fusion but also strengthen the feature information of important regions and weaken the interference of background noise before decoding starts. AGCNet not only allows richer contextual information transfer between two adjacent layers in the coding layer but also enhances the ability of features to suppress the background noise of high-dimensional semantic information, as shown in the sixth column of Fig.6.

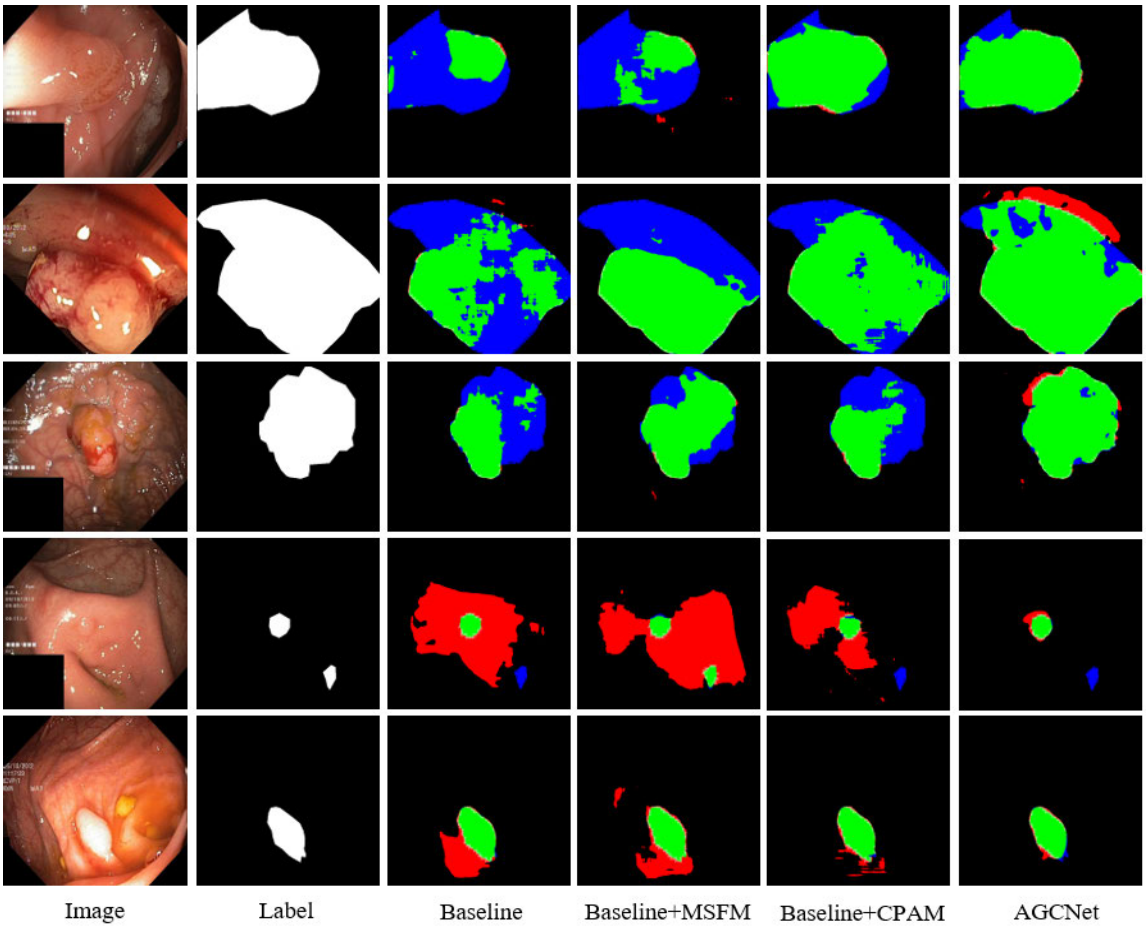


FIGURE 6. Ablation experiments between AGCNet internal modules.It is worth noting that the green, red and blue colors in the image indicate ture positive, false positive and false negative, respectively.

TABLE 3. Ablation experiments of the internal module of AGCNet on the Kvasir-SEG dataset.

Setting	IoU(%)	Dice(%)	Precision(%)
Baseline	70.24	78.07	89.51
Baseline+MSFM	77.69	82.06	94.88
Baseline+CPAM	81.53	79.25	89.66
AGCNet	**87.40**	**92.63**	**95.83**

C. COMPARISON WITH STATE-OF-THE-ARTS

To further validate the segmentation performance of the proposed AGCNet, we compared several current state-of-the-art models, including the U-Net [14], UNet++ [16], MRUNet [15], PraNet [6], CPFNet [17], and UACANet [28]. We set all models to the same computational environment and data augmentation for a fairer comparison. It is worth noting that all models are trained from scratch and are not loaded with pre-trained weights. We can see the main problems faced by the current polyp segmentation from the visualization of different competitors on the challenging cases (Fig.7 to Fig.10). The irregular and variable scale of polyp shapes and excessive background noise interfere with the model's sensitivity to foreground information. Simply stacking the depth of the network by some simple convolution and pooling operations, as shown in UNet, cannot cope with these challenging cases. MRUNet utilizes multi-scale and multi-parallel convolution to obtain richer contextual information and improve polyp segmentation results. By optimizing architecture connections to reduce the difference in semantic information between the encoding and decoding layers, UNet++ obtains more accurate segmentation results than UNet. Similarly, PraNet can effectively solve the problem of blurring between the target region and the background tissue by designing a parallel reverse attention mechanism that makes the boundaries of the polyp region more sensitive. CPFNet progressively develops and incorporates rich contextual information by modelling the global pyramid guide module and also obtains better segmentation results. UACANet constructs an improved version of the UNet architecture by using enhanced contextual information to capture those salient features that are easily overlooked. These methods described above remain deficient in polyp segmentation because they do not fully exploit the multi-scale contextual information in the feature extraction process,

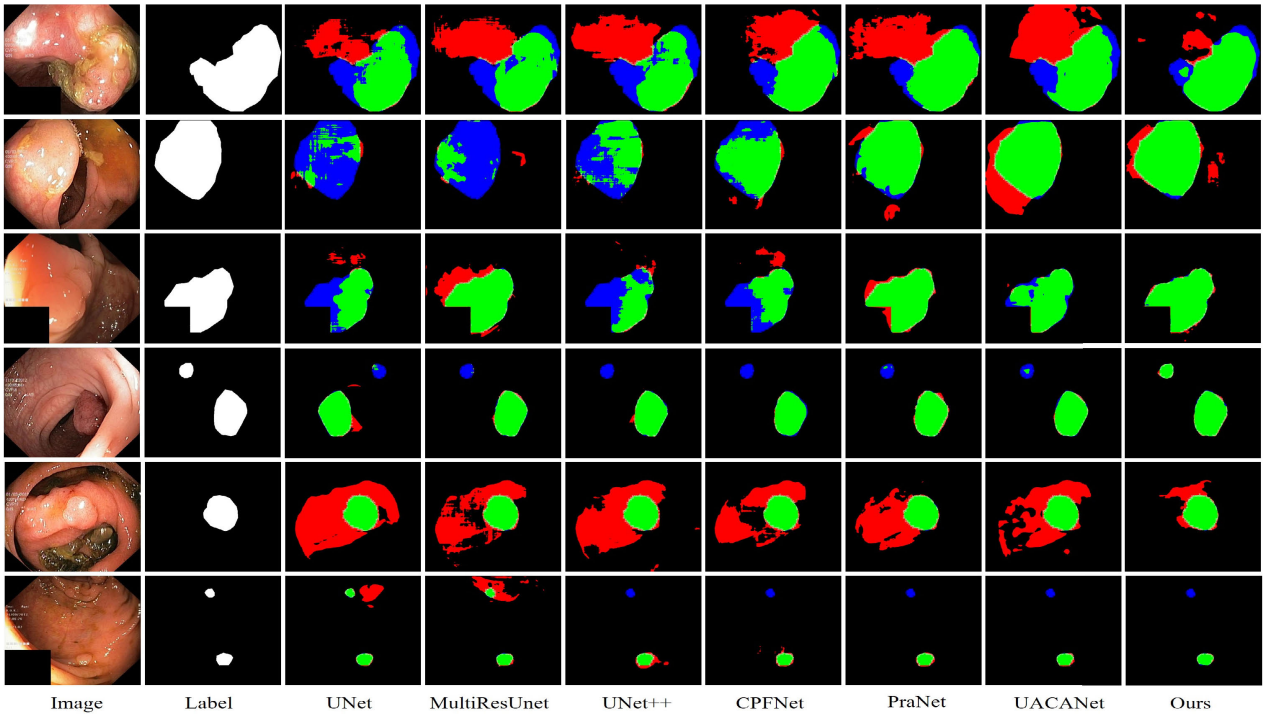


FIGURE 7. Results of the polyp segmentation images visualized on the dataset Kvasir-SEG, where the green, red and blue colors in the image indicate ture positive, false positive and false negative, respectively.

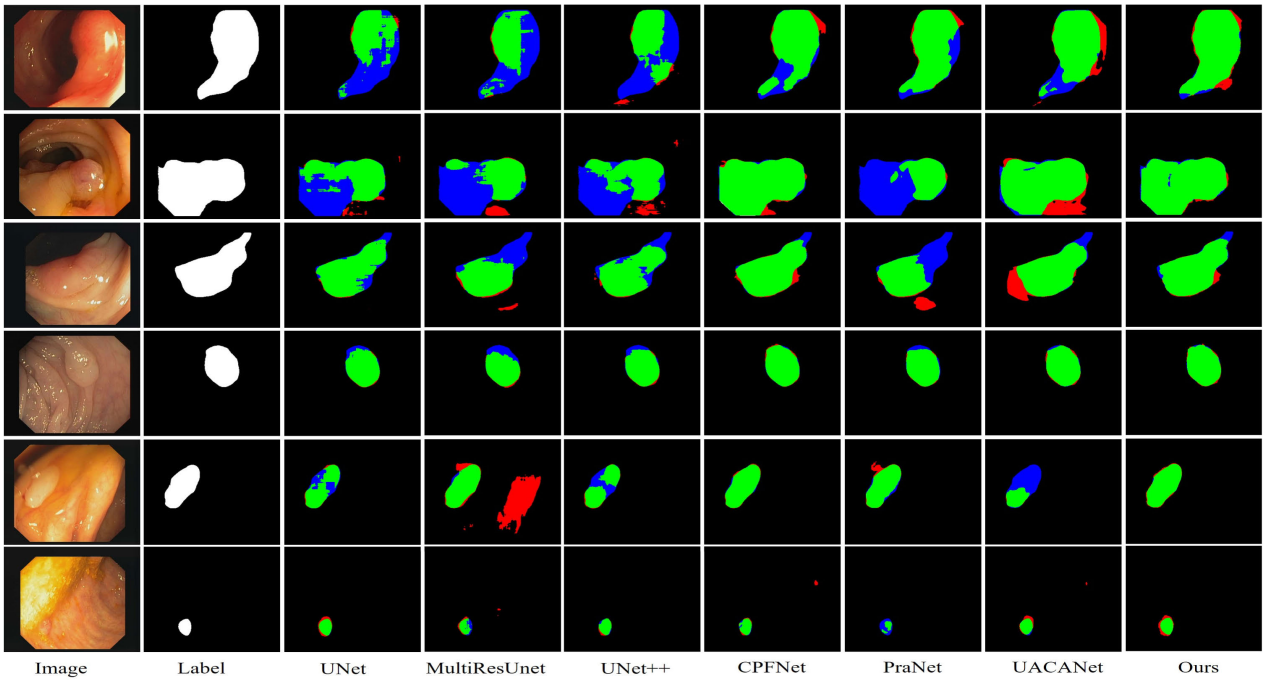


FIGURE 8. Results of the polyp segmentation images visualized on the dataset CVC-Clininc-DB.

which tends to result in the weak ability of the features to discriminate the target region from the Gut tissue. Our proposed AGCNet with MSFM and CPAM modules can fully solve the current problems, surpassing the abovementioned approaches. As shown in the ninth column of Fig.8, AGCNet produces segmentation results that are closest to ground truth, which can effectively cope with multi-scale polyps and suppress background noise. Since there is an inherent need for

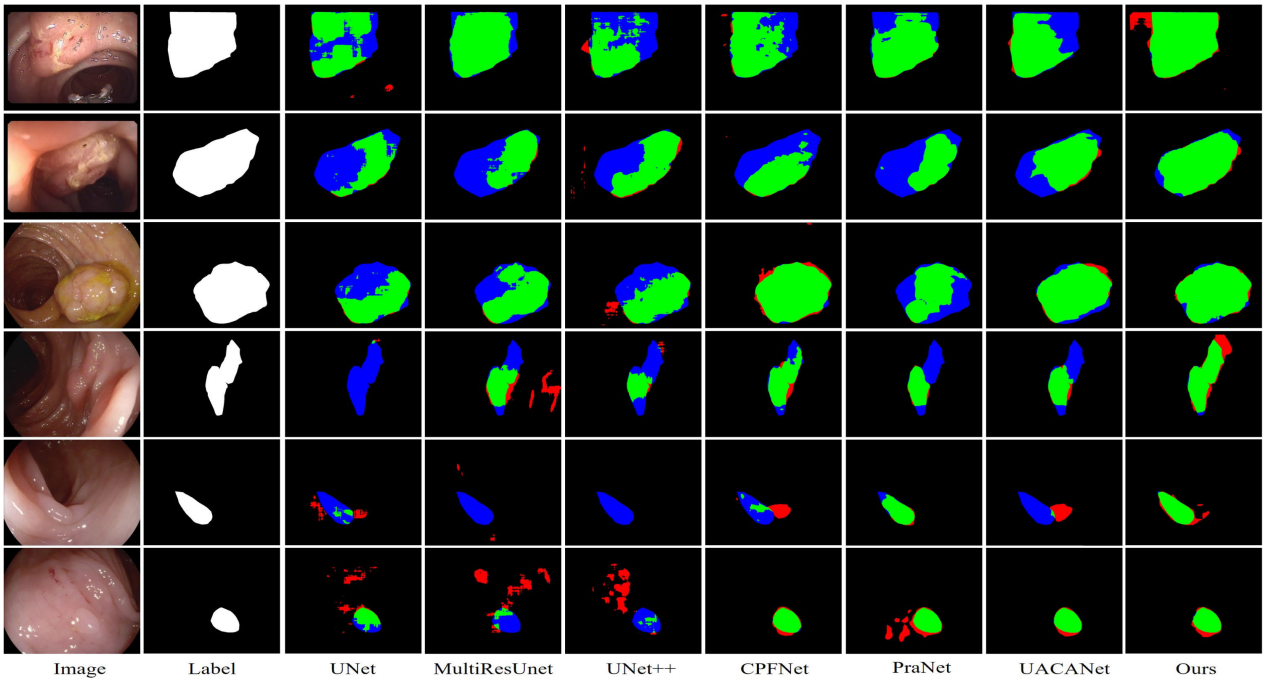


FIGURE 9. Results of the polyp segmentation images visualized on the dataset ETIS-LaribPolypDB.

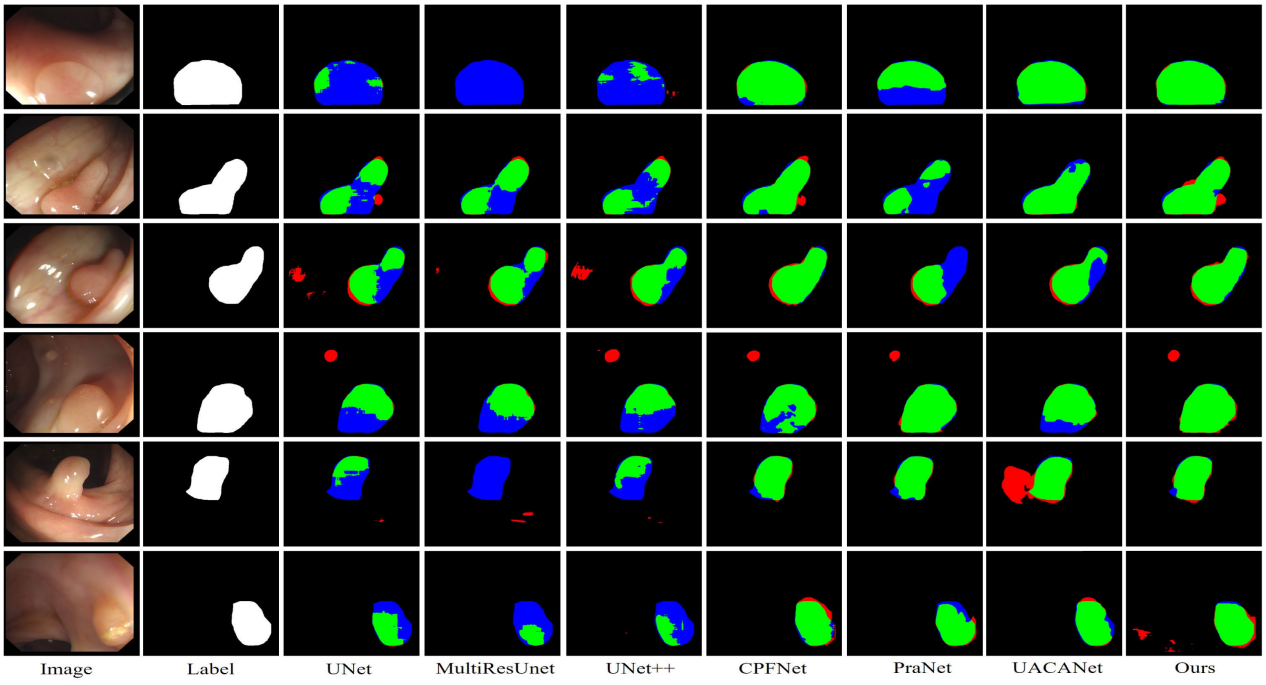


FIGURE 10. Results of the polyp segmentation images visualized on the dataset CVC-ColonDB.

real-time performance in the polyp segmentation task, AGCNet combines both accurate segmentation and guaranteed real-time performance.

D. RESULTS ON CROSS-DATASET

In addition, we conducted a statistical comparison for better quantitative analysis by collecting metric data for IoU, Dice, Precision, Recall, AUC, FPS, GFLOPS, Parameters(M) and Time(ms). As described above, we used two mixed datasets, Kvasir-SEG and CVC-ClinicDB, to train all the models during the execution of the experiments. Since different polyp datasets have different feature distributions, the model needs good generalization ability to obtain excellent segmentation results. In this part, we use Kvasir-SEG and CVC-ClinicDB

TABLE 4. Statistical comparison between different models on Kvasir-SEG dataset.

Method	IoU(%)	Dice(%)	Precision(%)	Recall(%)	AUC(%)	FPS	GPLOPS	Parameters(M)	Time(ms)
UNet	70.24	78.07	89.51	81.03	97.98	9.00	**223.21**	31.38	313.56
UNet++	77.71	80.67	90.23	78.45	97.67	7.00	140.5	9.16	374.01
MRUNet	72.33	76.96	85.95	83.94	97.11	10.00	74.24	**7.24**	307.67
CPFNet	82.47	80.11	95.31	90.74	99.38	26.00	32.18	42.76	178.34
PraNet	81.47	84.73	88.39	87.99	98.08	**28.00**	27.74	32.55	**166.09**
UAUCNet	82.39	86.44	**97.68**	**96.41**	99.11	19.00	25.48	26.90	202.73
Ours	**87.40**	**92.63**	95.83	95.26	**99.89**	18.00	93.50	110.80	226.16

TABLE 5. Statistical comparison between different models on CVC-ClinincDB dataset.

Method	IoU(%)	Dice(%)	Precision(%)	Recall(%)	AUC(%)	FPS	GPLOPS	Parameters(M)	Time(ms)
UNet	73.85	81.42	92.64	81.16	98.39	9.00	**223.21**	31.38	313.56
UNet++	75.19	78.16	90.56	76.91	97.15	7.00	140.5	9.16	374.01
MRUNet	65.17	74.08	76.02	76.65	97.58	10.00	74.24	**7.24**	307.67
CPFNet	80.29	84.94	87.16	86.02	98.81	26.00	32.18	42.76	178.34
PraNet	81.25	83.30	91.62	81.32	98.51	**28.00**	27.74	32.55	**166.09**
UAUCNet	77.53	78.90	86.06	80.12	98.23	19.00	25.48	26.90	202.73
Ours	**83.82**	**86.95**	**93.23**	**88.93**	**99.23**	18.00	93.50	110.80	226.16

TABLE 6. Statistical comparison between different models on ETIS-LaribPolypD dataset.

Method	IoU(%)	Dice(%)	Precision(%)	Recall(%)	AUC(%)	FPS	GPLOPS	Parameters(M)	Time(ms)
UNet	57.13	56.16	68.92	71.98	84.43	9.00	**223.21**	31.38	313.56
UNet++	55.98	57.83	59.24	64.32	80.16	7.00	140.5	9.16	374.01
MRUNet	54.61	54.14	61.79	55.79	74.87	10.00	74.24	**7.24**	307.67
CPFNet	56.50	60.51	65.21	63.98	75.98	26.00	32.18	42.76	178.34
PraNet	68.28	71.61	79.91	78.24	83.72	**28.00**	27.74	32.55	**166.09**
UAUCNet	76.39	76.64	**87.04**	81.99	89.11	19.00	25.48	26.90	202.73
Ours	**80.55**	**74.95**	87.03	**86.25**	**91.23**	18.00	93.50	110.80	226.16

TABLE 7. Statistical comparison between different models on CVC-ColonDB dataset

Method	IoU(%)	Dice(%)	Precision(%)	Recall(%)	AUC(%)	FPS	GPLOPS	Parameters(M)	Time(ms)
UNet	59.11	61.06	80.10	70.20	85.15	9.00	**223.21**	31.38	313.56
UNet++	53.88	56.03	55.14	61.17	78.14	7.00	140.5	9.16	374.01
MRUNet	55.45	66.02	57.53	58.41	85.21	10.00	74.24	**7.24**	307.67
CPFNet	61.09	70.85	69.72	68.94	86.59	26.00	32.18	42.76	178.34
PraNet	73.65	72.40	82.51	74.89	89.23	**28.00**	27.74	32.55	**166.09**
UAUCNet	75.80	77.20	**86.76**	80.70	**96.25**	19.00	25.48	26.90	202.73
Ours	**79.70**	**82.77**	85.67	**82.04**	96.50	18.00	93.50	110.80	226.16

datasets for testing and introduce ETIS-LaribPolypDB and CVC-ColonDB datasets to test the generalization ability of the model.

Table 4 and Table 5 show the comparisons of quantitative result on Kvasir-SEG and CVC-ClinicDB datasets. As shown in Table 4, our proposed AGCNet achieves excellent performance on the Kvasir-SEG dataset. It outperforms the second-best UAUCNet with an IoU of 87.40% and a Dice of 92.63%, leading by 5.01% and 6.19%,respectively. It is worth mentioning that our model surpasses the benchmark model UNet by a large margin. Specifically, it achieves 17.16% and 14.56% higher on IoU and Dice, respectively. The SOTA model PraNet achieves the highest FPS metric of 28 and the shortest execution time of 166.09ms, but Dice and Jaccard are unsatisfactory. In Table 5, our proposed AGCNet achieved scores of 83.82%, 86.95%, 93.23%, and 88.93% on IoU, Dice, Precision, and Recall, respectively. These surpass all state-of-the-art models reported in the Table 5. Through the extensive experiments described above, the validity of AGCNet was verified, and it was able to cope with the main problems faced by current polyp segmentation, including the irregular shape and multiple scales of polyps, the slight difference between foreground and background information, and the inherent real-time needs of clinical applications.

We test the model's generalization ability across different datasets. In Table 6, on ETIS-LaribPolypDB dataset, our proposed AGCNet outperforms the models in the table, achieving 80.55%, 74.95% and 86.25% on IoU, Dice and Recall, respectively. As shown in Table 7, PraNet and UACANet all lead U-Net by 11% to 16% on average in IoU and Dice metrics. However, our proposed AGCNet is still ahead of it in IoU and Dice. The above comparison results show that AGCNet can maintain high accuracy while ensuring strong generalization ability.

The superiority of AGCNet over current state-of-the-art models originates from its two inherent submodules: MSFM and CPAM, each fulfilling a distinct purpose. Specifically, MSFM enhances the representation of target features at a finer granularity level. Meanwhile, CPAM suppresses background noise in deep networks and highlights polyp features through a dual attention mechanism. AGCNet has the following advantages: 1. Multi-level pooling: AGCNet performs multi-level pooling on the input features multiple times to extract multi-scale contextual information from the image. 2. Dual attention mechanism: AGCNet utilizes two integrated attention mechanisms to selectively focus on the most informative regions of the image that are most relevant to the segmentation task. 3. Feature fusion: AGCNet fuses the multi-level pooled features with the attention maps to generate enhanced feature representations that capture global and local information. Overall, AGCNet first obtains richer semantic information at a finer level of granularity. Then it strengthens the targets while weakening the background, offering an optimal solution for addressing similar segmentation tasks with blurred boundaries between different categories.

IV. CONCLUSION

In this paper, we propose a novel convolutional neural network, AGCNet, which can extract multi-scale contextual information to bridge the semantic information gap between different layers and effectively suppress the interference of background noise in deep semantic information by using dynamic modelling of long-range dependency. The AGCNet includes two novel modules: a multi-scale semantic fusion module (MSFM) and a context-aware pyramid aggregation module (CPAM). The MSFM enhances the representation capability by collecting contextual information at different scales to adapt to the problem of large variation in polyp size. The CPAM aggregates feature information across different regions to boost the network's ability to utilize global context and model long-range dependency through dual attention. More importantly, AGCNet can maintain excellent real-time performance while considering segmentation accuracy, which is of great significance for clinical practice. We have also done extensive experiments on the datasets Kvasir-SEG, CVC-ClinicDB, ETIS-LaribPolypD and CVC-ColonDB to confirm the effectiveness of AGCNet. Our subsequent work focuses on using more extreme cases to train AGCNet to strengthen its learning yet ability, enhance its robustness, and integrate it into the procedure of colonoscopy.

REFERENCES

[1] J. Bernal, J. Sánchez, and F. Vilariño, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, Sep. 2012, doi: 10.1016/j.patcog.2012.03.002.

[2] G. C. Lou, J. M. Yang, Q. S. Xu, W. Huang, and S. G. Shi, "A retrospective study on endoscopic missing diagnosis of colorectal polyp and its related factors," *Turkish J. Gastroenterol.*, vol. 1, pp. 182–186, Dec. 2014, doi: 10.5152/tjg.2014.4664.

[3] D. K. Rex, "Colonoscopic withdrawal technique is associated with adenoma miss rates," *Gastrointestinal Endoscopy*, vol. 51, no. 1, pp. 33–36, 2000, doi: 10.1016/S0016-5107(00)70383-X.

[4] M. F. Kaminski, P. Wieszczy, M. Rupinski, U. Wojciechowska, J. Didkowska, E. Kraszewska, J. Kobiela, R. Franczyk, M. Rupinska, B. Kocot, A. Chaber-Ciopinska, J. Pachlewski, M. Polkowski, and J. Regula, "Increased rate of adenoma detection associates with reduced risk of colorectal cancer and death," *Gastroenterology*, vol. 153, no. 1, pp. 98–105, Jul. 2017, doi: 10.1053/j.gastro.2017.04.006.

[5] D. A. Corley, C. D. Jensen, A. R. Marks, W. K. Zhao, J. K. Lee, C. A. Doubeni, A. G. Zauber, J. de Boer, B. H. Fireman, J. E. Schottinger, V. P. Quinn, N. R. Ghai, T. R. Levin, and C. P. Quesenberry, "Adenoma detection rate and risk of colorectal cancer and death," *New England J. Med.*, vol. 370, no. 14, pp. 1298–1306, Apr. 2014, doi: 10.1056/NEJMoa1309086.

[6] D. P. Fan, G. P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Parallel reverse attention network for polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 263–273, doi: 10.48550/arXiv.1802.00368.

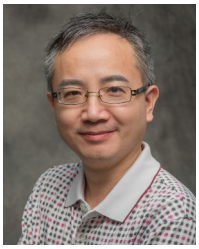
[7] C. M. Rutter, E. Johnson, D. L. Miglioretti, M. T. Mandelson, J. Inadomi, and D. S. M. Buist, "Adverse events after screening and follow-up colonoscopy," *Cancer Causes Control*, vol. 23, no. 2, pp. 289–296, Feb. 2012, doi: 10.1007/s10552-011-9878-5.

[8] B. Li and M. Q.-H. Meng, "Automatic polyp detection for wireless capsule endoscopy images," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10952–10958, Sep. 2012, doi: 10.1016/j.eswa.2012.03.029.

[9] P. Klare, C. Sander, M. Prinzen, B. Haller, S. Nowack, M. Abdelhafez, A. Poszler, H. Brown, D. Wilhelm, R. M. Schmid, S. von Delius, and T. Wittenberg, "Automated polyp detection in the colorectum: A prospective study (with videos)," *Gastrointestinal Endoscopy*, vol. 89, no. 3, pp. 576–582.e1, Mar. 2019, doi: 10.1016/j.gie.2018.09.042.

[10] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 65–75, Jan. 2017, doi: 10.1109/JBHI.2016.2637004.
[11] P. Wang, "Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): A double-blind randomized study," *Lancet Gastroenterol. Hepatol.*, vol. 5, no. 4, pp. 343–351, 2020, doi: 10.1016/S2468-1253(19)30411-X.
[12] R. Zhang, Y. Zheng, C. C. Y. Poon, D. Shen, and J. Y. W. Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern Recognit.*, vol. 83, pp. 209–219, Nov. 2018, doi: 10.1016/j.patcog.2018.05.026.
[13] P. Brandao, "Fully convolutional neural networks for polyp segmentation in colonoscopy," *Proc. SPIE*, vol. 10134, pp. 101–107, Mar. 2017, doi: 10.1117/12.2254361.
[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4-28.
[15] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020, doi: 10.1016/j.neunet.2019.08.025.
[16] Z. Zhou, "UNet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5-1.
[17] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, and X. Chen, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020, doi: 10.1109/TMI.2020.2983721.
[18] T. Kim, H. Lee, and D. Kim, "UACANet: Uncertainty augmented context attention for polyp segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2167–2175, doi: 10.1145/3474085.3475375.
[19] S. Gao, M. Cheng, K. Zhao, X. Zhang, M. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.
[20] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. (2019). *ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks*. [Online]. Available: http://arxiv.org/licenses/nonexclusive-distrib/1.0
[21] H. Fang and F. Lafarge, "Pyramid scene parsing network in 3D: Improving semantic segmentation of point clouds with multi-scale contextual information," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 246–258, Aug. 2019, doi: 10.1016/j.isprsjprs.2019.06.010.
[22] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020, doi: 10.1109/ACCESS.2019.2962617.
[23] C. H. Sudre, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 240–248, doi: 10.1007/978-3-319-67558-9-28.
[24] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015, doi: 10.1016/j.compmedimag.2015.02.007.
[25] D. Jha, "Kvasir-SEG: A segmented polyp dataset," in *Proc. Int. Conf. Multimedia Modeling*. Cham, Switzerland: Springer, 2020, pp. 451–462, doi: 10.1007/978-3-030-37734-2-37.
[26] C. Zhang, "Theory of deep learning IIb: Optimization properties of SGD," 2018, *arXiv:1801.02254*.
[27] D. P. Kingma and J. Ba. (2014). *Adam: A Method for Stochastic Optimization*. [Online]. Available: http://arxiv.org/licenses/nonexclusive-distrib/1.0
[28] J. Zhang and N. Tansu, "Optical gain and laser characteristics of InGaN quantum wells on ternary InGaN substrates," *IEEE Photon. J.*, vol. 5, no. 2, Apr. 2013, Art. no. 2600111.
[29] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, Mar. 2014, doi: 10.1007/s11548-013-0926-3.

LIANTAO SHI received the B.S. degree from Huaqiao University, in 2018, and the M.S. degree from the University of Science and Technology Liaoning, in 2022. His main research interests include embedded systems and computer vision semantic segmentation.



ZHENGGUO LI received the B.S. degree from the Wuhan University of Technology and the M.S. and Ph.D. degrees from Central South University. He is currently a Professor and a Master's Supervisor with the School of Electronics and Information Engineering, University of Science and Technology Liaoning, the Deputy Director of the Scientific Research Department, Shenzhen Polytechnic, and the Dean of the Research Institute for Carbon-Neutral Technology. His research interests include the simulation and study of electromagnetic radiation of new energy vehicle drive systems, EMC, and artificial intelligence.

JIANYANG LI received the B.S. degree from the Shandong University of Science and Technology, in 2019. He is currently pursuing the M.S. degree with the University of Science and Technology Liaoning. His main research interests include deep learning and electromagnetic compatibility.

YUFENG WANG received the B.S. and Ph.D. degrees from the Dalian University of Technology, in 2001 and 2007, respectively. He is currently an Associate Professor and a Master's Supervisor with the School of Electronics and Information Engineering, University of Science and Technology Liaoning. His main research interests include power quality, electromagnetic compatibility, and metallurgical automation.

HONGYU WANG received the B.S. and M.S. degrees from the University of Science and Technology Liaoning, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree with the Wuhan University of Technology. His main research interests include unsupervised learning, reinforcement learning, and robotics.

YUBAO GUO received the B.S. and M.S. degrees from the University of Science and Technology Liaoning, in 2020 and 2023, respectively. His main research interests include power electronics and intelligence algorithm.