Dual Attention Cascade Transformer for Polyp Segmentation

Yuan Ju
Department of Electrical and Electronic Engineering
Hanyang University
Ansan, South Korea
tangerine@hanyang.ac.kr

Bingxin Wei
Department of Electrical and Electronic Engineering
Hanyang University
Ansan, South Korea
wei0911@hanyang.ac.kr

Haewoon Nam
Department of Electrical and Electronic Engineering
Hanyang University
Ansan, South Korea
hnam@hanyang.ac.kr

***Abstract*—Polyp segmentation is of great importance for the prevention of colon cancer. Such a task remains highly challenging due to the high similarity of polyps to the background. Models based on convolutional neural networks and transformer have achieved promising results in this task, but their ability to combine local and global information remains limited. In this paper, we propose a novel network Dual Attention Cascade Transformer (DACFormer) that effectively combines local and global contextual information to suppress the effect of context on target recognition. The proposed method adopt the cascade structure of feature reuse, which effectively combines the semantic information of features at all levels and further exploits the potential of transformer and improves the generalization ability of the model is effectively improved. We conducted tests on four public datasets, CVC-ClinicDB, Kvasir, CVC-300, CVC-ColonDB. The results show that our network outperforms the current mainstream networks on the four benchmark datasets.**

***Index Terms*—deep learning, transformer, polyp segmentation**

I. Introduction

Colorectal cancer is one of the three most deadly cancers in the world [1]. And polyps are an important cause of colon cancer. Since polyps are very similar to human tissue, it requires great human resources. The automatic segmentation of polyps is a very rewarding task.

In medical image segmentation, most of the methods are inspired by UNet [2] and use the design structure of encoder-decoder. The existing methods can be divided into two categories depending on the encoder, Convolutional Neural Network (CNN) based and Transformer [3] based methods. Due to the lack of good global modeling capability of CNNs, CNN-based networks as encoders are dedicated to improve the perceptual wildness of the network. For example, CaraNet [4] uses Channel-wise feature pyramid module and axial attention to mine multi-scale information. LDNet uses different variants of self-attention mechanism to capture global information.

Transformer has good ability to model long-distance pixel dependencies. But transformer tends to ignore the local relationships of pixels, resulting in loss of detailed information of the results. So the network based on encoder as Transformer is dedicated to improve the local information mining ability of the network. SSFormer proposes an efficient scheme to mine local information and combine contextual information using convolutional and linear layers. But we think there is further room to exploit the potential of Transformer. In this paper, we propose a new Transformer-based network that efficiently combines global and local information in contextual relationships through an attention mechanism based on convolutional networks. The contributions of this paper are summarized as follows:

- A novel Transformer-based network for polyp segmentation is proposed. A special multiplexed cascade structure is used to enable the network to obtain accurate localization of the target and excellent generalization capability.
- Multi-scale information fusion attention and efficient multi-scale attention are employed to effectively fuse different scale features and adapt the network to polyp targets of different sizes.
- Experimental results on four challenging benchmark datasets show that the proposed DACFormer is better than other counterparts and achieves the new state-of-the-art performance.

II. METHODOLOGY

The overall structure is shown in Fig. 1, which is based on the encoder-decoder structure. The encoder is based on Transformer decoder, and the decoder includes the Multi-Scale Information Fusion Attention Module and Efficient Multi-Scale Attention Module. the four prediction maps generated in the four stages of the decoder are under deep supervision.

A. Transformer Encoder

Transformer has excellent performance in computer vision and requires variants. PVT v2 [5] uses convolution to give the encoder features a pyramidal structure, enriching the scale diversity of the features and achieving the best performance in computer vision. We use PVT v2 as our encoder to obtain four stages of coarse to fine multi-scale features $E_i, i \in \{1, 2, 3, 4\}$. These features are used as input to the decoder to obtain the final segmentation prediction map.

B. Multi-scale Information Fusion Attention Module

Deep features are rich in semantic information and shallow features are rich in detailed information. In order to better

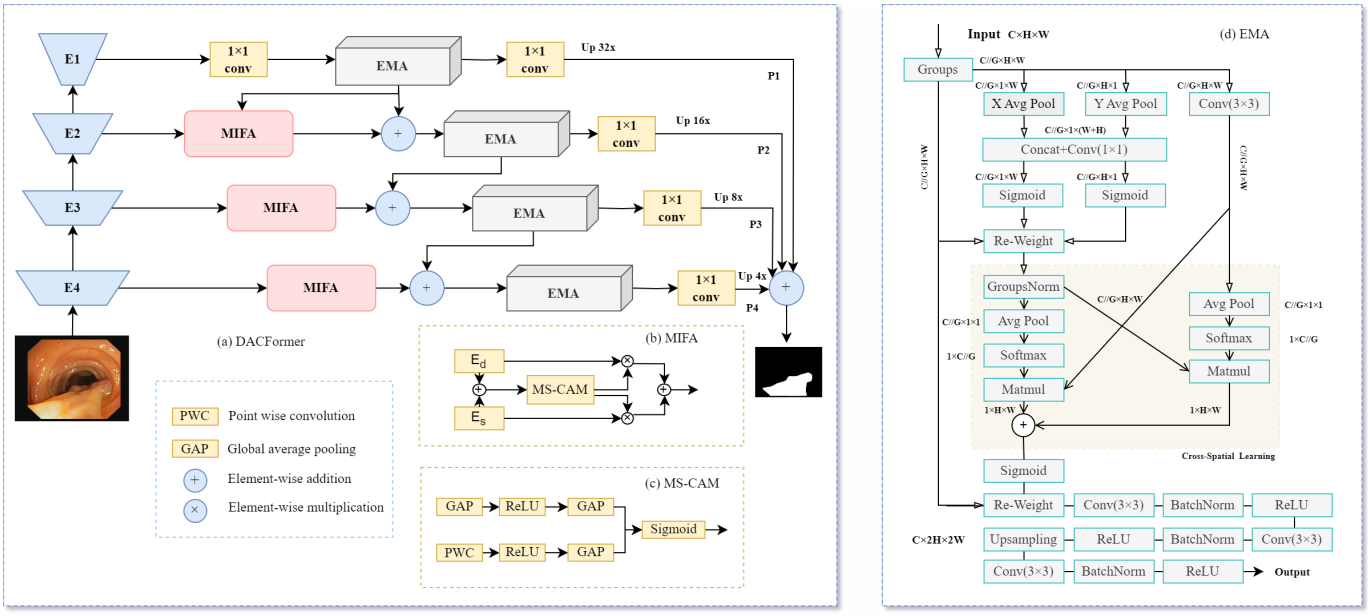

Fig. 1. Schematic diagram of the proposed DACFormer and its main building blocks: (b) Multi-scale Information Fusion Attention Module (MIFA), (c) Multi-scale Channel Attention Modul (MS-CAM), (d) Efficient Multi-Scale Attention Module (EMA).

combine semantic and scale inconsistent features [6], we use a multi-level information fusion attention module. For the deep features E_d, shallow features E_s, we first use upsampling to keep them at the same size. Then we enhance them by summation to reach a consensus region. Then the local features and global features are mined separately by two branches of Multi-scale Channel Attention Module (MS-CAM). The process follows the following equation:

$$M_o = Up(E_d) \otimes (1 - (M(Up(E_d) + E_s))) + E_s \otimes (1 + (M(Up(E_d) + E_s))), \quad (1)$$

$$M(E) = E \otimes \delta(L(E) + G(E)), \quad (2)$$

$$L(E) = BN(PW_2(\delta(BN(PW_1(E))))), \quad (3)$$

$$G(E) = BN(Gap(\delta(BN(Gap(E))))), \quad (4)$$

where Up refers to upsampling. δ refers to Sigmoid function. BN refers to Batch Normalization. PW refers to 1×1 convolution. Gap refers to global average pooling. $\otimes$ refers to point wise multiplication.

C. Efficient Multi-Scale Attention Module

Attentional mechanisms have been widely noticed in computer vision due to their flexible structural properties. However, past approaches have been implemented by downscaling convolution, and inspired by [7], we reshape feature partial channels into batch dimensions to achieve aggregated feature space information and channel information without downscaling. EMA divides the feature map into G sub-features across channel dimensional directions for learning different semantic information. It consists of three parallel branches, and the first two branches employ two 1D global averaging pooling operations to encode the channels along two spatial directions, respectively. The 3×3 convolution is stacked in the third branch to capture the multiscale feature representation. Considering the cross-space information aggregation approach, precise location information is embedded into the EMA while modeling remote dependencies. Contextual information at different scales is integrated.

D. Loss Function

We adopt a multi-stage loss function with prediction maps for all four stages under deep supervision. Our multi-stage loss function is shown as follows:

$$L_{total} = \sum L_i, i \in \{1, 2, 3, 4\}, \quad (5)$$

where the loss at each stage is a combination of f weighted intersection over union (IoU) loss and weighted binary cross-entropy (BCE) loss [8], which can be followed by the following equation:

$$L_i = L_{IoU}(P_i, G) + L_{BCE}(P_i, G), \quad (6)$$

where P_i refers to the prediction mask of the i th stage of the network, and G refers to the ground truth annotation.

III. Experiments

A. Experimental Setup

We use mean Dice [9](mDic), mean IoU (mIoU), mean absolute error (MAE), and E-measure [10] as evaluation metrics to verify the performance of the proposed framework. Specifically, we adopt Kvasir-SEG and ClinicDB datasets to verify the feature modeling capabilities of the proposed framework. And we test our network on unseen ColonDB and Endoscene datasets to verify the generalization performance. All experimental results are implemented using GPU-NVIDIA GeForce RTX 3090TI.

TABLE I
QUANTITATIVE RESULTS OF THE TEST DATASET KVASIR-SEG

method	meanDic	meanIoU	meanEm	mae
UNet	0.8183	0.7461	0.8807	0.0547
PraNet	0.8957	0.8422	0.9452	0.0268
ACSNet	0.8885	0.8312	0.9381	0.0338
UACANet	0.9111	0.8593	0.9596	0.0237
DCRNet	0.8954	0.8384	0.9392	0.0294
LDNet	0.9122	0.8578	0.9592	0.0227
Polyp-PVT	0.9129	0.8654	0.9589	0.0244
SSFormer	0.9149	0.8607	0.9567	0.0229
ColonFormer	0.9165	0.8639	0.9599	0.0243
Ours	**0.9227**	**0.8748**	**0.9602**	**0.0226**

TABLE II
QUANTITATIVE RESULTS OF THE TEST DATASET CLINICDB

method	meanDic	meanIoU	meanEm	mae
UNet	0.8226	0.7554	0.9133	0.0192
PraNet	0.9051	0.8548	0.9607	0.0093
ACSNet	0.9067	0.8545	0.9737	0.0132
UACANet	0.9111	0.864	0.9644	0.0078
DCRNet	0.9032	0.8494	0.9619	0.010
LDNet	0.921	0.8657	0.9769	0.0078
Polyp-PVT	0.9342	0.8878	0.9806	0.0063
SSFormer	0.903	0.8501	0.9634	0.0084
ColonFormer	0.9233	0.877	0.9726	0.0072
Ours	**0.9405**	**0.8945**	**0.9863**	**0.0060**

B. Experimental Results

This section compares the proposed method with past state-of-the-art methods, including UNet [2], PraNet [11], ACSNet [12], UACANet [13], DCRNet [14], LDNet [15], Polyp-PVT [16], SSFormer [17], ColonFormer [18]. TableI and Table II show the learning capabilities of the different models. We can observe that our proposed method obtains the highest performance for all four metrics on both datasets. For example, in the CVC-ClinicDB dataset, as opposed to ColonFormer, our method achieves a mean Dice score improvement of 1.72%, mean IoU score improvement of 1.75%.

TableIII and Table IV show the generalization ability of different models, and we can find that our model shows excellent generalization ability relative to the past methods. For example, in CVC-ColonDB dataset, as opposed to SSFormer, our method achieves a mean Dice score improvement of 2.76%, mean IoU score improvement of 2.3%. This verifies our proposed conclusion that feature reuse can effectively improve the generalization ability of the model for the model with the backbone network as Transformer.

Fig. 2 shows the comparison of the visualization results with different methods. We can clearly observe the precise localization of our method for polyps of different shapes. Also our method has a clearer boundary. As the image in the third row shows, in the case of large polyps with complex textures, our method is the only one with clear borders and no under-segmentation or over-segmentation.

TABLE III
QUANTITATIVE RESULTS OF THE TEST DATASET ENDOSCENE

method	meanDic	meanIoU	meanEm	mae
UNet	0.7099	0.6269	0.8475	0.0221
PraNet	0.888	0.8217	0.9635	0.006
ACSNet	0.8687	0.7941	0.943	0.0089
UACANet	0.8768	0.8098	0.9483	0.007
DCRNet	0.8601	0.7907	0.934	0.0094
LDNet	0.861	0.7823	0.9419	0.009
Polyp-PVT	0.8855	0.8156	0.9591	0.0091
SSFormer	0.8918	0.8223	0.9686	0.007
ColonFormer	0.8921	0.8245	0.9612	0.0084
Ours	**0.8967**	**0.8315**	**0.9646**	**0.0084**

TABLE IV
QUANTITATIVE RESULTS OF THE TEST DATASET COLONDB.

method	meanDic	meanIoU	meanEm	mae
UNet	0.5037	0.4357	0.6914	0.0586
PraNet	0.7244	0.6529	0.8487	0.0351
ACSNet	0.7602	0.6802	0.8785	0.0355
UACANet	0.7477	0.6731	0.8635	0.0399
DCRNet	0.7712	0.6871	0.876	0.0427
LDNet	0.7796	0.6986	0.8905	0.0327
Polyp-PVT	0.8146	0.7306	0.9201	0.026
SSFormer	0.7985	0.7161	0.9008	0.0314
ColonFormer	0.8033	0.7248	0.8983	0.0359
Ours	**0.8171**	**0.7391**	**0.9149**	**0.0288**

TABLE V
ABLATION STUDY FOR DACFOMER ON THE CLINICDB DATASET

Setting	meanDice	meanIoU	meanEm	MAE
w/o MIFA	0.9363	0.8892	0.9833	0.0067
w/o EMA	0.9344	0.8872	0.9858	0.0069
w/o Feature reuse	0.9337	0.8856	0.9796	0.0090
Ours	**0.9405**	**0.8945**	**0.9863**	**0.0060**

C. Ablation Studies

To further validate the effectiveness of the components and structure used in our network, we perform ablation experiments. As shown in Table V, we remove a certain design in the network and test their results on the CVC-ClinicDB dataset. For example, w/o MIFA means removing the MIFA module in the original network structure. We can observe that removing any component or changing the structure negatively affects the results of the network.

IV. CONCLUSION

In this paper, we propose a novel polyp segmentation network targeting the backbone network as Transformer. We use two attention mechanism based modules Multi-Scale Information Fusion Attention Module and Efficient Multi-Scale Attention Module to effectively fuse the features of encoder network at all levels. Contextual information is fully combined to effectively exploit the potential of Transformer. It is shown that we obtain excellent results on four challenging datasets,

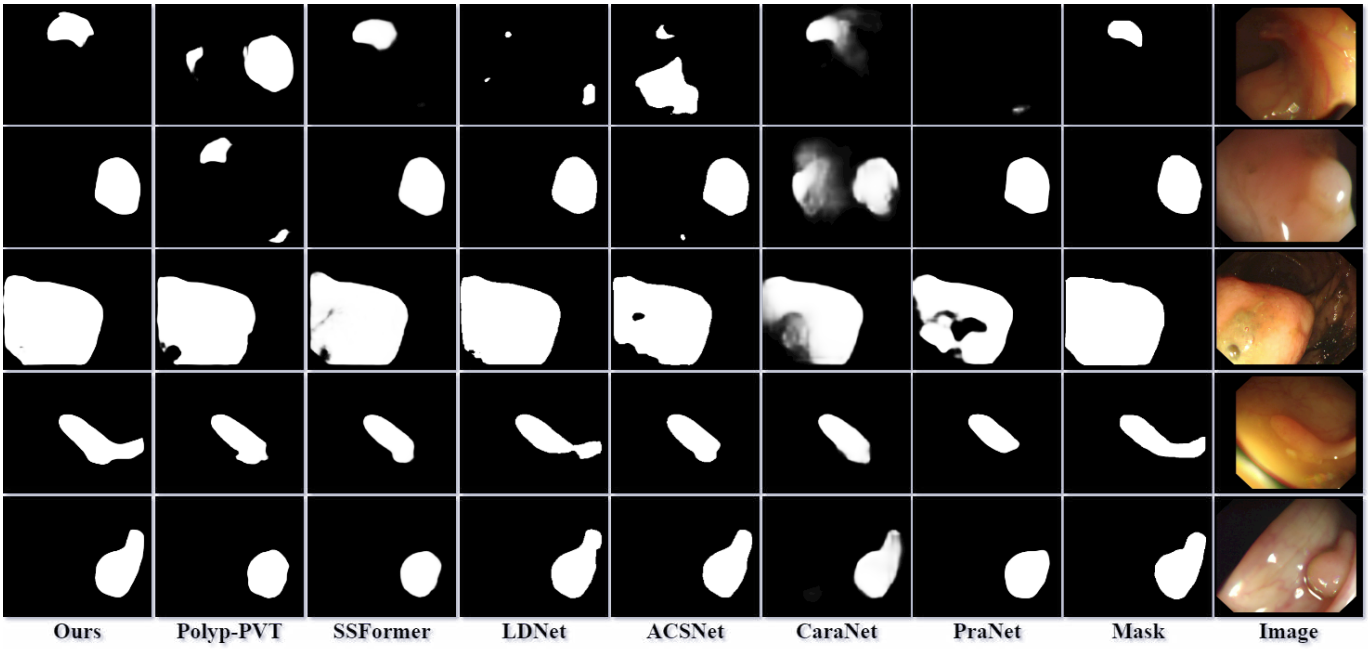

Fig. 2. Visualization results of different methods

which provide new ideas for Transformer-based approaches for polyp segmentation.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) through the Ministry of Science and Information and Communication Technology(MSIT) under Grant 2022R1A2C1011862.

REFERENCES

[1] R. L. Siegel, N. S. Wagle, A. Cercek, R. A. Smith, and A. Jemal, "Colorectal cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 3, pp. 233–254, 2023. [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21772

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[4] A. Lou, S. Guan, H. Ko, and M. H. Loew, "CaraNet: context axial reverse attention network for segmentation of small medical objects," in *Medical Imaging 2022: Image Processing*, O. Colliot and I. Išgum, Eds., vol. 12032, International Society for Optics and Photonics. SPIE, 2022, p. 120320D. [Online]. Available: https://doi.org/10.1117/12.2611802

[5] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, pp. 1–10, 03 2022.

[6] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3559–3568.

[7] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[8] J. Wei, S. Wang, and Q. Huang, "F^3net: fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 321–12 328.

[9] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[10] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.

[11] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2020, pp. 263–273.

[12] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, "Adaptive context selection for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*. Springer, 2020, pp. 253–262.

[13] T. Kim, H. Lee, and D. Kim, "Uacanet: Uncertainty augmented context attention for polyp segmentation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2167–2175.

[14] Z. Yin, K. Liang, Z. Ma, and J. Guo, "Duplex contextual relation network for polyp segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[15] R. Zhang, P. Lai, X. Wan, D.-J. Fan, F. Gao, X.-J. Wu, and G. Li, "Lesion-aware dynamic kernel for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 99–109.

[16] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-pvt: Polyp segmentation with pyramid vision transformers," *arXiv preprint arXiv:2108.06932*, 2021.

[17] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, "Stepwise feature fusion: Local guides global," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 110–120.

[18] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, and V. S. Dinh, "Colonformer: An efficient transformer based method for colon polyp segmentation," *IEEE Access*, vol. 10, pp. 80 575–80 586, 2022.