CM-MLP: Cascade Multi-scale MLP with Axial Context Relation Encoder for Edge Segmentation of Medical Image

Jinkai Lv[1†], Yuyong Hu[1†], Quanshui Fu[3†], Zhiwang Zhang[5], Yuqiang Hu[6], Lin Lv[7]
Guoqing Yang[3*], Jinpeng Li[4*], and Yi Zhao[1,2*]
[1]Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou, China
[2]The Research Center for Ubiquitous Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[3]Suining Central Hospital, Suining, China
[4]Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo ,China
[5]School of Electrical and Information Engineering, The University of Sydney, Sydney, Australia
[6]School of Mathematics and Computer Sciences, NanChang University, Nanchang, China
[7]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China
snszxyy@126.com, lijinpeng@ucas.ac.cn, biozy@ict.ac.cn

***Abstract*—The convolutional-based methods provide good segmentation performance in the medical image segmentation task. However, those methods have the following challenges when dealing with the edges of the medical images: (1) Previous convolutional-based methods do not focus on the boundary relationship between foreground and background around the segmentation edge, which leads to the degradation of segmentation performance when the edge changes complexly. (2) The inductive bias of the convolutional layer cannot be adapted to complex edge changes and the aggregation of multiple-segmented areas, resulting in its performance improvement mostly limited to segmenting the body of segmented areas instead of the edge. To address these challenges, we propose the CM-MLP framework on MFI (Multi-scale Feature Interaction) block and ACRE (axial context relation encoder) block for accurate segmentation of the edge of medical image. In the MFI block, we propose the cascade multi-scale MLP (Cascade MLP) to process all local information from the deeper layers of the network simultaneously and utilize a cascade multi-scale mechanism to fuse discrete local information gradually. Then, the ACRE block is used to make the deep supervision focus on exploring the boundary relationship between foreground and background to modify the edge of the medical image. The segmentation accuracy (Dice) of our proposed CM-MLP framework reaches 96.96%, 96.76%, and 82.54% on three benchmark datasets: CVC-ClinicDB dataset, sub-Kvasir dataset, and our in-house dataset, respectively, which significantly outperform the state-of-the-art method. The source code and trained models will be available at https://github.com/ProgrammerHyy/CM-MLP.**

***Index Terms*—MLP, medical image segmentation, semantic segmentation**

I. INTRODUCTION

In clinical diagnosis, medical image segmentation is a primary task, which has been extensively studied by the medical imaging community [1]–[3]. Compared with the traditional manual labelling method, the medical image segmentation algorithm [4], [5] can help doctors quickly find the location of lesions and reduce the workload. Therefore, the medical image segmentation algorithm is vital in medical image processing and analysis.

In recent years, more and more researchers have applied convolutional layers to medical image segmentation tasks [4]–[7]. Oktay *et al.* proposed a landmark work called UNet [8]. UNet [8] contains a U-shaped encoder-decoder architecture using a pyramid-like sampling process and skip connections to preserve the low-level semantic information. Following UNet, many different convolutional neural networks have been proposed, such as UNet++ [4], UNet3+ [9], 3D UNet [10], V-Net [11], Y-Net [12], and KiUNet [13]. However, these popular Unet-based structures focus on improving the overall segmentation performance rather than extracting the edge of medical image information, which is crucial for improving the segmentation performance of the network.

Recently, MLP-based (Multilayer Perceptron based) methods [14]–[20] provide promising results in computer vision tasks. MLP-Mixer [14] demonstrated that without convolution layers and self-attention mechanisms, it can still provide comparable performance to less computationally intensive transformers-based methods [21], [22]. Maxim [19] further applied MLPs to low-level vision tasks and achieved satisfactory performance. MLP-based methods [14]–[20] overcome the inductive bias of the weights and can process all the local information of the image at same time. Therefore, these MLP-based methods can naturally solve the problem of insufficient edge information extraction by the recent popular network.

Inspired by these MLP-based methods [14]–[20], we propose Cascade Multi-scale MLP (CM-MLP) framework, which is a denser design architecture with Multi-scale Feature Inter-

†These authors contributed equally to this work.
*Corresponding author.

action block (MFI) and axial context relation encoder (ACRE). Using the MFI block, CM-MLP can overcome the influence of inductive bias brought by convolution layers, simultaneously process all local information and gradually fuse the discrete local information. Using the ACRE block, CM-MLP can focus on the boundary relationship between foreground and background around the segmentation edge.

The contributions of this paper can be summarized as follows:

- We propose a novel CM-MLP framework to extract better edge information in medical image segmentation. In this framework, we proposed the MFI block to capture complex and dense edge (i.e., aggregated multiple segmented regions) information. MFI block can process all local information simultaneously and gradually fuse the discrete local information. In addition, we also proposed the ACRE block to make the CM-MLP framework focus on segmenting object edges rather than bodies. With the cooperation of the MFI block and ACRE block, CM-MLP overcomes the influence of inductive bias brought by convolution layers and neglect of the boundary relationship between foreground and background.
- Comparison results show that our proposed CM-MLP framework outperforms the previous state-of-the-art method on the CVC-ClinicDB dataset, sub-Kvasir dataset, and our in-house dataset.

II. RELATED WORK

A. Convolutional model

The emergence of medical image segmentation frameworks based on CNN (Convolutional Neural Network) models, especially Unet [8], pioneered segmentation networks with convolution as the main architecture. DUNet [23], based on the U-Net framework, uses a Deformable Convolution Block [24] as each unit of the encoder and decoder. The deformable convolution block simulates different shapes and scales by learning local, dense, and adaptive receptive fields. R2U-Net [25] combines residual connections and recurrent convolutions to replace the original submodules in U-Net. BIO-Net [26] proposed bi-directional skip connections to extract more spatial information by recurrently reusing the building blocks and then using the architecture optimization algorithm in BIX-NAS [27] to optimize the connection in a multi-stage network. PraNet [28] establish the relationship between areas and boundary cues using the reverse attention (RA) module, and then CaraNet [29] use Channel-wise Feature Pyramid (CFP) module with A-RA(Axial reverse attention) to further mine the boundary cues. However, in the reverse attention, only the pixels around the segmentation result are highlighted. The error pixels may be kept in the final result because the pixels have been wrong segmented in the previous operation while not fixing the error pixels.

With the continuous development of the convolutional network, the performance of which is increasingly affected by the complex edge information. However, all of the above attempts [26]–[29] are still based on convolutional-based architectures. It is inductive bias and neglect of the boundary relationship between foreground and background around the segmentation edge hindering the network's ability to extract edge information. Therefore, we take advantage of the ability of MLP to process all local information simultaneously and focus on the boundary relationship between foreground and background around the segmentation edge, thereby improving the network's ability to extract complex edge information.

B. MLP-based model

MLP-Mixer [14], an MLP-based architecture that replaces self-attention with simple token-mixing MLP, achieves competitive results in image classification. gMLP [15] demonstrated that self-attention is not necessary for NLP tasks through gated-based MLP. CYCLEMLP [16] achieves a linear computational complexity related to image size through CycleFC, enabling pure MLP architectures for object detection and segmentation in larger images. MLP-3D [17] leverages VISION PERMUTATOR [18] in the video classification task to encode feature representations with linear projections along the height and width dimensions, respectively. Then, giving token-mixing MLP a temporal modelling Ability by GTM (Grouped Time Mixing) in the temporal dimension. MAXIM [19], as a network based on a pure MLP architecture, utilizes a cross-gated module and a multi-axis gated MLP to achieve the mixing of local and global spatial information. RepMLPNet [20] merges the trained parameters of the parallel convolution kernels into the FC kernel and merges the local prior into the FC (Full Connect) layer utilizing local injection. Therefore, the RepMLPNet can capture local and global information and becomes a pure MLP-structured model in the inference stage.

When using MLP for vision tasks, we have noticed that some appropriate designs can make MLP even more potent than convolutions. Therefore, we propose the CM-MLP framework based on the MFI (Multi-scale Feature Interaction) block and ACRE (axial context relation encoder) block for accurate segmentation of the edge of the medical image. The MFI block can process all local information simultaneously by Cascade MLP (cascade Multi-scale MLP). Then the ACRE block can help our CM-MLP framework focus on the boundary relationship between foreground and background around the segmentation edge to better extract the edge information.

Note that although the computational complexity of the MFI block is linearly related to the H$\times$W same as [19], considering the size of the feature map and the amount of information contained, we only apply the MFI block in the last three layers of the network to further reduce the amount of parameters. The computational cost is negligible.

III. METHOD

We propose the CM-MLP framework for medical image segmentation and adopt the MFI and ACRE blocks. Unlike [30], which added MLP to the encoder and decoder of Unet, MLPs are in parallel with Unet in our proposed CM-MLP framework. In III-A, we briefly introduce the CM-MLP

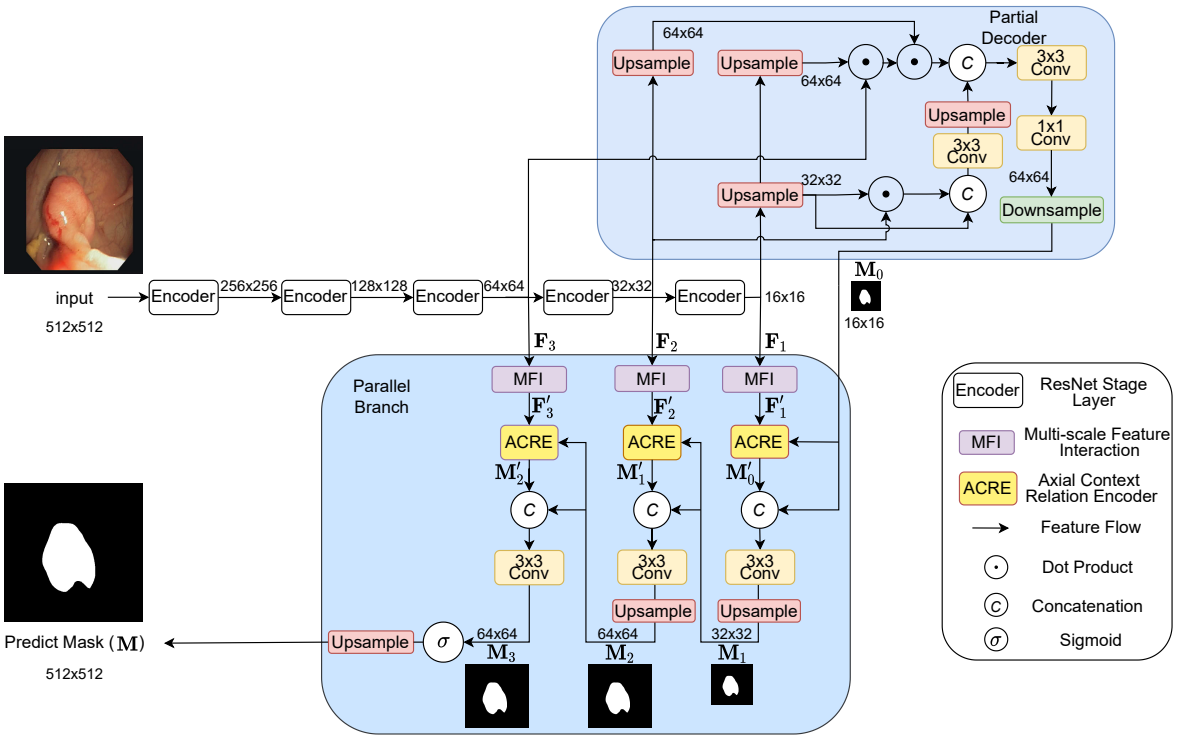

Fig. 1: Overview of the proposed CM-MLP framework, which consists of the 5-stage Encoder, the Partial Decoder, and the Parallel Branch. The 5-stage Encoder first encodes the input image. The feature maps from the last three stages (i.e., $\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3$) are decoded by the Partial Decoder to produce the original mask $\mathbf{M}_0$. In the Partial Decoder, the feature maps will dot products of each other in the same size by upsampling and generate the original mask $\mathbf{M}_0$ by concatenated operation, convolution layer, and downsampling operation. In the first branch of the Parallel Branch, the MFI block takes feature map $\mathbf{F}_1$ as the input to get feature map $\mathbf{F}'_1$. Then the ACRE block takes $\mathbf{F}'_1$ and $\mathbf{M}_0$ as input and generates the refined mask $\mathbf{M}'_0$. The higher resolution mask $\mathbf{M}_1$ is generated from refined mask $\mathbf{M}'_0$ and original mask $\mathbf{M}_0$ by concatenated operation, convolution layer, and upsampling operation. The other two branches go through the same operation. The final mask $\mathbf{M}$ is obtained by $\mathbf{M}_3$ using sigmoid and upsampling operations. In Deep Supervision, after having $\mathbf{M}_0$, $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$, we upsample those results to the same size as ground truth and calculate the total loss.

framework. In III-B, we introduce the principle of the MFI block, which can process all local information simultaneously and gradually fuse discrete local information. In III-C, we introduce the principle of the ACRE block, which can focus on the boundary relationship between foreground and background around the segmentation edge. In III-D, we introduce the loss function of our CM-MLP framework.

A. The main structure

Figure 1 shows our proposed CM-MLP framework, which consists of the 5-stage Encoder, the Partial Decoder, and the Parallel Branch. In the Parallel Branch, the MFI block can process all local information from the deeper layer of CNN (Convolutional Neural Network). The ACRE block can make the CM-MLP framework focus on exploring the boundary relationship between foreground and background. The MFI and ACRE block of the model is explained in detail below.

B. Multi-scale Feature Interaction (MFI) block

The application of MLP in visual tasks is mainly limited to image classification. By dividing the input image into non-overlapping patches and merging each patch in the spatial and channel dimensions to extract rich image information. Maxim's proposal [19] enables us to see that MLP has good performance on dense prediction tasks. It divides the original feature map into local and global branches aiming to extract information at different scales. Inspired by Maxim [19], we propose a cascade Multi-scale MLP (Cascade MLP) in MFI block to encode the information and then fuse the information into a larger receptive field through Local MLP and Global MLP of multi-scale.

Our proposed multi-scale Feature Interaction (MFI) block has two branches, in which the feature map $\mathbf{F}$ is channel-wisely split into $\mathbf{F}_{up}, \mathbf{F}_{bottom}$. As shown in Figure 2, three scales of Global MLP and Local MLP are connected in series to get the enriched features map gradually.

In the Global MLP block (i.e., red block of Figure 2), the input feature map (with size (H, W, C)) is grid into $(g \times g)$ non-overlapping patches of size $(H_g \times W_g)$ where $H = g \times H_g, W = g \times W_g$, to obtain the feature map $(g \times g,\ H_g \times W_g, C)$.

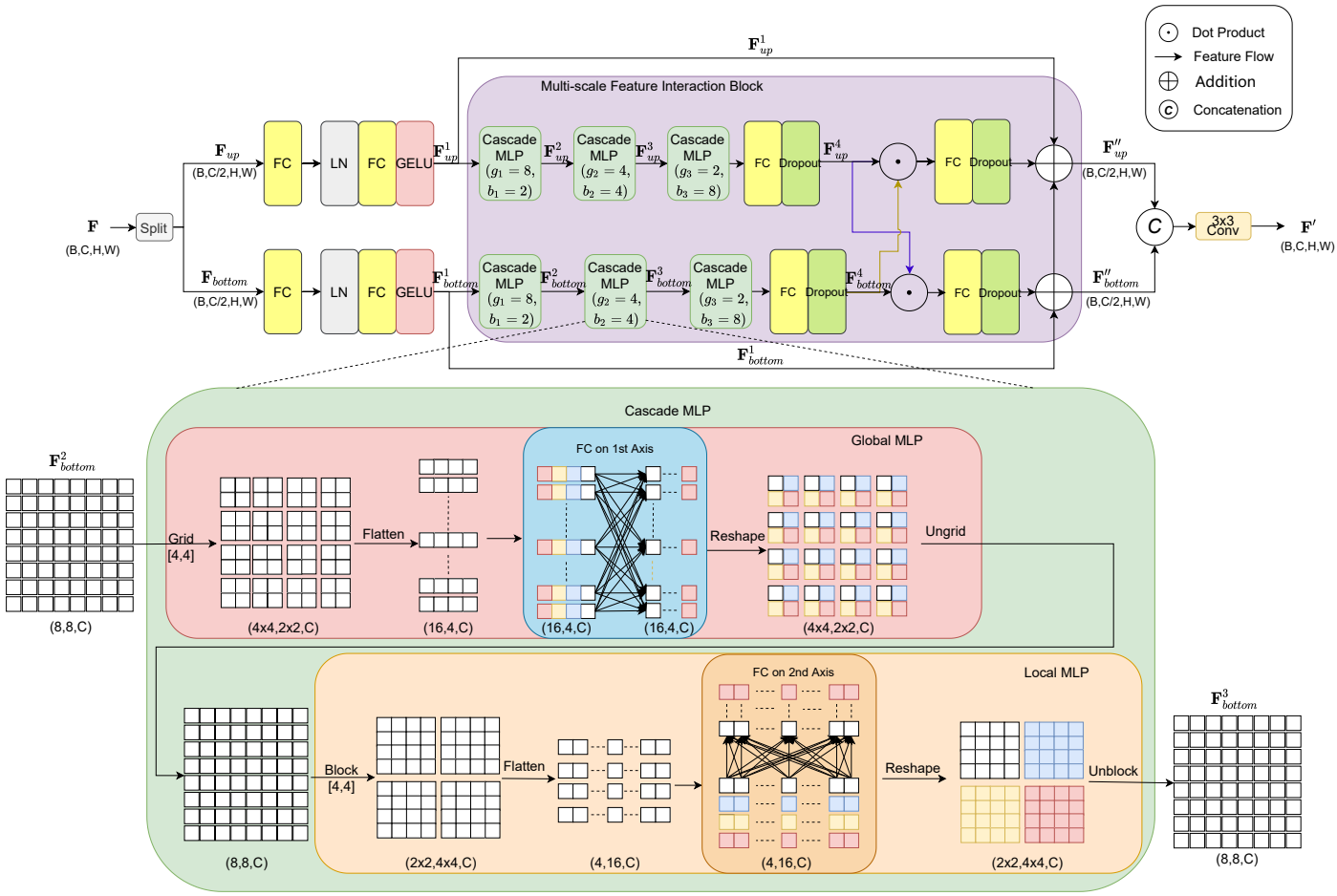

Fig. 2: Illustration of the Multi-scale Feature Interaction (MFI) block. The input feature map $\mathbf{F}$ is channel-wise split into two branches $\mathbf{F}_{up}$ and $\mathbf{F}_{bottom}$. After each branch processed by the multiple Cascade MLP blocks, it will be alternately multiplied to increase information interaction and added together ($\mathbf{F}''_{up} = \mathbf{F}^1_{up} + \mathbf{F}^4_{up} \times \mathbf{F}^4_{bottom} + \mathbf{F}''_{bottom}, \mathbf{F}''_{bottom} = \mathbf{F}^1_{bottom} + \mathbf{F}^4_{bottom} \times \mathbf{F}^4_{up}$). The output of MFI block $\mathbf{F}'$ is obtained by concatenating two branch features $\mathbf{F}''_{up}$ and $\mathbf{F}''_{bottom}$ and the convolution layer. For Cascade MLP, we take the second Cascade MLP in MFI block ($b_2 = 4$, $g_2 = 4$) as an example. For better understanding, we used $\mathbf{F}^2_{bottom}$ $(W = 8, H = 8, C)$ as input, where C is the size of the channel. Input feature $\mathbf{F}^2_{bottom}$ will be processed by Global MLP and Local MLP to obtain the output feature $\mathbf{F}^3_{bottom}$. In Global MLP, the feature map is first grid into 4×4 ($g_2 = 4$) non-overlapping patches which sizes is 2×2. After flattening, the FC layer is executed on the first axis (the same colour represents the FC layer's input and output vector) and then reshaped back and Ungrid to the original size. In Local MLP, the feature map is first blocked into 2×2 non-overlapping patches which sizes is 4×4 ($b_2 = 4$). After flattening, the FC layer is executed on the second axis and then reshaped back and unblock to the original size to get $\mathbf{F}^3_{bottom}$ feature map.

In the Local MLP block (i.e., orange block of Figure 2), the feature map (with size (H, W, C)) is blocked into $(H_b \times W_b)$ non-overlapping patches of size $(b \times b)$ where $H = b \times H_b, W = b \times W_b$, resulting in the feature map $(H_b \times W_b, b \times b, C)$.

It is noted that the size of patches in Global MLP and Local MLP (i.e., b and g) are not independent of each other. It is specified that $b \times g = W = H$ (for example: $H = W = 16$, $g_1 = 8$, $g_2 = 4$, $g_3 = 2$, and corresponding $b_1 = 2$, $b_2 = 4$, $b_3 = 8$). When the size of patches in Global MLP gradually decreases (g gradually shrinks), the size of patches in Local MLP (b gradually increases) increases. In other words, the distribution of the points in the FC input vector in Global MLP will be more sparse. In Local MLP, the number of points in the FC input vector in patches will be larger. Both of them will gradually expand the receptive field of the Cascade MLP. Therefore, The MFI block gradually fuses discrete local information and can get a gradually enriched features map.

C. Axial Context Relation Encoder (ACRE) block

In order to make the MFI block focus on mining edge information instead of the body, we propose Axial Context Relation Encoder (ACRE) block inspired by [31]. The ACRE block focuses on the distinction between foreground and background boundaries so that the MFI block will extract more edge information.

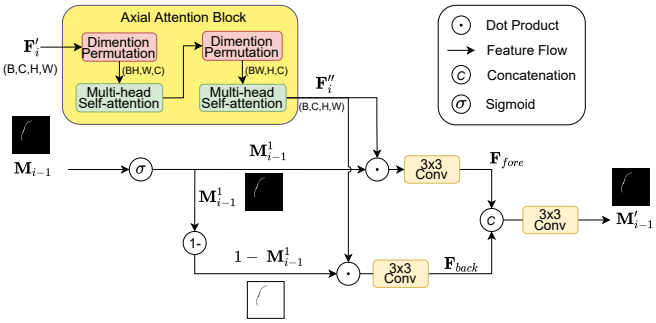

Fig. 3: Illustration of the Axial Context Relation Encoder (ACRE) block. The feature map $\mathbf{F}'_i$, $i = 1, 2, 3$ indicates the output of MFI block in the Figure 1. $\mathbf{F}'_i$ is first processed by Axial Attention Block, which contains two self-attention operations on different dimensions (H and W) through dimension permutation to get feature map $\mathbf{F}''_i$. The original mask $\mathbf{M}_{i-1}$ is processed by the sigmoid function and then reproduced in two copies, one kept as is and one processed as $1 - \mathbf{M}_{i-1}$. After that, we multiply the $\mathbf{F}''_i$ with $\mathbf{M}_{i-1}$ and $1-\mathbf{M}_{i-1}$ respectively to get foreground $\mathbf{F}_{fore}$ and background $\mathbf{F}_{back}$ and then get the refined information $\mathbf{M}'_{i-1}$ through concatenation of $\mathbf{F}_{fore}$ and $\mathbf{F}_{back}$ on channel level and convolution operation.

As shown in Figure 3, each ACRE block has three main steps: firstly, the input feature map $\mathbf{F}'_i$ is sent into the axial attention block, which contains two self-attention operations on different dimensions (H and W) through dimension permutation, to obtain the feature map $\mathbf{F}''_i$.

Secondly, $\mathbf{F}''_i$ will be masked by $\mathbf{M}_{i-1}$ to obtain the foreground feature $\mathbf{F}_{fore}$ and the background feature $\mathbf{F}_{back}$:

$$\mathbf{F}_{back} = \phi_{back}(\mathbf{F}''_i \odot (1 - \mathbf{M}_{i-1})), \tag{1}$$

$$\mathbf{F}_{fore} = \phi_{fore}(\mathbf{F}''_i \odot \mathbf{M}_{i-1}), \tag{2}$$

where $\phi_{back}(\cdot)$ and $\phi_{fore}(\cdot)$ represent the 3×3 convolutions and $\odot$ represents the dot product.

Finally, the output feature $\mathbf{M}'_{i-1}$ of the ACRE block is obtained through the concatenation of $\mathbf{F}_{fore}$ and $\mathbf{F}_{back}$ on the channel dimension and 3×3 convolution.

To ensure the effect of axial attention block on the output feature $\mathbf{M}'_{i-1}$, we do not apply the offset in [31] to the contextual feature extraction of spatial locations. Second, the ACRE block can also be regarded as a supplement to the axial reverse attention in CaraNet [29].

D. Deep Supervision

To better emphasize the segmentation task of each branch, we adopt a deep-supervised way to add the loss of each branch to the total loss.

The loss function can be represented as $\ell = \ell_{IOU} + \ell_{BCE}$ by applying weighted intersection over union (IoU) and weighted binary cross entropy (BCE). we apply deep supervision for the four branch results ($\mathbf{M}_0$, $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$). Before calculating the loss, we upsampled four branch results to the same size as ground truth $\mathbf{G}$ as ($\mathbf{M}_0^{up}$, $\mathbf{M}_1^{up}$, $\mathbf{M}_2^{up}$, $\mathbf{M}_3^{up}$). Thus, the total loss can be represented as:

$$\ell_{total} = \sum_{i=0}^{3} \ell(\mathbf{G}, \mathbf{M}_i^{up}) \tag{3}$$

IV. EXPERIMENTS AND RESULTS

A. Datasets and baselines

Experiments are performed on CVC-ClinicDB dataset [32], sub-Kvasir dataset [33], and our in-house dataset. The CVC-ClinicDB dataset [32] is a polyps segmentation dataset, which contains 612 open-access images from 31 colonoscopy clips. The sub-Kvasir dataset [33] is a polyps segmentation dataset, which contains 1,000 images selected from a sub-class (polyp class) of the Kvasir dataset. Our in-house dataset is a large subdural hematoma segmentation dataset comprising 1049 images from 65 patients. All the three segmentation dataset is divided into training, validation and testing sets with the ratio of 7:1:2. We compare our proposed CM-MLP framework with four the-state-of-art medical image segmentation methods: U-Net [8], U-Net++ [4], PraNet [28] and CaraNet [29].

B. Implementation details

We implement our model in PyTorch. Affine transformation, horizontal flip and vertical flip are used for data augmentation. All the input images are uniformly resized to 512×512. LookAhead [34] optimization algorithm is used to optimize the parameters. The entire network is trained in an end-to-end way. Following the work CaraNet [29], We employ three metrics (i.e., Mean Dice, Mean IoU and MAE) for quantitative evaluation and utilize MPA metrics to evaluate pixel-level accuracy.

TABLE I: COMPARISON OF SEGMENTATION RESULTS ON THE SUB-KVASIR DATASET.

Methods	Dice	mIoU	MAE	MPA
U-Net [8]	0.6321	0.7704	0.0608	0.8949
U-Net++ [4]	0.8139	0.7014	0.0372	0.8362
PraNet [28]	0.9454	0.8970	0.0120	0.9508
CaraNet [29]	0.9482	0.9027	0.0113	0.9595
CM-MLP (Ours)	**0.9676**	**0.9373**	**0.0087**	**0.9658**

TABLE II: COMPARISON OF SEGMENTATION RESULTS ON THE CVC-CLINICDB DATASET.

Methods	Dice	mIoU	MAE	MPA
U-Net [8]	0.6469	0.4858	0.0544	0.6585
U-Net++ [4]	0.7290	0.5917	0.0357	0.8443
PraNet [28]	0.9420	0.8951	0.0071	0.9582
CaraNet [29]	0.9611	0.9256	0.0060	0.9665
CM-MLP (Ours)	**0.9696**	**0.9412**	**0.0048**	**0.9758**

TABLE III: COMPARISON OF SEGMENTATION RESULTS ON OUR IN-HOUSE DATASET.

Methods	Dice	mIoU	MAE	MPA
U-Net [8]	0.7583	0.6231	0.0035	0.6707
U-Net++ [4]	0.6997	0.5595	0.0039	0.7175
PraNet [28]	0.7949	0.6649	0.0026	0.9315
CaraNet [29]	0.8155	0.6940	0.0025	0.9249
CM-MLP(Ours)	**0.8254**	**0.7087**	**0.0024**	**0.9378**

TABLE IV: ABLATION STUDY OF OUR PROPOSED CM-MLP FRAMEWORK ON THE CVC-CLINICDB, SUB-KVASIR AND OUR IN-HOUSE DATASET.

	settings	Dice	mIoU	MAE	MPA
sub-Kvasir	CM-MLP	**0.9676**	**0.9373**	0.0087	**0.9658**
	CM-MLP w/o MFI	0.9634	0.9295	0.0090	0.9653
	CM-MLP w/o Local	0.9631	0.9290	**0.0086**	0.9636
	CM-MLP w/o Global	0.9663	0.9350	0.0088	0.9647
	CM-MLP w/o ACRE	0.9644	0.9313	0.0093	0.9653
CVC-ClinicDB	CM-MLP	**0.9696**	**0.9410**	**0.0048**	0.9758
	CM-MLP w/o MFI	0.9683	0.9387	0.00054	**0.9759**
	CM-MLP w/o Local	0.9669	0.9361	0.0055	0.9753
	CM-MLP w/o Global	0.9682	0.9385	0.0053	0.9754
	CM-MLP w/o ACRE	0.9687	0.9393	0.0055	0.9725
In-house	CM-MLP	**0.8254**	**0.7087**	0.0024	**0.9378**
	CM-MLP w/o MFI	0.8148	0.6925	0.0024	0.9353
	CM-MLP w/o Local	0.8203	0.6992	0.0023	0.9348
	CM-MLP w/o Global	0.8164	0.6964	**0.0022**	**0.9378**
	CM-MLP w/o ACRE	0.8072	0.6811	0.0024	0.9235

C. Performance Comparison

We conduct the performance comparison of medical image segmentation task on Table I , Table II, Table III. In all three datasets, our proposed method outperforms the current state-of-the-art methods: U-Net [8], U-Net++ [4], PraNet [28], and CaraNet [29].

In Table I and Table II, our proposed CM-MLP framework outperforms the current state-of-the-art models in all metrics. In particular, the mIoU score of our proposed CM-MLP framework outperforms the previous method by 3.46% and 1.56% in the CVC-ClinicDB and sub-Kvasir datasets, respectively. The experiment results demonstrate that our proposed CM-MLP framework has a strong learning ability to segment polyps images with complex edges effectively.

In Table III, we report the results of current state-of-the-art methods with our proposed CM-MLP framework in a more challenging in-house dataset. Our proposed CM-MLP framework outperforms the current state-of-the-art models in all metrics. Particularly, the mIoU score of our proposed M-MLP framework outperforms the previous method by 1.47%. Therefore, our proposed CM-MLP can perform better on more complex segmentation tasks.

D. Component Analysis

We perform the following ablation studies on all used datasets to verify the effectiveness of each component in the CM-MLP framework: (1) **CM-MLP**: Our proposed CM-MLP framework (2) **CM-MLP w/o MFI**: Our proposed CM-MLP framework without MFI block; (3) **CM-MLP w/o Local**: Our proposed CM-MLP framework whose MFI block without local MLP; (4) **CM-MLP w/o Global**: Our proposed CM-MLP framework whose MFI block without global MLP; (5) **CM-MLP w/o ACRE**: Our proposed CM-MLP framework without ACRE block;

We have the following observations from the results in Table IV. First, our proposed approach **CM-MLP** outperforms the method **CM-MLP w/o MFI**, which demonstrates it is practical to use the MFI block to process all local information at the same time. Second, our approach **CM-MLP** also outperforms both **CM-MLP w/o Local** and **CM-MLP w/o Global**, which indicates that it is beneficial to use both Local MLP and Global MLP to catch the information from global to local. Third, our proposed approach **CM-MLP** is better than the method **CM-MLP w/o ACRE**, which demonstrates that it is helpful to focus on exploring the boundary relationship between foreground and background.

TABLE V: COMPARISON RESULTS OF DIFFERENT CONNECTIONS OF MLPS ON THE CVC-CLINICDB, SUB-KVASIR AND OUR IN-HOUSE DATASET.

	settings	Dice	mIoU	MAE	MPA
sub-Kvasir	CM-MLP	**0.9676**	**0.9373**	0.0087	0.9658
	MFI-PP	0.9674	0.9370	0.0088	**0.9725**
	MFI-CP	0.9653	0.9330	**0.0076**	0.9640
CVC-ClinicDB	CM-MLP	**0.9696**	**0.9410**	**0.0048**	**0.9758**
	MFI-PP	**0.9696**	**0.9410**	0.0050	0.9719
	MFI-CP	0.9678	0.9377	0.0055	0.9725
In-house Dataset	CM-MLP	**0.8254**	**0.7087**	0.0024	**0.9378**
	MFI-PP	0.8205	0.7002	**0.0022**	0.9291
	MFI-CP	0.8200	0.6986	0.0024	0.9357

TABLE VI: COMPARISON OF OUR PROPOSED MFI BLOCK WITH CFP BLOCK PROPOSED IN [29] ON THE CVC-CLINICDB, SUB-KVASIR AND OUR IN-HOUSE DATASET.

	settings	Dice	mIoU	MAE	MPA
sub-Kvasir	CM-MLP	**0.9676**	**0.9373**	0.0087	**0.9658**
	CFP	0.9644	0.9317	**0.0082**	0.9658
CVC-ClinicDB	CM-MLP	**0.9696**	**0.9410**	**0.0048**	**0.9758**
	CFP	0.9673	0.9368	0.0053	0.9370
In-house Dataset	CM-MLP	**0.8254**	**0.7087**	**0.0024**	0.9378
	CFP	0.8206	0.7011	0.0026	**0.9430**

TABLE VII: COMPARISON OF OUR PROPOSED ACRE BLOCK WITH A-RA BLOCK PROPOSED IN [29] ON THE CVC-CLINICDB, SUB-KVASIR AND OUR IN-HOUSE DATASET.

	settings	Dice	mIoU	MAE	MPA
sub-Kvasir	CM-MLP	**0.9676**	**0.9373**	**0.0087**	**0.9658**
	A-RA	0.9568	0.9181	0.0120	0.9623
CVC-ClinicDB	CM-MLP	**0.9696**	**0.9410**	**0.0048**	**0.9758**
	A-RA	0.9561	0.9169	0.0079	0.9623
In-house Dataset	CM-MLP	**0.8254**	**0.7087**	0.0024	**0.9378**
	A-RA	0.8226	0.7030	**0.0023**	0.9309

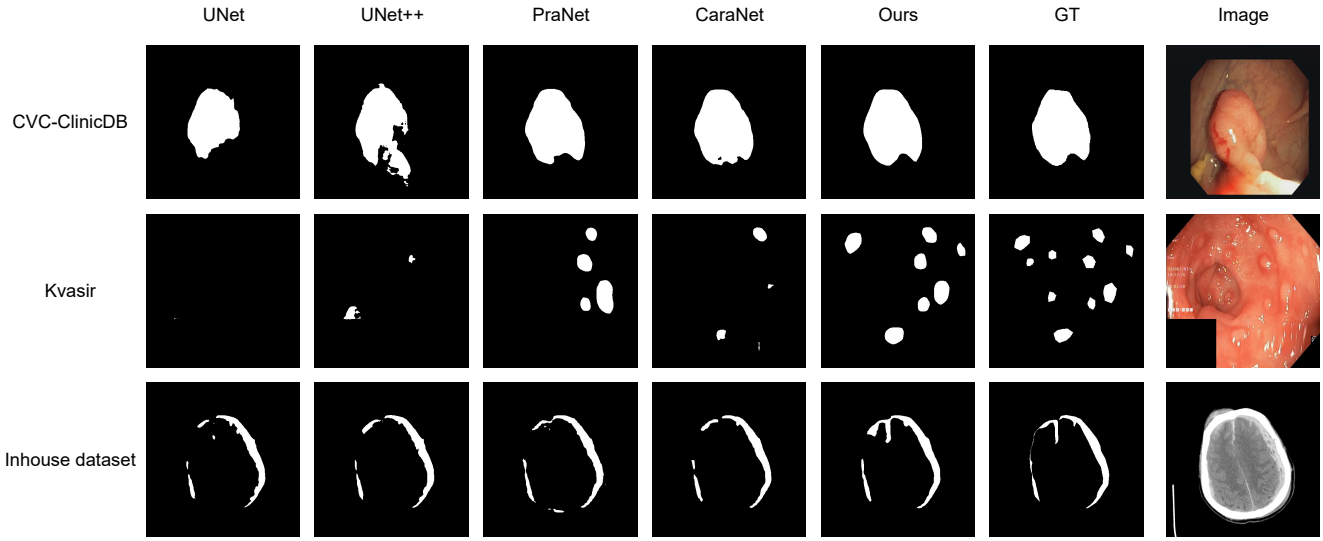

Fig. 4: Qualitative comparison of different methods in three datasets. From left to right: U-Net [8], U-Net++ [4], PraNet [28],CaraNet [29],CM-MLP (Ours), Ground Truth image, and input segmentation image.

In Table V, we report the results when comparing the different connections of Global MLP, Local MLP and Cascade MLP. Our proposed method **CM-MLP** use Cascade MLP in the MFI block in which Global MLP and Local MLP is connected in series. The method **MFI-PP** use parallel MLP between Local MLP and Global MLP proposed in Maxim [19] in Cascade MLP. For method **MFI-CP**, those three Cascade MLPs are paralleled with each other while keeping a series connection of Global MLP and Local MLP in each Cascade MLP. The comparison results demonstrate that the combined form of the Cascade MLP in the MFI block better captures the information when the segmentation tasks become difficult, which contains more complex edge of the segmented images.

In Table VI, we compare the effectiveness of our proposed MFI block with the **CFP** block proposed in [29]. The mIoU score of our proposed method is 0.5%, 0.4%, and 0.7% higher than using **CFP** block in the sub-Kvasir dataset, CVC-ClinicDB dataset, and our in-house dataset respectively. The experiments demonstrate that the MFI block can better capture the local information using MLP than the **CFP** block proposed by [29].

In Table VII, it can be seen that after replacing the ACRE block with the axial reverse attention module (**A-RA**) proposed in [29], the mIoU scores are significantly decreased by 1.9% and 2.4% in the sub-Kvasir dataset and CVC-ClinicDB dataset. This shows that the ACRE block is crucial to capturing edge information in sub-Kvasir and CVC-ClinicDB datasets.

E. Qualitative Analysis

Fig 4 shows the results of the qualitative comparison, where our proposed CM-MLP framework provides better performance.

In the CVC-ClinicDB dataset, we observe that the previous state-of-the-art methods cannot capture the bump changes when the edge of the segmented images is more complex. In addition, the previous state-of-the-art methods and our proposed CM-MLP framework can successfully capture the body of the segmented images. These observations demonstrate that our proposed CM-MLP framework provides better performance when the edge of the segmented images is more complex and keeps the ability to capture the body of the segmented images simultaneously.

In the sub-Kvasir dataset, we observe that the current state-of-the-art methods cannot capture the body of the segmented images when the aggregation of multiple-segmented areas occurs. This demonstrates that our proposed CM-MLP framework can perform better when the aggregation of multiple-segmented areas appear in the medical image. In addition, the observation demonstrates that our proposed MLP-based MFI block can process all local information at the same time.

In our in-house dataset, we can see that the edges of the segmented images are more complex and the aggregation of multiple-segmented areas. Our proposed CM-MLP framework still captures the vital part of edges in the segmented image.

V. CONCLUSION

We propose a general MLP-based framework called CM-MLP for medical image segmentation. This framework can process all the local information of the image simultaneously. Therefore, the CM-MLP can cope with the complex edge information of a segmented area and the aggregation of multiple-segmented areas. While improving network performance minimizes the amount of computation and complexity brought by MLP. Extensive experiments demonstrate that our proposed CM-MLP consistently outperforms state-of-the-art methods on three challenging medical segmentation datasets. In the future, we will conduct our proposed CM-MLP on more datasets and explore the impact of data types from different modalities and sizes.

REFERENCES

[1] C. N. Vasconcelos and B. N. Vasconcelos, "Experiments using deep learning for dermoscopy image analysis," *Pattern Recognition Letters*, vol. 139, pp. 95–103, 2020.

[2] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, *et al.*, "Nucleus segmentation across imaging experiments: the 2018 data science bowl," *Nature methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[3] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, *et al.*, "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *Medical Image Analysis*, vol. 70, p. 102002, 2021.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[5] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 327–331, 2018.

[6] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[7] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[8] 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), pp. 234–241, Springer International Publishing, 2015.

[9] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, *et al.*, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059, 2020.

[10] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, eds.), (Cham), pp. 424–432, Springer International Publishing, 2016.

[11] 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)*, pp. 565–571, 2016.

[12] S. Mehta, E. Mercan, J. Bartlett, D. Weaver, J. G. Elmore, and L. Shapiro, "Y-net: Joint segmentation and classification for diagnosis of breast biopsy images," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, eds.), (Cham), pp. 893–901, Springer International Publishing, 2018.

[13] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, eds.), (Cham), pp. 363–373, Springer International Publishing, 2020.

[14] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, *et al.*, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 24261–24272, Curran Associates, Inc., 2021.

[15] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 9204–9215, Curran Associates, Inc., 2021.

[16] S. Chen, E. Xie, C. Ge, D. Liang, and P. Luo, "Cyclemlp: A mlp-like architecture for dense prediction," *arXiv preprint arXiv:2107.10224*, 2021.

[17] Z. Qiu, T. Yao, C.-W. Ngo, and T. Mei, "Mlp-3d: A mlp-like 3d architecture with grouped time mixing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3062–3072, June 2022.

[18] Q. Hou, Z. Jiang, L. Yuan, M.-M. Cheng, S. Yan, and J. Feng, "Vision permutator: A permutable mlp-like architecture for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022.

[19] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5769–5780, June 2022.

[20] X. Ding, H. Chen, X. Zhang, J. Han, and G. Ding, "Repmlpnet: Hierarchical vision mlp with re-parameterized locality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 578–587, June 2022.

[21] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, *et al.*, "Big transfer (bit): General visual representation learning," in *Computer Vision – ECCV 2020* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), (Cham), pp. 491–507, Springer International Publishing, 2020.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[23] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "Dunet: A deformable network for retinal vessel segmentation," *Knowledge-Based Systems*, vol. 178, pp. 149–162, 2019.

[24] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[25] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *arXiv preprint arXiv:1802.06955*, 2018.

[26] T. Xiang, C. Zhang, D. Liu, Y. Song, H. Huang, and W. Cai, "Bio-net: Learning recurrent bi-directional connections for encoder-decoder architecture," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, eds.), (Cham), pp. 74–84, Springer International Publishing, 2020.

[27] X. Wang, T. Xiang, C. Zhang, Y. Song, D. Liu, H. Huang, *et al.*, "Bix-nas: Searching efficient bi-directional architecture for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* (M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, eds.), (Cham), pp. 229–238, Springer International Publishing, 2021.

[28] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, eds.), (Cham), pp. 263–273, Springer International Publishing, 2020.

[29] A. Lou, S. Guan, and M. Loew, "Caranet: Context axial reverse attention network for segmentation of small medical objects," *arXiv preprint arXiv:2108.07368*, 2021.

[30] J. M. J. Valanarasu and V. M. Patel, "Unext: Mlp-based rapid medical image segmentation network," *arXiv preprint arXiv:2203.04967*, 2022.

[31] H. Tang, X. Liu, S. Sun, X. Yan, and X. Xie, "Recurrent mask refinement for few-shot medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3918–3928, October 2021.

[32] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2016.

[33] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, *et al.*, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling* (Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, and W. De Neve, eds.), (Cham), pp. 451–462, Springer International Publishing, 2020.

[34] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton, "Lookahead optimizer: k steps forward, 1 step back," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.