Multi-scale in Multi-scale Subtraction Channel Spatial Fusion Network for Real-time Polyp Segmentation

Xin Wang, Zhigang Li*

Abstract—**Colonoscopy is considered the most important technique for detecting polyps and facilitating early screening and prevention of colorectal cancer. In a clinical setting, segmenting polyps from colonoscopy images is of critical importance, as it preserves key diagnostic and surgical information. Although deep learning has achieved significant success in polyp segmentation, existing models often suffer from performance degradation on unknown datasets due to the morphological diversity of polyps. Specifically, traditional architectures make handling multi-scale feature characterization and boundary ambiguity difficult. To address this problem, we propose a novel network architecture called M^{2}CSFormer to solve the challenges in polyp segmentation. The model uses a Pyramid Transformer to establish inter-layer multiscale associations through differentiated sensory fields to obtain rich multiscale disparity information. In addition, channel and spatial attention mechanisms are used to determine the location of polyps efficiently. In addition, we use shape blocks to enhance the edge segmentation accuracy through geometric constraints. According to M^{2}CSFormer experiments on five publicly available datasets, the present method achieves state-of-the-art performance. In cross-domain evaluation, our model achieves an average Dice coefficient of 0.935/0.949 on Kvasir-SEG and CVC-ClinicDB datasets, which is a 4% and 1.3% improvement over PraNet and SSFormer, respectively. The optimized architecture processes 256×256 images at 42 FPS on RTX 3090 GPU with 12.5% faster inference and higher accuracy than SSFormer.**

Index Terms—**colonoscopy, M^{2}CSFormer, multi-scale features, polyp segmentation.**

I. Introduction

COLORECTAL Cancer (CRC) is a significant cause of death worldwide. In the United States, it is the third most common cause of cancer-related fatalities, with approximately 151,030 new cases and 52,580 deaths in 2022[1]. In 2022, CRC was the third most prevalent disease in China, with 592,232 new cases and 309,114 deaths[2]. Currently, endoscopy is the most effective technique for the diagnosis and management of abnormalities; however, there are certain restrictions associated with this procedure. Initially, the effectiveness and precision of the colonoscopy examination can be influenced by the experience, skill, and attention of a specialized physician who is responsible for the operation and judgment. Secondly, colonoscopy is a lengthy, laborious, and costly process that is a burden for both patients and physicians. Third, colonoscopy is associated with the risk of underdiagnosis, misdiagnosis, and overdiagnosis, which may result in superfluous or delayed treatment.

The clinical acumen and experience of the endoscopist, as well as the reliance on manual manipulation, have been highlighted in recent reports, resulting in missed lesions in approximately 26% of colonoscopies[3]. This may result in treatment that is either unnecessary or delayed. The progression of colonoscopic polyp image segmentation can provide clinicians with supplementary diagnostic and decision-making support, improve the efficiency, accuracy, and objectivity of examinations, and promote early identification and management of colorectal carcinoma. Nevertheless, polyp image segmentation continues to encounter certain obstacles and complications[4]. There are three primary reasons why the automatic and precise segmentation of polyps remains a difficult undertaking. Initially, there are substantial variations in polyp size, color, texture, and other characteristics within each cohort. Secondly, there are minor interclass distinctions between polyp lesions and the surrounding tissue components, particularly folds. Lastly, the identification of polyps can be perplexing due to fluctuations in illumination, motion blur, low-contrast areas, and gastrointestinal contents during image acquisition[5]. These factors may exist not just within specific sections of the same polyp but also among various types of polyps, leading to fragmentation stability and unpredictability.

Traditional polyp segmentation methods predominantly depend on a small amount of characteristics, including texture[6], geometric features[7], and basic linear iterative clustering of hyperpixels[8], to address the aforementioned issues and challenges. Unfortunately, these techniques frequently result in subpar segmentation accuracy and restricted generalization capabilities. Deep learning techniques offer a precise and efficient solution for polyp segmentation through their implementation in medical image analysis. Numerous image division models utilizing convolutional neural networks (CNNs) have demonstrated outstanding performance in recent years. The U-shaped topology of U-Net[9] has been adopted as a classical network structure for medical image segmentation. It extracts feature information through convolutional layers in the encoding path and spatial information in the symmetric decoding path. Furthermore, to mitigate the semantic disparity between the encoder and decoder of the U-shaped network and generate satisfactory outcomes, U-Net++[10] and ResUNet++[11] have been implemented in polyp segmentation.

A variety of innovative methods have been suggested by researchers. For instance, ColonSegNet[12], which is comparable to UNet, constructs a lightweight model for

Manuscript received March 14, 2025; revised July 6, 2025.

Xin Wang is a postgraduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning 114051, PR China (e-mail: wangsanj11@163.com).

Zhigang Li is a professor of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning 114051, PR China (*Corresponding author, e-mail: li7275@163.com).

polyp image segmentation by employing skip connections, residual blocks, and transposed convolutions. It achieves the segmentation speed of about 180 FPS on the Kvasir dataset, providing an outstanding basis for actual time polyp segmentation of pictures tasks. To effectively localize polyp boundary regions, PraNet[13] employs a reverse attention mechanism. This is accomplished by incorporating boundary attention blocks and advanced feature aggregation into the network architecture, which aid in the alignment of unaligned predictions and the enhancement of the overall segmentation accuracy. Hardnetmseg[14] is a simplified encoder-decoder architecture that is based on PraNet. In particular, they replace the original Res2Net backbone with Hardnet[15] and eliminate the attention mechanism to facilitate speedier, more precise polyp segmentation. H and eliminate the attention mechanism to facilitate the more precise and rapid segmentation of polyps. The Information Context Enhancement (ICE) technique and the Adaptive Feature Aggregation (AFA) module, as well as the use of edge and structural coherence perceptual loss (ES-CLoss) for training, are introduced by HRENet[16]. This results in exceptional model performance. The primary method by which DeepLabv3+[17] enhances the accuracy is by altering the structure of the encoder-decoder. This modification introduces a decoder module that enables the reconstruction of segmentation results from the underlying features, thereby enhancing the detail and accuracy of the segmentation. The computational and memory footprint of the model is significantly reduced by utilizing depth-separable convolution, which enhances the model's practicality and real-time characteristics.

SANet[18] implements probabilistic correction and color migration strategies to resolve the obstacles of scale imbalance and color distribution that are a result of the objective size. MSNet[19] implements a multiscale subtraction network to mitigate redundancy and complementarity in multi-scale features. In the same vein, MSRFNet[20] implements a cross-scale fusion mechanism to disseminate both high-level and low-level features, as well as a shape-flow network to refine polyp boundaries. TGANet[[21] employs a text-guided approach to assimilate the distinctive characteristics of polyps of varying sizes, with the ultimate objective of improving the network's capacity to generalize across polyps of varying sizes.

The decoder is provided with information on the difference at the pixel and structure level by the basic intralayer multi-scale subtraction unit SU, which is designed by M^2SNet[22]. The method achieves interlayer multiscale feature aggregation and the acquisition of comprehensive multiscale disparity information by providing varying receptive fields to various levels of multiscale SUs. Another model, MCSF-Net, uses a multi-scale channel spatial fusion network[23]. The design suggested effectively fuses multidimensional multiscale features by combining a multiscale union section with positional and channel-focused attention mechanisms. A characteristic enhancement algorithm is also implemented for effectively obtaining outline signals about low-dimensional attributes, thereby ensuring computational complexity and improving segmentation speed. The same issue is resolved by a different approach, Polyp-PVT[24], which employs non-local modules to tacitly adjust the anticipated map with features at the lowest level. Furthermore, the accuracy of polyp segmentation is enhanced by the utilization of the GCN network for feature map closeness merging in this framework. The segmentation prediction maps of polyps are iteratively updated by LDNet[25], which employs a segmentation header that is derived from the general ambient traits of the input image. This is achieved by utilizing the taken-out lesion properties of the polyps.

A bottom-up model architecture was initially proposed in the Natural Language Processing (NLP) community as a Transformer[26]. A Vision Transformer (ViT)[27] was proposed by Dosovitskiy et al to optimize image classification tasks. The transformer adaptively extracts and blends features between all blocks by calculating the dot product between block vectors based on the similarity of all block pairings. This reduces the model's sensing bias and provides the Transformer with a global sense field that is effective. Consequently, Transformer possesses more potent generalization capabilities than multilayer perceptron architectures and CNNs[28]. SSformer[29] suggests a novel method for improving the encoder by enhancing the structure. The methodology comprises the execution of an individual attention mechanism and an ordered character consolidation mechanism that performs local detailed feature processing with efficiency.

Although these techniques exhibit encouraging outcomes in the segmentation of polyp images, they often neglect important factors that involve the intricate nature of the model and the simplicity of deployment in favor of enhancing segmentation accuracy. Precise error bounds enhance the reliability of these networks, especially in safety-critical applications where dependability is essential[30] . The model's capacity to localize polyp boundaries is improved by PraNet's use of reverse attention; however, its capacity to obscure polyp boundaries remains unsatisfactory. Nevertheless, CNNs frequently experience the loss of some critical information during downsampling and have restricted sensory domains, which impedes their capacity to establish global contextual semantic relations. Because of this, conventional CNN methods typically demonstrate restricted generalization capabilities when implemented on polyp images from various patients[31]. Multi-scale feature maps are processed by MCSF-Net to resolve this limitation; however, it fails to consider the semantic information that exists between pixels. TGANet utilizes text-directed attention to concentrate on the characteristics of polyps of varying diameters, there by improving its generalizability. Nevertheless, additional enhancements are required in the local detail segmentation of TGANet.

To resolve these concerns, we suggest the development of a novel polyp segmentation of pictures model, M^2CSFormer, that effectively balances immediate performance with accuracy. Initially, we introduce Transformer v2 (PVTv2) as an image encoder. This model, in contrast to conventional convolutional neural networks, employs a self-attention mechanism that allows it to encompass a broader effective receptive field. The Transformer can better capture global contextual semantic information by learning the dependencies between different positions in the sequence through self-attention for long sequence data. Nevertheless, the Transformer may be unable to effectively extricate local feature information and may also encounter the issue of distraction. Consequently,

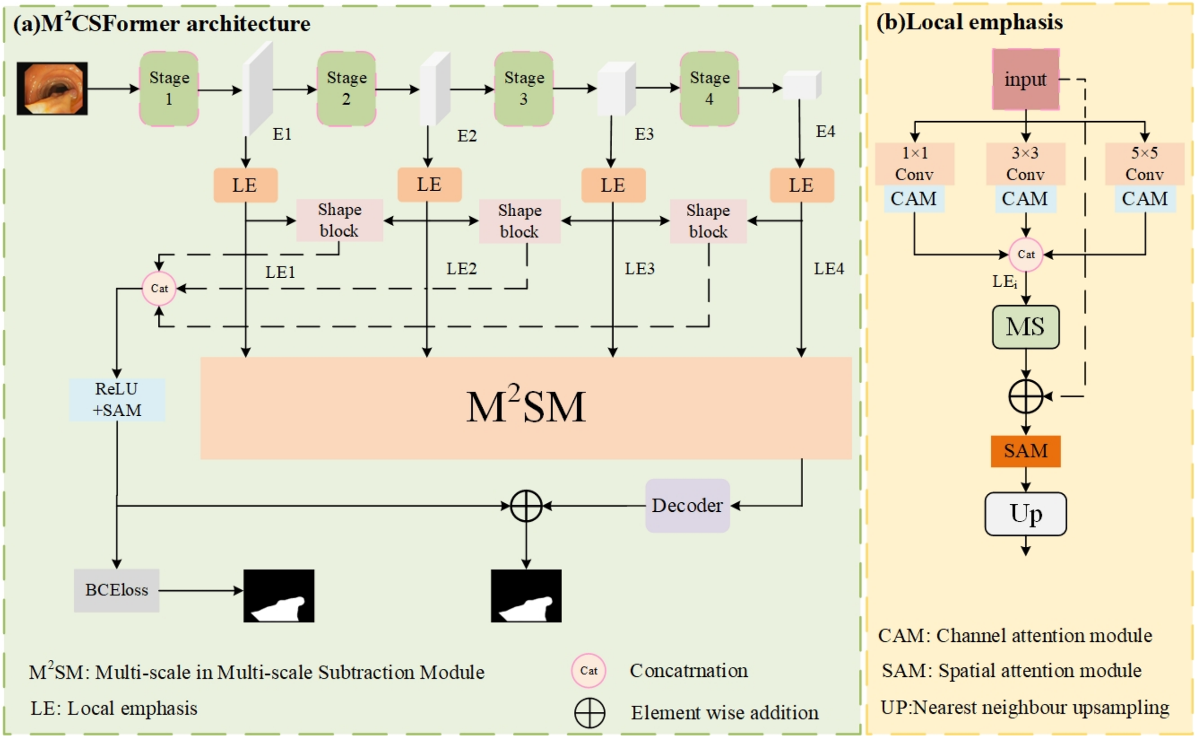

Fig. 1: The proposed M^2CSFormer architectures.

we have developed the Local Emphasis (LE) block, which employs three convolutional kernels of varying sizes, fusion channels, and spatial attention mechanisms to generate a variety of sensory field sizes that can capture features at varying scales and re-condense the disordered attention from new along key details such as contours and boundaries.

In this paper, a multi-scale module in multiscale reduction (M^{2}SM) is developed to enhance the textures and boundaries of polyps. This approach guarantees the precise maintenance of tumor limitations, minimizes the loss of critical information, and enhances the model's separation efficiency. Furthermore, we implemented a straightforward shape block to highlight the form of bounding data within the outcomes of segmentation. The lesion is distinguished from the background by utilizing the difference between various levels of characteristic data in this block, which is controlled by employing BinaryCross EntropyLoss (BCE Loss). M^{2}CSFormer is not just highly efficient but also highly accurate in the segmentation of polyp images. The main findings of this work have been as follows:

(1)We suggest the development of a novel immediate time segmentation method for papilloma pictures, known as M^{2}CSFormer. It comprises LE blocks and M^{2}SM blocks, as well as a further shape block that supervises the form and perimeter features inside the polyps to enhance the model's generalization and localized feature extraction capabilities.

(2)M^{2}CSFormer transfers the extracted multi-scale features, which are abundant in global information, through LE into M^{2}SM. The utilization of this block emphasizes the valuable difference information and eliminates the interference of the sunken sections, thereby enriching the representation of the local features and, in the end, generating a reliable segmented image that offers excellent current performance.

(3)The reliability of M^{2}CSFormer was successfully verified by means of thorough studies on five publicly available datasets: Kvasir-SEG[32], CVC-ClinicDB[33], Endoscene[34], CVC-ColonDB[35], and ETIS[36]. The findings indicate that our proposed M^{2}CSFormer outperforms the majority of the most recent advances in techniques in a variety of metrics used for evaluation.

II. Method

The M^{2}CSFormer introduces a novel framework for segmenting polyps in colonoscopy images by implementing multiscale fusion across channels and dimensions. This framework incorporates a pyramid Transformer encoder known for its strong generalization capabilities, enabling efficient feature extraction and processing with fewer parameters. Leveraging the inherent focus of M^{2}SM on extracting depth information, the model enhances the delineation of edge details and structural cues. Additionally, it integrates channel-attention and spatial-attention[37] mechanisms to further enhance the identification of these features. The overall architecture of M^{2}CSFormer, as depicted in Fig. 1(a), consists of an encoder module, an LE module, an M^{2}SM module, a shape module, and a decoder module. Detailed explanations of the functions and structures of these modules are provided below.

A. *Encoder module*

By employing the Pyramid Vision Transformer v2 (PVTv2) that has been pre-trained by ImageNet, the M^{2}CSFormer functions as an image encoder. The B3 variant of the PVTv2 is employed, and it contains 45.2M parameters. The PVTv2 backbone network extracts the feature maps from four distinct channels, which are denoted as {E1: 64,64×64}, {E2: 128,32×32}, {E3: 320,16×16}, and {E4: 512,8×8}. The data includes the resolution and number of channels.

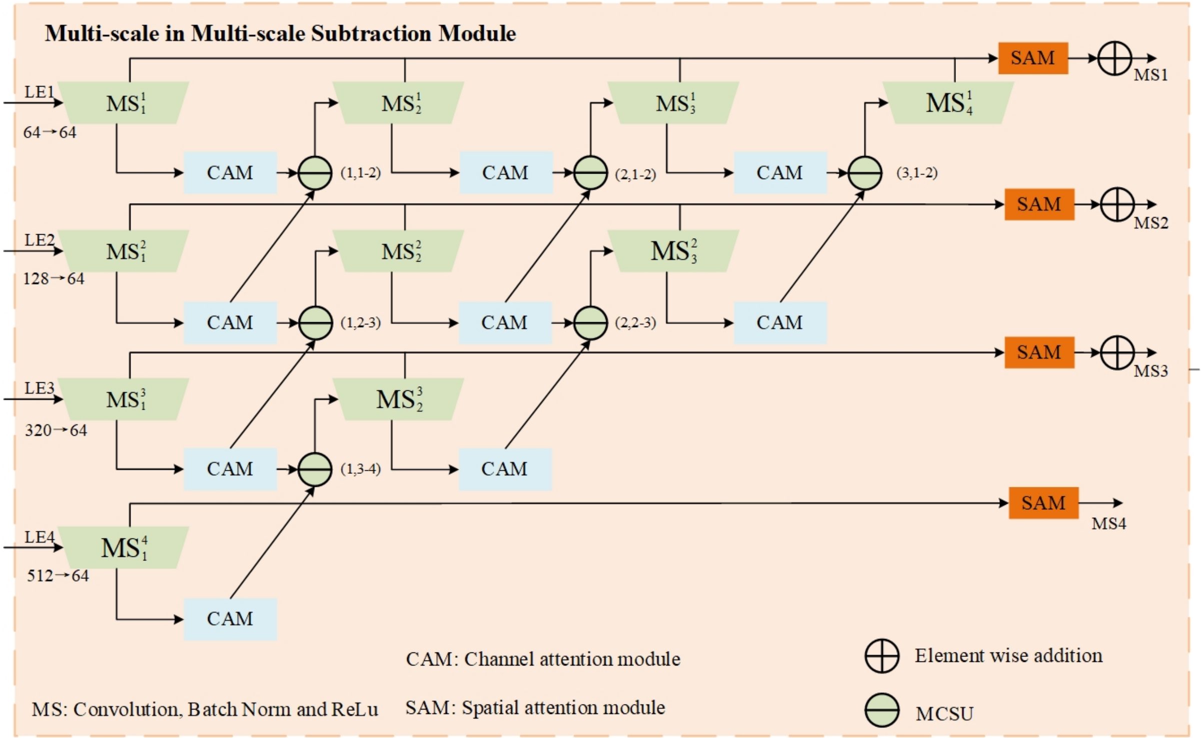

Fig. 2: The details of M^2SM block.

$$MCSU = \text{Conv}(\begin{array}{l} |\text{Filter}(\text{CAM}(F_A))_{1\times1} - \text{Filter}(\text{CAM}(F_B))_{1\times1}|+ \\ |\text{Filter}(\text{CAM}(F_A))_{3\times3} - \text{Filter}(\text{CAM}(F_B))_{3\times3}|+ \\ |\text{Filter}(\text{CAM}(F_A))_{5\times5} - \text{Filter}(\text{CAM}(F_B))_{5\times5}| \end{array}). \tag{1}$$

In order to effectively model the global context, PVTv2 possesses a multi-scale feature processing capability and a potent global receptive field. Furthermore, PVTv2 enhances parameter efficiency and can achieve comparable accuracy to other larger-scale models while retaining a reduced number of parameters. In particular, we employ features LE1-LE4 to improve the feature representation of M^{2}SM. This enhances the complementarity between various levels of features, allowing for more precise localization and the identification of tumor borders.

B. *M^{2}SM and LE block*

The characteristics of level features are weakened, resulting in the generation of redundant information and the failure to balance accurate localization and subtle boundary refinement. Typically, features at different levels contain rich local features, but the different information between different levels is not given the same level of attention. Consequently, we introduce M^{2}SM to enhance the representation of localized features by highlighting valuable difference information and eliminating the interference of sunken redundancy. The attention-enhanced features LE1, LE2, LE3, and LE4 are transmitted into M^{2}SM, as illustrated in the Fig. 2. MS represents the 3 × 3 convolution functioning, which follows routine normalization and ReLU, and CAM represents the channel of attentive mechanism. Initially, this module will be employed to perform feature mapping for each encoder block individually, thereby reducing the number of channels to 64 and, as a result, the number of parameters required for subsequent operations. Next, the features of adjacent layers will be processed by a potent intra-layer multiscale subtract space unit ($MCSU$). The feature mapping of neighboring layers is represented using F_A and F_B, as illustrated in the Fig. 3(b). This allows the $MCSU$ to be represented as follows Eq.(1).

Where $Filter(\cdot)_{n\times n}$ denotes a comprehensive one-dimensional filter of size n × n. We employ multi-scale convolutional filters with fixed all-1 weights of sizes 1 × 1, 3 × 3, and 5 × 5 to calculate detail and structural differences based on pixel-pixel and region-region patterns. Subsequently, the attention map is inferred together each of the channel lengths using a channel attention module. To automatically improve the significance of attributes, focus mapping are combined with feature maps. The MCSU can then capture the complementary information of $CAM(F_A)$ and $CAM(F_B)$ and emphasize the differences between them in terms of texture and structure, thereby providing the decoder with more comprehensive information. We compute a series of differential features with different orders and receptive fields by horizontally and vertically connecting multiple $MCSUs$ in order to obtain higher-order complementary information at multiple feature levels.

F_A and F_B are feature maps that match distinct widths and sizes. The primary components of narrow characteristics are localized information, including colors, textures, outlines, and edges. Nevertheless, the field of reception widens as the depth of the network increases, allowing the system to gather deeper features in the image, such as things, situa-

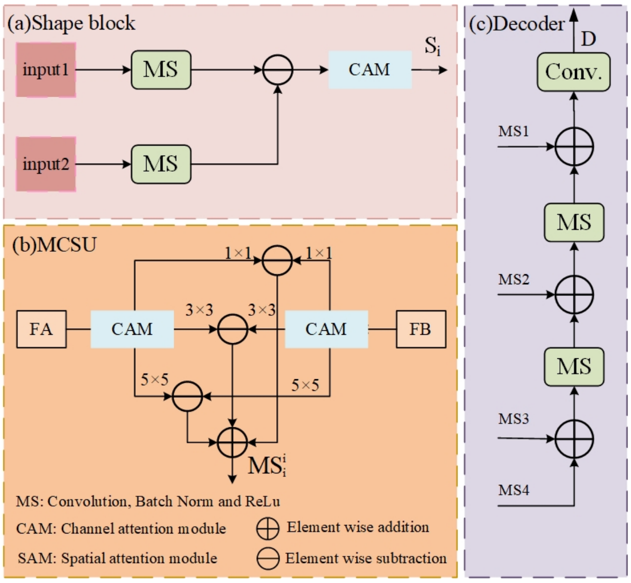

Fig. 3: The details of shape block, MCSU and decoder.

tions, and geographic data. Consequently, we calculate the discrepancies among the characteristics of the maps that were extracted at the proximal point in the network's structure. The separation of material that exists between colonies and the surroundings is emphasized by optimizing these distinctions employing multichannel concentration.

Fig. 2 contains the multiscale details of the multiscale reduction mechanism. We derive complementary enhancement characteristics (MSi) by combining scale-related features (MS_1^i) and cross-scale difference measures (MS_n^i) among the level in question and any other level. This procedure may be articulated as follows:

$$MSi = \text{Conv}(\left(\sum_{n=1}^{6-i} SAM(MS_n^i)\right), \quad i = 1,2,3,4). \quad (2)$$

SAM constitutes a spatially targeted attention generator that can infer interest models in the spatial domain by utilizing channel concentration with an emphasis on key feature mapping channels. The spatial geographical data of the mid-polyp is obtained by utilizing spatial attention between the same layers. Ultimately, the MSi n is supplemented with the feature disparity index through an additive fusion process. The disparity information between feature maps at different resolutions was captured and the critical information of feature maps at different scales was merged. This method enables the exchange of different information and produces more informative and accurate polyp feature maps to assist the decoder in predicting the final segmentation maps.

The LE module's detailed structure is illustrated in Fig. 1(b). The LE module is fed the features, and three distinct convolutional kernels are employed to enhance the local receptive fields. This boosts the macro-weights in the blocks surrounding the request, thereby refocusing the focus on the adjacent aspects and reducing the distraction. In the following manner, the LE is represented:

$$LE_i = Cat(CAM(\{Conv(input)_r, r = 1,3,5\})). \quad (3)$$

Where LE_i, $Conv$ denotes the temporal inversion the kernel r denotes the magnitude of the compression kernel, and $input$ denotes the features extracted via the encoder at varying scales of distraction. The information is combined utilizing a cascade functioning to enable the disordered attention to be attracted by the polyp location, denoted as LEi, after going through three different scales of convolution operations and utilizing the channel attention to focus on the important feature channels. This process is capable of efficiently removing clutter noise and emphasizing the critical local features. The disordered attention is recondensed around critical details, including outlines and limits, following the element streaming goes by the LE. Afterward, spatial attention is employed to further emphasize the precise location of the polyp, and additive fusion of LE and input is conducted so that all of the feature data is extracted from the encoder. LEi may be expressed in the following way, as illustrated in Fig. 1(b):

$$\text{LEi} = \{UP(SAM(\text{LEi} + \text{input})), i = 1,2,3,4\}. \quad (4)$$

C. shape block and decoder

A shape block is constructed in Fig. 3(a) to derive information about shapes to elements at varying dimensions. Characteristic reduction procedures take place across LE1 to LE4 to generate feature contrast vectors at varying scales, similar to the concept of multi-scale subtraction. The information about the local features of the polyp is distinguished from the background by these matrices. Following this, the various feature representations undergo processing using channel focus to emphasize the most significant channels in the characteristic maps. Then, a stitching procedure is conducted on all of the map features, and spatial focus is utilized to extract detailed positional information about the boundaries. This fusion facilitates the features extracting through channel and spatial attention strategy, augmenting the capacity to identify subtle details[38]. This leads to feature maps that are replete with detailed information regarding the morphologies and boundaries of polyps. The feature difference map may be depicted as follows:

$$S_i = \{CAM(|MS(E_i) - MS(E_{i-1})|), i = 4,3,2\}. \quad (5)$$

The shape boundary data of the polyps is supervised using the BCE loss. The resulting shape information is then combined with the decoder module's output to produce a more precise tumor split image.

The step-by-step summation operation is employed in the decoder in Fig. 3(c). This operation is basic and effective, and it effectively integrates the differential features extracted from each layer, thereby significantly reducing the computational expenses of the resulting model. The final result of encoder D is displayed as follows:

$$D = \text{Conv}(MS1 + MS(MS2 + MS(MS3 + MS4))). \quad (6)$$

D. Loss function

The application of the binary cross entropy (BCE) loss and dice loss is a component of our model supervision approach[39]. BCEloss is a loss function that is specifically

TABLE I: Datasets used in this study.

Dataset	Images	Input size	Train	Valid	Test
Kvasir-SEG	1000	Variable	800	100	100
CVC-ClinicDB	612	384 × 288	488	62	62
Endoscnene	912	574 × 500	—	—	60
CVC-ColonDB	380	574 × 500	—	—	380
ETIS	196	1225 × 966	—	—	19

designed to quantify the discrepancy between the objective and anticipated results in binary categorization tasks. Its calculation is as follows:

$$L_{BCE} = -w \left(Y \cdot \log(X) + (1 - Y) \cdot \log(1 - X) \right). \quad (7)$$

Where X represents the model's forecast value, Y represents the designation of the reality value, and w is the weighting value, which has a standard setting of 1.

The dice loss function was introduced to resolve the issue of a disparity between both negative and positive samples in the collected data. The BCE loss function only would generate a model that predominantly predicts every single type, as the colonoscopic polyp dataset's collecting approach is restricted and the positive and negative samples frequently exhibit substantial disparities. The dice loss function was employed to quantify the degree of resemblance among both examples, with scores extending from 0 to 1. Higher numbers suggest a greater degree of similarity between the data points. The precision and reliability of the cutoff are enhanced by the complementary use of the BCE loss and the dice loss. The dice loss algorithm works as follows:

$$L_{\text{Dice}} = 1 - \frac{2|X \cap Y|}{|X| + |Y|}. \quad (8)$$

Therefore, the last loss function employed for the decoder outcome is as follows:

$$Loss = L_{BCE} + L_{Dice}. \quad (9)$$

III. Experiments

A. Datasets

Kvasir-SEG, CVC-ClinicDB, Endoscene, CVC-ColonDB, and ETIS were the five openly accessible colonoscopic tumor databases on which the recommended M^{2}CSFormer model was assessed. The following is a comprehensive description of the five public datasets.

(1) Kvasir-SEG: The Kvasir-SEG dataset comprises 1000 coral pictures and their associated comments. The present set is distinguished from the others by the variability of the polyps' size and morphology in the images. The dimensions of the pictures vary from 332 × 487 to 1920 × 1072. 48 tiny colonies fewer than 64 × 64, 700 giant colonies larger compared to 160 × 160, and 323 medium in size colonies are included in the dataset. 900 images were utilized for training and validation, while 100 images were employed for assessment.

(2)CVC-ClinicDB: The dataset CVC-ClinicDB comprises 612 pictures taken from 25 endoscopy films, from which 29 segments were chosen. The dimensions of the image are 384 × 288. 550 images are employed for validation and training purposes, while 62 images are employed for assessment.

TABLE II: Quantitative results on Kvasir-SEG datasets.

Method	Backbone	mDice	mIoU	Recall	Precision
Dataset: Kvasir-SEG					
U-Net[9]	–	0.821	0.753	0.816	0.895
U-Net++[10]	–	0.832	0.767	0.862	0.896
DeepLabV3+[17]	Xception	0.891	0.832	0.887	0.917
PraNet[13]	Res2Net50	0.899	0.849	0.897	0.922
SANet[18]	Res2Net50	0.905	0.852	0.897	0.928
TGANet[21]	Res2Net50	0.900	0.843	0.895	0.932
MCSF-Net[23]	ResNet101	0.911	0.861	0.908	0.936
SSFormer[29]	Transformer	0.923	0.867	0.915	0.935
M^{2}CSFormer(Ours)	Transformer	**0.935**	**0.873**	**0.926**	**0.940**

TABLE III: Quantitative results on CVC-ClinicDB datasets.

Method	Backbone	mDice	mIoU	Recall	Precision
Dataset: CVC-ClinicDB					
U-Net[9]	–	0.837	0.786	0.861	0.889
U-Net++[10]	–	0.850	0.807	0.904	0.884
DeepLabV3+[17]	Xception	0.891	0.843	0.893	0.924
PraNet[13]	Res2Net50	0.898	0.854	0.911	0.890
SANet[18]	Res2Net50	0.915	0.862	0.933	0.915
TGANet[21]	Res2Net50	0.926	0.874	0.936	0.922
MCSF-Net[23]	ResNet101	0.941	**0.895**	**0.956**	0.932
SSFormer[29]	Transformer	0.932	0.870	0.944	0.942
M^{2}CSFormer(Ours)	Transformer	**0.949**	0.893	0.949	**0.950**

(3)Endoscene: The Endoscene dataset comprises 912 pictures taken from 44 endoscopic segments of 36 individuals. We employed CVC-300 as the test set, which comprises an overall 60 pictures, as the Endoscene dataset is an amalgam of CVC-ClinicDB and CVC-300.

(4)CVC-ColonDB: A total of 380 images were acquired from 15 distinct colorectal segments. Testing was conducted on all 380 pictures.

(5)ETIS: The ETIS dataset comprises 196 images that were gathered from 34 endoscopy recordings. The dimensions of the image are 1225 × 966. This dataset presents a significant challenge due to the fact that the tumor forms are more varied than those within the remaining datasets, and the majority of them are tiny and tricky to identify. Testing was conducted on all 196 images in the data set.

The initial training configuration was identical to that of PraNet, and 80% of the pictures from Kvasir-SEG and CVC-ClinicDB were selected at random for training objectives. Furthermore, 10% of the images were employed for validation purposes. The remaining 10% of the images, as well as Endoscene, CVC-ColonDB, and ETIS, were utilized for testing. Table I illustrates the precise data division.

TABLE IV: Quantitative results on endoscene datasets.

Method	Backbone	mDice	mIoU	Recall	Precision
Dataset: Endoscopy					
U-Net[9]	—	0.709	0.630	0.709	0.878
U-Net++[10]	—	0.761	0.691	0.784	0.861
DeepLabV3+[17]	Xception	0.862	0.789	0.925	0.850
PraNet[13]	Res2Net50	0.868	0.796	0.903	**0.882**
SANet[18]	Res2Net50	0.879	0.809	0.936	0.851
TGANet[21]	Res2Net50	0.872	0.798	0.963	0.820
MCSF-Net[23]	ResNet101	0.901	0.834	**0.966**	0.859
SSFormer[29]	Transformer	0.895	0.838	0.960	0.862
M^{2}CSFormer(Ours)	Transformer	**0.910**	**0.851**	0.959	0.878

TABLE V: Quantitative results on CVC-ColonDB datasets.

Method	Backbone	mDice	mIoU	Recall	Precision
Dataset: CVC-ColonDB					
U-Net[9]	—	0.629	0.547	0.654	0.804
U-Net++[10]	—	0.628	0.567	0.720	0.726
DeepLabV3+[17]	Xception	0.732	0.651	0.757	0.824
PraNet[13]	Res2Net50	0.676	0.610	0.676	0.754
SANet[18]	Res2Net50	0.745	0.680	0.775	0.845
TGANet[21]	Res2Net50	0.752	0.674	0.790	0.790
MCSF-Net[23]	ResNet101	0.765	0.692	0.782	**0.848**
SSFormer[29]	Transformer	0.788	0.708	0.820	0.828
M^{2}CSFormer(Ours)	Transformer	**0.812**	**0.890**	**0.831**	0.837

TABLE VI: Quantitative results on ETIS datasets.

Method	Backbone	mDice	mIoU	Recall	Precision
Dataset: ETIS					
U-Net[9]	—	0.629	0.547	0.654	0.804
U-Net++[10]	—	0.628	0.567	0.720	0.726
DeepLabV3+[17]	Xception	0.732	0.651	0.757	0.824
PraNet[13]	Res2Net50	0.676	0.610	0.676	0.754
SANet[18]	Res2Net50	0.745	0.680	0.775	0.845
TGANet[21]	Res2Net50	0.752	0.674	0.790	**0.790**
MCSF-Net[23]	ResNet101	0.765	0.690	0.831	0.756
SSFormer[29]	Transformer	0.770	0.685	**0.851**	0.763
M^{2}CSFormer(Ours)	Transformer	**0.781**	**0.702**	0.841	0.771

B. Evaluation metrics and implementation details

The model was constructed with PyTorch and accelerated with an NVIDIA RTX3090 GPU. The AdamW optimizer was employed with a batch size of 16 and an initial learning rate of 0.0001. A grand total of 300 periods of training were conducted. Our loss function is a hybrid of BCE loss and dice loss. We supplement the data with arbitrary vertical and horizontal turns, rotations, and cutting operations, and resize the pictures to 256 × 256 during the process of training. We employ several standard metrics for quantitative evaluation, including recall, precision, average IoU, FPS, and average dice.

C. Experiments on the public polyp benchmarks

TIn recent years, a variety of methods have been applied to the adenoma division, such as UNet, UNet++, DeepLabV 3+, PraNet, SANet, TGANet, MCSF-Net, and SSFormer. We conducted a comparison between the M^{2}CSFormer and these methods.

D. Quantitative analysis

Tables II-IX present the efficacy evaluation of each method across various metrics. The results in Tables II-IX were derived by reapplying the openly accessible algorithms of these methods or algorithms utilizing the same dataset segments as our methods. For clarity, we have highlighted the most favorable results for each metric. Our M^{2}CSFormer demonstrates superior performance across the majority of efficiency metrics compared to the other methods.

The findings of M^{2}CSFormer along with additional methods applied to the Kvasir-SEG and ClinicDB datasets are presented in Tables II and III. M^{2}CSFormer obtains 0.935 mDice and 0.873 IoU with the Kvasir-SEG dataset, which are 1.2% and 0.6% larger than SSFormer concerning of mDice and mIoU, accordingly. Furthermore, in comparison to MCSF-Net, M^{2}CSFormer has a 2.4% and 1.2% greater cost and mIoU, separately. In the ClinicDB dataset, M^{2}CSFormer obtains the second-best mIoU and increases mice by 0.8%, surpassing the most competitive MCSF-Net. M^{2}CSFormer's mDice and mIoU numbers are 2.3% as well as 1.9% higher than those of TGANet, as well.

TABLE VII: The average speed of different methods (FPS).

Method	Average speed	Method	Average speed
DeepLabV3+	62	MCSF-Net	45
SSFormer	48	TGANet	34
PraNet	40	M^{2}CSFormer(Ours)	42

Our approach demonstrates superior precision in tumor separation compared to existing methods, as evidenced by the presented results. This advantage can be attributed to the integration of M^{2}SM and LE blocks within the M^{2}CSFormer architecture. The incorporation of the LE block effectively highlights polyp boundaries, enhancing the representation of diverse feature levels. To improve the accuracy of feature maps, we have introduced channel attention and spatial attention mechanisms to emphasize crucial channel characteristics and tumor localization details. Additionally, the M^{2}SM enhances the representation of local features by mitigating the impact of recessed regions and emphasizing significant differentiation information.

E. In the performance tests on the unseen dataset

This test demonstrates our model's capacity for accurate prediction and generalization on uncertain datasets. The efficacy of neural network-based segmentation methods in clinical applications may be limited by variations in image acquisition systems and individual cases. Comparative results for various methodologies are presented in Tables IV-VI. On the Endoscene dataset, M^{2}CSFormer outperforms the highly competitive SSFormer, increasing mDice and mIoU by 1.5% and 1.3%, respectively. Furthermore, it surpasses MCSF-Net with improvements of 0.9% in mDice and 1.7% in mIoU. Our method also exhibits superior generalization on the CVC-ColonDB dataset compared to other approaches. Specifically, against SSFormer, it achieves gains of 2.4% in mDice and 1.8% in mIoU.

The ETIS dataset poses significant challenges due to the morphological distinctiveness of the majority of parasite images, rendering them imperceptible to the model. Our M^{2}CSFormer model yielded the highest mDice and mIoU values, achieving 0.781 and 0.702, respectively, as shown in Table VI. The MCSF-Net model demonstrated the highest precision levels. Compared to alternative methodologies, the M^{2}CSFormer model exhibited notable enhancements in segmentation outcomes when dealing with imperceptible datasets. This improvement can be attributed to the encoder module in the M^{2}CSFormer model, which integrates a Transformer mechanism capable of robust global context comprehension and multi-scale feature processing to effectively capture global contextual information. Additionally, the Shape block within the M^{2}CSFormer model enhances the discrimination of colonies from the background by encoding details regarding the shape of polyp boundaries, thereby enhancing precision in the final fragmentation prediction.

Table VII presents the frames per second (FPS) evaluation of various competing techniques. All models underwent eval-

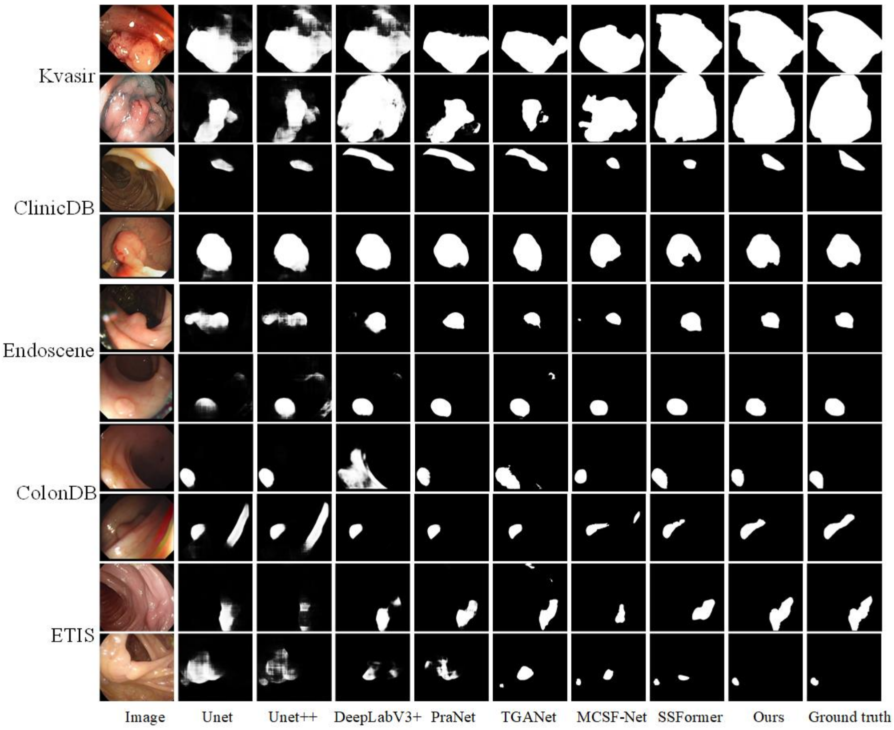

Fig. 4: Visual comparison of polyp segmentation results.

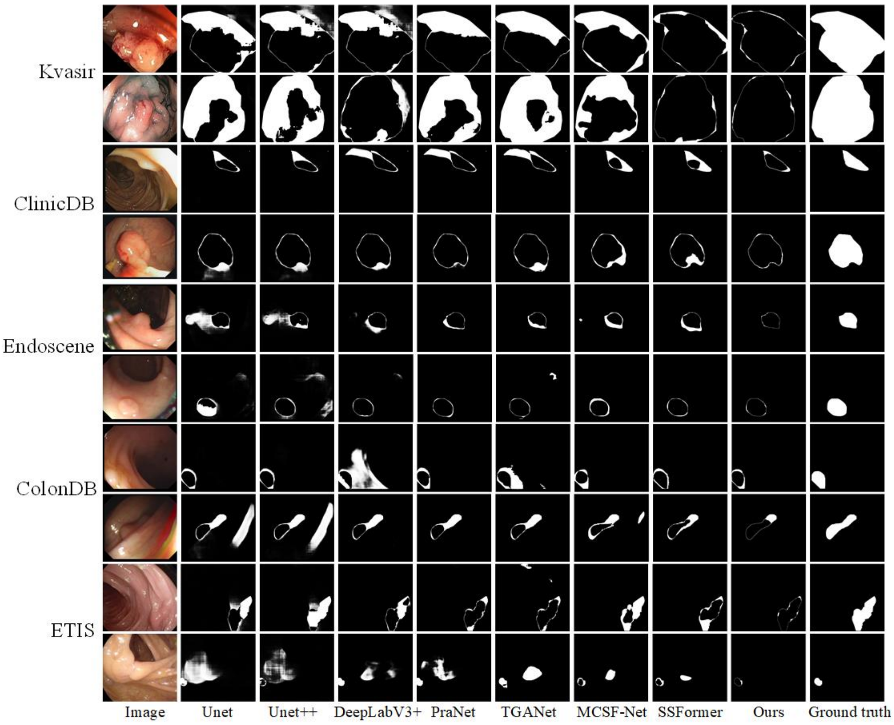

Fig. 5: Visual comparison of differences between results and labels for different methods.

TABLE VIII: Comparison of cross-dataset segmentation results.

Method	Backbone	mDice	mIoU	Recall	Precision
Training dataset: Kvasir-SEG Test dataset: CVC-ClinioDB					
U-Net[9]	—	0.647	0.628	0.701	0.785
U-Net++[10]	—	0.656	0.632	0.724	0.801
DeepLabV3+[17]	Xception	0.803	0.746	0.827	0.872
PraNet[13]	Res2Net50	0.815	0.732	0.814	0.883
SANet[18]	Res2Net50	0.846	0.780	0.861	0.840
TGANet[21]	Res2Net50	0.821	0.754	0.831	0.862
MCSF-Net[23]	ResNet101	0.865	0.792	0.887	**0.887**
SSFormer[29]	Transformer	0.831	0.760	0.845	0.879
M^{2}CSFormer(Ours)	Transformer	**0.878**	**0.804**	**0.905**	0.874

TABLE IX: Comparison of cross-dataset segmentation results.

Method	Backbone	mDice	mIoU	Recall	Precision
Training dataset:CVC-ClinioDB Test dataset:Kvasir-SEG					
U-Net[9]	—	0.464	0.549	0.533	0.655
U-Net++[10]	—	0.593	0.670	0.780	0.659
DeepLabV3+[17]	Xception	0.667	0.741	0.765	0.800
PraNet[13]	Res2Net50	0.712	0.640	0.627	0.567
SANet[18]	Res2Net50	0.724	0.804	0.812	0.851
TGANet[21]	Res2Net50	0.818	0.882	0.937	0.852
MCSF-Net[23]	ResNet101	0.826	0.890	**0.949**	0.865
SSFormer[29]	Transformer	0.802	0.721	0.796	0.720
M^{2}CSFormer(Ours)	Transformer	**0.840**	**0.892**	0.907	**0.909**

uation under identical experimental conditions. The methods were tested in precisely the same experimental setup. It is apparent that our M^{2}CSFormer outperforms the other techniques in terms of real-time performance, segmentation accuracy, and scalability.

F. Generalization ability

Due to the colon polyp segmentation task requiring the model to possess both accurate prediction and strong generalization capabilities, it is essential to separately evaluate its performance on experimental and unseen benchmark datasets. To this end, three distinct training-test configurations were employed to conduct cross-dataset evaluations of the model's prediction accuracy and learning capabilities. In the first experiment, the generalization ability of the M^{2}CSFormer was assessed by testing a model trained on CVC-ClinicDB against the Kvasir-SEG dataset. As presented in Table VIII, the test results indicate that M^{2}CSFormer significantly outperforms other methods. Specifically, M^{2}CSFormer's mDice and mIoU scores are 1.5% higher than those of the most competitive MCSF-Net. In the second experiment, the CVC-ClinicDB dataset was exclusively used for model training, followed by testing on the entire Kvasir-SEG dataset. As shown in Table IX, our model achieved superior performance under this testing scheme, with M^{2}CSFormer's mDice and mIoU scores being 1.7% and 0.2% higher than those of the most competitive MCSF-Net, respectively. In the third experiment, the CVC-ClinicDB dataset was again used solely for model training, and the model was subsequently tested on the entire CVC-ColonDB dataset. As demonstrated in Table X, M^{2}CSFormer's Recall score is 0.5% higher than that of the most competitive MCSF-Net. Through the results of these three experiments, we have conclusively demonstrated that M^{2}CSFormer exhibits robust generalization capabilities and high prediction accuracy.

TABLE X: Comparison of cross-dataset segmentation results.

Method	Backbone	mDice	mIoU	Recall	Precision
Training dataset:CVC-ClinioDB Test dataset:CVC-ColonDB					
U-Net[9]	—	0.334	0.409	0.422	0.545
U-Net++[10]	—	0.353	0.447	0.357	0.596
DeepLabV3+[17]	Xception	0.650	0.761	0.652	0.780
PraNet[13]	Res2Net50	0.738	0.646	0.752	0.831
SANet[18]	Res2Net50	0.724	0.804	0.812	0.851
TGANet[21]	Res2Net50	0.804	0.831	0.925	0.884
MCSF-Net[23]	ResNet101	**0.815**	**0.887**	0.918	**0.911**
SSFormer[29]	Transformer	0.797	0.869	0.784	0.782
M^{2}CSFormer(Ours)	Transformer	0.810	0.871	**0.923**	0.905

G. Qualitative analysis

The segmentation outcomes of all methods are presented in Fig. 4, while Fig. 5 illustrates the disparities between the resulting segmentations from each method and the ground truth labels. In Fig. 5, white pixels indicate areas where the generated output differs from the true labels, with a higher concentration of black pixels indicating closer alignment with the actual labels. M^{2}CSFormer shows notable improvements in segmenting polyp samples of various sizes when utilizing LE blocks, M^{2}SM, and Shape blocks, as depicted in Fig. 5. Specifically, while other approaches tend to oversegment large polyp samples in the Kvasir-SEG dataset, M^{2}CSFormer accurately delineates polyp boundaries in images containing large polyps.

Alternative methods in ClinicDB are less effective for segmenting massive and small to medium-sized tumors due to inaccurate segmentation and unclear boundaries. In contrast, our approach excels in accurately identifying and efficiently utilizing diverse polyps with varying features. The M^{2}CSFormer exhibits robust segmentation capabilities for capturing structured data in Endoscene and reliably dividing small and medium-sized tumor samples through shape blocks. Its robust detection capabilities are evident in outperforming other methods in ColonDB and ETIS datasets, particularly in accurately segmenting extremely small polyps. By leveraging the LE block and M^{2}SM, the M^{2}CSFormer accurately locates colonies and distinguishes them from the background, thereby preventing polyp omission and image over-segmentation.

Fig. 6 illustrates a case of segmentation failure in the M^{2}CSFormer model. The first row demonstrates the model's tendency to overlook important regions of the tumor when the features in the image closely resemble the surrounding folds. Severe distortions or reflections, as seen in the second and third panels, notably hinder the model's ability to accurately identify polyp areas. In the fourth row, the model incorrectly identifies normal tissue due to the resemblance between the polyp and the prominent surrounding striations.

H. Ablation study

In order to verify the efficacy of M^{2}SM, LE, and shape block, we conducted erasure operations on the Kvasir-SEG and CVC-ClinicDB datasets to examine the role of every

TABLE XI: Ablation study for M^{2}CSFormer on the Kvasir and CVC-ClinicDB datasets.

Experiment description	Kvasir-SEG		CVC-ClinicDB	
	mDice	MIoU	mDice	MIoU
M^{2}CSFormer	0.935	0.873	0.949	0.893
Without M^{2}SM	0.911	0.855	0.925	0.877
Without LE	0.919	0.861	0.931	0.886
Without shape block	0.923	0.865	0.938	0.880

element in the M^{2}CSFormer. We trained the model and watched its impact on its performance by systematically removing these blocks from the M^{2}CSFormer while keeping the others intact. Table XI displays the quantitative outcomes of our ablation studies.

Initially, we'll eliminate an M^{2}SM module, which led to a 2.4% and 1.8% reduction in mDice and mIoU scores within the Kvasir-SEG dataset, correspondingly. In the same vein, the mDice and mIoU scores on the CVC-ClinicDB dataset declined by 2.4% and 1.6%, accordingly. These results indicate that M^{2}SM integrates disparate information from feature maps of varying decisions, and the model's classification effectiveness is typically enhanced by the presence of rich local features in various layers of features. M^{2}SM efficiently minimizes the parameter case of the operation, improves the representation of feature information, and accurately integrates the local feature information of polyps.

Consequently, we eliminated the LE blocks and incorporated the encoder-extracted features into M^{2}SM. It is evident that the mDice and mIoU scores experienced a substantial decline in the Kvasir-SEG and CVC-ClinicDB datasets. In the Kvasir-SEG dataset, the mDice and mIoU scores decreased by 1.6% and 1.2%, respectively, whereas these values declined by 1.8% and 0.7% on the CVC-ClinicDB dataset.

It is evident from Table XI that the model's performance is also impacted by the removal of the shape block, which is responsible for maintaining high accuracy when interacting with data from invisible sources. In the Kvasir-SEG dataset, the mDice and mIoU scores declined by 1.2% and 0.8%, respectively, whereas in the CVC-ClinicDB dataset, they dropped by 1.1% and 1.3%.

The heatmap that shows all of the properties of our proposal block both before and after the insertion of every element is depicted in Fig. 7. It is evident that the polyp region is the primary focus of attention following the LE block. Inclusion of the M^2 SM block improves the representation of local features, decreases the influence of other chaotic information, and more clearly captures boundary information. The polyp boundary information signals are captured by the shape block, which aids the model in generating a more precise division map.

IV. Discussion and conclusion

Challenges in segmenting colonies in images include unclear boundaries between colonies and tissue, numerous anatomical variations, and the tumor's close resemblance to the background color. Machine learning models often result in either over-segmentation or under-segmentation. U-Net ++, an enhanced version of U-Net, emphasizes short-range connections over long-range connections. While it integrates features from multiple levels, it struggles to establish semantic relationships between pixels, leading to the exclusion of multiple polyps. PraNet, on the other hand, utilizes an attention mechanism to indirectly infer contour cues. However, it lags behind Falls in accurately capturing fine division details and effectively handling the segmentation of multiple polyps. Despite its ability to improve polyp segmentation accuracy, PraNet is susceptible to missing targets when confronted with numerous polyps.

In contrast, TGANet utilizes text-guided attention to address the challenge of detecting polyps with varying sizes and quantities. This approach enables the network to capture additional features that aid in distinguishing between polyps of different diameters. However, TGANet's ability to extract detailed features at the pixel level is limited by its lack of consideration for the semantic correlations between pixels. In contrast, M^{2}CSFormer differs from TGANet by leveraging Transformer to capture global dependencies within an image. It demonstrates superior generalization and learning capabilities and incorporates the LE module to mitigate distractions, thereby refocusing attention on crucial details such as contours and boundaries. By employing M^{2}SM to extract valuable disparity information and eliminating background noise, the model enhances the representation of local features. Additionally, it integrates channel and spatial attention mechanisms along with a multi-scale merging approach to enhance feature generation.

Our M^{2}CSFormer model is designed to mitigate incomplete feature loss in polyp data resulting from variations in polyp dimensions and shapes by incorporating shape blocks. These shape blocks enhance the model's ability to generalize across polyps with diverse features by capturing unique boundary information signals specific to polyps. The improved blocks, as detailed in Tables II-X, demonstrate the superior efficiency of M^{2}CSFormer compared to contemporary techniques on various datasets. Notably, the model achieves the highest mDice scores on the Kvasir and CVC-ClinicDB datasets.

Furthermore, M^{2}CSFormer demonstrated superior performance in terms of mDice and mIoU scores across the Endoscene, CVC-ColonDB, and ETIS datasets. To assess the model's predictive accuracy and generalizability, we conducted three distinct experiments. Initially, we trained the model on the complete Kavsir-SEG dataset and evaluated its performance on the entire CVC-ClinicDB dataset. The outcomes, detailed in Table VIII, underscored M^{2}CSFormer's top-ranking mDice and mIoU scores. Subsequently, as illustrated in Table IX, training on the full CVC-ClinicDB dataset and testing on the complete Kavsir-SEG dataset reaffirmed M^{2}CSFormer's superior performance in terms of mDice and mIoU scores. Lastly, when the model was exclusively trained on the CVC-ClinicDB dataset and assessed on the entire CVC-ColonDB dataset, M^{2}CSFormer exhibited a superior Recall score compared to the leading competitor, MCSF-Net, as indicated in Table X. Overall, M^{2}CSFormer demonstrates excellent robustness when processing images from various sources, consistently exhibiting strong segmentation capabilities and superior generalization ability. These results lay a solid foundation for its application in real-world medical scenarios.

Table VII displays an analysis of FPS (frames per sec-

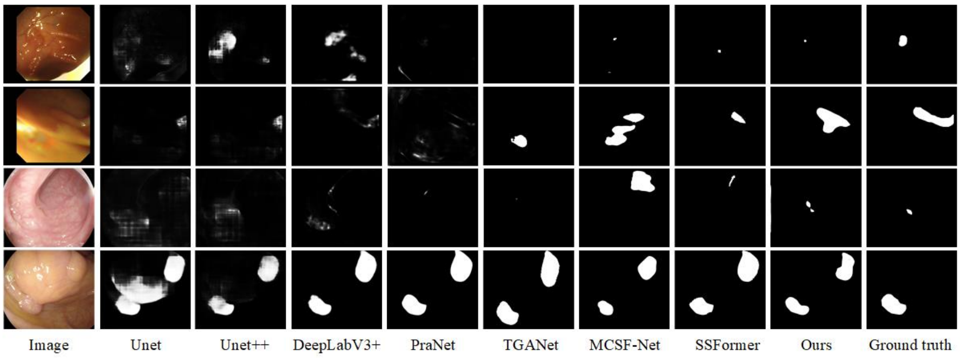

Fig. 6: Sample of segmentation failures.

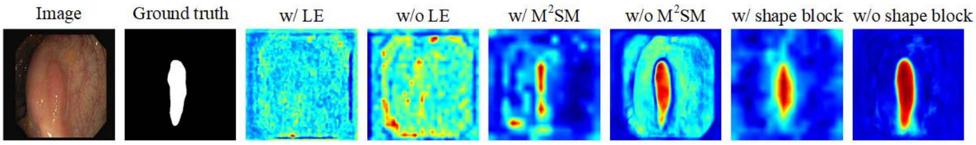

Fig. 7: The feature heatmap of different modules.

ond) metrics for various methodologies implemented under identical experimental conditions. The results indicate that M^{2}CSFormer demonstrates exceptional segmentation precision, excellent generalization capabilities, and competitive FPS values. Comparative visualization of classification graphs generated by different segmentation methods can be found in Fig. 4 and 5. M^{2}CSFormer consistently outperforms other methods across all evaluated datasets, providing enhanced segmented maps for polyp cases of varying sizes. Notably, M^{2}CSFormer accurately segments the majority of polyps with different diameters, in contrast to other methods that often exhibit issues such as over-segmentation or under-segmentation, thereby compromising the overall accuracy of localization maps.

Several instances of segmentation failure by M^{2}CSFormer are demonstrated in Fig. 6. The validation framework of this study imposes limitations, warranting further clinical investigations to validate the technique's efficacy in practical settings. Moreover, M^{2}CSFormer's limited imaging capabilities lead to compromised image quality under specific environmental conditions, as depicted in Fig. 6. Addressing this challenge may involve integrating specialized imaging strategies and refining data enhancement techniques during the training phase.

Ablation experiments were conducted to validate all components of the M^{2}CSFormer model. The modified version, excluding the M^{2}SM, LE block, and shape block, was trained while preserving the remaining elements' integrity (see Table XI). Removal of the M^{2}SM resulted in a notable decrease in the model's performance on Kvasir-SEG and CVC-ClinicDB datasets. This indicates that the decoder accurately predicts classifications and improves segmentation precision by leveraging the comprehensive spatial features obtained from the M^{2}SM. Furthermore, eliminating the LE block also led to a reduction in the model's effectiveness. The model demonstrated accurate polyp separation across various levels of complexity, with the LE block aiding in consolidating fragmented attention by offering boundary information.

The shape block, initially designed to aid the model in identifying polyp shape features, was subsequently removed. This removal led to a decline in the model's performance, as demonstrated in the table above. Shape blocks play a crucial role in distinguishing polyps from surrounding tissue by leveraging characteristic differences across different scales. They facilitate the extraction of detailed boundaries of polyps, enabling the M^{2}CSFormer to accurately segment polyps of diverse shapes across various polyp datasets. These results emphasize the significance of the proposed module in improving the overall segmentation efficacy of the model.

Our proposed module effectively mitigates attention dispersion in the Transformer and eliminates submerged residual information to enhance the accuracy and effectiveness of features, as demonstrated by the feature heatmap produced by each module (Fig. 6). The M^{2}CSFormer model provides a viable solution for medical applications, showcasing notable segmentation capabilities and exceptional real-time performance. This study is expected to introduce innovative concepts for polyp image segmentation. We are ready to refine the network's foundational architecture to address more complex scenarios in future research endeavors.

REFERENCES

[1] T. B. Kratzer, A. Jemal, K. D. Miller, S. Nash, C. Wiggins, D. Redwood, R. Smith, and R. L. Siegel, "Cancer statistics for american indian and alaska native individuals, 2022: Including increasing disparities in early onset colorectal cancer," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 2, pp. 120–146, 2023.

[2] C. Xia, X. Dong, H. Li, M. Cao, D. Sun, S. He, F. Yang, X. Yan, S. Zhang, N. Li, W. Chen, and J. Ni, "Cancer statistics in china and united states, 2022: Profiles, trends, and determinants," *Chinese Medical Journal*, vol. 135, no. 05, pp. 584–590, 2022.

[3] S. Zhao, S. Wang, P. Pan, T. Xia, X. Chang, X. Yang, L. Guo, Q. Meng, F. Yang, W. Qian *et al.*, "Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: A systematic review and meta-analysis," *Gastroenterology*, vol. 156, no. 6, pp. 1661–1674, 2019.

[4] P. Favoriti, G. Carbone, M. Greco, F. Pirozzi, R. E. M. Pirozzi, and F. Corcione, "Worldwide burden of colorectal cancer: a review," *Updates in Surgery*, vol. 68, no. 6, pp. 7–11, 2016.

[5] Y. Lin, X. Han, K. Chen, W. Zhang, and Q. Liu, "Cswindoubleunet: A double u-shaped network combined with convolution and swin transformer for colorectal polyp segmentation," *Biomedical Signal Processing and Control*, vol. 89, p. 105749, 2024.

[6] M. Fiori, P. Musé, and G. Sapiro, "A complete system for candidate polyps detection in virtual colonoscopy," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 07, p. 1460014, 2014.

[7] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1488–1502, 2014.

[8] O. H. Maghsoudi, "Superpixel based segmentation and classification of polyps in wireless capsule endoscopy," in *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. IEEE, 2017, pp. 1–4.

[9] 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*. Cham: Springer International Publishing, 2015, pp. 234–241.

[10] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "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: Springer International Publishing, 2018, pp. 3–11.

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

[12] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *Ieee Access*, vol. 9, pp. 40 496–40 510, 2021.

[13] 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.

[14] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps," *ArXiv Preprint ArXiv:2101.07172*, 2021.

[15] P. Chao, C.-Y. Kao, Y.-S. Ruan, C.-H. Huang, and Y.-L. Lin, "Hardnet: A low memory traffic network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[16] Y. Shen, X. Jia, and M. Q.-H. Meng, "Hrenet: A hard region enhancement network for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 559–568.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[18] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 699–708.

[19] X. Zhao, L. Zhang, and H. Lu, "Automatic polyp segmentation via multi-scale subtraction network," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 120–130.

[20] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. D. Johansen, D. Johansen, M. A. Riegler, S. Ali, and P. Halvorsen, "Msrf-net: a multi-scale residual fusion network for biomedical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2252–2263, 2021.

[21] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "Tganet: Text-guided attention for improved polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 151–160.

[22] X. Zhao, H. Jia, Y. Pang, L. Lv, F. Tian, L. Zhang, W. Sun, and H. Lu, "M^{2}snet: Multi-scale in multi-scale subtraction network for medical image segmentation," *Arxiv Preprint Arxiv:2303.10894*, 2023.

[23] X. Zheng, S. Chen, S. Wang, X. Huang, Y. Chen, J. Li, and W. Han, "Mscfnet: A multi-scale spatial and channel fusion network for geological environment remote sensing interpreting," *Springer Nature Singapore*, pp. 18–30, 2024.

[24] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-pvt: Polyp segmentation with pyramid vision transformers," *CAAI Artificial Intelligence Research*, vol. 2, p. 9150015, 2023.

[25] 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 *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Cham: Springer Nature Switzerland, 2022, pp. 99–109.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc, 2017.

[27] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv Preprint ArXiv:2010.11929*, 2020.

[28] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, "Intriguing properties of vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 296–23 308, 2021.

[29] 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.

[30] R. Katende, H. Kasumba, G. Kakuba, and J. Mango, "On the error bounds for relu neural networks," *IAENG International Journal of Applied Mathematics*, vol. 54, no. 12, pp. 2602–2611, 2024.

[31] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran, "On the limitation of convolutional neural networks in recognizing negative images," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 352–358.

[32] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling*. Cham: Springer International Publishing, 2020, pp. 451–462.

[33] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. R. de Miguel, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Medical Imaging Graph.*, vol. 43, pp. 99–111, 2015.

[34] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, vol. 2017, no. 1, p. 4037190, 2017.

[35] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognition*, vol. 45, no. 9, pp. 3166–3182, 2012.

[36] 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," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, pp. 283–293, 2014.

[37] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[38] M. Chi, W. Shaofan, and H. Hui, "Underwater image enhancement using dual regression u-structure network." *IAENG International Journal of Applied Mathematics*, vol. 54, no. 11, pp. 2459–2469, 2024.

[39] 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)*, 2016, pp. 565–571.