This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

MSD-EMA: Multi-Scale Decoupled Expectation-Maximization Attention for Polyp Segmentation

Xiaogang Du, *Member, IEEE*, Yibin Zou, Tao Lei*, *Senior Member, IEEE*, Dongxin Gu, Xuejun Zhang, *Member, IEEE*, Asoke K. Nandi, *Fellow, IEEE*

***Abstract*—Automatic polyp segmentation is a crucial technique of computer aided clinical diagnosis. However, some current polyp segmentation methods cannot accurately extract polyps from colonoscopy images due to the diversity of polyp shapes and sizes, as well as the blurry boundaries caused by the adhesion between polyps and surrounding tissues. To address this issue, we propose a multi-scale decoupled Expectation-Maximization attention, namely MSD-EMA. There are two advantages of MSD-EMA. Firstly, we design the decoupled Expectation-Maximization attention, which decouples attention weights into the sum of pairwise term representing inter regional features and unary term representing salient boundary features, thereby extracting boundary features between polyps and surrounding tissues while reducing computational complexity. Secondly, we propose the parallel collaborative strategy, which enables MSD-EMA to simultaneously extract sparse and dense feature maps using lower computational complexity. Sparse features are suitable for segmenting small polyps due to filtering out noise interference. Dense features are suitable for capturing large polyps that contain more location information. Comparative experiments are conducted with currently excellent polyp segmentation networks on five publicly available datasets, and the experimental results demonstrate that MSD-EMA can effectively improve polyp segmentation performance. Moreover, MSD-EMA is a plug-and-play module that can be applied to other types of segmentation tasks. The source code is available at *https://github.com/EmarkZOU/MSD-EMA*.**

***Index Terms*—Deep learning, Medical image segmentation, Attention mechanism, Multi-scale features, Neural networks, Automatic Polyp Segmentation.**

I. Introduction

COLORECTAL cancer is one of the most common and deadly cancers in the world [1]. Generally, the patient has small polyps inside colon, some of which may transform into colorectal cancer over time. Therefore, it is crucial to discover and remove the polyps to reduce the risk of colorectal cancer in the early stages. Polyp segmentation, as a research task for accurately locating polyps, is of great significance for clinical diagnosis and treatment of colorectal cancer [2].

In the past decade, with the development of deep learning, the polyp segmentation method based on Convolutional Neural Networks (CNNs) has gradually become a very important technology of the aided polyp diagnosis [3], [4], [5]. However, there are still two drawbacks. First, since the shapes, textures and colors of polyps and surrounding normal tissues are very similar, it is difficult to distinguish the features between polyps and the different features between polyps and similar surrounding tissues through the limited receptive field of convolution, resulting in the loss of small polyps. Secondly, due to the different shapes and sizes of polyps, the limited training dataset makes it challenging for segmentation models to accurately locate polyps, and insufficient generalization ability of some models leads to inaccurate segmentation results.

It is of great significance for improving the segmentation accuracy of polyps to solve the above two issues. For the first issue, to extract the inherent features between polyps and the boundary features between polyps and similar surrounding tissues, firstly, by introducing the Expectation-Maximization (EM) attention [6] to capture long-distance dependencies in polyp images, we improve the feature representation capacity within the polyp region while reducing computational complexity. Secondly, the attention weights are decoupled to obtain the salient boundary features of the image. For the second issue, when locating small polyps, we use a small number of image pixels as key points in the self-attention mechanism to avoid introducing unnecessary noise. When locating large polyps, the opposite is true. Therefore, to address these issues, inspired by EM attention, we propose the multi-scale decoupled EM attention, namely MSD-EMA. The contributions of this work are summarized as follows:

- We propose a decoupled EM attention branch, called D-EMA, by introducing the disentangled non-local operation in the EM attention. The proposed D-EMA can decouple attention weights into the sum of the pairwise and unary terms. The pairwise term represents features between polyps and the unary term represents salient

This work is partly supported by National Natural Science Foundation of China (Nos. 61861024, 62271296, and 62201334), Scientific Research Program Funded by Education Department of Shaanxi Provincial Government (Nos. 23JP022, and 23JP014), Key Research and Development Program of Shaanxi (No. 2021ZDLGY08-07), and General Project of Key Research and Development Programs in Shaanxi Province, China, Social Development Area (No. 2022SF-105). (Corresponding author: Tao Lei.)

X. Du, Y. Zou and D. Gu are with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China (E-mail: duxiaogang@sust.edu.cn, emarkzou@gmail.com, gdx1517253@163.com).

T. Lei is with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, and also with the Department of Geriatric Surgery, First Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710021, China (E-mail: leitao@sust.edu.cn).

X. Zhang is with the School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China (E-mail: xuejunzhang@mail.lzjtu.cn).

A. K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University of London, Uxbridge UB8 3PH, U.K., and also with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China. (E-mail: asoke.nandi@brunel.ac.uk)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

boundary features. Therefore, the proposed D-EMA can extract these two non-interfering features, effectively distinguishing polyps from their similar surroundings.

- We propose a Parallel Collaborative Strategy (PCS) to construct multi-scale attention feature maps. Each D-EMA branch in MSD-EMA generates attention feature maps with different sparsities from different numbers of key pixels. The sparse attention feature maps reduce noise interference and are suitable for segmenting small polyps. The dense attention feature maps contain more location information and are suitable for capturing large polyps. Through the integration of these distinct attention feature maps, the PCS improves the generalization capacity of the segmentation network for polyps of different shapes and sizes.

The rest of this paper is organized as follows. Section II reviews the related work of attention mechanism and multi-scale feature fusion for polyp segmentation. Section III proposes the MSD-EMA and introduces the structure of the D-EMA and the PCS. In Section IV, we demonstrate the ablation studies and comparative experimental results, and provide reliability and generalization evaluations. In Section V, we discuss the applicability of MSD-EMA. Finally, we conclude this work in Section VI.

II. Related work

With the development of deep learning, image segmentation methods based on CNN have gradually become the main development trend of polyp segmentation task [8], [9], [10], [11], [12]. However, because the receptive field of the convolution operation is limited, there are limitations in learning the long distance dependencies between pixels. In recent years, many excellent polyp segmentation networks have employed attention mechanisms [13] and multi-scale feature fusion [14] to overcome these limitations. Therefore, in this section, we review related work from two aspects: attention mechanism and multi-scale feature fusion.

A. Attention mechanism

In recent years, scholars have proposed many polyp segmentation methods based on deep learning [15], [16], [17]. Some current popular polyp segmentation methods employ specific attention mechanisms [18], [19], [20] to address the problem of polyps being difficult to accurately segment due to low contrast between polyps and surrounding tissues and different shapes and sizes of polyps.

To address this issue, Fan et al. [3] proposed a reverse attention module to improve the ability of network to extract boundary features. This module can establish the relationships between polyps and boundaries, thereby improving the segmentation accuracy. Besides, due to high similarity between some polyps and surrounding tissues, it is challenging to segment the boundaries of these polyps. To solve this drawback, Nguyen et al. [21] proposed the cascading context module and attention balance module to better integrate local and global features, thereby effectively focusing on the boundaries of polyps. The above methods primarily focus on feature selection rather than network structure design, thus limiting their generalization capacity. Zhang et al. [22] used lesion-aware cross-attention to enhance the feature contrast between polyps and background regions, and designed an efficient self-attention module to capture long-distance contextual relationships, further improving segmentation accuracy.

Zhang et al. [23] argued that the segmentation of polyps with different sizes relies on different local and global contextual information for regional comparative analysis. Therefore, to accurately segment polyps of different sizes, they proposed the local context attention to transfer local contextual features from the encoder to the decoder, increasing attention to the key regions in the previous prediction map. However, this method focuses on easily segmented regions while ignoring those that are difficult to segment. To address this issue, Shen et al. [5] designed an information context enhancement module to improve feature representation capacity under the guidance of the difficulty-ware attention module, thereby improving segmentation accuracy. However, these methods improve the performance of the network by using additional network layers, but increase the model complexity. To reduce model complexity, Tomar et al. [24] used a text-guided attention to encode attributes such as the number and size of polyps using simple bytes. Wei et al. [25] proposed a shallow attention module, which filters out background noise from shallow features and preserves the features of small polyps.

B. Multi-scale feature fusion

Multi-scale feature fusion is the process of fusing and exchanging information between low-level and high-level features during encoding and decoding. Multi-scale feature fusion is also an important solution for accurately segmenting objects with different shapes and sizes. In many classical image segmentation networks, scholars have employed different multi-scale feature fusion modules to improve segmentation performance [26], [27], [28]. For the polyp segmentation task, scholars have used multi-scale feature fusion to effectively improve the features encoding and decoding prediction capabilities of polyp segmentation networks.

To improve the feature encoding of networks for polyps, Srivastava et al. [14] introduced the Dual-Scale Dense Fusion (DSDF) block in the multi-scale residual fusion network. The DSDF block uses different scale features from two encoders as inputs, establishing a fusion strategy between high-level and low-level features, which helps to enhance shallow features using high-level features. To supplement specific boundary information during the encoding process, Qiu et al. [29] proposed BDG-Net to generate boundary distribution maps, which are as supplementary spatial information and sent to the decoder to guide polyp segmentation. Inspired by the idea of enhancing boundary features in BDG-Net, Cheng et al. [30] first calculated eight directional derivatives for each pixel, and then selected pixels with large directional derivatives to form candidate boundary regions for polyps. Finally, boundary features and high-level semantic features were fused to improve the segmentation accuracy of polyp boundaries. To enhance the multi-scale context feature representations, Zhong et al. [31]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

proposed Adaptive Scale Context (ASC) module and Semantic Global Context (SGC) module. The ASC aggregates multi-scale contextual information to focus on the target region. The SGC filters out background noise in low-level features by fusing high-level and low-level features in the decoder.

To improve the predictive capacity of the decoder, Wang et al. [32] proposed a selective feature aggregation module and inserted it into the convolutional layer between the encoder and decoder, which can adaptively extract features using kernels of different sizes. However, the above methods directly use element-wise addition or concatenation to fuse the features of different levels of the encoder. These operations do not pay more attention to the differential information between different levels, which not only generates redundant information but also weakens the characteristics of specific level features. To address this issue, Liu et al. [33] utilized the domain specific batch normalization layer units in the encoder and decoder to preserve the feature differences between adjacent levels, thereby preserving the localization information and subtle boundary information of polyps.

III. Multi-Scale Decoupled EM Attention

A. The overall structure

The convolution operation with limited receptive field cannot effectively extract key features of polyps from similar surroundings. We propose MSD-EMA to improve the performance of CNN in segmenting polyps with different sizes from similar surrounding tissues. The overall structure of the proposed MSD-EMA is shown in the figure 1. In the figure 1, MSD-EMA is a multi-scale attention consisted of multiple D-EMA branches and a residual connection. We design D-EMA to extract distinct features intra- and inter-classes, and propose PCS to parallelize multiple D-EMA branches for generating multi-scale attention feature maps.

Initially, we design a D-EMA module to accurately segment polyps from surrounding tissues. D-EMA establishes long-distance dependency relationships in polyp images through efficient self-attention. D-EMA has two advantages as follows.

(1) D-EMA utilizes the EM algorithm to extract attention feature maps, thereby capturing the long-distance dependency relationships between polyps, which can reduce the original computational complexity $O(N^2)$ to $O(NK)$, $K \ll N$, where K is the size of the compact subset and N is the number of pixels in the original input feature map. Specifically, the key idea of D-EMA is to find a compact base subset that can represent all pixels, rather than directly using all pixels themselves. Therefore, EMA uses this compact base subset to calculate attention maps, effectively reducing computational complexity. Assuming that the input feature map has N pixels, the size of the compact base subset is K ($K \ll N$), and the number of iterations is T, the computational complexity of D-EMA is $O(NKT)$. Due to the small number of iterations T, usually $T = 3$, the computational complexity of EMA can be approximated as $O(NK)$. Furthermore, due to $K \ll N$, $O(NK)$ is much smaller than $O(N^2)$.

(2) We introduce the Disentangled Non-Local (DNL) block [7] into the EM framework of D-EMA, which decomposes the original attention weights into the sum of the pairwise and unary terms, avoiding the degradation of attention feature representations. The pairwise term can learn the relationships between the features of the object regions in the image, while the unary term can learn the salient boundary information in the image. Therefore, the attention feature maps calculated by D-EMA include both polyp features and salient boundary features, which can more accurately locate and segment polyps.

To better segment polyps of varying shapes and sizes, we propose the PCS to construct multi-scale attention feature maps. Parallel D-EMAs generate attention feature maps with different sparsities by initializing matrices of different spatial dimensions as subsets μ_i, $i \in \{0, 1, 2\}$. Sparse attention feature maps are low rank and filter out input noise information, making them suitable for capturing small polyp features. On the contrary, a dense attention feature maps include as many features of large polyps as possible. When performing small polyp segmentation, the value of K is correlated with N, and the computational complexity of MSD-EMA will degrade to $O(N^2)$. Compared with the original Non-Local block, there is still no increase in computational complexity under the premise of extracting multi-scale long-range dependencies.

B. Decoupled EM Attention

Due to the similarity between polyps and surrounding tissues, the segmentation results of polyps in colonoscopy images often include non-polyp areas, leading to the problem of over-segmentation. To solve this problem, the proposed D-EMA is capable of simultaneously extracting intra-class correlation features within polyps and inter-class differential features between polyps and boundaries. The structure of D-EMA is illustrated in the figure 2.

In the step E of D-EMA, D-EMA randomly initializes a compact subset μ of size (B, C, K), where B represents the batch size and C represents the number of channels in the feature map. Here, K is the size of the compact subset, which is much smaller than the number of pixels N in the original input feature maps $X \in \mathbb{R}^{B\times C\times K}$. The subset μ is inputted into the $A_E^{Decoupled}$ and $Unary$ modules for calculation, and the attention weights as hidden variables in the EM algorithm are decoupled into the sum of pairwise and unary terms. These two terms are added pixel by pixel to obtain new attention weights, as shown in (1):

$$\omega = \omega_p + \omega_u, \tag{1}$$

where $\omega \in \mathbb{R}^{B\times N\times K}$ represents the attention weights, $\omega_p \in \mathbb{R}^{B\times N\times K}$ represents the values of pairwise term, and $\omega_u \in \mathbb{R}^{B\times N\times K}$ represents the values of unary term. The $A_E^{Decoupled}$ and $Unary$ modules are separately illustrated in the figure 3. In the figure 3, the $A_E^{Decoupled}$ module in the left box generates and outputs the pairwise term, while the $Unary$ module in the right box generates and outputs the unary term.

In the $A_E^{Decoupled}$ module, the feature maps X and a randomly initialized subset μ are used as inputs to the $A_E^{Decoupled}$ module. The feature maps X are mapped to the matrix W_q to obtain the feature matrix $Query \in \mathbb{R}^{B\times N\times C}$, and the subset μ is mapped to the matrix W_k to obtain the feature matrix

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

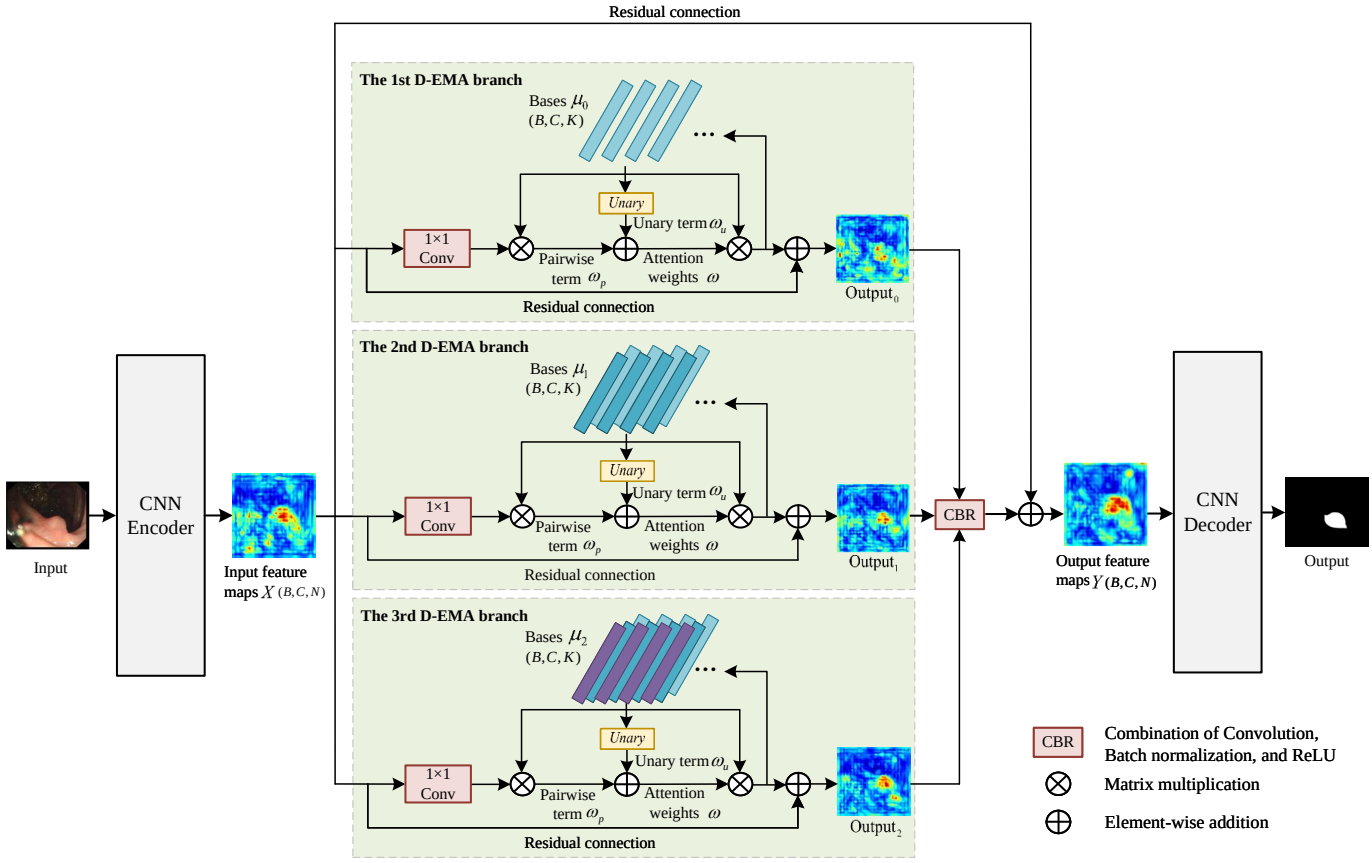


Fig. 1. The structure of MSD-EMA. MSD-EMA is a multi-scale attention mechanism consisting of three D-EMA branches and a residual connection. We design three D-EMA branches with different numbers of compact subset μ to extract feature maps with different sparsities. Then through PCS, the feature maps of the three branches are finally integrated to capture polyps of different sizes. Notably, MSD-EMA is a plug-and-play module that can be inserted into the appropriate position of segmentation networks, not just between the encoder and decoder.

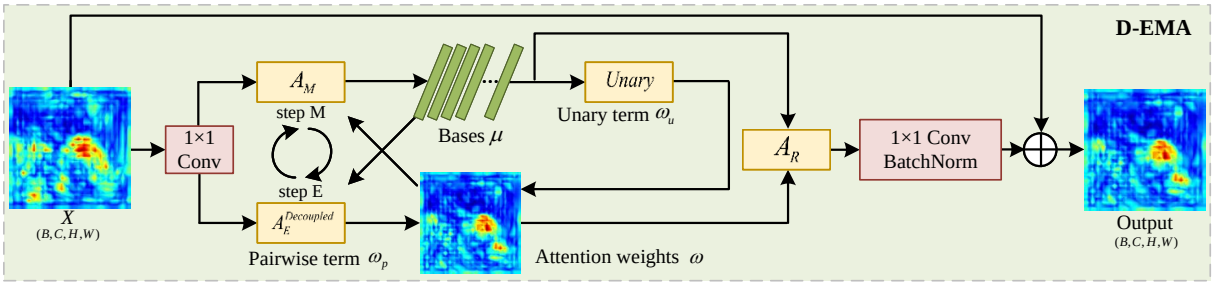


Fig. 2. The structure of D-EMA. D-EMA effectively extracts internal and boundary features of polyps by decoupling the attention weights into pairwise and unary terms, thereby improving segmentation accuracy.

$Key \in \mathbb{R}^{B\times C\times K}$. After the whitening without the mean, matrix multiplication, and Softmax normalization operations, the feature matrices Key and $Query$ are output as the pairwise term of size (B, N, K). The calculation process of the pairwise term are shown in (2):

$$\omega_p(X_i, \mu^j) = \sigma((q_i - \overline{q})^T(k_i - \overline{k})), \tag{2}$$

where X_i represents the feature value at the position i of the input feature maps, and μ^j represents the feature value at the subset position j, and $j \in \{0, 1, ..., K-1\}$. q_i and k_i are the feature values at the corresponding positions of the matrices $Query$ and Key obtained by using 1×1 convolution for X_i and μ^j, respectively. $\overline{q}$ and $\overline{k}$ represent the mean values of features on the spatial dimensions of matrices $Query$ and Key, respectively. $\sigma(\cdot)$ represents Softmax normalization operation. $q_i - \overline{q}$ and $k_i - \overline{k}$ represent whiten operations that remove the mean, respectively.

In the $Unary$ module, we first reduce the channel dimension of the subset μ from C to 1, then expand the channel of the subset μ to N through channel expansion, and finally normalize to obtain the unary term of size (B, N, K). The calculation process is shown in (3):

$$\omega_u(X_i, \mu^j) = \sigma(W_m\mu^j), \tag{3}$$

where W_m is essentially a 1×1 convolution used to fuse

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

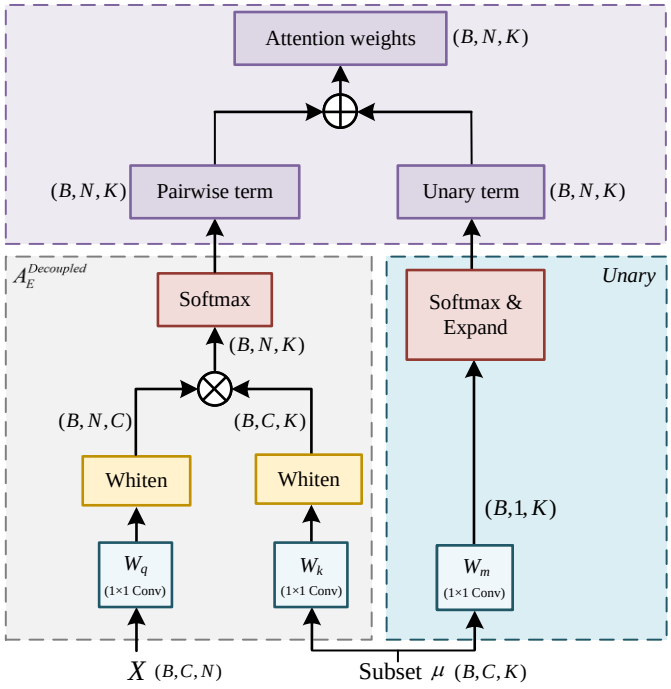


Fig. 3. The structure of $A_E^{Decoupled}$ and $Unary$ modules. $A_E^{Decoupled}$ generates the pairwise term to identify the relationships between polyps. $Unary$ generates the unary term to highlight distinct boundary features. The combination of these two terms results in the attention weights, which represents both the internal features of regions and the boundary features.

features in the channel dimension of subsets μ, allowing unary term to focus more on spatial features and effectively capture boundary features of polyps.

Therefore, by incorporating (2) and (3) into (1), it can be concluded that:

$$\omega\left(X_i, \mu^j\right) = \underbrace{\sigma\left(\left(q_i - \overline{q}\right)^T \left(k_i - \overline{k}\right)\right)}_{\text{pairwise term}} + \underbrace{\sigma\left(W_m \mu^j\right)}_{\text{unary term}}. \tag{4}$$

In (4), by matrix multiplication and normalization calculation, $\sigma\left(\left(q_i - \overline{q}\right)^T \left(k_j - \overline{k}\right)\right)$ can maximize the normalized difference between pixels in the feature matrix $Query$ and pixels in the Key to determine the whitening dot product. Therefore, the pairwise term tends to learn pixel relationships within the polyp region. $\sigma\left(W_m \mu^j\right)$ tends to learn the influence of boundary pixels on all pixels, so the unary term can capture the boundary features. Therefore, by decoupling the original attention weights into the pairwise and unary terms, the attention weights contain two different types of feature weights.

In the step M of D-EMA, the attention weights and X are inputted to the A_M, thereby performing matrix multiplication. As a result, the subset μ gradually converges to a subspace of the input feature maps X, as shown in the (5):

$$\mu = \omega \times X. \tag{5}$$

Finally, after T iterations of the E and M steps, both the subset μ and attention weights converge to obtain the optimal solution. The converged attention weights and subset μ are input into the A_R to generate an attention feature map with rich long-distance dependencies. Besides, the input feature maps X transmitted through residual connections are added to obtain an output of the same size as the input feature maps. The specific process is shown in (6):

$$\text{Output} = \omega \times \mu + X. \tag{6}$$

C. Parallel Collaborative Strategy

Due to the decoupling operation in D-EMA, the extracted attention feature map contains feature relationships within the target region and salient boundary features. However, since the element number of subset μ in D-EMA is initialized to a fixed size, the generated attention feature maps cannot adequately meet the segmentation requirements for polyps with different shapes and sizes. To address this issue, we propose the PCS for constructing multi-scale attention feature maps based on a parallel multi-branch architecture, as shown in the figure 1.

In the figure 1, MSD-EMA is consisted of three parallel D-EMAs and a residual connection. Each D-EMA randomly initializes a different element number of subset μ_i. For example, the sizes of subsets μ_0, μ_1, and μ_2 are $(B, C, \frac{1}{8}N)$, $(B, C, \frac{1}{4}N)$, and $(B, C, \frac{1}{2}N)$, respectively. After three iterations of each D-EMA, both the subset μ_i and attention weights reach convergence. The converged subset μ_i is the subspace of different sizes in the input feature maps X, which is multiplied by the corresponding attention weight matrix. Each D-EMA generates attention feature maps with different sparsities. The D-EMA that initializes a small number of subsets μ_i generates a sparse attention feature map, which is also low rank and filters out most of the noise in the input feature maps, making it suitable for extracting features of small polyps. The D-EMA that initializes a large number of subsets μ_i generates a dense attention feature map, which includes as many features of large polyps as possible. After fusing all attention feature maps and adding the original input feature maps X transmitted through residual connections, MSD-EMA can construct and output the multi-scale attention feature maps Y. The calculation process is shown in (7):

$$Y = \text{CBR}(\text{Output}_0 \oplus \text{Output}_1 \oplus \text{Output}_2) + X, \tag{7}$$

where CBR denotes a combination of convolution, batch normalization, and ReLU. $\oplus$ denotes the channel concatenation.

The multi-scale attention feature maps Y contain the features enhanced by attention weights with different sparsities, which can effectively enhance the generalization capacity for the polyp segmentation task. In addition, in MSD-EMA, the computational complexity of the attention weights generated by the three D-EMAs are $O(\frac{1}{8}TN^2)$, $O(\frac{1}{4}TN^2)$ and $O(\frac{1}{2}TN^2)$, respectively, where T is the number of iterations and $T = 3$.

IV. Experiments

In this section, we first introduce five publicly available datasets and data preprocessing. Secondly, we describe the experimental environment and parameter settings. Thirdly, we provide evaluation metrics for the experimental results. Finally, we provide the detailed results and analysis of the ablation study, comparative experiment, reliability and generalization evaluation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

A. Dataset and data preprocessing

To verify the effectiveness of MSD-EMA in the polyp segmentation task, ablation studies and comparative experiments are conducted on five datasets: Kvasir [34], ETIS [35], CVC-ColonDB [36], CVC-ClinicDB [37], and CVC-300 [38]. The details of these datasets are presented as follows.

(1) Kvasir: This dataset is part of the Medical Multimedia Challenge and contains 1000 images of varying sizes, ranging from 332×487 to 1920×1072. We use 900 images for training and 100 images for testing, each image contains at least one polyp. In the testing dataset, there are 13 images containing multiple polyps.

(2) CVC-ClinicDB: This dataset comes from 31 colonoscopy sequences and contains 612 images with a size of 384×288. We use 550 images for training and 62 for testing, each image contains polyps. In the testing sample, there are 5 images containing multiple polyps.

(3) CVC-300: This dataset comes from 44 colonoscopy sequences and contains 912 images with a size of 574×500. We use 60 images for testing. Each image contains only one polyp. These images are all used for testing.

(4) CVC-ColonDB: This dataset comes from 15 different colonoscopy sequences, with a total of 380 images and an image size of 574×500 pixels. We use all 380 images for testing. All images contain a single polyp, and the size of the polyp varies greatly in the image.

(5) ETIS: This dataset comes from 34 colonoscopy videos, containing 196 images with a size of 1225×966. We use all 196 images for testing. Although all images contain a single polyp, there are significant differences in the size of polyps in these images.

To ensure fairness in the experiments, the dataset partitioning is consistent with MS-Net [33]. This data partitioning is currently the popular approach in the field of polyp segmentation [3], [15], [22], [33]. Specifically, 900 samples are randomly selected from the Kvasir, and 550 samples are randomly selected from the CVC-ClinicDB, totaling 1450 samples to form the training dataset. The testing dataset consists of the remaining 798 images from these five datasets.

In the data preprocessing, the data augmentation operations include: random scaling, horizontal flipping, vertical flipping, and random rotation of 90 degrees. We apply four data augmentation operations to each sample in the training dataset, and use the enhanced images as training samples. The above data augmentation methods are used for the training dataset in the ablation study. Therefore, UNet [8], as a benchmark network, performed better in the ablation study than in the comparative experiment, which proves the effectiveness of the data preprocessing.

B. Experimental environment and parameter settings

The main environment configuration for these experiments is: Intel Xeon Gold 6226R CPU, 32GB memory, Nvidia Geforce RTX 3090 GPU, and 24GB graphics memory. These experiments are conducted on the Ubuntu 16.04.10. The deep learning framework is PyTorch 1.7.

We employ the binary cross-entropy and Dice coefficient as the loss function to guide model training. We used the Adam optimizer in the training process of our model. The maximum number of training epochs is set to 150. The batch size is 48. The initial learning rate is 0.08. The learning rate adjustment strategy is $lr = lr \times (1 - (\frac{epoch}{Epoch})^{0.9})$, where $epoch$ is the current number of epochs, $Epoch$ is the maximum number of epochs, and $Epoch = 150$. The weight decay is 0.0005, and the momentum is 0.9.

C. Evaluation metrics

We choose IoU and $Dice$ as objective metrics to evaluate the segmentation performance of MSD-EMA. IoU and $Dice$ are widely used evaluation metrics, both of which can measure the similarity between segmentation results and Ground Truth. IoU and $Dice$ are calculated in (8) and (9), respectively:

$$IoU = \frac{TP}{TP + FP + FN}, \tag{8}$$

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN}, \tag{9}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

D. Ablation study

To validate the effectiveness of the D-EMA and PCS for the proposed MSD-EMA, ablation studies are conducted on five publicly available polyp datasets. In addition, to demonstrate that MSD-EMA can be used as a universal plug-and-play component in the segmentation networks, we inserted MSD-EMA behind the classical encoders of UNet [8] and ResNet101 [39] and conducted two sets of ablation studies.

(1) *The effectiveness of the D-EMA*. To demonstrate the effectiveness of D-EMA, we compare it with EM Attention and DNL modules, and the visualization results are shown in the figure 4. The figure 4 shows the feature maps and segmentation results of UNet, UNet+EM, UNet+DNL, and UNet+D-EMA on the Kvasir dataset, respectively. Firstly, after introducing the EM attention, the generated feature maps focus more effectively on polyp regions and represent the global relationships of polyps. However, some feature maps mistakenly focus on noisy areas due to similar surrounding interference. Secondly, after introducing the DNL module, the attention feature maps extract and retain more boundary features from the original image. Although the segmentation results are improved compared to UNet, there is an over-segmentation phenomenon due to the fact that the DNL module only focuses on the local features of polyps. Finally, after introducing D-EMA, it is evident that the generated attention feature maps focus more accurately on the segmented polyps, and the segmentation results are also closer to the Ground Truth.

To further evaluate the performance of D-EMA, the quantitative results are shown in Table I. As shown in the 4th and 7th rows of Table I, it can be seen that D-EMA significantly improves the segmentation accuracy compared to the original

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

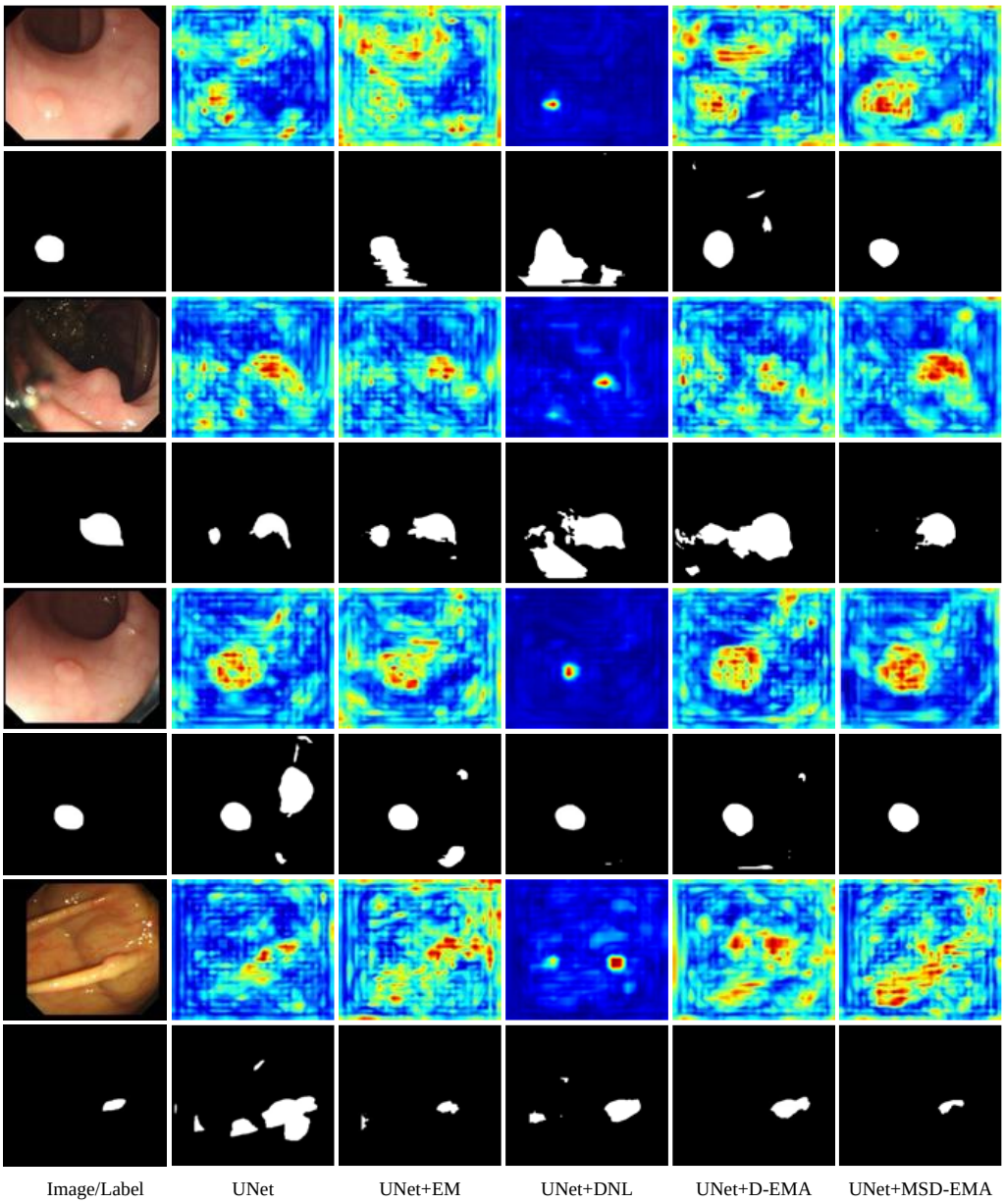


Fig. 4. Visualization comparison of feature maps on the Kvasir dataset. In the ablation study, the incremental components allows the model to focus more effectively on the critical regions. In the eighth row, the UNet+MSD-EMA exhibits a superior capacity to accurately capture the features of small polyps.

UNet and ResNet on all five polyp datasets. These results prove that the D-EMA can not only focus on the polyps through the established global relationships, but also extract the boundaries of polyps, effectively improving segmentation accuracy.

(2) *The effectiveness of the PCS.* To verify the effectiveness of the PCS, the visualization of the multi-scale attention feature maps and segmentation results after the introduction of MSD-EMA are shown in the last column of the figure 4. In the figure 4, compared to introducing D-EMA, the constructed multi-scale attention feature maps can significantly increase the attention in the polyp regions, reduce the attention in unrelated regions, thereby avoiding noise interference. Therefore, it can be found that the segmentation results are also closer to the Ground Truth.

In addition, since the downsampling process of the UNet causes the resolution of feature maps to decrease, this may make it difficult for the network to capture small polyps at the end of the encoder. However, as depicted in the eighth row of the figure 4, UNet+MSD-EMA demonstrates a superior capacity to accurately capture the features of small polyps and concentrate the attention on these regions. The main reason is that MSD-EMA is able to generate multi-scale feature maps with different sparsity levels through PCS, which can capture both small and large polyps simultaneously. Meanwhile, MSD-EMA utilizes self-attention to capture long-range dependencies, which enables the network to capture features of small polyps within a larger receptive field, even if these polyps occupy fewer pixels in the downsampled feature map.

In Table I, while UNet is chosen as baseline, compared

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

TABLE I
RESULTS OF ABLATION STUDIES ON THE POLYP DATASET. THE BEST VALUES ARE IN BOLD.

Methods \ Datasets	CVC-ColonDB		ETIS		Kvasir		CVC-300		CVC-ClinicDB	
	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$
UNet	0.525	0.614	0.344	0.430	0.747	0.827	0.688	0.774	0.811	0.823
UNet+EM	0.517	0.610	0.389	0.481	0.757	0.834	0.633	0.747	0.812	0.860
UNet+DNL	0.547	0.623	0.356	0.423	0.765	0.840	0.662	0.758	0.792	0.851
UNet+D-EMA	0.553	0.636	0.417	0.489	0.771	0.846	0.716	0.804	0.802	0.862
UNet+MSD-EMA	**0.560**	**0.639**	**0.449**	**0.530**	**0.774**	**0.850**	**0.746**	**0.828**	**0.814**	**0.880**
Res101+EMA	0.554	0.648	0.481	0.577	0.767	0.843	0.760	0.853	0.731	0.812
Res101+D-EMA	0.567	0.663	0.511	0.606	0.783	0.856	0.738	0.818	0.757	0.829
Res101+MSD-EMA	**0.590**	**0.671**	**0.538**	**0.626**	**0.786**	**0.857**	**0.761**	**0.853**	**0.780**	**0.853**

to a single D-EMA, PCS can significantly improve the segmentation performance on the ETIS, CVC-300, and CVC-ClinicDB. Concretely, on the ETIS, PCS increases $Dice$ by 4.1% and IoU by 3.2% compared to a single D-EMA. On the CVC-300, compared to a single D-EMA, PCS increases $Dice$ by 2.4% and IoU by 3%. On the CVC-ClinicDB, compared to a single D-EMA, PCS increases $Dice$ by 1.8% and IoU by 1.2%. PCS utilizes three different D-EMA branches to capture features of different sparsity levels. Notably, polyp segmentation is a very challenging task due to the diversity of polyp shapes and sizes, as well as the blurry boundaries. It is significant to improve $Dice$ and IoU by more than 2-4% using UNet+MSD-EMA on these datasets. Therefore, in the presence of polyps of different sizes, PCS effectively captures details of large polyps while enhancing boundary detection for small polyps. In the ETIS, CVC-300, and CVC-ClinicDB, the sizes of polyps are more diverse, which can fully utilize the feature extraction capacity of PCS. Additionally, on the CVC-ColonDB and Kvasir, PCS shows relatively little improvement in segmentation performance. Compared to a single D-EMA, PCS increases $Dice$ by 0.3% and IoU by 0.7% on the CVC-ColonDB, $Dice$ by 0.4% and IoU by 0.3% on the Kvasir. Therefore, when using UNet as a baseline, PCS can effectively improve the performance of polyp segmentation task.

In Table I, when Res101 is used as the baseline, PCS achieves good performance gains on the CVC-ColonDB, ETIS, CVC-300, and CVC-ClinicDB, significantly improving the segmentation performance. On the CVC-ColonDB, compared to a single D-EMA, PCS increases $Dice$ by 0.8% and IoU by 2.3%. On the ETIS, compared to a single D-EMA, PCS increases $Dice$ by 2% and IoU by 2.7%. On the CVC-300, compared to a single D-EMA, PCS increases $Dice$ by 3.5% and IoU by 2.3%. On the CVC-ClinicDB, compared to a single D-EMA, PCS increases $Dice$ by 2.4% and IoU by 2.3%. PCS can fully utilize the D-EMA to effectively extract features of polyps and their surrounding tissues, and can better handle blurry boundaries. Therefore, PCS can achieve good performance improvement on these four datasets. On the Kvasir, PCS can not significantly improve the segmentation performance, with $Dice$ increasing by 0.1% and IoU increasing by 0.3%. The reason may be that there are more polyps in the test samples on this dataset. However, PCS is less effective in handling samples with multiple polyps, resulting in limited performance improvement. To summarize, when using Res101 as the baseline, PCS can effectively improve the performance of polyp segmentation task.

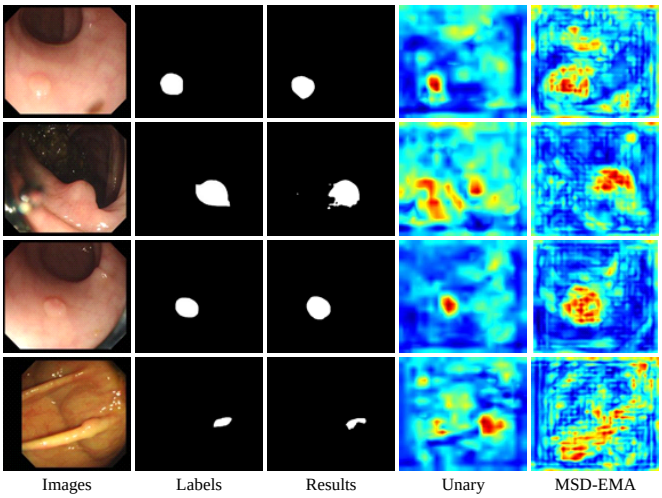


Fig. 5. The visualization results for extracting boundary features.

(3) *The effectiveness of extracting boundary features.* To verify that MSD-EMA can independently extract the salient boundary features of the objects, we first separately extract the unary term through the $Unary$ module in D-EMA. Secondly, we apply it to the input images to generate the salient boundary feature maps. Thirdly, we apply the attention weights, which are finally extracted by MSD-EMA, to the input features to generate the attention feature maps. To demonstrate that MSD-EMA can effectively extract boundary features, we analyze the results using heat map visualization, as illustrated in the figure 5. It is evident that MSD-EMA can accurately extract boundary features of polyps, thereby improving the localization and segmentation accuracy of the polyps adhered to surrounding tissues.

E. Comparative experiments

To evaluate the segmentation performance of MSD-EMA on polyp datasets, we conduct experiments and compare it with the popular polyp segmentation methods in recent years, including UNet [8], UNet++ [9], EMA-Net [6], MS-Net [33], PraNet [3], TGA-Net [24], ResUNet++ [15], and SA-Net [25]. To facilitate experimental comparison, we insert MSD-EMA into the backend of the shallow attention module of SA-Net to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

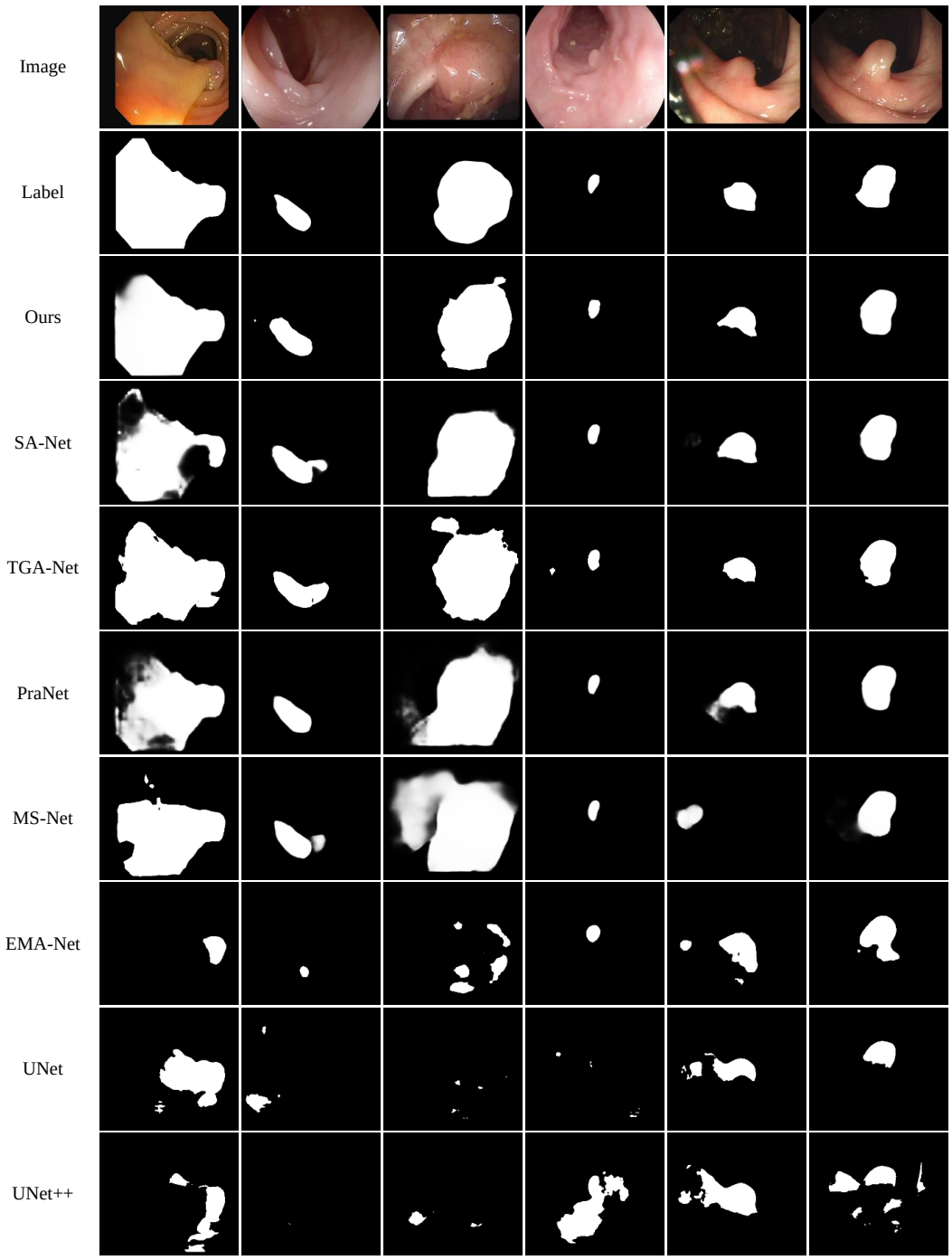


Fig. 6. The segmentation results of popular networks on the Kvasir dataset.

construct a new model for polyp segmentation, which is called our model. In next, we demonstrate the segmentation results of our model and current state-of-the-art models on five public available polyp datasets, which are shown in the figure 6.

In the figure 6, there are some significant differences between the segmentation results of UNet and the Ground Truth. The main reason is that the encoder structure of UNet is relatively simple, which leads to insufficient feature extraction for complex polyps. UNet++ slightly improves the segmentation accuracy compared to UNet due to the addition of dense skip connections between the encoder and decoder. MS-Net captures the differential information between different levels of features by adding dense differential modules between the encoder and decoder, resulting in a significant improvement in segmentation results compared to UNet++. However, from the segmentation results in columns 1 and 5 of the figure 6, it can be seen that MS-Net is very sensitive to noise and prone to incorrect segmentation. EMA-Net reduces the computational complexity of self-attention by initializing a small element number of subset μ. But as a result, it loses many key shape features of polyps, resulting in under-segmentation of large polyps with complex shapes. TGA-Net enhances the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

TABLE II
COMPARATIVE RESULTS WITH EXCELLENT SEGMENTATION NETWORKS ON THE FIVE POLYP DATASETS. THE BEST VALUES ARE IN BOLD.

Methods \ Datasets	CVC-ColonDB		ETIS		Kvasir		CVC-300		CVC-ClinicDB	
	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$
UNet(MICCAI'15) [8]	0.454	0.547	0.248	0.319	0.665	0.766	0.621	0.727	0.756	0.824
UNet++(TMI'19) [9]	0.483	0.410	0.335	0.398	0.743	0.821	0.627	0.710	0.729	0.794
EMA-Net(CVPR'19) [6]	0.554	0.648	0.481	0.577	0.767	0.843	0.760	0.853	0.731	0.812
ResUNet++(ISM'19) [15]	-	-	-	-	0.793	0.813	-	-	0.796	0.796
PraNet(MICCAI'20) [3]	0.645	0.716	0.664	0.719	0.840	0.898	0.804	0.873	0.858	0.902
MS-Net(TMI'20) [33]	0.651	0.722	0.585	0.644	0.839	0.894	0.788	0.841	**0.879**	**0.921**
SA-Net(MICCAI'21) [25]	0.675	0.754	0.683	0.763	0.841	0.897	0.831	0.893	0.861	0.913
TGANet(MICCAI'22) [24]	0.633	0.707	0.578	0.653	0.839	0.894	0.819	0.886	0.855	0.907
APCNet(TIM'23) [40]	0.679	0.758	0.648	0.726	0.842	0.899	0.827	0.893	0.859	0.911
Ours	**0.684**	**0.763**	**0.691**	**0.775**	**0.846**	**0.903**	**0.849**	**0.912**	0.863	0.914

extraction of features related to the size and quantity of polyps by introducing text-guided attention, resulting in a significant improvement in segmentation results compared to EMA-Net. PraNet gradually integrates the deep contextual semantic features into the decoder, guiding the decoding process and restoring local detail features, and utilizes a reverse attention module to extract the boundary features of polyps. Therefore, PraNet improves segmentation results for small polyps compared to EMA-Net. Compared to MS-Net, PraNet has more accurate segmentation results under noise interference such as lighting. However, the segmentation results are still inaccurate when polyps are similar to surrounding tissues. SA-Net uses deep contextual semantic features to continuously guide the decoder to recover detailed information, improving the accuracy of segmenting polyps from similar surrounding tissues. After inserting MSD-EMA into the backend of SA-Net, our model can better extract the features of polyp regions and boundaries. From the visualization results in the figure 6, it can be seen that MSD-EMA further improves the accuracy of polyp segmentation.

The quantitative results of the comparative experiment are shown in Table II. On the CVC-ColonDB, ETIS, Kvasir, and CVC-300, the IoU scores of our model are 0.684, 0.691, 0.846, and 0.849, respectively, and the $Dice$ scores of our model are 0.763, 0.775, 0.903, and 0.912, respectively. Our model achieves the best segmentation performance on these four datasets. On the CVC-ClinicDB, the IoU and $Dice$ scores of our model are 0.863 and 0.914, respectively, which are only slightly lower than the results of MS-Net, but still better than the results of other methods. Therefore, the proposed MSD-EMA can improve the performance of the polyp segmentation network and demonstrate excellent segmentation accuracy on five polyp datasets.

F. Reliability evaluation

To further verify the stability and reliability of our model, we conduct a 5-fold cross validation experiment. Specifically, we randomly split the entire dataset into five equal sized subsets. One subset is retained as the test dataset, while the remaining four subsets are used to train our model. In addition, while conducting cross validation, we retain the original five independent test datasets, which are not used throughout the entire training and validation process and are used to ultimately evaluate reliability of our model. The evaluation results are shown in Table III.

In Table III, the evaluation results for each fold are very close, indicating that our model performs consistently across different data partitions. Additionally, the average values of the evaluation metrics across all folds are very close to the results obtained from training on the entire dataset (Compare with the last row of Table II), further verifying the good reliability of our model.

G. Generalization evaluation

We also insert MSD-EMA between the encoder and decoder of PMR-Net [41] and conduct generalization evaluation on nine medical segmentation datasets, which include the five polyp datasets mentioned above, as well as retinal vessel dataset DRIVE [42], macular retinal vessel dataset STARE [43], skin lesion dataset ISIC2018 [44], and cell nucleus dataset DSB2018 [45]. The experimental visualization results are shown in the figure 7. In the figure 7, the segmentation results of introducing MSD-EMA into PMR-Net are better than PMR-Net on these datasets. Therefore, the proposed MSD-EMA can improve segmentation performance.

The experimental results of inserting MSD-EMA into PMR-Net and quantitatively comparing it with PMR-Net on five polyp datasets are shown in Table IV. In Table IV, it can be seen that inserting MSD-EMA into PMR-Net resulted in improved segmentation results compared to PMR-Net on all five polyp datasets. Therefore, MSD-EMA is a plug-and-play module that can improve segmentation accuracy.

Furthermore, MSD-EMA is inserted into PMR-Net and comparative experiments are conducted on four publicly available medical image datasets, DRIVE, STARE, ISIC2018, and DSB2018. The experimental results are shown in Table V. Therefore, MSD-EMA can be applied to many medical image segmentation tasks and has good generalization capacity.

V. DISCUSSION

In the above experiments, we attempt different insertion positions of MSD-EMA in other segmentation networks. For example, inserting MSD-EMA into the encoder of UNet, inserting MSD-EMA after the shallow attention, and inserting

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/TIM.2025.3547131, IEEE Transactions on Instrumentation and Measurement

TABLE III
THE EVALUATION RESULTS USING 5-FOLD CROSS VALIDATION. THE AVERAGE RESULTS ARE IN BOLD.

Methods \ Datasets	CVC-ColonDB		ETIS		Kvasir		CVC-300		CVC-ClinicDB	
	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$
Fold 1	0.673	0.758	0.686	0.769	0.841	0.895	0.848	0.913	0.862	0.915
Fold 2	0.685	0.762	0.692	0.774	0.837	0.892	0.841	0.907	0.864	0.913
Fold 3	0.682	0.765	0.689	0.777	0.848	0.901	0.847	0.914	0.860	0.908
Fold 4	0.686	0.761	0.694	0.781	0.844	0.905	0.851	0.910	0.865	0.911
Fold 5	0.681	0.757	0.688	0.776	0.849	0.907	0.846	0.915	0.859	0.917
Average	**0.681**	**0.760**	**0.690**	**0.776**	**0.844**	**0.899**	**0.846**	**0.912**	**0.861**	**0.912**

TABLE IV
THE GENERALIZATION EVALUATION ON THE FIVE POLYP DATASETS. THE BEST VALUES ARE IN BOLD.

Methods \ Datasets	CVC-ColonDB		ETIS		Kvasir		CVC-300		CVC-ClinicDB	
	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$
PMR-Net [41]	0.493	0.578	0.351	0.414	0.736	0.814	0.666	0.755	0.763	0.797
PMR-Net+MSD-EMA	**0.520**	**0.602**	**0.369**	**0.420**	**0.752**	**0.827**	**0.671**	**0.768**	**0.786**	**0.838**

TABLE V
THE GENERALIZATION EVALUATION ON FOUR AVAILABLE MEDICAL IMAGE DATASETS. THE BEST VALUES ARE IN BOLD.

Methods \ Datasets	ISIC 2018		DRIVE		STARE		DSB2018	
	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$	$IoU\uparrow$	$Dice\uparrow$
PMR-Net [41]	0.811	0.886	0.650	0.787	0.606	0.749	0.817	0.888
PMR-Net+MSD-EMA	**0.823**	**0.893**	**0.665**	**0.797**	**0.614**	**0.756**	**0.823**	**0.894**

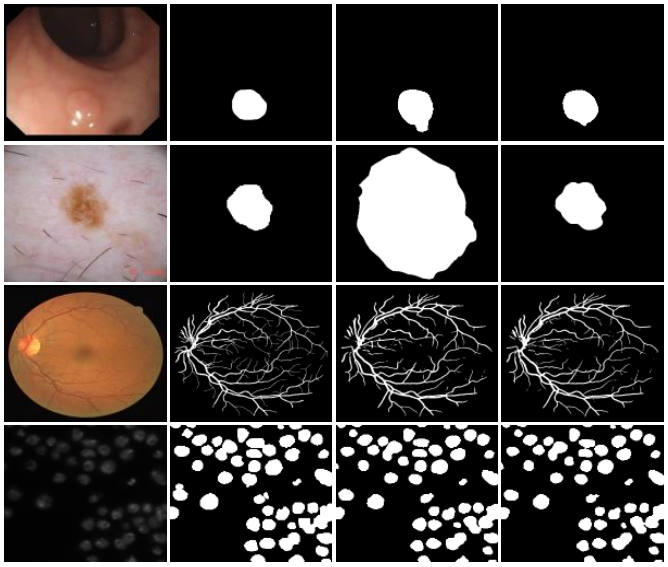


Fig. 7. The visualization results of generalization evaluation of MSD-EMA.

MSD-EMA between the encoder and decoder of PMR-Net. These different insertion positions bring different improvements to the segmentation performance of the network. Therefore, MSD-EMA has good plug-and-play performance.

The MSD-EMA lies in its excellent plug-and-play performance, providing tremendous convenience for existing polyp segmentation networks. MSD-EMA can be easily embedded into existing polyp segmentation networks without structural adjustments. Researchers can fine-tune MSD-EMA according to specific task requirements to achieve better performance without delving into the details of the module. This enables researchers to achieve performance improvement with minimal effort while maintaining the original network framework.

MSD-EMA is suitable for various polyp segmentation tasks, whether targeting specific disease types or diverse datasets. This generalization ensures that the MSD-EMA can demonstrate its excellent performance in different polyp application scenarios, making it a widely applicable tool.

More importantly, as a plug-and-play module, MSD-EMA can achieve high accuracy in the polyp segmentation task and be widely applied to other segmentation tasks such as retinal vessel, skin lesion, and cell nuclei, etc.

VI. CONCLUSION

In this work, we have proposed MSD-EMA and applied it to the polyp segmentation task. Firstly, we have designed the D-EMA module, which represents attention weights as the sum of the pairwise term representing the intra-class relationships of polyps and the unary term representing boundary information of polyps. The new attention feature maps include both polyp features as well as features between polyp and surrounding tissues, thus effectively extracting polyps from similar surroundings. Secondly, we have proposed the PCS for constructing multi-scale attention feature maps based on multi-branch D-EMA. By fusing attention feature maps with different sparsities, the attention on segmented objects with different shapes and sizes has been improved, while the attention on unrelated regions has been reduced, resulting in accurately localizing and segmenting objects. We have applied the proposed MSD-EMA to classical segmentation networks and achieve more accurate segmentation results on publicly available polyp segmentation datasets. More importantly, MSD-EMA is a universal plug-and-play attention

module that can be widely applied in many popular medical image segmentation networks.

In future work, we will further investigate the reduction of the parameters of MSD-EMA as a plug-and-play lightweight module without sacrificing performance. In addition, we will continue to explore the application of MSD-EMA in other image segmentation tasks.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, et al. "Cancer statistics," *CA: A Cancer Journal for Clinicians*, 72(1): 7-33, 2022.

[2] N. H. Kim, Y. S. Jung, W. S. Jeong, et al. "Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies," *Intestinal research*, 15(3): 411, 2022.

[3] D. Fan, G. Ji, T. Zhou, et al. "PraNet: parallel reverse attention network for polyp segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 263-273, 2020.

[4] D. Jha, S. Ali, N. K. Tomar, et al. "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *IEEE Access*, 9, 40496-40510, 2021.

[5] Y. Shen, X. Jia, M. Meng. "HRENet: a hard region enhancement network for polyp segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 559-568, 2021.

[6] X. Li, Z. Zhong, J. Wu, et al. "Expectation-maximization attention networks for semantic segmentation," *IEEE International Conference on Computer Vision*, 9167-9176, 2019.

[7] M. Yin, Z. Yao, Y. Cao, et al. "Disentangled non-local neural networks," *European Conference on Computer Vision*, 191-207, 2020.

[8] O. Ronneberger, P. Fischer, T. Brox. "U-Net: convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234-241, 2015.

[9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, et al. "UNet++: a nested U-Net architecture for medical image segmentation," *IEEE Transactions on Medical Imaging*, 39(6): 1856-1867, 2019.

[10] R. Wang, T. Lei, Y. Wan, et al. 'Medical image segmentation using deep learning: a survey," *IET Image Processing*, 16(5): 1243-1267, 2022.

[11] T. Lei, D. Zhang, X. Du, et al. "Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network," *IEEE Transactions on Medical Imaging*, 42(5): 1265-1277, 2023.

[12] T. Lei, R. Sun, X. Du, et al. "SGU-Net: shape-guided ultralight network for abdominal image segmentation," *IEEE Journal of Biomedical and Health Informatics*, 27(3): 1431-1442, 2023.

[13] X. Wang, R. Girshick, A. Gupta, et al. "Non-local neural networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 7794-7803, 2018.

[14] A. Srivastava, D. Jha, S. Chanda, et al. "MSRF-Net: a multi-scale residual fusion network for biomedical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, 26(5): 2252-2263, 2021.

[15] D. Jha, P. H. Smedsrud, M. A. Riegler, et al. "ResUNet++: an advanced architecture for medical image segmentation," *IEEE International Symposium on Multimedia*, 225-230, 2019.

[16] G. Ji, Y. Chou, D. Fan, et al. "Progressively normalized self-attention network for video polyp segmentation," *International Conference on Medical Image Computing and Computer Assisted Intervention*, 142-152, 2021.

[17] K. Park, J. Lee. "SwinE-Net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer," *Journal of Computational Design and Engineering*, 9(2): 616-632, 2022.

[18] M. M. Rahman, R. Marculescu. "Medical image segmentation via cascaded attention decoding," *IEEE Winter Conference on Applications of Computer Vision*, 6222-6231, 2023.

[19] T. Kim, H. Lee, D. Kim. "UACANet: uncertainty augmented context attention for polyp segmentation," *ACM International Conference on Multimedia*, 2167-2175, 2021.

[20] A. Lou, S. Guan, H. Ko, et al. "CaraNet: context axial reverse attention network for segmentation of small medical objects," *Medical Imaging: Image Processing*, 81-92, 2022.

[21] T. C. Nguyen, T. P. Nguyen, G. H. Diep, et al. "CCBANet: cascading context and balancing attention for polyp segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 633-643, 2021.

[22] R. Zhang, P. Lai, X. Wan, et al. "Lesion-aware dynamic kernel for polyp segmentation," *International Conference on Medical Image Computing and Computer Assisted Intervention*, 18-22, 2022.

[23] R. Zhang, G. Li, Z. Li, et al. "Adaptive context selection for polyp segmentation," *International Conference on Medical Image Computing and Computer Assisted Intervention*, 253-262, 2020.

[24] N. K. Tomar, D. Jha, U. Bagci, et al. "TGANet: text-guided attention for improved polyp segmentation," *International Conference on Medical Image Computing and Computer Assisted Intervention*, 151-160, 2022.

[25] J. Wei, Y. Hu, R. Zhang, et al. "Shallow attention network for polyp segmentation," *International Conference on Medical Image Computing and Computer Assisted Intervention*, 699-708, 2021.

[26] Z. Gu, J. Cheng, H. Fu, et al. "CE-Net: context encoder network for 2D medical image segmentation," *IEEE Transactions on Medical Imaging*, 38(10): 2281-2292, 2019.

[27] C. Huang, H. Wu, Y. Lin. "HarDNet-MSEG: a simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps," *arXiv*, 2101.07172, 2021.

[28] M. Yang, K. Yu, C. Zhang, et al. "DenseASPP for semantic segmentation in street scenes," *IEEE Conference on Computer Vision and Pattern Recognition*, 3684-3692, 2018.

[29] Z. Qiu, Z. Wang, M. Zhang, et al. "BDG-Net: boundary distribution guided network for accurate polyp segmentation," *Medical Imaging: Image Processing*, 792-799, 2022.

[30] M. Cheng, Z. Kong, G. Song, et al. "Learnable oriented-derivative network for polyp segmentation," *International Conference on Medical Image Computing and Computer Assisted Intervention*, 720-730, 2021.

[31] J. Zhong, W. Wang, H. Wu, et al. "PolypSeg: An efficient context-aware network for polyp segmentation from colonoscopy videos," *International Conference on Medical Image Computing and Computer Assisted Intervention*, 285-294, 2020.

[32] J. Wang, Q. Huang, F. Tang, et al. "Stepwise feature fusion: local guides global," *International Conference on Medical Image Computing and Computer Assisted Intervention*, 110-120, 2022.

[33] Q. Liu, Q. Dou, L. Yu, et al. "MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data," *IEEE Transactions on Medical Imaging*, 39(9): 2713-2724, 2020.

[34] K. Pogorelov, K. R. Randel, C. Griwodz, et al. "Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection," *ACM Multimedia Systems Conference*, 164-169, 2017.

[35] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, et al. "Wm-DOVA maps for accurate polyp highlighting in colonoscopy validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, 43: 99-111, 2015.

[36] J. Silva, A. Histace, O. Romain, et al. "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, 9(2): 283-293, 2014.

[37] N. Tajbakhsh, S. R. Gurudu, J. Liang. "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, 35(2): 630-644, 2015.

[38] D. Vázquez, J. Bernal, F. J. Sánchez, et al. "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, 2017(1), 2017.

[39] K. He, X. Zhang, S. Ren, et al. "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, 770-778, 2016.

[40] G. Yue, S. Li, R. Cong, et al. "Attention-guided pyramid context network for polyp segmentation in colonoscopy images," *IEEE Transactions on Instrumentation and Measurement*, 72: 1-13, 2023.

[41] X. Du, D. Gu, T. Lei, et al. "PMR-Net: parallel multi-resolution encoder-decoder network framework for medical image segmentation," *arXiv*, 2409.12678, 2024.

[42] J. Staal, M. D. Abramoff, M. Niemeijer, et al. "Ridge-based vessel segmentation in color images of the retina," *IEEE Transactions on Medical Imaging*, 23(4): 501-509, 2004.

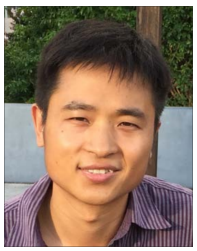
[43] A. D. Hoover, V. Kouznetsova, M. Goldbaum. "Locating blood vessels in retinal images by piece-wise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging*, 19(3): 203-210, 2000.

[44] N. Codella, V. Rotemberg, P. Tschandl, et al. "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," *arXiv*, 1902.03368, 2018.

[45] J. C. Caicedo, A. Goodman, K. W. Karhohs, et al. "Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl," *Nature Methods*, 16(12): 1247-1253, 2019.

Xiaogang Du received the Ph.D. degree from Lanzhou Jiaotong University, China, in 2017. He is currently an Associate Professor with the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology. He has authored or coauthored more than 30 research articles, including IEEE Transactions on Medical Imaging (TMI), IEEE Journal of Biomedical and Health Informatics (JBHI), etc. His research interests include computer vision, machine learning, and deep learning.

Yibin Zou received his B.Eng. degree in Computer Science and Technology from Shaanxi University of Science and Technology, Xi'an, China. He is currently pursuing an M.Eng. degree in Computer Technology at the same institution. His research interests include medical image processing, deep learning, and machine learning.



Tao Lei (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2011. From 2012 to 2014, he was a Post-Doctoral Research Fellow with the School of Electronics and Information, Northwestern Polytechnical University. From 2015 to 2016, he was a Visiting Scholar with the Quantum Computation and Intelligent Systems Group, University of Technology Sydney, Sydney, NSW, Australia. He is currently a Professor with the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an. He has authored or coauthored more than 80 research articles, including IEEE Transactions on Image Processing (TIP), IEEE Transactions on Fuzzy Systems (TFS), IEEE Transactions on Geoscience and Remote Sensing (TGRS). His research interests include image processing, pattern recognition, and machine learning.

Dongxin Gu obtained a master's degree from the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, from 2020 to 2023. His primary research focused on medical image segmentation based on deep learning. After graduation, he joined the Radar Systems Department of Shaanxi Yellow River Group Co., Ltd., where he is engaged in the development and application of radar systems. His research interests include medical image processing, deep learning, and intelligent system integration.

Xuejun Zhang received the PhD degree in computer science and technology from Xi'an Jiaotong University, China, in 2016. He is currently a professor with the School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, China. He is a Senior Member of CCF and an ACM Member. His main research interests include data privacy and machine learning, network security, and edge computing.

Asoke K. Nandi (Life Fellow, IEEE) received the degree of Ph.D. in Physics from the University of Cambridge (Trinity College), Cambridge (UK), in 1978.

He held academic positions in several universities, including Oxford (UK), Imperial College London (UK), Strathclyde (UK), and Liverpool (UK) as well as Finland Distinguished Professorship in Jyvaskyla (Finland). In 2013, he moved to Brunel University London (UK), to become the Chair and Head of Electronic and Computer Engineering. Professor Nandi is a Distinguished Visiting Professor at Xi'an Jiaotong University (China). In 1983 Professor Nandi co-discovered the three fundamental particles known as W^+, W^- and Z^0 (by the UA1 team at CERN), providing the evidence for the unification of the electromagnetic and weak forces, for which the Nobel Committee for Physics in 1984 awarded the prize to his two team leaders for their decisive contributions. His current research interests lie in signal processing and machine learning, with applications to communications, image segmentations, biomedical data, etc. He has made many fundamental theoretical and algorithmic contributions to many aspects of signal processing and machine learning. He has much expertise in "Big and Heterogeneous Data", dealing with modelling, classification, estimation, and prediction. He has authored over 600 technical publications, including 270 journal papers as well as five books, entitled Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines (Wiley, 2020), Automatic Modulation Classification: Principles, Algorithms and Applications (Wiley, 2015), Integrative Cluster Analysis in Bioinformatics (Wiley, 2015), Blind Estimation Using Higher-Order Statistics (Springer, 1999), and Automatic Modulation Recognition of Communications Signals (Springer, 1996). The H-index of his publications is 80 (Google Scholar) and his ERDOS number is 2.

Professor Nandi is a Fellow of the Royal Academy of Engineering (UK) as well as a Fellow of seven other institutions, including the IEEE and the IET. Among the many awards he received are the Institute of Electrical and Electronics Engineers (USA) Heinrich Hertz Award in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers (UK) in 1999, and the Mountbatten Premium, Division Award of the Electronics and Communications Division, of the Institution of Electrical Engineers (UK) in 1998. Professor Nandi is an IEEE EMBS Distinguished Lecturer (2018-19).