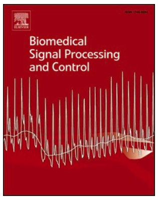
Biomedical Signal Processing and Control 96 (2024) 106487

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



DEMF-Net: A dual encoder multi-scale feature fusion network for polyp segmentation

Xiaorui Cao [a], He Yu [a,b,*], Kang Yan [a], Rong Cui [a], Jinming Guo [a], Xuan Li [a], Xiaoxue Xing [a,b,c], Tao Huang [d,*]

[a] *College of Electronic and Information Engineering, Changchun University, Changchun 130022, China*
[b] *Key Laboratory of Intelligent Rehabilitation and Barrier-free for the Disabled Ministry of Education, Changchun University, Changchun 130022, China*
[c] *Jilin Provincial Key Laboratory of Human Health Status Identification and Function Enhancement, Changchun University, Changchun 130022, China*
[d] *College of Science and Engineering, James Cook University, Cairns, 4870 QLD, Australia*

ARTICLE INFO

Keywords:
Polyp segmentation
Feature fusion
Swin Transformer
Medical image segmentation

ABSTRACT

Colorectal cancer is a common malignant tumour of the gastrointestinal tract. Studies have shown that colonoscopy can be an effective screening method for detecting colon polyps and removing them to prevent the development of colorectal cancer. In this study, we propose a new approach called the Dual Encoder Multi-Scale Feature Fusion Network (DEMF-Net). This approach uses a dual-scale Swin Transformer and CNN as an encoder to extract semantic features at different scales. In order to enhance the marginal characteristics of irregular polyps and improve the polyp detection rate, we propose a Dual-Branch Attention Fusion Module (DAF) that captures different shapes of target features through the attention mechanism and assigns higher weights to feature channels with high contributions. Additionally, we use an Advanced Feature Fusion Module (AFFM) to establish long-range dependencies and strengthen the target region to ensure that the high-level semantic features of polyps are not lost. We also propose Characterization Supplementary Blocks (CSB) for colorectal polyp images with irregular shapes and unclear boundaries to capture the structure and details of images and enhance model accuracy. We conducted experiments on five widely adopted polyp datasets and showed that our method achieved superior results in terms of both segmentation accuracy and edge details.

1. Introduction

Colorectal cancer (CRC), a malignancy originating from colorectal polyps, has garnered significant attention due to its prevalence and potential preventability [1]. The primary preventive measure against CRC is timely colorectal endoscopy, which facilitates the detection and removal of precancerous polyps [2,3]. However, traditionally, polyp detection depended solely on manual observation by endoscopists, which is a process that could result in errors and oversights [4]. Fortunately, the integration of computer-aided diagnostic systems in polyp detection promises enhanced accuracy in segmentation, thereby assisting physicians in diagnosis and augmenting the polyp detection rate during endoscopies, which can make effective planning for subsequent treatment possible [5].

Traditional polyp segmentation methods rely on shallow features like shape [6], texture [7], and Simple Linear Iterative Clustering (SILC) [8] for initial image segmentation, yet these techniques often suffer from low accuracy. The advent and progression of Deep Learning (DL) [9] have ushered in Convolutional Neural Networks (CNNs) with robust feature representation capabilities, markedly improving colorectal polyp segmentation. Ronneberger *et al.* [10] introduced U-Net, a pivotal model in medical image segmentation. U-Net's architecture comprises an encoding component for feature extraction and a decoding component employing deconvolution for upsampling, ensuring the final segmentation matches the input image size for precise pixel-level segmentation. Building on U-Net, several variants have been developed, including UNet++ [11], Res-Unet [12] and nnUNet [13], each enhancing polyp segmentation. However, despite CNNs' decade-long success, they exhibit inherent limitations: their convolutional kernels effectively capture local features [14,15] but struggle with long-range dependencies, limiting their ability to model global contexts effectively.

The Transformer framework [16], initially prominent in natural language processing (NLP), has recently made a significant impact as an alternative to CNNs, particularly due to its Multi-Head Self-Attention

* Corresponding authors.
E-mail addresses: yuh82@ccu.edu.cn (H. Yu), tao.huang1@jcu.edu.au (T. Huang).

https://doi.org/10.1016/j.bspc.2024.106487
Received 10 January 2024; Received in revised form 9 May 2024; Accepted 23 May 2024
Available online 31 May 2024

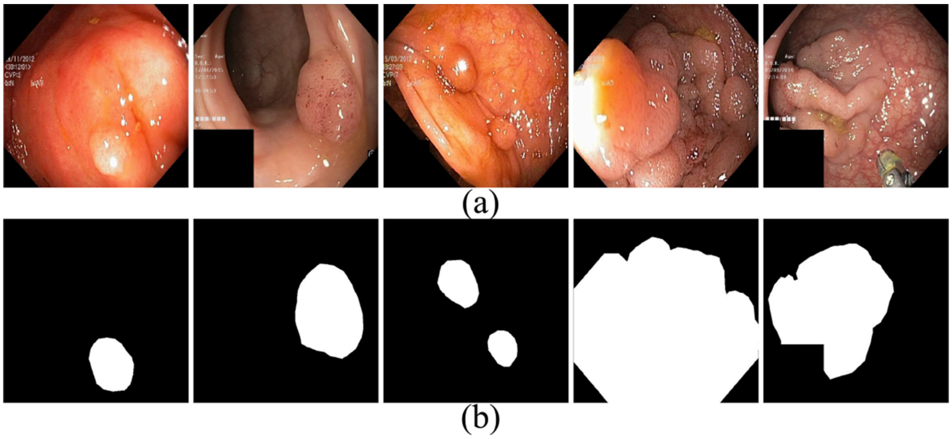


Fig. 1. (a) The original polyp images exhibit varying sizes, blurred boundaries, and weak contrast between the lesion and the background. (b) Ground truth map for polyp segmentation.

mechanism (MSA), which excels in capturing long-range dependencies [17]. This framework's adaptation to medical image segmentation marks a noteworthy trend. The Vision Transformer (ViT) [18] initially showcased its efficacy in image classification by using the Transformer as an encoder, thus facilitating the framework's shift from NLP to image processing tasks. Presently, Transformer-based models are demonstrating considerable promise in medical image segmentation.

One such model, TransUNet [19], integrates Transformer blocks into the U-Net [10] structure, marking its first application in medical segmentation and yielding enhanced results in multi-organ and cardiac segmentation. TransFuse [20] takes a dual-fusion approach, amalgamating features from both Transformer and CNN pathways, and has achieved notable success in polyp segmentation tasks. Swin-UNet [21], built on the Swin Transformer [22], employs Transformers exclusively in both encoder and decoder sections, showing significant efficacy in multi-organ and cardiac segmentation. Lastly, DSTransUNet [23] implements a dual-branch Swin Transformer for encoding, enabling feature extraction at varying scales, further advancing the field of medical image segmentation.

Additionally, multiscale representation has been proven to play a crucial role in Vision Transformers (ViT), as it helps improve the accuracy and robustness of segmentation results by extracting feature information from images at different scales. CrossViT [24] proposed a novel dual-branch Transformer architecture to combine image patches (tokens in a Transformer) of different sizes, producing stronger image features. Zhang *et al.*[25] employed two techniques for encoding high-resolution images. One is the multiscale model structure, which provides multiscale image encoding at a controllable computational cost. Multiscale Vision Transformer (MViT) [26] is used for video and image recognition, pioneering the connection of multiscale feature hierarchies with Transformer models.

Despite the Transformer's [16] adeptness at capturing long-range dependencies in images, its methodology of transforming image sequences into one-dimensional tokens tends to diminish emphasis on positional information. This is a crucial limitation, particularly for two-dimensional medical image segmentation, where local feature extraction is vital. Furthermore, the inherent weak inductive bias of Transformers may impede their effectiveness when applied to large-scale datasets. This poses a significant challenge in the medical domain, where datasets are often diverse yet limited in scale, necessitating models that can efficiently learn from smaller data.

This paper introduces the Dual Encoder Multi-Scale Feature Fusion Network (DEMF-Net), a new network that uses CNN and Transformer for colorectal polyp segmentation and extracts multi-scale features, especially for enhanced segmentation of irregularly shaped colorectal polyps with poorly defined boundaries, as exemplified in Fig. 1. It is worth noting that DEMF-Net has significant advantages compared to previous methods: 1) The dual-branch parallel encoder structure of DEMF-Net simultaneously extracts both local and global features of polyp images, enhancing the network's ability to capture both local and global information. 2) Multi-scale feature extraction enhances the network's scale-awareness, allowing it to capture the overall shape and contours of targets at larger scales, while providing more detailed local information at smaller scales, thereby improving segmentation accuracy. Additionally, using multi-scale features also enhances the algorithm's robustness. 3) The proposed Dual-Branch Attention Fusion Module (DAF) naturally integrates attention mechanisms to fuse the implicitly extracted long-range relationships by the multi-scale Swin Transformer. Inspired by the Squeeze-and-Excitation (SE) network, DAF improves feature fusion quality by assigning larger weights to channels containing richer features, strengthening the fuzzy edges of irregular polyps, and increasing the detection rate of smaller polyps on the same image. 4) The Advanced Feature Fusion Module (AFFM) aims to balance the network's horizontal and vertical dimensions, enhancing the network's ability to handle spatial information of polyps and ensuring that high-level semantic information of polyp images is not lost. 5) The Feature Supplement Block (CSB) is designed to prevent the loss of polyp edge features during upsampling, thereby improving the accuracy of polyp detection. These advantages collectively contribute to the effectiveness of DEMF-Net in polyp segmentation tasks.

Our contributions are as follows:

- **Parallel Multi-Scale Encoding Structure**: We propose a parallel encoding structure combining CNN and a multi-scale Swin Transformer. This approach mitigates CNN's limitations in modeling long-range dependencies for global context while also addressing the Swin Transformer's shortcomings in local feature extraction.
- **Dual-Branch Attention Fusion Module (DAF)**: The DAF is designed to fuse features extracted by the dual-branch Swin Transformer using self-attention mechanisms, effectively enhancing the contrast between lesions and background.
- **Advanced Feature Fusion Module (AFFM)**: AFFM leverages different kernel sizes to implicitly establish long-distance dependencies at the CNN level. This deepens the network and augments model performance. Additionally, the Characterization Supplementary Blocks (CSB) are implemented to replenish the lost polyp edge features and detail features during upsampling, thus refining the segmentation results.
- **Empirical Evaluation**: The proposed model was rigorously evaluated on five widely recognized polyp datasets. The outcomes demonstrate that our method attains superior segmentation accuracy

and edge detail, indicating its efficacy in colorectal polyp segmentation.

2. Related works

This section provides an in-depth exploration of colorectal polyp segmentation, covering a spectrum of methodologies from traditional image processing to advanced machine learning and deep learning approaches. It emphasizes the evolution of techniques, particularly the role of CNNs and Transformer architectures, in enhancing segmentation accuracy and efficiency, offering a concise overview of the significant technological and methodological advancements in the field.

2.1. Traditional and machine learning approaches in polyp segmentation

Polyp segmentation techniques have traditionally been categorized into two main approaches: traditional image algorithms and machine learning methods. The former includes threshold-based segmentation [27], edge detection [6,28], and region segmentation [29,30], which predominantly utilize characteristics like color, texture, and shape. On the other hand, machine learning-based methods are adept at extracting color and texture features from images for polyp segmentation. Li *et al.* [31,32] have advanced this approach by mapping polyp image features to higher dimensions, facilitating segmentation through machine-driven learning. Similarly, Maghsoudi *et al.* [8] have developed a method to transform pixels into vector form, grouping segments with similar features for more effective segmentation.

Despite these advancements, traditional methods for colorectal polyp segmentation often depend heavily on personal expertise and manual feature selection. This reliance introduces considerable uncertainty and generally results in lower accuracy, highlighting the need for more robust and automated segmentation techniques.

2.2. Advancements in deep learning for polyp segmentation

The rapid progress of deep learning has transformed medical image segmentation, notably polyp segmentation [33]. Akbari *et al.* [34] developed a fully convolutional neural network (FCN) that outperformed previous models in polyp segmentation, marking a substantial advancement in this field. Building upon this, Ronneberger *et al.* [10] introduced the U-Net model, notable for its incorporation of skip connections. This model has gained widespread adoption in medical image segmentation due to its exceptional performance, especially in polyp segmentation. Extending the capabilities of the Unet, Li *et al.* [35] developed DenseUNet, which integrates hybrid dense connections to further enhance segmentation.

In a similar vein, Ozan Oktay *et al.* [36] added an attention mechanism to the U-Net framework, utilizing gating signals to more precisely control the extraction of spatial position features and adjust output features, thereby refining segmentation performance. The UNet++ [37] model further evolved this approach, proposing nested dense skip connections to improve image segmentation performance. PraNet [38] introduced a reverse attention mechanism to enhance polyp segmentation accuracy by mining edge cues and establishing connections between regions and boundary cues. Brandao *et al.* [39] also contributed to this evolution by introducing a fully convolutional neural network that surpassed existing polyp detection methods in segmentation accuracy.

Fang *et al.* [40] proposed a shared encoder and mutually constrained decoder, characterized by area and boundary constraints, for predicting polyp regions and boundaries. Murugesan *et al.* [41] introduced Psi-Net, featuring a single encoder and three parallel decoders for mask prediction, contour detection, and distance map estimation, aiding in capturing boundary and shape information of polyps. Wei J. *et al.* [42] developed SANet, designing shallow attention modules and color-swapping operations to focus more on the contour and structure of polyps, thus filtering shallow features and enhancing segmentation quality. Pan *et al.*[43] designed GLSNet, which fully considered the negative impact of background noise in different levels of features, and effectively improved the accuracy of polyp segmentation through feature enhancement and feature fusion modules. Finally, Taehun Kim *et al.* [2] introduced UACANet, employing additional encoders and decoders to compute saliency maps in each module, predicting the flow from bottom to top and propagating them to subsequent prediction modules, effectively improving the accuracy of polyp segmentation.

2.3. Transformer encoding architectures in medical image segmentation

The Transformer architecture has recently exhibited exceptional performance in various computer vision tasks [18,22,44,45]. Pioneering this advancement, Dosovitskiy [18] and others researchers [19,20] effectively transitioned the Transformer from NLP to computer vision applications, introducing the Vision Transformer (ViT). The ViT approach involves dividing images into multiple patches, each projected into a fixed length and subsequently processed by the Transformer for image classification. The Swin Transformer [22] further innovates with its Window-based Multi-Head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA), which facilitate reduced computational load and enhanced information transmission across adjacent windows, respectively, thereby significantly enhancing visual task performance.

In the context of image segmentation, Polyp-PVT [46] employs Transformer as an encoder to learn features and additionally designs three standard modules (cascade fusion module, camouflage recognition module, similarity aggregation module) to effectively suppress noise in features and enhance the model’s expressive power.

TransUNet [19] integrates Transformer blocks with the UNet architecture, yielding impressive results. nnFormer [47] introduces a novel 3D Transformer, utilizing skip attentions to replace traditional operations in skip connections. HTC-Net [48] excels in lesion segmentation by designing the trident multi-layer fusion module and united attention. Swin UNet [21] also employs the Swin Transformer as an encoder and UNet as a decoder for effective feature information restoration. Despite these innovations, a common limitation of Transformers is their tendency to neglect the crucial role of local spatial information in medical image segmentation.

2.4. Transformer and CNN hybrid encoding architectures in medical image segmentation

Although CNN is widely used in polyp segmentation tasks, it is difficult to make greater breakthroughs in image segmentation because it is good at capturing local information and ignoring spatial context information and global information. However, Transformer can effectively capture long-distance dependencies due to its self-attention mechanism, but lacks local feature extraction. Therefore, with the in-depth understanding of Transformer and CNN knowledge in the field of image segmentation, a new type of polyp segmentation framework has emerged that combines Transformer and CNN in various settings.

TransFuse [20] amalgamates CNN and Transformer branches, achieving superior performance in medical image segmentation. SwinE-Net [49] effectively combines the CNN-based EfficientNet and ViT-based Swin Transformer models, complements the spatial features and global semantic features, maintains the global semantics without sacrificing the low-level features of CNN, and effectively improves the segmentation robustness and accuracy. MIA-Net [50] is a multi-information aggregation network that combines Transformer and convolutional features, combining the advantages of long-distance Transformer modeling with the advantages of convolutional attention to adjacent pixel relationships, and the model also recognizes features through feature extraction module and multi-information fusion module, effectively improving the segmentation accuracy. Hybrid Semantic Networks (HSNet) [51], which combine CNNs and Transformer to capture local

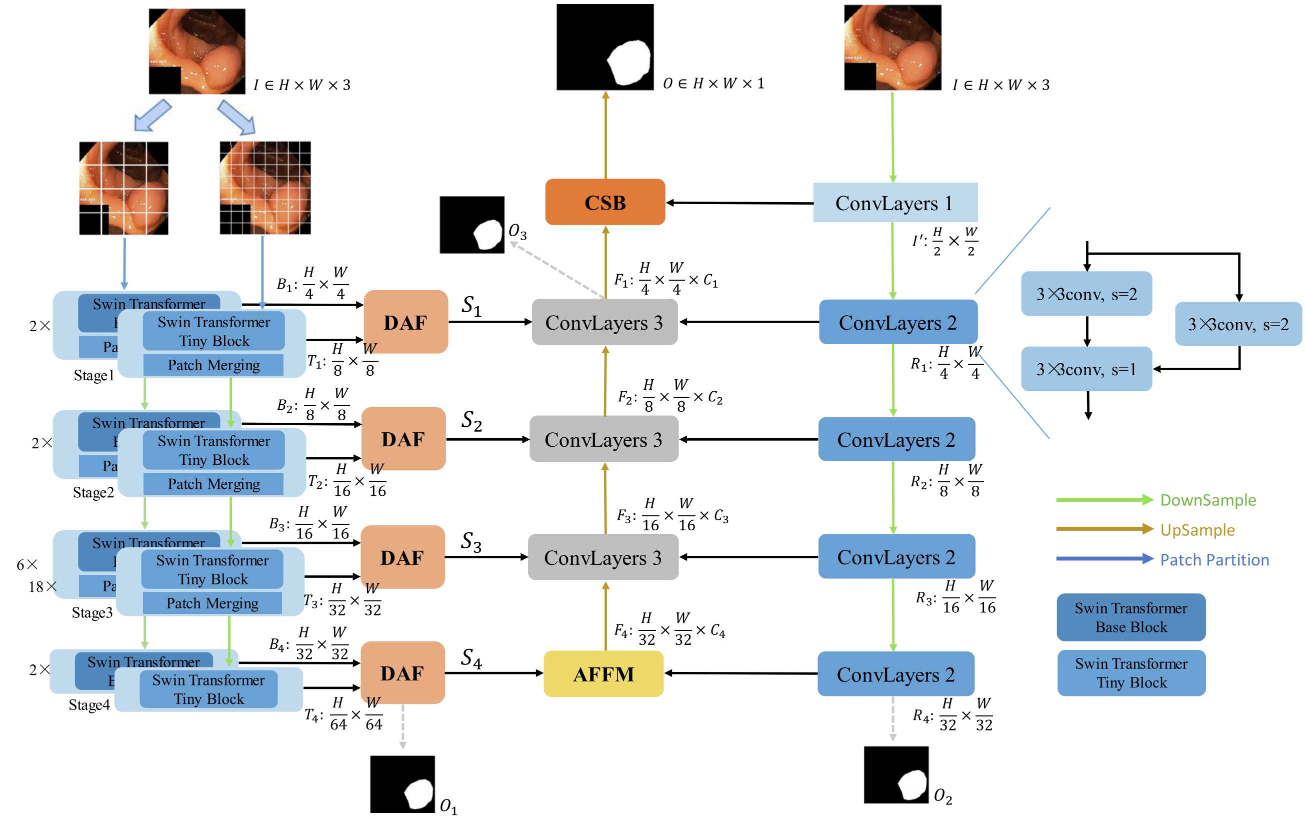

$I \in H \times W \times 3$
$O \in H \times W \times 1$
$I \in H \times W \times 3$
CSB
ConvLayers 1
O_3
I': $\frac{H}{2} \times \frac{W}{2}$
B_1: $\frac{H}{4} \times \frac{W}{4}$
S_1
F_1: $\frac{H}{4} \times \frac{W}{4} \times C_1$
2×
Swin Transformer
Swin Transformer Tiny Block
Patch Merging
Stage1
DAF
ConvLayers 3
ConvLayers 2
T_1: $\frac{H}{8} \times \frac{W}{8}$
R_1: $\frac{H}{4} \times \frac{W}{4}$
3×3conv, s=2
3×3conv, s=2
3×3conv, s=1
B_2: $\frac{H}{8} \times \frac{W}{8}$
S_2
F_2: $\frac{H}{8} \times \frac{W}{8} \times C_2$
Stage2
T_2: $\frac{H}{16} \times \frac{W}{16}$
R_2: $\frac{H}{8} \times \frac{W}{8}$
DownSample
UpSample
Patch Partition
B_3: $\frac{H}{16} \times \frac{W}{16}$
S_3
F_3: $\frac{H}{16} \times \frac{W}{16} \times C_3$
6×
18×
Stage3
T_3: $\frac{H}{32} \times \frac{W}{32}$
R_3: $\frac{H}{16} \times \frac{W}{16}$
Swin Transformer Base Block
Swin Transformer Tiny Block
B_4: $\frac{H}{32} \times \frac{W}{32}$
S_4
F_4: $\frac{H}{32} \times \frac{W}{32} \times C_4$
Stage4
AFFM
T_4: $\frac{H}{64} \times \frac{W}{64}$
R_4: $\frac{H}{32} \times \frac{W}{32}$
O_1
O_2

Fig. 2. The overall architecture of DEMF-Net, including the dual encoders of Swin Transformer and CNN for feature extraction, the Advanced Feature Fusion Module (AFFM) for extracting advanced features, the Dual-Branch Attention Fusion Module (DAF) for multi-scale feature fusion, and the Characterization Supplementary Blocks (CSB) for feature supplementation.

and long-distance features in polyps, respectively. ColonFormer [52] used a lightweight Transformer as an encoder to learn multi-scale features, and CNN as a decoder to learn to extract feature maps from different scales and regions, which improved the segmentation accuracy of small polyps.

2.5. *Analysis of previous work*

Previous methods in the field of medical image segmentation have achieved advanced results, but there are still some significant limitations. Firstly, traditional medical image segmentation methods such as Sobel, Prewitt, and Laplacian are mainly suitable for images with certain edge and texture information, lacking sufficient representational power and efficiency [53]. It is noteworthy that when segmenting images with significant differences, simply using one algorithm often cannot achieve satisfactory results, requiring the combination of multiple algorithms for comprehensive processing, which makes it difficult to generalize extensively to other datasets. Secondly, for CNNs, the inherent limitations of convolutional kernels result in a very limited receptive field, making CNNs proficient only in capturing local information of images, rather than effectively modeling long-range dependencies in images. Thirdly, for Transformers, although they have successfully addressed the issues of long-range dependencies and parallel training in traditional sequence models, they neglect the local information of images, leading to the loss of crucial detailed information during segmentation. Additionally, Transformers excel on large-scale datasets, while medical datasets are often diverse and limited in size. Finally, for hybrid models combining Transformers and CNNs, they mostly focus on the advantages brought by the methods themselves, without fully considering the interactions during feature extraction, often relying on single-scale or single-method feature extraction approaches.

Although previous methods have achieved decent results, there has been a lack of proposed medical image segmentation methods that integrate multi-scale feature extraction and attention mechanism-based fusion. To solve this problem, our study introduces a dual-encoder multi-scale feature fusion network (DEMF-Net), which is specifically used for polyp segmentation.

3. Methods

3.1. *Overall architecture*

The proposed DEMF-Net architecture, depicted in Fig. 2, comprises four integral modules: the Encoder, Decoder, Dual-Branch Attention Fusion Module (DAF), Advanced Feature Fusion Module (AFFM), and the Characterization Supplementary Blocks (CSB). The Encoder employs a dual mechanism, integrating Swin Transformer and CNN, to extract global dependencies and local features from polyp images concurrently. The Decoder, using CNN, reconstructs the image to its original size. DAF utilizes self-attention mechanisms to amalgamate features of varying scales extracted by the dual-branch Swin Transformer at each layer. AFFM enhances multi-scale feature extraction by simultaneously employing convolutional blocks of various sizes, merging advanced features from both encoders to capture the structural and detailed aspects of the image. Lastly, CSB is designed to replenish polyp edge features lost in the upsampling process, bolstering the model's representational efficacy.

Given an input $I \in \mathbb{R}^{H \times W \times C}$, C represents the default channel number of the image, which is set to 3. Firstly, we utilize the Swin-B and Swin-T [22] backbone networks to extract eight features in total, two for each of the four layers, denoted as $B_i \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i}$, $T_i \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times \frac{3C_i}{4}}$,

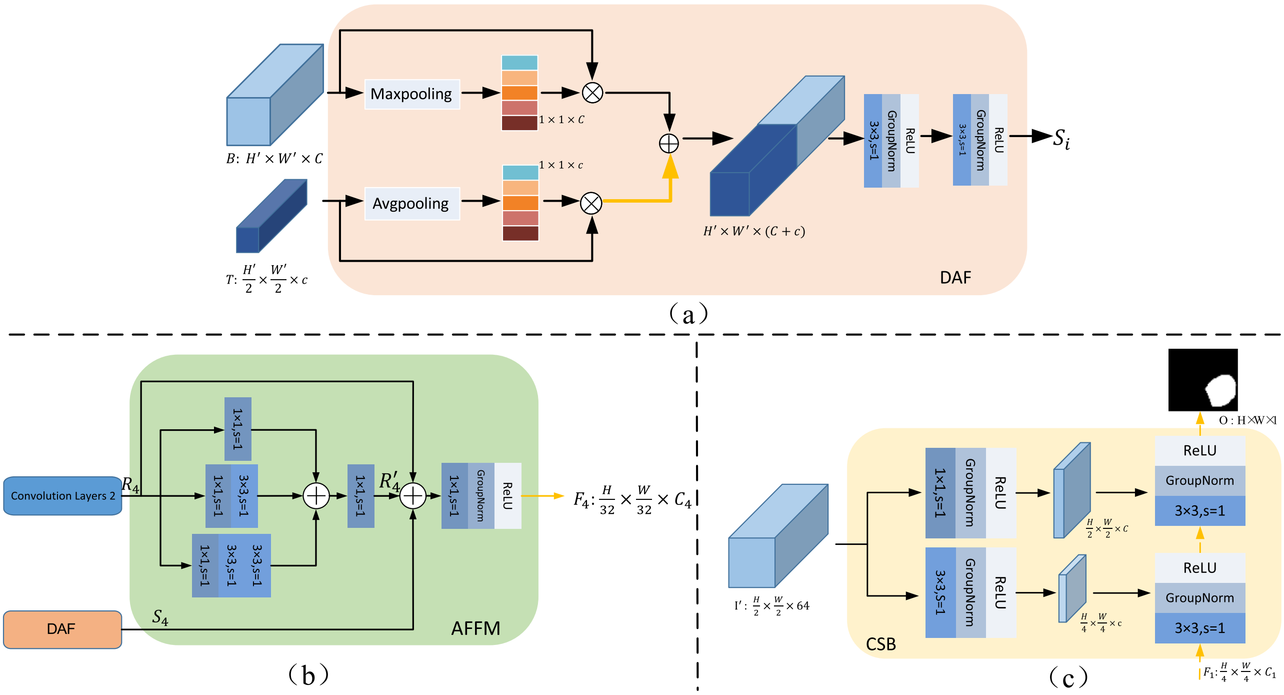


Fig. 3. Details of the introduced Dual-Branch Attention Fusion (DAF) module. (b) Details of the introduced Advanced Feature Fusion Module (AFFM). (c) Details of the introduced Characterization Supplementary Blocks (CSB).

where $C_i \in \{128, 256, 512, 1024\}$, and $i \in \{1, 2, 3, 4\}$. Secondly, we fuse the extracted features B_i and T_i from each layer using DAF, resulting in S_i. Meanwhile, another CNN's encoding branch outputs R_i through convolution. The advanced characteristics extracted by the last layer of Swin Transformer and CNN are fused through AFFM to obtain F_4. Other layer features S_i and R_i are aligned and fused to generate feature F_i. The high-level feature F_4 is combined with the previous level feature map through upsampling and convolution modules, and so on, until we obtain $F_1 \in \mathbb{R}^{16\times16\times448}$. F_1 is processed by CSB to yield the final predicted polyp segmentation result $O \in \mathbb{R}^{H\times W\times1}$. Detailed network module design is described in the following section.

3.2. Encoder

3.2.1. Swin Transformer encoder

The Swin Transformer [22] effectively establishes long-range connections that transcend the limitations of conventional convolutional kernels. However, its approach to patch segmentation overlooks the intricate inter-pixel structures within each patch. This segmentation strategy presents a trade-off: larger patches encapsulate more detailed information but at the expense of increased computational demand, whereas smaller patches enhance computational efficiency but risk losing detail. To navigate this dichotomy, we employ a multi-scale Swin Transformer for feature extraction. This approach allows patches of varying sizes to synergistically complement each other, effectively balancing the capture of detailed information with computational efficiency.

3.2.2. Convolutional encoder

The Convolutional Encoder in our architecture is founded on the principles of ResNet [54]. It efficiently captures local spatial features of polyp images by producing four layers of feature maps at varying scales. However, a challenge in deep CNNs, such as ResNet, is the rapid diminution of distant voxels, leading to the loss of vital local information. Moreover, using deep networks on small-scale medical image datasets frequently results in instability and a propensity for overfitting. To mitigate these issues, we have judiciously chosen a 9-layer convolutional operation for the convolutional encoding segment, striking a balance between depth and the preservation of crucial local details in the images.

To match the feature size extracted by the Swin Transformer encoding part, the input image $I \in \mathbb{R}^{H\times W\times C}$ is first downsampled by a stride of 2 using a 3 × 3 convolutional kernel in the CNN encoder, resulting in output $I' \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times64}$. This process is represented by the function F, which can be formulated as Eqn. (1):

$$I' = F(I) \tag{1}$$

The first-layer features output by the CNN are denoted as $R_1 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times128}$:

$$R_1 = F(I') + Conv_1(I') \tag{2}$$

The $Conv_2$ is composed of two 3×3 convolutional blocks. The specific outputs of the other layers can be formulated as Eqn. (3):

$$R_{i+1} = F(R_i) + Conv_1(R_{i+1}), i \in \{1, 2, 3\} \tag{3}$$

3.3. Decoder

A streamlined convolutional decoder is employed to reduce the computational demands of the model while effectively integrating features extracted by the parallel branches of the Swin Transformer and CNN. This approach ensures efficient feature fusion and yields robust segmentation outcomes. The success of this simplified decoder underscores the dual-branch encoder's proficiency in extracting rich and comprehensive features, affirming its effectiveness in the context of complex image segmentation tasks.

The fused feature F_i is obtained through size alignment and convolution of $S_i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$ obtained by DAF and $R_i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$ obtained by the convolutional encoding module:

$$F_i = Conv_{1\times1}(Conv_{3\times3}(concat(S_i + R_i))), i \in \{1, 2, 3\} \tag{4}$$

The fusion is required when i = 4, through the AFFM fusion.

3.4. Dual-Branch Attention Fusion Module

Upon extracting multi-scale features from the Dual-Branch Swin Transformer Encoder, it becomes evident that mere concatenation fails to harness these dual-scale features effectively. Additionally, the long-range dependency relations, implicitly modeled by the Swin Transformer, may not always produce optimal outcomes. Drawing inspiration from Squeeze-and-Excitation Networks (SE) [55], we introduce the Dual-Branch Attention Fusion (DAF) module. This module employs an attention mechanism to process the output features selectively. DAF enhances the overall feature quality by allocating greater weights to channels that encapsulate richer features. Subsequently, these high-quality features are amalgamated through convolution, ensuring a more effective fusion and improving the segmentation results.

As shown in Fig. 3(a), We apply max pooling to the features B_i from the Swin-B output before the Squeeze-and-Excitation Networks (SE) operation, and average pooling to the features T_i from the Swin-T output. After concatenating the two features, we effectively fuse the concatenated features using two convolutional kernels with a stride of 1 and a size of 3 × 3, resulting in S_i. The function P represents the process of attention selection, which includes Linear, ReLU, and Sigmoid operations. This process can be expressed as follows:

$$\begin{gathered} M = \mathrm{P}_{Maxpool}(B_i) + \mathrm{P}_{Avgpool}(T_i) \\ S_i = Conv_{3\times3}(Conv_{3\times3}(M_i)\,), i \in \{1,2,3,4\} \end{gathered} \tag{5}$$

3.5. Advanced Feature Fusion Module

Fig. 3(b) illustrates the strategic implementation of a 9-layer convolutional operation sequence in the Convolutional Encoder, designed to avert the loss of the effective receptive field, a common issue with shallow CNNs that are less capable of explicitly modeling long-range spatial information. Drawing inspiration from the Inception architecture [56], we incorporate a multi-scale convolution module following the final CNN output. This addition is pivotal for capturing the image's structural and detailed elements across various scales. It effectively balances the network's horizontal and vertical dimensions and synergistically merges the output with the high-level features derived from the DAF module. This integration enhances the network's ability to process and interpret complex spatial information, ensuring a more comprehensive and detailed feature representation.

First, the Convolutional Encoder outputs the final layer, $R_4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 512}$. It is simultaneously input into a three-branch network consisting of 1 × 1 and 3 × 3 convolutions.

The AFFM operation can be formulated as Eqn. (6):

3.6. Characterization Supplementary Blocks

Upon acquiring F_1 from the decoder, directly performing upsampling to generate the segmentation map would result in a significant loss of edge detail information. To address this problem, we have developed the CSB, tailored to enhance the segmentation image produced during the final decoding stage. This approach effectively preserves and supplements edge details, ensuring that the critical features of the segmentation map are maintained and accurately represented.

As shown in Fig. 3(c), After obtaining I' in the Convolutional Encoder, two convolutional blocks are cascaded in parallel to supplement the edge details of the segmented image, resulting in $A_1 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times c}$ and $A_2 \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$. Concatenating with the output F_1 of the decoder are processed through a convolutional layer to obtain O_1. Convolution of O_1 and A_2 to restore the segmentation image to an image with one channel. This process results in the final predicted polyp segmentation result, $O \in \mathbb{R}^{H\times W\times 1}$. This process can be expressed as follows:

$$\begin{gathered} O = Conv_{1\times1}(Conv_{3\times3}(concat(A_2 + O_1))) \\ O_1 = Conv_{3\times3}(concat(A_1 + F_1)) \end{gathered} \tag{7}$$

4. Experiments and results

We conducted experiments on five publicly available polyp datasets and compared them with other advanced models to evaluate the effectiveness of DEMF-Net.

4.1. Datasets

For our study, we utilized five publicly available polyp datasets, each with distinct characteristics:

- **CVC-ClinicDB**[57]**:** This dataset includes 612 polyp images with corresponding annotations, sourced from 29 different endoscopic video clips. The images have a resolution of 384 × 288 pixels.
- **Kvasir-SEG**[57]**:** Comprising 1000 polyp images, this dataset features a wide range of resolutions, from 332 × 487 to 1920 × 1072 pixels. The polyp regions within these images vary significantly in size and shape.
- **CVC-ColonDB**[58]**:** Collected from 15 different colonoscopic examination sequences, this dataset contains 380 polyp images, each with corresponding annotations, and uniform image dimensions of 574 × 500 pixels.

$$\begin{gathered} R_4' = Conv_{1\times1}[Conv_{1\times1}(R_4) + Conv_{3\times3}(Conv_{1\times1}(R_4)) + Conv_{3\times3}(Conv_{3\times3}(Conv_{1\times1}(R_4)))] + R_4, \\ F_4 = Conv_{1\times1}(concat(R_4' + S_4)) \end{gathered} \tag{6}$$

Table 1

Network parameters of Encoder.each module. The parameters for each Swin Transformer layer remain unchanged without any modifications. The parameters are [in-channel, out-channel, kernel size, stride].

Swin Transformer Encoder			CNN Encoder	
	Swin-Tiny	Swin-Base		
Patch_size	8	4	ConvLayer1	[3,64,3,2]
Embed_dims	96	128	ConvLayer2	[64,128,3,2]; [128,128,3,1]
num heads	[3,6,12,24]	[4,8,16,32]	ConvLayer2	[128,256,3,2]; [256,256,3,1]
Window size	7	7	ConvLayer2	[256,512,3,2]; [512,512,3,1]
depths	[2,2,6,2]	[2,2,18,2]	ConvLayer2	[512,512,3,2]; [512,512,3,1]
Drop_path_rate	0.2	0.5		

Table 2
Network parameters of each module.

AFFM		DAF		CSB	
Conv2d	[512,256,1,1]	Pooling	AvgPool2d	Conv2d	[64,32,1,1]
Conv2d	[256,256,3,1]	Pooling	Maxpool2d	Conv2d	[64,64,3,1]
Conv2d	[768,512,1,1]	Attention	SE	Upsample	Nearest
Conv2d	[736,128,1,1]	Conv2d	[726,224,3,1]	Concat	
UpSample	nearest	Conv2d	[224,224,3,1]	Conv2d	[32,1,3,1]

Table 3
Quantitative results of the learning capabilities, comparing the CVC-ClinicDB dataset and Kvasir-SEG dataset with previous state-of-the-art methods.

Encoder type	model	CVC-ClinicDB[57]				Kvasir-SEG [62]			
		Dice	mIoU	Precision	Recall	Dice	mIoU	Precision	Recall
CNN	U-Net [10]	0.832	0.750	0.887	0.851	0.812	0.722	0.855	0.835
	UNet++ [11]	0.715	0.602	0.814	0.773	0.724	0.633	0.836	0.769
	ResUNet++ [61]	0.795	0.796	0.878	0.702	0.813	0.793	0.877	0.706
	Focus U-Net [63]	0.941	0.893	0.956	0.930	0.910	0.845	0.917	0.916
	PraNet [38]	0.899	0.849	0.961	0.912	0.898	0.840	0.903	0.910
	DDANet [64]	0.941	0.893	0.951	0.936	0.857	0.780	0.864	0.888
	PolypSegNet [65]	0.915	0.862	0.961	0.911	0.887	0.826	0.917	0.925
	ACSNet[66]	0.882	0.826	0.923	0.911	0.898	0.838	0.902	0.930
	HarDNet-MSEG [67]	0.932	0.882	0.960	0.923	0.904	0.848	0.923	0.907
	SANet [68]	0.899	0.859	0.925	0.914	0.904	0.847	0.916	0.923
Transformer	nnFormer [47]	0.923	0.876	0.924	0.913	0.892	0.830	0.959	0.912
	Polyp-PVT [46]	0.937	0.889	N/A	N/A	0.917	0.864	N/A	N/A
	TransUnet [19]	0.935	0.887	N/A	N/A	0.913	0.856	N/A	N/A
Hybrid	TranSEFusionNet [69]	0.865	0.791	N/A	N/A	0.845	0.781	N/A	N/A
	CSwinDoubleU-Net [70]	0.931	0.883	**0.965**	0.872	0.910	0.844	0.903	0.854
	DS-TransUNet [23]	0.942	0.894	0.937	0.950	0.913	0.859	0.914	0.936
	MGCBFormer [71]	**0.955**	0.915	0.963	0.949	**0.931**	**0.885**	**0.955**	0.919
	Ens2[72]	0.935	0.893	N/A	N/A	0.927	0.883	N/A	N/A
	DEMF-Net (Ours)	0.953	**0.919**	0.962	**0.953**	0.915	0.860	0.913	**0.937**

- **EndoScene**[59]**:** Encompassing a total of 912 images with annotations, EndoScene is an amalgamation of the CVC-ClinicDB and CVC300 datasets, providing a diverse range of image types.
- **ETIS**[60]**:** This dataset includes 192 polyp images, along with their annotations, obtained from 29 different sequences of colonoscopic examinations. Each image in this dataset maintains a consistent size of 1225 × 996 pixels.

The selection of these datasets, with their varying resolutions and diverse image characteristics, allows for a comprehensive evaluation of the segmentation and training settings in our study.

4.2. *Implementation details and loss function*

4.2.1. *Implementation details*

Our methodology is implemented using the PyTorch framework, leveraging the computational power of the RTX 3080Ti GPU. The Swin Transformer Encoder, a key component of our approach, is initialized with pre-trained weights [22], ensuring a robust starting point for feature extraction. To accommodate the diverse dimensions of each polyp image, we adopted a multi-scale training strategy, opting for this approach over traditional data augmentation methods [38].

For the optimization process, the Stochastic Gradient Descent (SGD) optimizer is utilized to fine-tune the model parameters. Strategies such as early stopping and cosine annealing are employed to further enhance model performance. The learning rate is set at 0.001, and the model is trained over 60 epochs with a batch size of 2.

The effectiveness of our model is assessed using a suite of evaluation metrics: Dice coefficient, mean Intersection over Union (mIoU), Precision, and Recall. These metrics comprehensively evaluate the model's accuracy and reliability in segmenting polyp images. Unless otherwise specified, the hyperparameter settings for our experiments adhere to the parameters outlined above. Detailed network parameters are presented in Table 1 and 2.

4.2.2. *Loss function*

The loss function can be described as:

$$\begin{gathered} L_{total} = \alpha L(G, O_1) + \beta L(G, O_2) + \delta L(G, O_3) + \gamma L(G, O) \\ L = L_{IoU}^{W} + L_{BCE}^{W} \end{gathered} \tag{8}$$

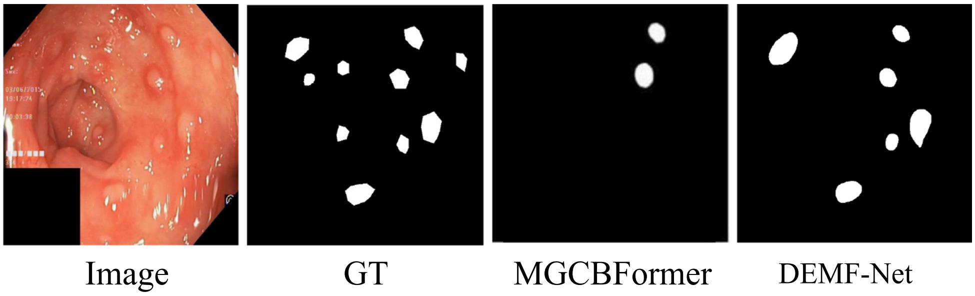


Fig. 4. Detail drawing of a small polyp.

Table 4
Quantitative results of generalization ability, with the best results highlighted in bold for each column.

model	CVC-ColonDB [58]		ETIS [60]		EndoScence [59]	
	Dice	mIoU	Dice	mIoU	Dice	mIoU
U-Net [10]	0.471	0.390	0.401	0.340	0.627	0.535
UNet++ [11]	0.318	0.253	0.297	0.247	0.428	0.357
ResUNet++ [61]	0.300	0.222	0.121	0.081	0.834	0.777
PraNet [38]	0.709	0.640	0.628	0.567	0.871	0.797
DDANet [64]	0.520	0.439	0.247	0.195	0.834	0.777
ACSNet [66]	0.716	0.649	0.578	0.509	0.863	0.787
HarDNet-MSEG [67]	0.731	0.660	0.677	0.613	0.887	0.821
SANet [68]	0.745	0.665	**0.750**	**0.654**	0.888	0.815
DEMF-Net (Ours)	**0.751**	**0.668**	0.736	0.635	**0.908**	**0.882**

Through both loss functions L^{W}_{IoU} and L^{W}_{BCE}, the training process is supervised to balance global constraints and pixel-level constraints, $\alpha, \beta, \delta, \gamma$ are defined as 0.1, 0.1, 0.2, and 0.6. Using a weighted loss function can optimize the model and improve its predictive capabilities. O_1 represents the output after the fusion of the fourth level of the dual-scale Swin Transformer, O_2 represents the output of the last level of the CNN encoder, O_3 represents the output of the first-level feature fusion module, and O represents the final predicted image.

4.3. Analysis of learning ability

Settings: To ascertain the learning efficacy of our model, we performed experiments on the CVC-ClinicDB and Kvasir-SEG datasets. For the CVC-ClinicDB dataset, images were resized to 384 × 384 pixels, with 90 % designated for training and the remaining 10 % for testing. The Kvasir-SEG dataset images were resized to 512 × 512 pixels, utilizing 900 images for training and 100 for testing.

Results: Our model was benchmarked against several leading open-source polyp segmentation models, including U-Net [10], UNet++ [11], ResUNet++ [61], and others, up to DS-TransUNet [23]. The comparative results are tabulated in Table 3. DEMF-Net notably outperformed the classical model PraNet in all evaluated metrics. On the CVC-ClinicDB

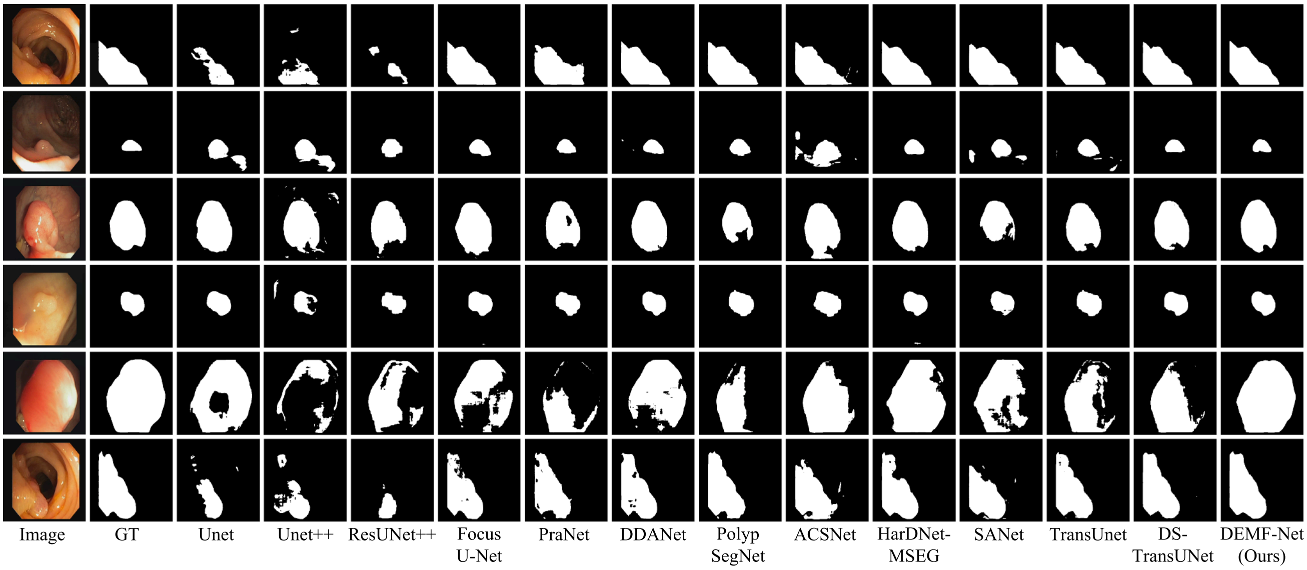


Fig. 5. Partial qualitative results on CVC-ClinicDB.

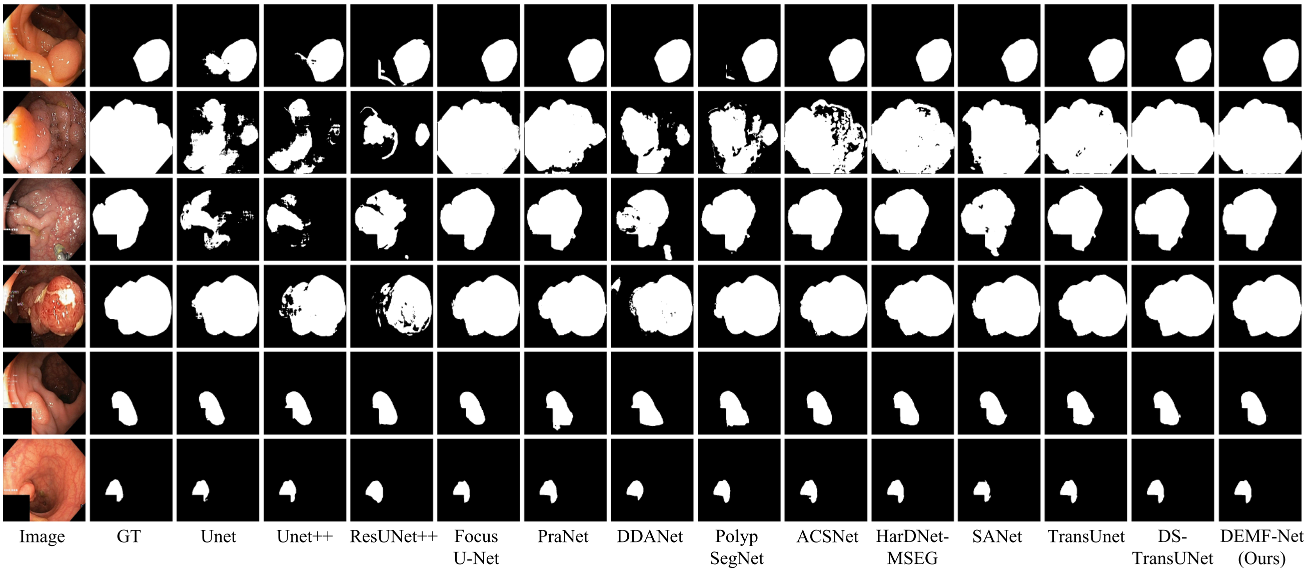


Fig. 6. Partial qualitative results on Kvasir-SEG.

Table 5
Quantitative results of the ablation experiments.

Index	Setting				Kvasir-SEG [62]				CVC-ColonDB [58]			
	Backbone	AFFM	CSB	DAF	Dice	mIoU	Precision	Recall	Dice	mIoU	Precision	Recall
Med.1	√				0.881	0.811	0.916	0.884	0.861	0.774	0.857	0.879
Med.2	√	√			0.891	0.822	0.919	0.892	0.892	0.817	0.876	0.903
Med.3	√		√		0.897	0.831	0.928	0.894	0.916	0.542	0.918	0.909
Med.4	√			√	0.892	0.824	0.919	0.895	0.916	0.854	0.918	0.909
Med.5	√	√	√		0.894	0.835	0.932	0.893	0.918	0.856	0.914	0.933
Med.6	√	√		√	0.898	0.835	0.931	0.894	0.925	0.866	0.920	0.930
Med.7	√		√	√	0.901	0.836	0.935	0.892	0.919	0.861	0.910	0.934
DEMF-Net	√	√	√	√	**0.911**	**0.853**	**0.941**	**0.905**	**0.926**	**0.867**	**0.922**	**0.934**

dataset, most of the metrics of DEMF-Net outperformed the existing methods, with mIoU and Recall at 0.919 and 0.953, respectively. In the Kvasir-SEG dataset, DEMF-Net recorded a Dice coefficient of 0.915 and a mIoU of 0.86, exceeding most current SOTAs.Additionally, DEMF-Net showed advancements over the classical polyp segmentation algorithm, HarDNet-MSEG, with gains of 1.1 %, 1.2 %, and 3 % in Dice, mIoU, and Recall metrics, respectively. These results affirm the superior learning capabilities of DEMF-Net in polyp segmentation. In the Kvasir-SEG dataset, most of the polyp images have only one polyp block, and the evaluation index of DEMF-Net is slightly lower than that of MGCBFormer. In fact, as shown in Fig. 4, the detection rate and accuracy of polyps on small and multi-image polyps exceeded that of MGCBFormer, indicating that DEMF-Net can detect small polyps in a targeted manner, which will be discussed in Qualitative Analysis.

4.4. *Analysis of generalization ability*

Settings: In assessing the model's generalization capabilities, we conducted experiments on the CVC-ColonDB, EndoScene, and ETIS datasets. Notably, the EndoScene dataset comprises 612 images from both the CVC-ClinicDB and CVC300 datasets. Despite minimal exposure to these datasets during training, we evaluated them using our optimally trained model.

Results: As detailed in Table 4, DEMF-Net exhibits commendable generalization performance compared to existing models, effectively adapting to unseen data across various domains/distributions. In the CVC-ColonDB dataset, DEMF-Net's Dice score surpasses that of SANet and PraNet by 0.6 % and 4.2 %, respectively, and its mIoU is ahead by 0.3 % and 2.8 %. On the ETIS dataset, the model's Dice score outperforms HarDNet-MSEG by 5.9 %, although it is marginally lower than SANet. Moreover, in the EndoScene dataset, DEMF-Net's Dice score

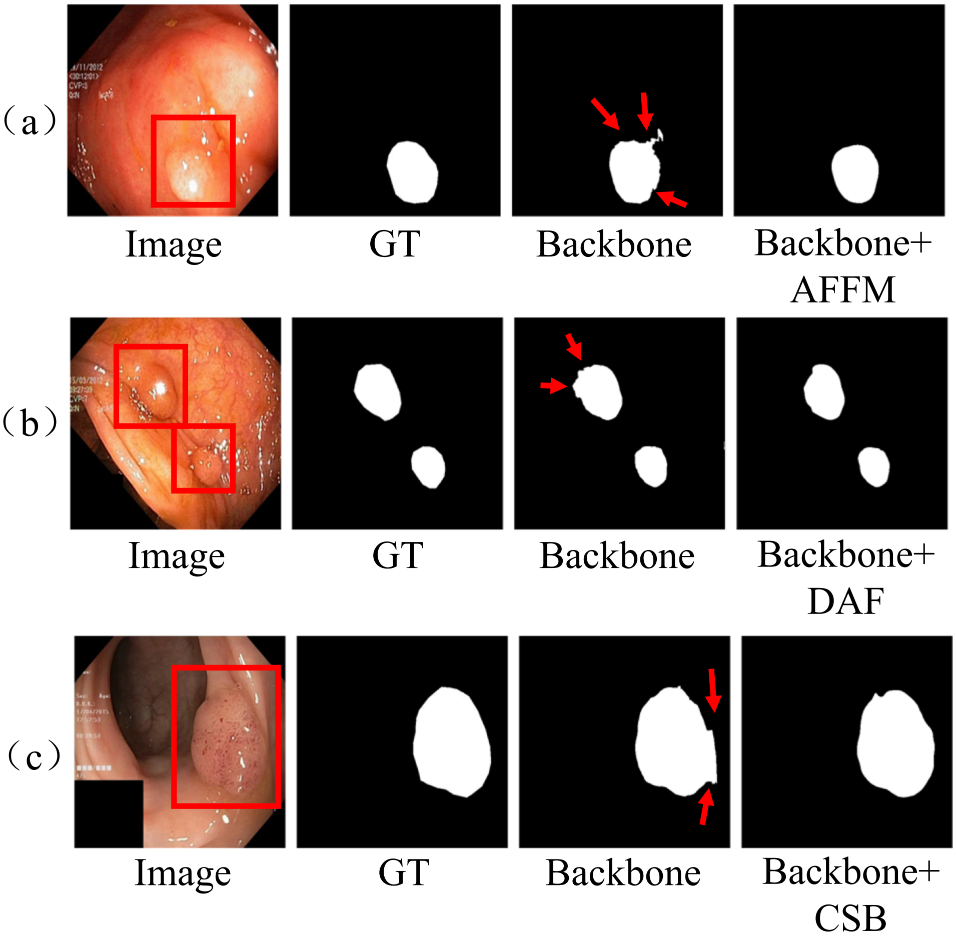


Fig. 7. The detailed visual results of the ablation experiments. The target polyps outlined in red.

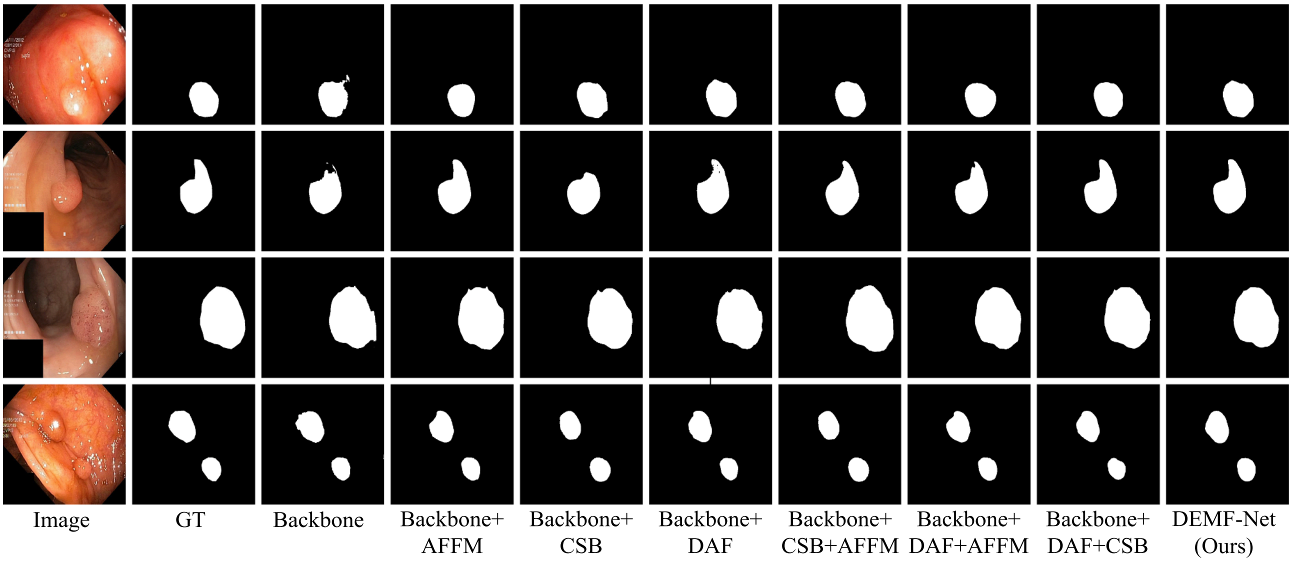


Fig. 8. The overall visual results of the ablation experiments.

exceeds SANet and PraNet by 2 % and 3.7 %, respectively, with its mIoU leading by 6.7 % and 8.5 %. These findings highlight DEMF-Net's superior ability to generalize across different datasets, demonstrating its robustness and effectiveness in diverse segmentation scenarios.

4.5. Qualitative analysis

Figs. 5 and 6 present the visual segmentation results of DEMF-Net alongside comparative models on the CVC-ClinicDB and Kvasir-SEG datasets. From these visualizations, DEMF-Net's proficiency in polyp segmentation is evident, showcasing three notable advantages:

- **Versatility in Various Scenarios**: DEMF-Net consistently produces segmentation results with a bright foreground and a clear background, closely aligning with the ground truth labels. This demonstrates the model's ability to adapt to diverse segmentation scenarios effectively.
- **Precision in Targeting Small and Irregular Polyps**: The model excels in delineating clear and accurate boundaries, especially for small and irregularly shaped polyps. It precisely locates small polyp targets and discerns the blurred boundaries of polyp regions, thereby ensuring accurate segmentation of both the polyp bodies and their edges.
- **Adaptability to Different Acquisition Environments**: DEMF-Net shows remarkable adaptability to varying conditions in image acquisition, such as changes in brightness, reflections, and shadows. This adaptability is crucial for handling large polyp targets, with the model effectively capturing polyp characteristics and maintaining internal consistency in the segmentation results. These attributes underline DEMF-Net's robustness and effectiveness across different imaging conditions, reinforcing its utility in practical medical applications.

4.6. Ablation experiments

We conducted ablation studies on four model variants using the Kvasir-SEG and CVC-ColonDB datasets to assess our model's segmentation performance on colorectal polyp images and elucidate the efficiency and advancement contributed by each component. These experiments adhered to the training, testing, and hyperparameter frameworks specified in Section 4.2.

The foundational structure of the model includes the Swin Transformer's dual-branch encoder, a CNN encoder, and the previously described decoder. The model's effectiveness was gauged by alternately incorporating or omitting the AFFM, CSB, and DAF. These variants were compared against our standard model configuration, denoted as "Backbone + DAF + CSB + AFFM"(DEMF-Net). The quantitative results of the ablation experiments are shown in the table, while the overall visual results are depicted in the Fig. 8.

Effectiveness of AFFM. The efficacy of the AFFM is substantiated through the results presented in Table 5. Incorporating AFFM in the model, referred to as "Med.2″, yielded substantial improvements on both the Kvasir-SEG and CVC-ColonDB datasets, with a 1 % and 3.1 % increase in Dice scores, respectively. This enhancement extends to mIoU, Precision, and Recall, indicating a comprehensive improvement across various metrics. Additionally, the critical role of AFFM is further highlighted in the ″Med.7″ configuration. On the Kvasir-SEG dataset, the exclusion of AFFM resulted in decreased performance across all metrics – Dice, mIoU, Precision, and Recall – by 1 %, 1.7 %, 0.6 %, and 1.3 %, respectively. A similar trend in performance reduction was observed on the CVC-ColonDB dataset, underscoring the integral contribution of AFFM to the overall model's effectiveness. The detailed results of adding the AFFM module are shown in Fig. 7 **(a)**. In the baseline "Backbone" model, there are misclassifications of the background part of the polyps, resulting in unclear segmentation edges. With the inclusion of the AFFM in the "Med.2″ model, the network's ability to handle spatial information of the polyps is enhanced, leading to a clearer distinction between the polyps and background regions, as well as more accurate edge segmentation.

Effectiveness of DAF. The DAF module's impact is evident in the enhanced performance metrics of the "Med.4″ model variant. As Table 5 delineates, this model configuration achieves a notable increase in Dice scores – 1.1 % on the Kvasir-SEG dataset and 5.5 % on the CVC-ColonDB dataset – compared to the baseline ″Backbone“ model. Further emphasizing DAF's role, the ″Med.5″ configuration underscores the decrement in performance metrics in its absence. Specifically, the omission of DAF leads to a reduction in Dice and mean Intersection over Union (mIoU) by 1.7 % and 1.8 %, respectively, on the Kvasir-SEG dataset when compared to the DEMF-Net. The detailed results of adding the DAF module are shown in Fig. 7 **(b)**. The fuzzy edges of irregular polyps are enhanced, improving the detection rate and accuracy of small and multiple polyps on the same image. These findings indicate the critical function of DAF in both preserving and effectively integrating the dual-scale features extracted by the Swin Transformer, thereby enhancing the model's overall segmentation capability.

Table 6
The complexity comparison in different methods.

Model	Dataset	Input size	Epoch	Training time	Modelsize	Inference time ↓
DEMF-Net	CVC-ClinicDB	384 × 384	50	2.5 h	514 MB	**0.084 s**
	Kvasir-SEG	512 × 512	50	4.16 h	514 MB	0.112 s
MGCBFormer [71]	Kvasir-SEG	352 × 352	/	/	/	0.1183 s
Polyp-PVT [46]	Kvasir-SEG	352 × 352	/	/	/	0.0299 s

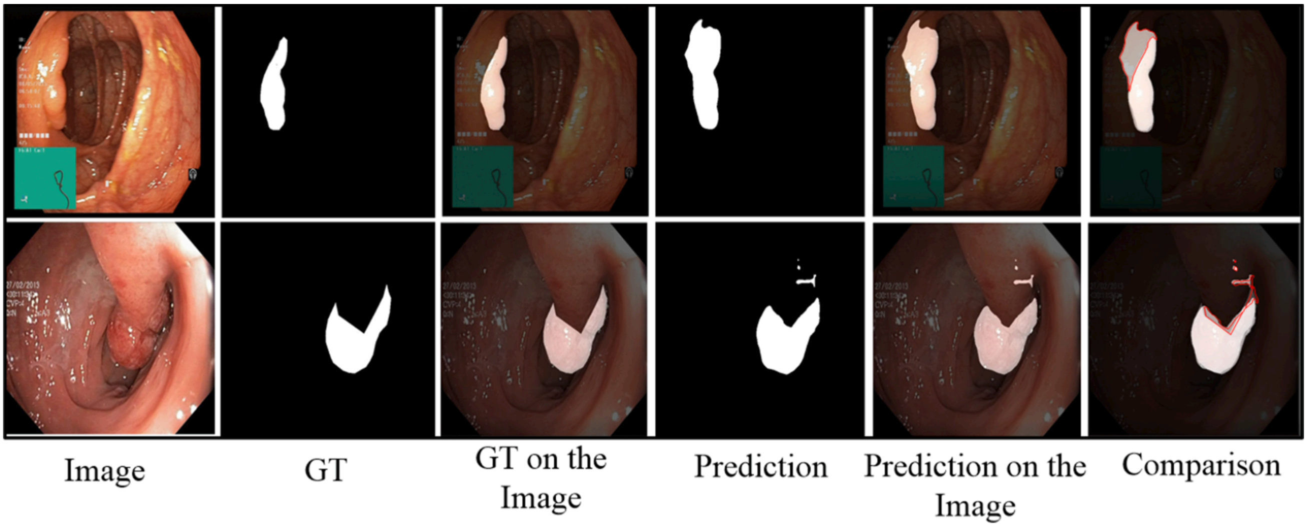


Fig. 9. The prediction result of DEMF-Net.

Effectiveness of CSB. The CSB module significantly enhances segmentation accuracy, as evidenced by the improved Dice scores in the "Med.3″ configuration. This variant demonstrates a Dice score increase of 1.6 % on the Kvasir-SEG dataset and 5.5 % on the CVC-ColonDB dataset, compared to the baseline ″Backbone" model. The effectiveness of CSB is further corroborated in the ″Med.6″ setup, where the removal of CSB from the DEMF-Net results in a reduction of 0.5 % in Dice and 0.1 % in mIoU on the CVC-ColonDB dataset. This impact of CSB is visually depicted in Fig. 7(c). The segmentation images without CSB show a clear background and a bright foreground, yet lack sharpness in the details and edges of the polyp images. The inclusion of the CSB module leads to a marked enhancement in the definition and clarity of the polyp edges, highlighting the module's vital role in refining the segmentation output's overall quality and precision.

5. Discussion

This chapter primarily discusses the computational complexity and limitations.

5.1. Analysis of computational complexity

We listed the Training time, Model size, and Inference time of DEMF-Net on the CVC-ClinicDB and Kvasir-SEG datasets in Table 6. The quality of the images directly affects the effectiveness of endoscopy. Real endoscopes use high-definition cameras [73], so we maximally retained the maximum pixel provided by the original dataset as the input size of the training set, rather than resizing the dataset to 384 × 384 like other methods [71]. This high-resolution feature extraction incurs high computational costs and relatively long training times. However, it is fortunate that our Inference time is faster compared to the current SOTA. Although our model has a longer training time, it still has an advantage in polyp image prediction segmentation Inference time.

5.2. Limitations analysis

Firstly, upon meticulous analysis of DEMF-Net's segmentation outcomes on polyp datasets, we observed that its accuracy is diminished when faced with polyp images against complex backgrounds. As depicted in the Fig. 9, DEMF-Net erroneously identified background regions as polyps. This can be attributed to the relatively small scale and inadequate annotation quality of currently available colorectal polyp datasets. Consequently, the network fails to learn diverse polyp features from these limited datasets, posing a significant challenge to current research on colorectal polyp image segmentation.

Furthermore, the majority of current studies indicate progress in deep learning for colorectal polyp image segmentation. However, this technology has yet to undergo large-scale clinical validation, both domestically and internationally, and lacks validated technical and scientific methodologies [74,75]. At present, the training speed of networks is not optimal, and they have not been applied to real-time clinical diagnostics in medical practice.

6. Conclusions

In this study, we introduce DEMF-Net, an innovative model designed for colorectal polyp segmentation. DEMF-Net capitalizes on a dual encoding mechanism, combining the multiscale Swin Transformer and CNN to extract comprehensive and effective features. The model integrates the Dual-Branch Attention Fusion Module to fuse features effectively obtained by the dual-scale Swin Transformer. Concurrently, the Advanced Feature Fusion Module is employed to establish long-distance dependencies within the CNN architecture. Additionally, the Characterization Supplementary Blocks are incorporated to enrich the structural and detailed aspects of the polyp image. This addresses the prevalent challenges in polyp segmentation, such as large variations in lesion area scales, irregular shapes, and unclear boundaries. Empirical evaluations demonstrate that DEMF-Net outperforms existing methods in polyp segmentation across five challenging colorectal polyp datasets.

Moving forward, our objective is to enhance the generalization capabilities of DEMF-Net further and extend its application to other image domains. We are actively promoting the adoption of the proposed method in clinical settings, aiming to expand the clinical value boundaries of medical image segmentation. This will ultimately increase its practicality in medical image analysis.

CRediT authorship contribution statement

Xiaorui Cao: Writing – original draft, Validation, Software, Methodology, Conceptualization. **He Yu:** Supervision, Project administration, Funding acquisition. **Kang Yan:** Writing – original draft, Validation, Data curation. **Rong Cui:** Visualization, Data curation. **Jinming Guo:** Visualization, Data curation. **Xuan Li:** Validation, Data curation. **Xiaoxue Xing:** Supervision, Investigation. **Tao Huang:** Writing – review & editing, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported by the Science and Technology Program Projects of Jilin Provincial Department of Education (No: JJKH20230677KJ).

References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 71 (2021) 209–249, https://doi.org/10.3322/caac.21660.

[2] T. Kim, H. Lee, D. Kim, Uacanet: Uncertainty augmented context attention for polyp segmentation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2167–2175, 10.1145/3474085.3475375.

[3] P. Sharma, D.R. Nayak, B.K. Balabantaray, M. Tanveer, R. Nayak, A survey on cancer detection via convolutional neural networks: Current challenges and future directions, Neural Netw. (2023), https://doi.org/10.1016/j.neunet.2023.11.006.

[4] S.Y. Quan, M.T. Wei, J. Lee, R. Mohi-Ud-Din, R. Mostaghim, R. Sachdev, D. Siegel, Y. Friedlander, S. Friedland, Clinical evaluation of a real-time artificial intelligence-based polyp detection system: a US multi-center pilot study, Sci. Rep. 12 (2022) 6598, https://doi.org/10.1038/s41598-022-10597-y.

[5] H. Xiao, L. Li, Q. Liu, X. Zhu, Q. Zhang, Transformers in medical image segmentation: A review, Biomed. Signal Process. Control 84 (2023) 104791, https://doi.org/10.1016/j.bspc.2023.104791.

[6] N. Tajbakhsh, S.R. Gurudu, J. Liang, Automatic polyp detection using global geometric constraints and local intensity variation patterns, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14–18, 2014, Proceedings, Part II 17, Springer, 2014, pp. 179–187, 10.1007/978-3-319-10470-6_23.

[7] M. Fiori, P. Musé, G. Sapiro, A complete system for candidate polyps detection in virtual colonoscopy, Int. J. Pattern Recognit Artif Intell. 28 (2014) 1460014, https://doi.org/10.1142/S0218001414600143.

[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, https://doi.org/10.1109/SPMB.2017.8257027.

[9] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.

[10] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, Medical Image Computing and Computer-Assisted Intervention–MICCAI, in: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241, 10.1007/978-3-319-24574-4_28.

[11] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, 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: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11, 10.1007/978-3-030-00889-5_1.

[12] X. Xiao, S. Lian, Z. Luo, S. Li, Weighted res-unet for high-quality retina vessel segmentation, in: 2018 9th international conference on information technology in medicine and education (ITME), IEEE, 2018, pp. 327–331, 10.1109/ITME.2018.00080.

[13] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2021) 203–211, https://doi.org/10.1038/s41592-020-01008-z.

[14] H.-Y. Zhou, C. Lu, S. Yang, Y. Yu, Convnets vs. transformers: Whose visual representations are more transferable?, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2230–2238, 10.1109/iccvw54120.2021.00252.

[15] M.M. Naseer, K. Ranasinghe, S.H. Khan, M. Hayat, F. Shahbaz Khan, M.-H. Yang, Intriguing properties of vision transformers, Adv. Neural Inf. Proces. Syst. 34 (2021) 23296–23308.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, Attention is all you need in Advances in Neural Information Processing Systems, 2017, Search PubMed, 5998-6008.

[17] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, A.K. Nandi, Medical image segmentation using deep learning: A survey, IET Image Proc. 16 (2022) 1243–1267, https://doi.org/10.1049/ipr2.12419.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, (2020).

[19] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, arXiv preprint arXiv:2102.04306, (2021).

[20] Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, Medical Image Computing and Computer Assisted Intervention–MICCAI 2021, in: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 14–24, 10.1007/978-3-030-87193-2_2.

[21] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, European conference on computer vision, Springer, 2022, pp. 205–218, 10.1007/978-3-031-25066-8_9.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022, 10.1109/iccv48922.2021.00986.

[23] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, D. Zhang, Ds-transunet: Dual swin transformer u-net for medical image segmentation, IEEE Trans. Instrum. Meas. 71 (2022) 1–15, https://doi.org/10.1109/TIM.2022.3178991.

[24] C.-F.-R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366.

[25] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, J. Gao, Multi-scale vision longformer: A new vision transformer for high-resolution image encoding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2998–3008.

[26] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, C. Feichtenhofer, Multiscale vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6824–6835.

[27] S. Xia, S.M. Krishnan, M.P. Tjoa, P.M. Goh, A novel methodology for extracting colon's lumen from colonoscopic images, Journal of Systemics, Cybernetics and Informatics 1 (2003) 7–12.

[28] Z. Wang, L. Li, J. Anderson, D.P. Harrington, Z. Liang, Computer-aided detection and diagnosis of colon polyps with morphological and texture features, in: Medical Imaging 2004: Image Processing, SPIE, 2004, pp. 972–979, 10.1117/12.535664.

[29] S. Hwang, J. Oh, W. Tavanapong, J. Wong, P.C. De Groen, Polyp detection in colonoscopy video using elliptical shape feature. 2007 IEEE International Conference on Image Processing, 2007 pp. II-465-II-468. 10.1109/ICIP.2007.4379193.

[30] T.A. Chowdhury, O. Ghita, P.F. Whelan, A statistical approach for robust polyp detection in CT colonography, in: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, IEEE, 2006, pp. 2523–2526, 10.1109/IEMBS.2005.1616982.

[31] P. Li, K.L. Chan, S.M. Krishnan, Learning a multi-size patch-based hybrid kernel machine ensemble for abnormal region detection in colonoscopic images, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 670–675, 10.1109/CVPR.2005.201.

[32] D.K. Iakovidis, A. Koulaouzidis, Automatic lesion detection in wireless capsule endoscopy—a simple solution for a complex problem, in: 2014 IEEE International Conference on Image Processing (ICIP), 2014, pp. 2236–2240, https://doi.org/10.1109/ICIP.2014.7025453.

[33] S. Abbasi, M. Tavakoli, H.R. Boveiri, M.A.M. Shirazi, R. Khayami, H. Khorasani, R. Javidan, A. Mehdizadeh, Medical image registration using unsupervised deep neural network: A scoping literature review, Biomed. Signal Process. Control 73 (2022) 103444, https://doi.org/10.1016/j.bspc.2021.103444.

[34] M. Akbari, M. Mohrekesh, E. Nasr-Esfahani, S.R. Soroushmehr, N. Karimi, S. Samavi, K. Najarian, Polyp segmentation in colonoscopy images using fully convolutional network, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 69–72, 10.1109/embc.2018.8512197.

[35] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.-A. Heng, H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, IEEE Trans. Med. Imaging 37 (2018) 2663–2674, https://doi.org/10.1109/TMI.2018.2845918.

[36] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999, (2018). Doi: 10.48550/arXiv.1804.03999.

[37] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (2019) 1856–1867, https://doi.org/10.1109/TMI.2019.2959609.

[38] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, International conference on medical image computing and computer-assisted intervention, Springer (2020) 263–273, https://doi.org/10.1007/978-3-030-59725-2_26.

[39] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, D. Stoyanov, Fully convolutional neural networks for polyp segmentation in colonoscopy, Medical Imaging, in: 2017 Computer-Aided Diagnosis, SPIE, 2017, pp. 101–107, https://doi.org/10.1117/12.2254361.

[40] Y. Fang, C. Chen, Y. Yuan, K.-Y. Tong, Selective feature aggregation network with area-boundary constraints for polyp segmentation, Medical Image Computing and Computer Assisted Intervention–MICCAI, in: 2019 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22, Springer, 2019, pp. 302–310, 10.1007/978-3-030-32239-7_34.

[41] B. Murugesan, K. Sarveswaran, S.M. Shankaranarayana, K. Ram, J. Joseph, M. Sivaprakasam, Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 7223–7226, 10.1109/EMBC.2019.8857339.

[42] J. Wei, Y. Hu, R. Zhang, Z. Li, S.K. Zhou, 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, 10.1007/978-3-030-87193-2_66.

[43] X. Pan, C. Ma, Y. Mu, M. Bi, GLSNet: A Global Guided Local Feature Stepwise Aggregation Network for polyp segmentation, Biomed. Signal Process. Control 87 (2024) 105528, https://doi.org/10.1016/j.bspc.2023.105528.

[44] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, Cvt: Introducing convolutions to vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 22–31, https://doi.org/10.1109/iccv48922.2021.00009.

[45] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, B. Guo, Cswin transformer: A general vision transformer backbone with cross-shaped windows, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12124–12134, https://doi.org/10.1109/cvpr52688.2022.01181.

[46] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, L. Shao, Polyp-pvt: Polyp segmentation with pyramid vision transformers, arXiv preprint arXiv:2108.06932, (2021). Doi: 10.26599/AIR.2023.9150015.

[47] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, nnformer: Interleaved transformer for volumetric segmentation, arXiv preprint arXiv:2109.03201, (2021).

[48] H. Tang, Y. Chen, T. Wang, Y. Zhou, L. Zhao, Q. Gao, M. Du, T. Tan, X. Zhang, T. Tong, HTC-Net: A hybrid CNN-transformer framework for medical image segmentation, Biomed. Signal Process. Control 88 (2024) 105605, https://doi.org/10.1016/j.bspc.2023.105605.

[49] K.-B. Park, J.Y. Lee, SwinE-Net: Hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer, J. Comput. Des. Eng. 9 (2022) 616–632, https://doi.org/10.1093/jcde/qwac018.

[50] W. Li, Y. Zhao, F. Li, L. Wang, MIA-Net: Multi-information aggregation network combining transformers and convolutional feature learning for polyp segmentation, Knowl.-Based Syst. 247 (2022) 108824, https://doi.org/10.1016/j.knosys.2022.108824.

[51] W. Zhang, C. Fu, Y. Zheng, F. Zhang, Y. Zhao, C.-W. Sham, HSNet: A hybrid semantic network for polyp segmentation, Comput. Biol. Med. 150 (2022) 106173, https://doi.org/10.1016/j.compbiomed.2022.106173.

[52] N.T. Duc, N.T. Oanh, N.T. Thuy, T.M. Triet, V.S. Dinh, Colonformer: An efficient transformer based method for colon polyp segmentation, IEEE Access 10 (2022) 80575–80586, https://doi.org/10.1109/ACCESS.2022.3195241.

[53] K. Ramesh, G.K. Kumar, K. Swapna, D. Datta, S.S. Rajest, A review of medical image segmentation algorithms, EAI Endorsed Trans. Pervasive Health Technol. 7 (2021) e6–e, https://doi.org/10.4108/eai.12-4-2021.169184.

[54] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, 10.1109/cvpr.2016.90.

[55] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, https://doi.org/10.1109/cvpr.2018.00745.

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826, 10.1109/cvpr.2016.308.

[57] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado, Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer, Int. J. Comput. Assist. Radiol. Surg. 9 (2014) 283–293, https://doi.org/10.1007/s11548-013-0926-3.

[58] J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Comput. Med. Imaging Graph. 43 (2015) 99–111, https://doi.org/10.1016/j.compmedimag.2015.02.007.

[59] N. Tajbakhsh, S.R. Gurudu, J. Liang, Automated polyp detection in colonoscopy videos using shape and context information, IEEE Trans. Med. Imaging 35 (2015) 630–644, https://doi.org/10.1109/TMI.2015.2487997.

[60] D. Vázquez, J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, A.M. López, A. Romero, M. Drozdzal, A. Courville, A benchmark for endoluminal scene segmentation of colonoscopy images, Journal of Healthcare Engineering 2017 (2017), https://doi.org/10.1155/2017/4037190.

[61] D. Jha, P.H. Smedsrud, D. Johansen, T. de Lange, H.D. Johansen, P. Halvorsen, M. A. Riegler, A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation, IEEE J. Biomed. Health Inform. 25 (2021) 2029–2040, https://doi.org/10.1109/JBHI.2021.3049304.

[62] D. Jha, P.H. Smedsrud, M.A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26, Springer, 2020, pp. 451–462, 10.1007/978-3-030-37734-2_37.

[63] M. Yeung, E. Sala, C.-B. Schönlieb, L. Rundo, Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy, Comput. Biol. Med. 137 (2021) 104815, https://doi.org/10.1016/j.compbiomed.2021.104815.

[64] N.K. Tomar, D. Jha, S. Ali, H.D. Johansen, D. Johansen, M.A. Riegler, P. Halvorsen, DDANet: Dual decoder attention network for automatic polyp segmentation, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021 Proceedings Part VIII, Springer, 2021, pp. 307–314, 10.1007/978-3-030-68793-9_23.

[65] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495, https://doi.org/10.1109/TPAMI.2016.2644615.

[66] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, Y. Yu, Adaptive context selection for polyp segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23, Springer, 2020, pp. 253–262, 10.1007/978-3-030-59725-2_25.

[67] C.-H. Huang, H.-Y. Wu, 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).

[68] J. Mei, M.-M. Cheng, G. Xu, L.-R. Wan, H. Zhang, SANet: A slice-aware network for pulmonary nodule detection, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2021) 4374–4387, https://doi.org/10.1109/TPAMI.2021.3065086.

[69] Y. Zhang, L. Liu, Z. Han, F. Meng, Y. Zhang, Y. Zhao, TranSEFusionNet: Deep fusion network for colorectal polyp segmentation, Biomed. Signal Process. Control 86 (2023) 105133, https://doi.org/10.1016/j.displa.2023.102600.

[70] Y. Lin, X. Han, K. Chen, W. Zhang, Q. Liu, CSwinDoubleU-Net: A double U-shaped network combined with convolution and Swin Transformer for colorectal polyp segmentation, Biomed. Signal Process. Control 89 (2024) 105749, https://doi.org/10.1016/j.bspc.2023.105749.

[71] Y. Xia, H. Yun, Y. Liu, J. Luan, M. Li, MGCBFormer: The multiscale grid-prior and class-inter boundary-aware transformer for polyp segmentation, Comput. Biol. Med. 167 (2023) 107600, https://doi.org/10.1016/j.compbiomed.2023.107600.

[72] L. Nanni, A. Lumini, C. Fantozzi, Exploring the potential of ensembles of deep learning networks for image segmentation, Information 14 (2023) 657, https://doi.org/10.3390/info14120657.

[73] H. Yamashita, H. Aoki, K. Tanioka, T. Mori, T. Chiba, Ultra-high definition (8K UHD) endoscope: our first clinical success, Springerplus 5 (2016) 1–5.

[74] J. Bernal, N. Tajkbaksh, F.J. Sanchez, B.J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge, IEEE Trans. Med. Imaging 36 (2017) 1231–1249, https://doi.org/10.1109/TMI.2017.2664042.

[75] Y. Wang, X. He, H. Nie, J. Zhou, P. Cao, C. Ou, Application of artificial intelligence to the diagnosis and therapy of colorectal cancer, Am. J. Cancer Res. 10 (2020) 3575.