Focusing the View: Enhancing U-Net with Convolutional Block Attention for Superior Medical Image Segmentation

Nhu-Tai Do[1†], Dat Nguyen Khanh[2†], Tram-Tran Nguyen-Quynh[2], Quoc-Huy Nguyen[1*]

[1]Saigon University, Vietnam.
[2]Ho Chi Minh City University of Foreign Language-Information Technology, Vietnam.

*Corresponding author(s). E-mail(s): nqhuy@sgu.edu.vn;
Contributing authors: dntai@sgu.edu.vn;
nguyenkhanhdat12.7@gmail.com; tramtnq@huflit.edu.vn;
†These authors contributed equally to this work.

Abstract

Detecting and segmenting polyps from endoscopic images is a significant challenge in the medical field, aimed at enhancing the early diagnosis rate of potentially cancerous conditions. To address the limitations of current methods in identifying the complex features of polyps, we have developed an improved version of U-Net, integrating the Convolutional Block Attention Module (CBAM). This combination improves the feature extraction, focuses on critical aspects of polyps, and minimizes unnecessary noise and errors. Testing on multiple datasets has shown that our model significantly outperforms traditional U-Net versions, particularly in detecting small and deformed polyps, while also delivering high computational efficiency suitable for real-world medical applications. This study not only opens new avenues in medical imaging segmentation technology but also has the potential to improve diagnostic and treatment procedures significantly.

Keywords: U-Net, CBAM, Multi-Level Context, Deep learning, CBAM, Polyp segmentation, Medical image segmentation

1 Introduction

In the field of medical diagnostic imaging, the U-Net architecture has been groundbreaking, advancing image segmentation significantly with its encoder-decoder framework [1, 2]. This model has become indispensable in computational biology, allowing for the precise segmentation of complex images like those of polyps in endoscopic examinations [3]. Despite its widespread success, the dynamic and complex nature of medical images, where each pixel might contain critical diagnostic information, necessitates continuous advancements.

Recently, Transformer-based models have been introduced to medical imaging, using global dependencies within images to enhance segmentation [4, 5]. These models excel in capturing broad contextual information but often come with substantial computational demands. Furthermore, their emphasis on global features may lead to a neglect of local, detail-oriented features that are crucial for accurate medical analysis.

In contrast, our U-Net Focus model, enhanced with the Convolutional Block Attention Module (CBAM), provides a targeted and efficient approach [6]. CBAM refines our model by intensifying important features while diminishing less relevant ones, akin to a detective focusing on crucial clues. This addition ensures that U-Net Focus not only retains but also enhances the granularity that the classic U-Net is known for, making it highly effective at capturing the delicate nuances in medical images.

Our U-Net Focus with CBAM distinguishes itself from Transformer-based methods by combining the dependable feature extraction capabilities of U-Net with precise, localized attention. This strategy avoids the heavy computational overhead associated with Transformers and focuses on enhancing the discernment of local features without compromising efficiency.

The effectiveness of U-Net Focus with CBAM is demonstrated through our extensive testing. In evaluations on CVC-ClinicDB [7] and Kvasir-SEG [8], our model surpassed both standard U-Net and other advanced segmentation models in terms of Dice Loss and mIOU scores. These results not only confirm the enhanced performance of our model but also highlight the significant benefits of incorporating focused attention mechanisms into neural network architectures for medical imaging.

This paper explores the integration of CBAM with U-Net Focus, detailing its transformative impact on medical image segmentation. Our comprehensive analysis across multiple datasets shows that our model outperforms transformer-based solutions, offering more precise local feature enhancement and greater computational efficiency.

2 Proposed Method

2.1 Problem Overview

Given a colonoscopy image I in $R^{h\times w\times 3}$, where each pixel $p = (x, y)$, I represents intensity values in RGB, the primary objective of our study is to classify each pixel p as belonging to a polyp or non-polyp area. This classification is achieved through a binary segmentation mask $S_{\text{seg}} \in \mathbb{R}^{h\times w\times 1}$.

To enhance segmentation accuracy and specificity, we incorporate multi-level distance masks $S_{\text{dis}} \in \mathbb{R}^{h\times w\times 1}$, where d denotes the number of distance levels. These

masks extend the areas adjacent to the identified polyp region, enhancing the local receptive field around the polyp and improving segmentation by capturing detailed nuances of polyp boundaries.

The process begins with the input image I, which undergoes preprocessing to enhance feature visibility and prepare it for segmentation. The binary segmentation mask S_{seg} is then generated through our enhanced U-Net Focus model, which has been modified to integrate the CBAM for better focus on relevant features.

The multi-level distance masks S_{dis} are subsequently created from the segmentation mask. These masks provide a spatial context by highlighting areas progressively farther from the boundaries of the polyp, aiding in capturing a broader context around the detected regions. It helps handle polyps with diverse shapes and sizes, ensuring that the segmentation accurately captures the core and peripheral areas.

2.2 Proposed model

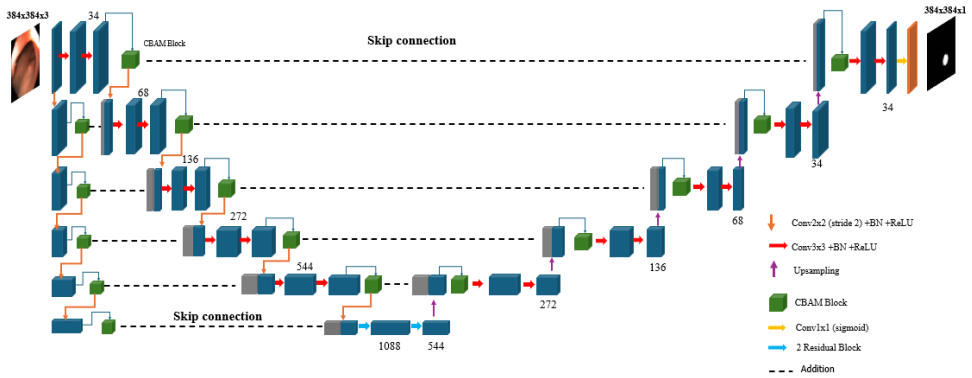

Fig. 1: Our proposed method

Our enhanced U-Net Focus architecture in Fig. 1 incorporates the Convolutional Block Attention Module (CBAM), optimizing it for in-depth feature extraction across extensive network layers. The model begins with an input layer followed by convolutional layers. Each convolutional layer is paired with a CBAM unit, sequentially increasing the filters to enhance the model's feature refinement capabilities. This setup ensures that each layer captures essential features and focuses on the most relevant aspects through channel and spatial attention. As the network deepens, two Residual Blocks are integrated at the bottleneck. These blocks utilize identity shortcuts, facilitating adequate gradient flow during training and avoiding the degradation accompanying increased depth.

In the decoder path, the architecture uses up-sampling to reverse the encoder's operations, carefully combining these unsampled outputs with corresponding feature maps from the encoder through skip connections. This technique is crucial for restoring lost spatial details during the down-sampling process. The output from the decoder

is processed through a 1x1 convolutional layer, which classifies each pixel to generate precise segmentation maps. This final step is essential for producing accurate medical diagnostics, making the network highly effective for analyzing complex medical images.

A distinctive feature of our model's architecture is the integration of CBAM in the decoder path, which is not typically seen in standard U-Net models [9, 10]. In the decoder, each upsampling step is followed by a skip connection that merges the upsampled output with the corresponding feature maps from the encoder, crucial for restoring spatial details lost during downsampling. Immediately after this skip connection and before any further processing, a CBAM is applied. Additionally, our model introduces an innovative adaptation within the 'Focus' branch, where CBAM is strategically deployed after each downsampling step.

2.3 Implementation Details

Channel Attention Module. It plays a crucial role in CBAM by selectively emphasizing the most relevant feature channels shown in Fig. 2. This process helps the neural network focus on channels containing crucial information for the task. To achieve this, MC applies two global pooling operations on the input feature map $F \in \mathbb{R}^{h \times w \times c}$:

$$f_{\text{avg}} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{ij} \tag{1}$$

$$f_{max} = \max_{i,j} F_{ij} \tag{2}$$

where f_{avg} is the global average pooling to compute the mean intensity of each channel across all spatial locations, providing an overall representation of the features present in each channel. f_{max} is the global max pooling to capture the maximum intensity in each channel, highlighting the most significant features.

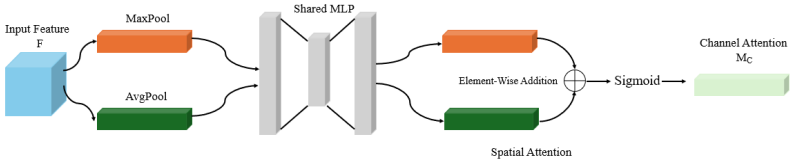

Fig. 2: Channel Attention Module

Fig.2 shows the structure of the Channel Attention Module. Both descriptors are passed through a shared multi-layer perceptron (MLP), composed of two fully connected layers. The MLP reduces the descriptor dimensionality by a factor of r before restoring it to the original channel dimension C. The two outputs are then combined via element-wise addition and activated using a sigmoid function, resulting in a channel attention weight vector:

$$w_c = \sigma\left(\text{MLP}(f_{avg}) + \text{MLP}(f_{max})\right) \tag{3}$$

$$F_{c'} = F \cdot w_c \tag{4}$$

w_c is then applied to the original input feature map F via element-wise multiplication to amplify or suppress channels based on their importance. This refined feature map $F_{c'}$ allows subsequent network layers to concentrate on the most critical channels.

Spatial Attention Module. This module determines which regions of the feature map to emphasize, shown in Fig. 3.

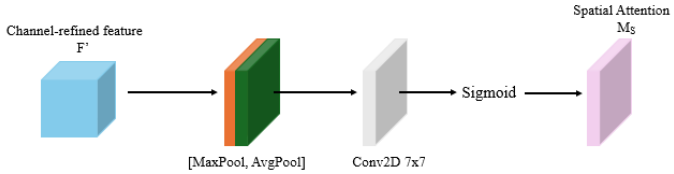

Fig. 3: Spatial Attention Module

The output from MC, $F_{c'}$ undergoes two pooling operations to produce two spatial descriptors as follows:

$$f_{\text{avg}}^{\text{spatial}} = \text{AvgPooling}(F_{c'}) \tag{5}$$

$$f_{\text{max}}^{\text{spatial}} = \text{MaxPooling}(F_{c'}) \tag{6}$$

$$w_s = \sigma\left(\text{Conv7x7}\left[\left(f_{\text{avg}}^{\text{spatial}}, f_{\text{max}}^{\text{spatial}}\right)\right]\right) \tag{7}$$

where $f_{\text{avg}}^{\text{spatial}}/f_{\text{max}}^{\text{spatial}}$ computes the mean/maximum intensity of features across all channels at each spatial location. This spatial attention map w_s is multiplied with the feature map F_c' to refine the network's focus on relevant spatial regions. This step ensures the network focuses on specific regions, enabling it to focus on pertinent spatial areas:

$$F_s' = F_c' \cdot w_s \tag{8}$$

Convolutional Block Attention Module (CBAM). It combines MC and MS into a single framework to refine feature representation across channel and spatial dimensions shown in Fig. 4.

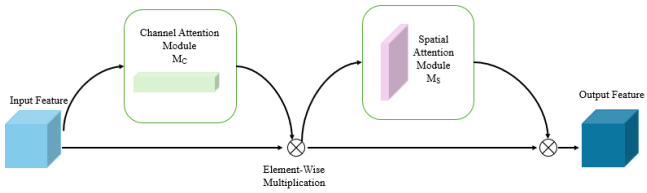

Fig. 4: Convolutional Block Attention Module

It enables the network to identify critical channels and spatial regions by incorporating both attention mechanisms, significantly improving its performance in various computer vision tasks. The resulting feature map provides a comprehensive focus on 'what' and 'where' to look in a given image, thus enhancing the network's ability to discern complex patterns and features effectively.

3 Experiment and Results

3.1 Experiments Setup

Datasets. Our study utilized two prominent datasets widely recognized in the medical imaging community to evaluate polyp segmentation algorithms: CVC-ClinicDB [7] and Kvasir-SEG [8], summarized in Table 1. Each dataset provides unique challenges and features critical for training and testing our enhanced U-Net architecture with the Convolutional Block Attention Module (CBAM).

Table 1: Kvasir-SEG and ClinicDB Datasets for Polyp Segmentation

Dataset	Total Images	Resolution Images	Details
Kvasir-SEG [8]	1000	332x487 to 1920x1072	Images, segmentation masks
CVC-ClinicDB [7]	612	384×288	Images, segmentation masks

CVC-ClinicDB and Kvasir-SEG are employed in our experiments to train and validate the CBAM-enhanced U-Net Focus model. These datasets' comprehensive and diverse nature ensures that our model learns to effectively handle a wide range of real-world scenarios, thus enhancing its applicability and reliability in clinical settings. The ground truth provided with these datasets also allows for precise evaluation of segmentation accuracy, facilitating detailed analysis and comparison of our model's performance against existing methods.

Evaluation Metrics. In image segmentation, the Mean Intersection over Union (mIOU) and Dice Similarity Coefficient (DSC) are widely used metrics. These metrics evaluate the model's performance by comparing the predicted segmentation maps against ground truth annotations. mIOU calculates the ratio between the intersection and the union of the predicted and the true regions:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|} \tag{9}$$

where $\boldsymbol{N}$ is the number of classes, Y_i is the ground truth, and $\hat{Y}_i$ is the predicted segmentation for class i. The Dice Similarity Coefficient, is another critical metric used primarily to gauge the similarity between two samples:

$$DSC = \frac{2 \times |Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \tag{10}$$

where $|Y \cap \hat{Y}|$ represents the number of pixels correctly predicted as the foreground (true positives), while $|Y|$ and $|\hat{Y}|$ are the total number of pixels in the actual and predicted foregrounds, respectively.

More details. The datasets are divided into three subsets to ensure robust training, validation, and testing of the model. Specifically, 80% of the data is used for training, allowing the model to learn and adapt to the complexity of medical images. The remaining 20% is equally split, with 10% used for validation and 10% for testing. This division helps in fine-tuning the model parameters during the validation phase and assessing the model's generalizability and performance during the testing phase. To enhance the model's ability to generalize and reduce overfitting, we employ a comprehensive data augmentation strategy using the Albumentations library.

The models are trained on an NVIDIA Tesla P100 GPU with a batch size of 4. This setup balances the computational load and training efficiency. We utilize the AdamW optimizer, which combines the benefits of Adam and weight decay regularization with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. This configuration helps in stabilizing the training process and improving convergence properties. AdamW helps better handle sparse gradients and provides an adaptive learning rate mechanism crucial for deep learning models in medical image analysis.

3.2 Results

Results on ablation studies of attaching CBAM. Firstly, we conducted experiments on attaching CBAM at the encoder and decoder paths to evaluate the performance of our proposed method. Interestingly, our results shown in Table 2 consistently improve segmentation performance by incorporating CBAM into the U-Net architecture. Specifically, models with CBAM integration exhibit higher Dice scores and mIOU values than their counterparts without CBAM. Notably, the most comprehensive configuration, where CBAM is utilized in both the encoder and decoder, achieves the highest segmentation accuracy, indicating the synergistic benefits of leveraging attention mechanisms at multiple network stages.

Table 2: Results on ablation studies of attaching CBAM

Models	**CVC-ClinicDB**		**Kvasir-SEG**	
	DSC	**mIOU**	**DSC**	**mIOU**
U-Net [1]	0.710	0.627	0.818	0.746
U-Net Focus	0.882	0.793	0.857	0.758
U-Net Focus + CBAM Encoder	0.911	0.836	0.862	0.762
U-Net Focus + CBAM Decoder	0.919	0.852	0.883	0.801
U-Net Focus + CBAM Encoder/Decoder	**0.935**	**0.877**	**0.891**	**0.803**

The best performance on the dice score was achieved at U-Net Focus + CBAM encoder/decoder with the dice score values of 0.935 on CVC-ClinicDB [7] and 0.877

on Kvasir-SEG [8]. The focus view helps the dice score increase by 17% on CVC-ClinicDB and 4% on Kvasir-SEG. Besides, the CBAM module at encoder/decoder improved the dice score by 2% on CVC-ClinicDB and 1% on Kvasir-SEG.

Comparision on the state-of-the-art methods. We conducted comparative analyses against several state-of-the-art architectures in Table 3, including pure U-Net [1], ResUnet++ [11], HRNetV2 [12], DCRNet [13], and MSRF-Net [14]. We opted for a focused evaluation using the ClinicDB and Kvasir-SEG datasets to ensure a stringent assessment of our approach's effectiveness.

Table 3: Comparision on the state-of-the-art methods

Models	Year	CVC-ClinicDB		Kvasir-SEG	
		DSC	mIOU	DSC	mIOU
U-net [1]	2015	0.710	0.627	0.818	0.746
ResUNet++ [11]	2019	0.763	0.701	0.813	0.793
HRNetV2 [12]	2019	0.778	0.636	0.853	0.744
DCRNet [13]	2022	0.856	0.788	0.886	**0.825**
MSRF-Net [14]	2022	0.906	0.828	0.851	0.740
Our proposed method		**0.935**	**0.877**	**0.891**	0.803

The segmentation results showed our proposed method achieved the best result with the dice score values of 0.935 on ClinicDB [7] and 0.891 on Kvasir-SEG [8]. The table reveals that our method consistently outperforms the standard U-Net and exhibits competitive or superior performance compared to more complex architectures like ResUNet++ and DCRNet. This demonstrates the substantial enhancements that CBAM provides, particularly in terms of segmentation precision and the ability to handle challenging imaging scenarios.

3.3 Quality evaluation on feature maps

The visualization of Grad-CAM in Fig. 5 effectively illustrates the performance difference between U-Net and our model. These images provide a direct visual comparison of each model's focus regions during the performance of the segmentation task.

The superior performance of our model is evident through its concentrated heatmaps, which closely align with the polyp boundaries, demonstrating a more precise and effective segmentation capability. This visualization highlights the impact of integrating CBAM, which directs the network's attention more accurately toward relevant features, thereby enhancing the precision of the segmentation output. Such visual evidence underscores the technological advancements brought by CBAM in improving the interpretative capabilities and operational efficiency of U-Net architectures in medical imaging tasks.

Our model demonstrates the enhanced ability to segment closely packed and overlapping polyps, a common challenge in medical imaging, thanks to the targeted

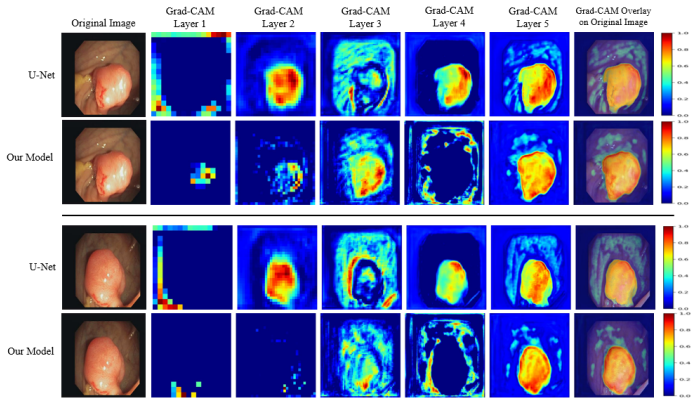

Fig. 5: Grad-CAM Comparison Highlighting Enhanced Focus by Our Model

attention mechanisms of CBAM. This improves clinical utility as the segmented images provide more transparent and usable diagnostic information.

4 Conclusions

This study has demonstrated that integrating the CBAM into the U-Net Focus architecture is not merely a technical enhancement but a breakthrough in addressing the challenges of medical image segmentation. By enabling the model to focus more precisely on critical areas of an image, CBAM has significantly improved the accuracy and sensitivity of detecting and segmenting medical imaging objects.

Our tests on the CVC-ClinicDB [7] and Kvasir-SEG [8] datasets have yielded impressive results, with our U-Net Focus model outperforming the traditional U-Net significantly and showing competitive capabilities against other advanced models. These outcomes affirm the effectiveness of CBAM and open new avenues for further exploitation and enhancement of attention mechanisms in practical applications.

References

[1] O. Ronneberger, P. Fischer, T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (Springer, 2015), pp. 234–241

[2] R. Azad, E.K. Aghdam, A. Rauland, Y. Jia, A.H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J.P. Cohen, E. Adeli, D. Merhof, Medical image segmentation review: The success of u-net. arXiv preprint arXiv:2211.14830 (2022)

[3] J. Mei, T. Zhou, K. Huang, Y. Zhang, Y. Zhou, Y. Wu, H. Fu, A survey on deep learning for polyp segmentation: Techniques, challenges and future trends.

Journal of Biomedical Informatics **110**, 103532 (2020)

[4] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, Y. Xie, From cnn to transformer: A review of medical image segmentation models. Journal of Imaging Informatics in Medicine pp. 1–19 (2024)

[5] X. Li, H. Ding, H. Yuan, W. Zhang, J. Pang, G. Cheng, K. Chen, Z. Liu, C.C. Loy, Transformer-based visual segmentation: A survey. arXiv preprint arXiv:2304.09854 (2023)

[6] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module. European Conference on Computer Vision pp. 3–19 (2018)

[7] J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño. Cvc-clinicdb dataset (2015)

[8] D. Jha, P.H. Smedsrud, M.A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H.D. Johansen. Kvasir-seg dataset (2019)

[9] X. Fang, *Research on the Application of Unet with Convolutional Block Attention Module to Semantic Segmentation Task*, in *Proceedings of the 2022 5th International Conference on Sensors, Signal and Image Processing* (2022), pp. 13–16

[10] Z. Song, H. Yao, Segmentation method of u-net sheet metal engineering drawing based on cbam attention mechanism. arXiv preprint arXiv:2209.14102 (2022)

[11] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H.D. Johansen, *Resunet++: An advanced architecture for medical image segmentation*, in *2019 IEEE international symposium on multimedia (ISM)* (IEEE, 2019), pp. 225–2255

[12] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, J. Wang, High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019)

[13] L. Qin, W. Che, Y. Li, M. Ni, T. Liu, *Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification*, in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34 (2020), pp. 8665–8672

[14] A. Srivastava, D. Jha, S. Chanda, U. Pal, H.D. Johansen, D. Johansen, M.A. Riegler, S. Ali, P. Halvorsen, Msrf-net: a multi-scale residual fusion network for biomedical image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(5), 2252–2263 (2021)