Mathematical Biosciences
and Engineering

http://www.aimspress.com/journal/MBE

MBE, 21(2): 2024–2049.
DOI: 10.3934/mbe.2024090
Received: 05 October 2023
Revised: 23 December 2023
Accepted: 26 December 2023
Published: 08 January 2024

Research article

Simultaneous segmentation and classification of colon cancer polyp images using a dual branch multi-task learning network

Chenqian Li[1,2], **Jun Liu**[1,2,*] **and Jinshan Tang**[3,*]

[1] School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-Time Industrial System, Wuhan 430065, China
[3] Department of Health Administration and Policy, College of Public Health, George Mason University, Fairfax, VA 22030, USA

* **Correspondence:** Email: ljwhcn@qq.com, jtang25@gmu.edu.

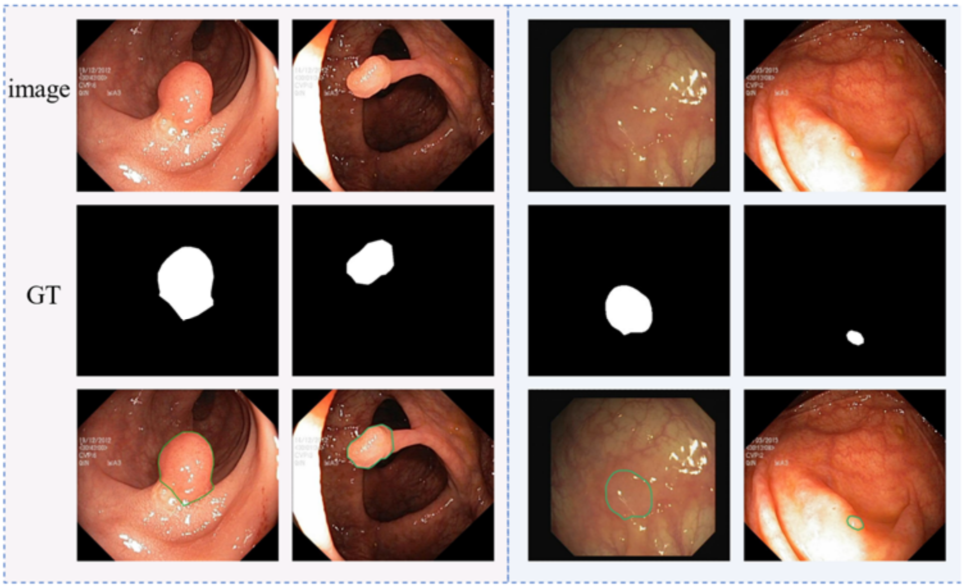
Abstract: Accurate classification and segmentation of polyps are two important tasks in the diagnosis and treatment of colorectal cancers. Existing models perform segmentation and classification separately and do not fully make use of the correlation between the two tasks. Furthermore, polyps exhibit random regions and varying shapes and sizes, and they often share similar boundaries and backgrounds. However, existing models fail to consider these factors and thus are not robust because of their inherent limitations. To address these issues, we developed a multi-task network that performs both segmentation and classification simultaneously and can cope with the aforementioned factors effectively. Our proposed network possesses a dual-branch structure, comprising a transformer branch and a convolutional neural network (CNN) branch. This approach enhances local details within the global representation, improving both local feature awareness and global contextual understanding, thus contributing to the improved preservation of polyp-related information. Additionally, we have designed a feature interaction module (FIM) aimed at bridging the semantic gap between the two branches and facilitating the integration of diverse semantic information from both branches. This integration enables the full capture of global context information and local details related to polyps. To prevent the loss of edge detail information crucial for polyp identification, we have introduced a reverse attention boundary enhancement (RABE) module to gradually enhance edge structures and detailed information within polyp regions. Finally, we conducted extensive experiments on five publicly available datasets to evaluate the performance of our method in both polyp segmentation and classification tasks. The experimental results confirm that our proposed method outperforms other state-of-the-art methods.

Keywords: polyps; colonoscopy; dual branch network; FIM; RABE; multi-task network; attention mechanism

1. Introduction

Colorectal cancers are malignant tumors that commonly occur in the colon and rectum. They make up one of the most prevalent cancer types globally, ranking third in terms of cancer incidence and being the leading cause of cancer-related deaths in the United States. Advanced rectal cancers are difficult to cure, and thus how to improve survival efficiency is key. Early detection and diagnosis play crucial roles in improving survival efficiency.

Colon polyps are lumps in the lining of the colon. Colon polyps have a high possibility to turn into cancers and are a leading cause of colon cancers. Thus, the detection and removal of polyps are important in preventing polyps from developing into colon cancers. The primary tool for screening for colon cancers is colonoscopy. Studies show that the prevalence of rectal cancers can be reduced by as much as 30 percent with regular colonoscopies.



(a) Pedunculated Polyps

(b) Sessile Polyps

Figure 1. Two basic shapes of colorectal polyps: (a) pedunculated polyps and (b) sessile polyps. GT represents ground truth.

Polyps have two types, based on their shapes: pedunculated polyps and sessile polyps (as shown in Figure 1). The latter only account for 15%, and they are difficult to detect in images. Incorrect diagnosis carries the risk of bleeding and perforation, thus accurate identification of the type of polyps and treatment requires a high degree of concentration and experience of physicians. However, according to [1], about 25% of polyps are missed during routine colonoscopy. Thus, we need an accurate and efficient method for polyp classification and segmentation. In recent years, deep learning has made significant progress in processing medical images. Chen et al. [2] proposed a gait pattern recognition method for lower limb exoskeleton based on long short-term memory (LSTM) and convolutional neural network (CNN) to improve the recognition accuracy. Tian [3] proposed a new artificial neural network model evaluation strategy, which has been experimentally proved to be closer to the actual biological nervous system. Xu [4] proposed using deep learning methods to predict new cases of the new coronavirus, and the experiment achieved high prediction performance. Xie [5] proposed a Physically Constrained Deep Active Learning (P-DAL) framework to model spatiotemporal cardiac electrodynamics. The results showed that the proposed P-DAL method is significantly better than the traditional modeling methods. Guan [6] proposed a texture-constrained multi-channel asymptotic generative adversarial network (TMP-GAN), which adopts multi-channel joint training, which effectively avoids the typical shortcomings of current generative methods. Because of the good performance of deep learning for image analysis, deep learning was also used to help endoscopists to improve accuracy and efficiency of diagnosis. In a CAD system for colorectal cancers, polyp segmentation and classification are two important tasks. With the development of CNN based methods, research on polyp segmentation and classification has made some progress. Zhang [7] presented a migration learning algorithm to perform classification of colorectal polyps and achieved excellent results. Bourne et al. [8] proposed a real-time evaluation model to classify polyps into two classes: adenomatous polyps and hyperplastic polyps. Their experiments on polyp videos showed that their model obtained an accuracy of 94%. Younas et al. [9] combined the strengths of individual weak learners to form a weighted integrated model for polyp multiclass classification. In addition to polyp classification, some work is also about polyp segmentation. U-type architectures are the baseline for most medical image segmentation, and thus they are also studied in polyp segmentation. Inspired by U-Net [10], U-Net++ [11] improved U-Net by employing multi-scale nested jump connections and showed high accuracy in polyp segmentation. Jha et al. [12] extended the ResUnet++ [13] by incorporating temporal random fields for polyp segmentation. To solve the edge blurring problem caused by high similarity of polyps to the background, Fan et al. [14] developed a method using a parallel reverse attention network. The proposed network aggregates high-level information through a parallel partial decoder to generate a global mapping. The global map is combined with a reverse attention module to extract boundary information. Zhang [15] designed a network with an attention module aiming at adaptively focusing on different background information. The proposed network alleviates intra-class inconsistency. Experimental results on two datasets validated the accuracy of the method. Ji et al. [16] first proposed the study of video polyp segmentation, introduced a high-quality frame-by-frame annotated VPS dataset, designed a simple and efficient model (PNS+), and demonstrated the effectiveness and high performance of the baseline through many experiments. Lin et al. [17] proposed a new bit-slice contextual attention network for polyp segmentation to improve the ability to extract boundary information, and they proposed a dual-path attention link encoder to further improve the segmentation performance for polyps. Many experiments proved that this method can effectively improve the performance of polyp segmentation. Zhang et al. [18] built a parallel architecture by adding a transformer to CNN for polyp segmentation. Their network can capture both long-term dependency and local information. Experiments on several datasets substantiated the effectiveness of the method.

Although the forementioned methods show improvements in polyp segmentation or classification with comparison to traditional methods, they still face several challenges: 1) The past methods performed polyp segmentation and classification separately, ignoring the intrinsic correlation information between the two tasks. However, multi-task learning allows the network to share feature representations such as image texture, shape, and boundary, which improves the learning efficiency and representation of features. In addition, multi-task learning allows information interaction and co-learning between polyp segmentation and classification tasks, which improves the robustness and generalization of the model. 2) Existing multi-stage methods [19–21] combining the two tasks are based on CNNs. Due to the limitations of convolutional operations, they can only establish short-distance dependency relationships and cannot establish relationships between target pixels and global pixels. This often leads to the neglect of a significant amount of global information crucial for detecting the location of the targets. However, these global features are necessary to achieve more accurate classification and segmentation of polyps. For this reason, some researchers have taken steps to enhance CNNs by extracting global contextual information. Nonetheless, in most cases, this method fails to yield satisfactory results.

To overcome the limitations of existing methods, we propose a multi-task network that enhances the model's performance by concurrently training it for both classification and segmentation. Furthermore, we develop a dual-branch network structure that combines a transformer encoder and a CNN encoder, inheriting the advantages of both the transformer and CNN. The transformer branch enhances the model's ability to capture global context information by learning long-term dependencies among inter-pixel features, aiding in the localization of polyp regions. Meanwhile, the CNN branch excels at capturing feature representations with spatial information (especially local information), such as edge information, which is more beneficial for the segmentation of small targets. In addition, to fully use the advantages of the two branches, we propose a feature interaction module (FIM) for information fusion and a RABE module to enhance the extraction of fuzzy boundaries. In summary, our contributions are as follows:

1) We propose a multi-task model for simultaneous segmentation and classification of colon polyps. The proposed network utilizes an end-to-end architecture, employing a task-sharing encoder to enhance the correlation between different task networks more effectively. In this network, we adopt a dual-branch structure that incorporates a transformer to extract global features from colon images, thus combining the advantages of CNN into the proposed model. This approach enables the network to learn more meaningful feature information and significantly improves the segmentation and classification results of polyps.

2) We propose a feature interaction module that serves to eliminate the semantic gap between the transformer and the CNN. It also fully integrates the global contextual information of polyps extracted by both, along with local detail information. This approach reduces the loss of polyp location and detailed information.

3) We design a RABE module to further extract boundary information by establishing relationships between the targets and the boundaries. This enhancement improves the network's performance in detecting polyps with ambiguous target boundaries.

4) Our proposed method has undergone extensive evaluation on several benchmark datasets, and a significant number of experimental results demonstrate that our approach outperforms other state-of-the-art methods. It exhibits superior performance in both polyp segmentation and classification.

The paper is structured into five distinct sections. The introductory section provides background knowledge relevant to the paper's focus. The second section delves into research pertinent to the methodology employed in this study. Following that, the third section offers a comprehensive exploration of the principles and structure underlying the chosen method. The fourth section substantiates the method's effectiveness through a series of experiments. Lastly, the fifth section conducts an analysis and summary of the paper's findings.

2. Related works

2.1. Multitask learning

Multi-task learning (MTL) is a learning paradigm in machine learning that aims to utilize useful information contained in multiple related tasks to help improve the performance of all tasks, and it has had a great impact in many fields, such as natural language processing and computer vision directions. Compared with single-task learning, it can share the common features of multiple tasks, achieve multiple tasks at the same time, and have good generalization ability, which is an important application of deep learning. Due to the time-consuming and labor-intensive nature of radiologists' annotation work in the field of medical images as well as the label-intensive nature of the images, it is necessary to analyze the medical images comprehensively by means of multiple related tasks. Chen et al. [22] proposed a multi-task learning network for segmentation and classification of atria, and the results showed that by sharing features between related tasks, the multi-task network can obtain additional anatomical information about the atria and achieve more accurate segmentation of atria. Zhang et al. [23] proposed a multi-task relational learning network for segmentation, localization, and identification of vertebrae, which utilized the relationship between vertebrae and the correlation of three tasks to train the network and finally proved the effectiveness of the network on an MRI dataset. Zhou et al. [24] proposed a multi-task learning framework for joint classification and segmentation of tumors in ultrasound images. The framework includes a network for segmentation and a multi-scale network for classification. Experiments were conducted on three clinical datasets using an iterative training strategy. The experimental results demonstrated that the proposed multi-task framework has better performance than the single task learning framework. Liu et al. [25] proposed a multi-task learning method for processing data stored in different locations. This method transformed the original centralized computing framework into a distributed framework that can be computed in parallel, thereby enhancing both learning performance and efficiency.

In summary, previous studies have demonstrated the effectiveness of multi-task learning networks. However, the multi-task models still overlook the importance of global features. Therefore, we propose to leverage the transformer architecture to construct a two-branch network for capturing global features.

2.2. Transformer

In earlier studies, various CNN-based network models were developed for polyp classification or segmentation, and they achieved some level of effectiveness. However, these methods often overlook the global features of the targets due to the limitations of convolution operations, hindering the improvement of experimental results. In recent years, transformers have been proven to be an excellent model for extracting global features from targets, primarily through the self-attention mechanism. A large body of research, even before its emergence [26,27], has confirmed that self-attention can enhance the performance of CNNs in many applications. Inspired by self-attention, a lot of models on transformers have been proposed. Dosovitskiy [28] applied a transformer to image classification and achieved good performance. Carion et al. [29] proposed DETR, a model for object detection. Experimental results on the Coco dataset outperformed Faster-Rcnn. Due to the superior performance of transformers, many studies combining transformers with other models have been applied to the vision direction. Chen [30] proposed to combine a CNN and a transformer for medical image segmentation and achieved promising results. Transformer-based methods have also shown great potential in colon polyp detection and classification. Wang et al. [31] proposed a multilayer fusion network using a hierarchical guided strategy to aggregate information. The proposed network combined a transformer encoder and CNN encoder to extract deep semantic information and shallow localized spatial features for polyp detection and yielded reliable results. Huang et al. [32] explored the potential of using a joint technique that combines transformers and CNNs to address the challenges of polyp segmentation. They introduced interaction modules for the identification and fusion of information from both sources, resulting in a more robust model compared to existing methods. Park et al. [33] proposed the SwineE-Net network for polyp segmentation, and extensive experiments on five public datasets demonstrated the model's generalizability and scalability. In contrast to the tasks mentioned above, our goal is to develop a multi-task model that combines transformers and CNNs for diagnosing colon cancer from colonoscopy images.

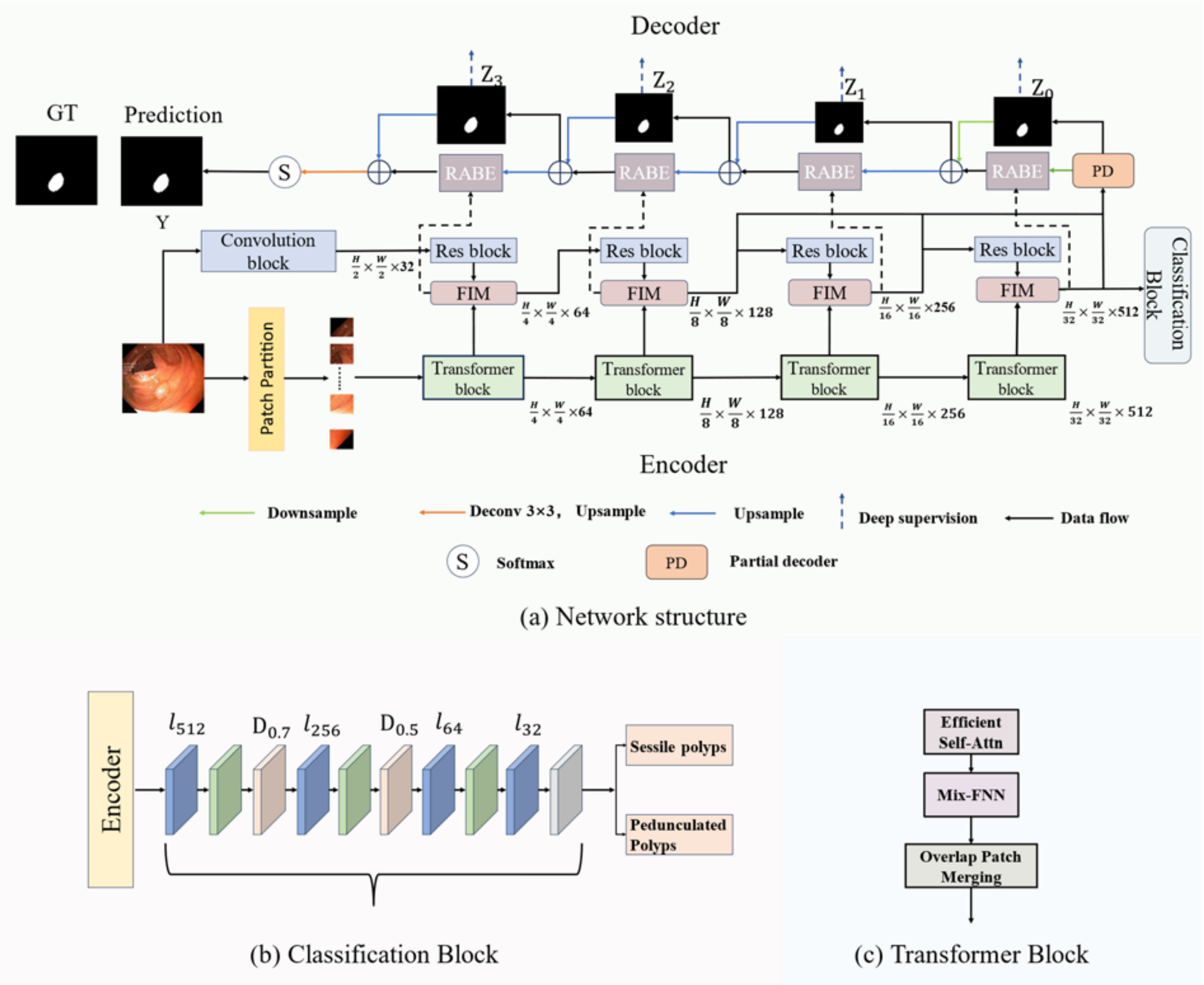


Figure 2. (a) Overall structure of the network. (b) Classification block, where the blue blocks represent the fully connected layers, the green block represents the Normalization operation, and the yellow and gray blocks represent Dropout and sigmoid, respectively. (c) Transformer block.

3. Methods

3.1. Network architecture

The structure of our proposed network for the segmentation and classification of polyps is shown in Figure 2(a). The network is composed of a dual-branch encoder, a classification module, and a decoder. The dual-branch encoder consists of a transformer branch and a CNN branch. The transformer branch is responsible for gathering high-level semantic information of polyps, aiming at capturing long-term dependent features. The CNN branch is used to learn the localized detailed texture features of polyps. We designed a feature interaction module, which can better fuse global and local features by eliminating the semantic gap between the two branches and learning more useful information from the fused features. The decoder consists of a partial decoder (PD) as well as a RABE module. The PD combines features from multiple levels for decoding and initially aggregates a pre-segmentation result. This combined information is then used in conjunction with the RABE module for level-by-level decoding, resulting in a series of feature mappings $Z_j,\ j \in \{0,1,2,3\}$. The mappings are used to facilitate network learning. The decoding process is to extract boundary information of the polyps, aiming at capturing structural details, minimizing the segmentation errors at the boundaries. In addition, our segmentation and classification tasks share a two-branch encoder. The classification task consists of the shared encoder as in Figure 2(a) and a classification module as in Figure 2(b) to classify the polyp images. The classification module consists of four fully connected layers, three normalized layers, two dropout layers, and a sigmoid activation function. In the classification module, we set the first dropout rate as 0.7 and the second set to 0.5. The results were labeled as two different colonoscopy polyp images.

3.2. Dual branch encoder structure

The shapes, sizes, and locations of polyps vary significantly in different images, especially within large and small target regions. Consequently, inaccurate segmentation and classification may occur, primarily due to the absence of contextual information. Many existing segmentation and classification networks utilize encoding and decoding structures, with the encoder playing a pivotal role in information extraction. The encoder is responsible for learning the mapping relationships between pixels and their corresponding topology and projecting the learned salient features onto the pixel space. Hence, the design of the encoder holds significant importance as it directly impacts the robustness of the extracted features, thereby influencing the overall performance of the network. A traditional CNN-based encoder learns through convolutional parameter sharing, making the encoder more sensitive to noise from the input. To address this issue, we propose to integrate transformers in the network because transformers can obtain more robust information through remote dependency modeling. The combination of the two possesses the advantages of both CNNs and transformer, which can provide richer coded information and semantic features for polyp segmentation and classification. For the transformer branch, we use patch partition to divide the input polyp image $x \in R^{H \times W \times 3}$ into a set of non-overlapping image patches. The feature dimension of each patch is $H/4 \times W/4 \times 3$. The number is 4×4, and then these patches are used as input to the transformer branch. We employ a mix transformer (MIT) [34] (as shown in Figure 2(c)). Compared with the models in [29], MIT generates multi-scale features and can improve the performance of semantic segmentation. MIT extracts multi-scale features, preserving both coarse- and fine-grained features, enabling a more accurate region classification and a more complete set of edges. MIT demonstrates several advantages through its unique processes. First, to reduce the complexity of self-attention computation, MIT additionally employs a sequence reduction operation known as efficient self-attention. This operation reduces the computational complexity by decreasing the sequence length. Specifically, in this process, for each input patch, its Q, K, and V values are calculated through linear transformation, and then the Attention weight of the multi-head attention is calculated based on these three vector values. Then, the Attention of each head is spliced to obtain the final Attention representation to update the model. In the original multi-head self-attention process, The Q, K, and V of each head are the same dimensions N × C, where represents the length of the sequence. This process is expressed by the formula

$$Attention(Q,K,V) = Softmax\left(\frac{QK^{T}}{\sqrt{d_{head}}}\right)V \tag{1}$$

To reduce the amount of calculation, the K matrix with input dimension N × C is transformed as follows:

$$\breve{K} = Linear(C \cdot R, C)\left(Reshape\left(\frac{N}{R}, C \cdot R\right)(K)\right) \tag{2}$$

where K is the sequence to be reduced, $\breve{K}$ is the reduced sequence, its dimension is $\frac{N}{R} \times C$, d_{head} represents a scalar value used to scale the attention weights, used to solve the gradient disappearance problem of the softmax function when the inner product value is too large. We set the value of d_{head} to 4 and set the R values for the four transformer blocks to $[64, 16, 4, 1]$.

Second, to enhance the representation capability of the model, Mix-FFN is added as a technique used to improve the self-attention model. A depth-separable convolution and a multilayer perceptron are employed to convey position information to Mix-FFN to ensure local continuity. This not only reduces computational complexity and parameter requirements but also greatly aids in localizing the position of the polyp region. This process is expressed by the formula

$$\mathrm{X_{out}} = \mathrm{MLP}\left(\mathrm{GELU}\left(\mathrm{Conv_{3\times3}}(\mathrm{MLP}(\mathrm{X_{in}}))\right)\right) + \mathrm{X_{in}} \tag{3}$$

where MLP represents the Multilayer Perceptron, GELU () represents the activation function, and X_{in} represents the output of efficient self-attention.

Finally, MIT includes an overlap patch merging module, which serves to reduce the feature map size while increasing the number of channels in the feature map. However, MIT still exhibits some shortcomings; it does not effectively handle continuity information between blocks, potentially leading to segmentation results with boundary or detail loss consequently. The above issues can be mitigated by the CNN branch. The convolutional encoder in the CNN branch can preserve shallow high-resolution features for better characterization of local information. The CNN branch employs ResNet18 as the backbone and adopts small 3 × 3 convolution kernels. The small size kernels can learn relative relationship between neighboring pixel points effectively, thereby extracting texture and detail information effectively.

The process in the network is described as follows: For a given input image $x \in R^{H\times W\times 3}$, the CNN branch initially performs a convolution operation to obtain the feature map $A \in R^{\frac{H}{2}\times\frac{W}{2}\times 16}$. Subsequently, it undergoes four ResBlock operations, resulting in feature maps $F_i \in R^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$, where $C_i \in \{64, 128, 256, 512\}$ and $i \in \{1, 2, 3, 4\}$. In contrast, the transformer branch extracts four layers of feature maps $T_{\mathrm{i}} \in R^{\frac{H}{2^{\mathrm{i}}} \times \frac{W}{2^{\mathrm{i}}} \times D_{\mathrm{i}}}$, where $D_{\mathrm{i}} \in \{64, 128, 256, 512\}$. Finally, the outputs of each layer from the two branches $\{\mathrm{F}_i, T_{\mathrm{i}}\}$ are jointly input into the feature interaction module to combine the information of the two branches.

3.3. Feature interaction module

Given that the learning mechanisms and semantic information acquired by the transformer branch and the CNN branch are distinct, seamless fusion of information from both branches becomes crucial. This allows us to leverage the integration advantages offered by both branch encoders effectively. We propose employing feature interaction module (FIM) to achieve the goal. The FIM employs an interactive fusion approach to integrate the local features with the global representation, effectively eliminating the semantic gap between the two branches. The proposed FIM is shown in Figure 3.

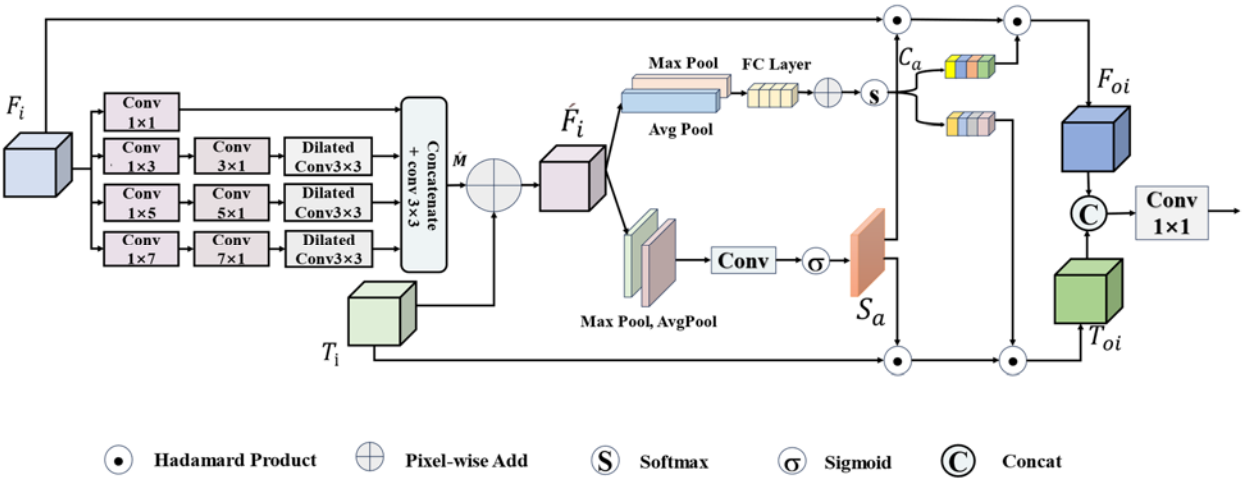


Figure 3. FIM.

In the proposed FIM, to narrow the semantic gap between the CNN and the transformer, we begin by employing convolutional kernels of various sizes to expand the receptive field of the feature map F_{i} obtained from the CNN branch. This enables the capture of more contextual information. The results after various convolutional operations are then catenated into a new feature map denoted as $\acute{M}$. Subsequently, we conduct an element-wise summation operation between $\acute{M}$ and the feature T_{i} produced by the transformer branch. The whole process can be represented by the following equation:

$$\acute{F}_i = \mathrm{Add}(T_{\mathrm{i}}, \acute{M}\,) \tag{4}$$

where

$$\acute{M} = \mathrm{f}(F_i)$$

where $f()$ denotes various convolutional operations on feature map F_{i} to enlarge the receptive field and the information catenation operation. The convolutional operations on feature map F_{i} is composed of four convolutional branches. The first level of each convolutional branch employs $1 \times t$ conventional kernel, where $t \in \{1,3,5,7\}$. Except for the first convolutional branch, each other conventional branch is composed of three conventional levels. For the second level of the second to fourth branches, the convolution operation employs $3 \times 1, 5 \times 1, 7 \times 1$ convolution kernel, respectively. For the third level of the second to fourth branches, 3×3 dilated convolution operations are employed. Each conventional branch output a feature map M_k (k = 1,2,3,4). The feature maps M_1 and M_k, which have different receptive fields, are fused together to obtain $\acute{M}$.

$$\acute{M} = Conv_{3\times3}(Concat(M_1, M_3, M_5, M_7)) \tag{5}$$

where

$$M_1 = Conv_{1\times1}(F_i)$$

$$M_k = DConv_{3\times3}\left(Conv_{k\times1}\big(Conv_{1\times k}(F_i)\big)\right)$$

In addition, we hope that F_i and T_i can learn useful features from their fused features $\acute{F_i}$ while retaining their respective original features. Inspired by the attention mechanism [35], we learned that spatial attention operation is used to extract the spatial relationship of features, focusing on the regions with key information in the image, thus improving the perception of local details. Meanwhile, channel attention focuses on learning the relationship between feature channels and ultimately selecting effective features. Therefore, we first obtain the important information of the fusion feature $\acute{F_i}$ in channel and spatial dimensions: the spatial attention map S_a and the channel attention map C_a. Then, we pass the information in S_a and C_a to the original input features F_i and T_i to make them learn the effective features of $\acute{F_i}$. Meanwhile, to maintain the original features of each of the two branches in this process, we perform a Split operation on the channel attention map to select different channel weights for F_i and T_i. After that, the features F_i and T_i of the two branches are first subjected to spatial level multiplication operation with the spatial attention map, and then subjected to channel level multiplication operation with different channel weights respectively. Finally, the final fused feature map $\tilde{F_i}$ is obtained by combining the results F_{oi} and T_{oi} produced by the two branches. This process can be expressed as equations:

$$F_{oi} = Split(C_a) \odot S_a \odot F_\mathrm{i} \tag{6}$$

$$T_{oi} = Split(C_a) \odot S_a \odot T_\mathrm{i} \tag{7}$$

$$\tilde{F_i} = Conv(Concat(F_{oi},\ T_{oi})) \tag{8}$$

where

$$C_a = s\left(FC\big(\mathrm{MaxPool}(\acute{F_i})\big) \oplus FC\big(\mathrm{AvgPool}(\acute{F_i})\big)\right)$$

$$S_a = \sigma\left(Conv\left(Concat\big(\mathrm{MaxPool}(\acute{F_i}), \mathrm{AvgPool}(\acute{F_i})\big)\right)\right)$$

where FC represents fully connected operation.

3.4. The partial decoder

The good performance of Unet-based image segmentation relies on the aggregation of multilevel features extracted from the encoder. For example, Unet aggregates all the hierarchical features extracted from the encoder, and there are many network variants that utilize Unet, such as Unet++ [11] and ResUnet [36]. However, research found that low-level features contribute less to the performance in comparison with high-level features, while the computational cost is high when both low-level and high-level features are used. Thus, to achieve a more efficient use of the features and reduction of computational cost, we developed a PD module, as shown in Figure 4 in the decoder path. In the PD module, we only use three high-level feature mappings $\{\tilde{F}_n,\ n = 2, 3, 4\}$. The specific steps are as follows: We first reshape the three feature mappings to the same channel size using 1 × 1 convolution. Then, we resample the resulting feature mappings to the same spatial resolution and concatenate them together. The final feature map $Z_0 = PD(\tilde{F}_2, \tilde{F}_3, \tilde{F}_4)$ is obtained using convolution, batch normalization and ReLu operations. Our PD module uses a small number of parameters to preserve multi-scale contextual information for localizing the approximate location of polyps.

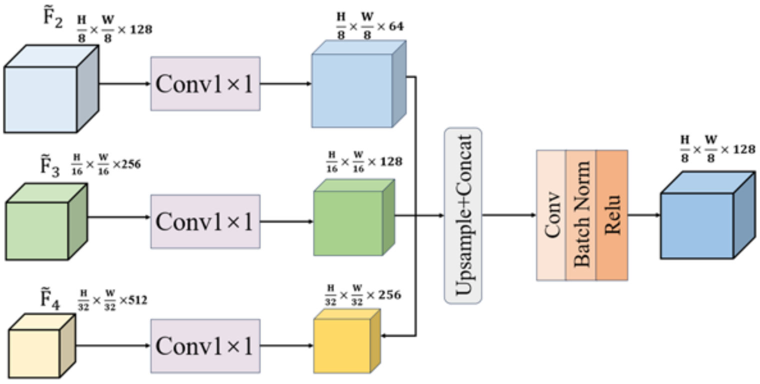

$\tilde{F}_2$ $\frac{H}{8}\times\frac{W}{8}\times128$
$\frac{H}{8}\times\frac{W}{8}\times64$
Conv1×1
$\tilde{F}_3$ $\frac{H}{16}\times\frac{W}{16}\times256$
$\frac{H}{16}\times\frac{W}{16}\times128$
Conv1×1
$\tilde{F}_4$ $\frac{H}{32}\times\frac{W}{32}\times512$
$\frac{H}{32}\times\frac{W}{32}\times256$
Conv1×1
Upsample+Concat
Conv
Batch Norm
Relu
$\frac{H}{8}\times\frac{W}{8}\times128$

Figure 4. The partial decoder.

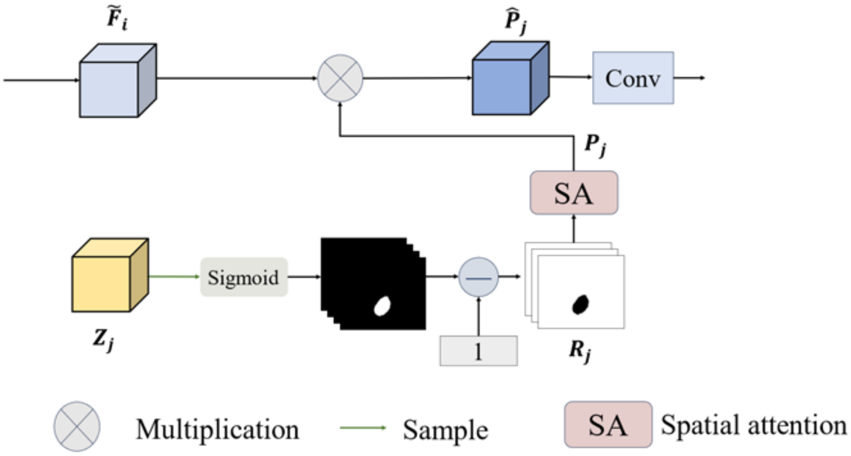


Figure 5. The reverse attention boundary enhancement (RABE) module.

3.5. The reverse attention boundary enhancement module

The network with dual-branch encoder and FIM module can only obtain the approximate positions of various polyps, lacking the refinement of polyp structure. Conversely, the general network for polyp segmentation lacks detailed boundary information and structural details. To address this issue, [14] introduced the inverse attention module for polyp segmentation and obtained some improvements in preserving the boundaries of polyps. Inspired by [14], we added a reverse attention boundary enhancement (RABE) module in the decoder section, as shown in Figure 5. The RABE module consists of reverse attention and spatial attention. We first use the reverse attention mechanism to focus on the details of the polyp boundary. The network can better identify the edge information between the target and the background, thereby making the boundary of the segmentation result clearer and more accurate, and then gradually incorporate it into the decoder to obtain the global segmentation feature map, specifically expressed by the following formula:

$$R_j = 1 - Sigmoid\big(Sample(Z_j)\big) \tag{9}$$

where Z_j denotes the feature map obtained from the decoder, and R_j denotes the output reverse attentional feature. Here, $j \in \{0, 1, 2, 3\}$. The operation *Sample* () denotes the sampling operation corresponding to the input of Z_j to the module, as shown in Figure 1(a).

We employed spatial attention to extract polyp location information from the initially segmented feature maps. When a segmentation network locates polyp boundaries in the feature map, extreme binarization of polyp regions and other regions could easily lose boundary details. To address this issue, we put more weight on the initially located non-polyp regions in the spatial direction while reducing the weights of the polyp regions. The spatial attention module can capture rich boundary information of polyps and enhance the performance of polyp segmentation. Let the output of the spatial attention module be P_j, and it can be computed from R_j by

$$P_j = R_j \times Sigmoid\left(Conv2\left(RELU\left(Conv1(R_j)\right)\right)\right) \tag{10}$$

where Conv represents a convolution operation, and RELU represents an activation function operation.

Because P_j can capture detailed boundary information, we combine it with the feature map $\tilde{F}_i$ to enrich the boundary details of polyps in the initial predicted segmentation map and finally get the feature map $\hat{P}_j$.

$$\hat{P}_j = \tilde{F}_i \otimes P_j \tag{11}$$

3.6. Loss function

We employed cross-entropy loss for the classification task, which is defined as follows:

$$L_{class} = -\frac{1}{Q}\sum_{Q=1}^{Q} y_Q\big(\log \hat{y}_Q\big) + \big(1 - y_Q\big)\log\big(1 - \hat{y}_Q\big) \tag{12}$$

where Q is the number of classes, y_Q represents true labels of a class, and $\hat{y}_Q$ represents the predicted labels.

For colonoscopy polyp segmentation, Dice loss is usually used. However, when the polyps are small, it could make significant changes in the network gradients. One way to mitigate the issue is to use binary cross entropy (BCE) loss to guide Dice loss to make the gradient reasonably small. Therefore, we combine the two losses for the segmentation task. These losses are defined as in [37]. In addition, we performed deep supervision on the outputs Z_j of the four decoders. Before calculating the deeply supervised losses, we up-sampled them to the same size as GT. Therefore, the total segmentation loss is

$$L_{seg} = \sum_{j=0}^{3} L\left(\hat{G}, Z_j^{up}\right) + L\left(\hat{G}, Y\right) \tag{13}$$

where

$$L = L_{BCE} + L_{DSC}$$

where $\hat{G}$ denotes GT, Z_j^{up} denotes the results after up-sampling to the original image size, which are used for deep supervision, and Y denotes the final prediction result.

The total loss for classification and segmentation is as follows:

$$L_{total} = L_{class} + L_{seg} \tag{14}$$

4. Experiments and results

The proposed segmentation and classification network was implemented using the PyTorch framework with NVIDIA RTX 3090Ti graphics environment. For the experiments, we trained the model using the Adam optimizer with a momentum of 0.9 and a weight decay of 1e-4. The initial learning rate was set to 0.01 and then reduced by half every 30 cycles. The batch size was set to 8, and the learning period was set to 100. Our network was pre-trained on the ImageNet dataset to accelerate network training. To evaluate the effectiveness of the proposed method, we conducted experiments on segmentation and classification tasks using five publicly available datasets: KvasirSEG [38], CVC-ClinicDB [39], CVC-ColonDB [40], ETIS-LaribPolypDB [41] and CVC-EndoSceneStill [42]. In these datasets, images were annotated by a specialized endoscopist. We adopted dataset division criterion by [9]: 900 images from KvasirSEG and 550 images from CVC-ClinicDB were used as the training set while the remaining images from KvasirSEG and CVC-ClinicDB plus all the images from the other three datasets (ETIS, CVC-ColonDB, and CVC-300) were used as the test set. Many existing methods have utilized this criterion for their experiments, so to be fair, we also used this division criterion in our experiments. Tables 1 and 2 show the dataset divisions for the two tasks, respectively.

Table 1. Division of the data set for the polyp segmentation task.

Datasets		Image
Training Set	KvasirSEG [38]	900
	CVC-ClinicDB [39]	550
Testing Set	KvasirSEG [38]	100
	CVC-ClinicDB [39]	62
	CVC- ColonDB [40]	380
	ETIS [41]	196
	CVC-300 [42]	60
All		2248

Table 2. Division of the data set for the polyp classification task.

Datasets		Pedunculated	Sessile
Training Set	KvasirSEG [38] CVC-ClinicDB [39]	1067	383
Testing Set	KvasirSEG [38] CVC-ClinicDB [39] CVC- ColonDB [40] ETIS [41] CVC-300 [42]	300	498
All		1367	881

4.1. Evaluation metric

For the classification task, we evaluated the performance using four metrics: specificity (Spe), recall (Rec), accuracy (ACC), and the area under the curve (AUC). According to [43], sessile polyps are at a higher risk of complications such as perforation or hemorrhage during the treatment process. Therefore, we considered the class of sessile polyps as positive samples, and the class of pedunculated polyps as negative samples. TP denotes the instances where the class of sessile polyps is correctly predicted. So, these four indicators can be represented by the following formulas:

$$Spesitivity = \frac{TP}{TP+FP} \tag{15}$$

$$Recall = \frac{TP}{TP+FN} \tag{16}$$

$$ACC \ = \frac{TP+TN}{TP+TN+FP+FN} \tag{17}$$

For the segmentation task, we used three evaluation metrics: Dice similarity coefficient (DSC) which was used to evaluate the accuracy of the segmentation; intersection over union (IoU), which was used to assess the internal consistency of the segmented objects; and Hausdorff distance (HD), which was used as a similarity metric. These three metrics are follows:

$$\mathrm{DSC} = \frac{2*TP}{2*TP+FN+FP} \tag{18}$$

$$IoU = \frac{TP}{TP+FP+FN} \tag{19}$$

$$HD = max(h(C,D), h(D,C)) \tag{20}$$

where

$$h(C,D) = \underset{c\in C}{max}\, \underset{d\in D}{min}\, ||(c-d)||$$

$$h(D,C) = \underset{d\in D}{max}\, \underset{c\in C}{min}\, ||(d-c)||$$

where TP, FP, FN, and TN represent true positive, false positive, false negative, and true negative, respectively. True positive means that the model correctly labels the pixels or regions in the image that belong to the polyps to be segmented. The DSC and IoU metrics have a range of [0, 1], with higher values representing better segmentation or classification results. On the other hand, for the HD metric, lower values represent better results.

4.2. Comparison experiment

4.2.1. Segmentation results

For segmentation of colon polyps, we compared the proposed network with several state-of-the-art models. These comparison models include CNN networks that are widely used for segmentation tasks, including Unet [10] and Unet++ [11]. We also compared the proposed network with the networks that are specifically designed for polyp segmentation, including ParNet [14], EMS-Net [44], BDG-Net [45] and BSCA-Net [17]. In addition, we also compared with some transformer-based segmentation models, including TransUnet [30] and TransFuse [18].

Table 3. Segmentation results on the KvasirSEG dataset.

Methods	DSC	IoU	HD (mm)	F_β^ω	S_α	E_ξ
Unet [10]	0.811	0.726	26.498	0.780	0.848	0.893
Unet++ [11]	0.821	0.738	23.458	0.797	0.856	0.900
ParNet [14]	0.898	0.840	14.537	0.885	0.915	0.948
TransUNet [30]	0.913	0.856	13.767	0.887	0.918	0.960
EMS-Net [44]	0.897	0.842	10.421	0.889	0.915	0.949
TransFuse [18]	0.918	0.868	**7.103**	0.902	0.917	0.962
BDG-Net [45]	0.915	0.863	7.235	0.906	0.920	0.964
BSCA-Net [17]	0.910	0.855	8.340	0.900	0.913	0.957
Ours	**0.932**	**0.882**	7.212	**0.918**	**0.934**	**0.975**

Table 4. Segmentation results on the CVC-ClinicDB dataset.

Methods	DSC	IoU	HD (mm)	F_β^ω	S_α	E_ξ
Unet [10]	0.876	0.818	16.498	0.870	0.915	0.943
Unet++ [11]	0.886	0.830	15.458	0.881	0.921	0.953
ParNet [14]	0.899	0.849	13.537	0.896	0.936	0.979
TransUNet [30]	0.929	0.887	7.167	0.913	0.942	0.978
EMS-Net [44]	0.923	0.874	7.021	0.923	0.949	**0.980**
TransFuse [18]	0.934	0.886	**4.235**	0.926	0.941	0.977
BDG-Net [45]	0.915	0.863	6.074	0.902	0.930	0.968
BSCA-Net [17]	0.926	0.887	5.387	0.912	0.940	0.973
Ours	**0.958**	**0.914**	4.732	**0.933**	**0.950**	**0.980**

Tables 3–7 presents the results of the proposed model for each metric on each of the five test datasets. In addition to the standard evaluation metrics, we also included three additional metrics inspired by ParNet to provide a more comprehensive evaluation of the model's performance. These additional metrics are a weighted measure (F_β^ω) [46] that combines recall and precision, an S-measure (S_α) [47] that evaluates the similarity between predicted and true values, and an E-measure (E_ξ) [48] that assesses similarity at both the pixel and global level. In the table, the optimal results were highlighted in bold, while the second-best results were highlighted in blue font. Except for HD, the

proposed network achieved the best segmentation results evaluated by other metrics on both the KvasirSEG and CVC-ClinicDB datasets. Specifically, for the KvasirSEG dataset, as shown in Table 3, our model outperforms the next best results by 1.4% in Dice, 1.4% in IoU, 1.2% in F_β^ω, 1.4% in S_α, and 1.1% in E_ξ. For the CVC-ClinicDB dataset, as shown in Table 4, our model achieved results of 0.958, 0.914, 4.732, 0.933, 0.950 and 0.980 for DSC, IoU, HD, F_β^ω, S_α, E_ξ, respectively.

For the ETIS dataset, the small size of polyps in the image makes segmentation challenging, and as seen in Table 5, most of the models performed poorly on this dataset. Compared with other models, our model showed better performance than other methods. Specifically, it achieved optimal results in five metrics, reaching 0.786, 22.49, 0.766, 0.902, and 0.921 in DSC, HD, F_β^ω, S_α *and* E_ξ, and is higher than the inferior results in each of the metrics by 2%, 1.6, 1%, 1% and 1.1% respectively.

Table 5. Segmentation results on the ETIS dataset.

Method	DSC	IoU	HD (mm)	F_β^ω	S_α	E_ξ
Unet [10]	0.398	0.335	57.35	0.357	0.662	0.673
Unet++ [11]	0.418	0.356	46.21	0.357	0.682	0.635
ParNet [14]	0.628	0.567	36.74	0.600	0.794	0.841
TransUNet [30]	0.718	0.672	27.95	0.735	0.842	0.894
EMS-Net [44]	0.682	0.611	29.83	0.660	0.820	0.876
TransFuse [18]	0.737	0.659	25.48	0.744	0.892	0.905
BDG-Net [45]	0.756	0.679	24.36	0.719	0.860	0.910
BSCA-Net [17]	0.768	**0.714**	24.04	0.753	0.886	0.908
Ours	**0.786**	0.713	**22.49**	**0.766**	**0.902**	**0.921**

Table 6. Segmentation results on the CVC- ColonDB dataset.

Methods	DSC	IoU	HD (mm)	F_β^ω	S_α	E_ξ
Unet [10]	0.584	0.493	42.54	0.559	0.740	0.773
Unet++ [11]	0.618	0.538	44.67	0.602	0.764	0.790
ParNet [14]	0.709	0.640	28.87	0.696	0.819	0.869
TransUNet [30]	0.779	0.683	26.55	0.728	0.827	0.903
EMS-Net [44]	0.715	0.642	27.93	0.707	0.822	0.891
TransFuse [18]	0.773	0.696	23.05	0.783	0.859	0.907
BDG-Net [45]	0.802	0.723	21.17	0.781	0.870	**0.912**
BSCA-Net [17]	0.783	0.720	20.46	0.775	0.869	0.904
Ours	**0.829**	**0.752**	**19.12**	**0.793**	**0.897**	0.907

For the CVC-ColonDB dataset and CVC-300, our method was also efficient because it outperformed other models in most of the metrics (see Table 6). Our method achieved the best results on five of these metrics, The results on the DSC, IoU, HD, F_β^ω, and S_α metrics are all optimal at 0.829, 0.752, 19.12, 0.793, 0.897, and 0.907, respectively. The CVC-300 dataset, as shown in Table 7, contained fewer images, and there was a great deal of inter-image variability among the images. Most of the models achieved stable performance on this dataset. Our proposed model achieved optimal results on two measures and suboptimal results on two other measures.

Table 7. Segmentation results on the CVC-300 dataset.

Methods	DSC	IoU	HD (mm)	F_β^ω	S_α	E_ξ
Unet [10]	0.743	0.648	21.35	0.708	0.840	0.877
Unet++ [11]	0.773	0.687	19.68	0.760	0.861	0.882
ParNet [14]	0.871	0.797	16.45	0.843	0.925	0.972
TransUNet [30]	0.893	0.824	13.13	0.879	0.939	0.971
EMS-Net [44]	0.900	0.834	10.98	0.885	0.943	0.978
TransFuse [18]	0.904	0.838	10.37	0.882	0.944	0.979
BDG-Net [45]	0.899	0.831	10.88	0.881	0.935	0.975
BSCA-Net [17]	**0.927**	**0.875**	**9.81**	**0.912**	0.950	0.985
Ours	0.908	0.833	11.04	0.897	**0.953**	**0.987**

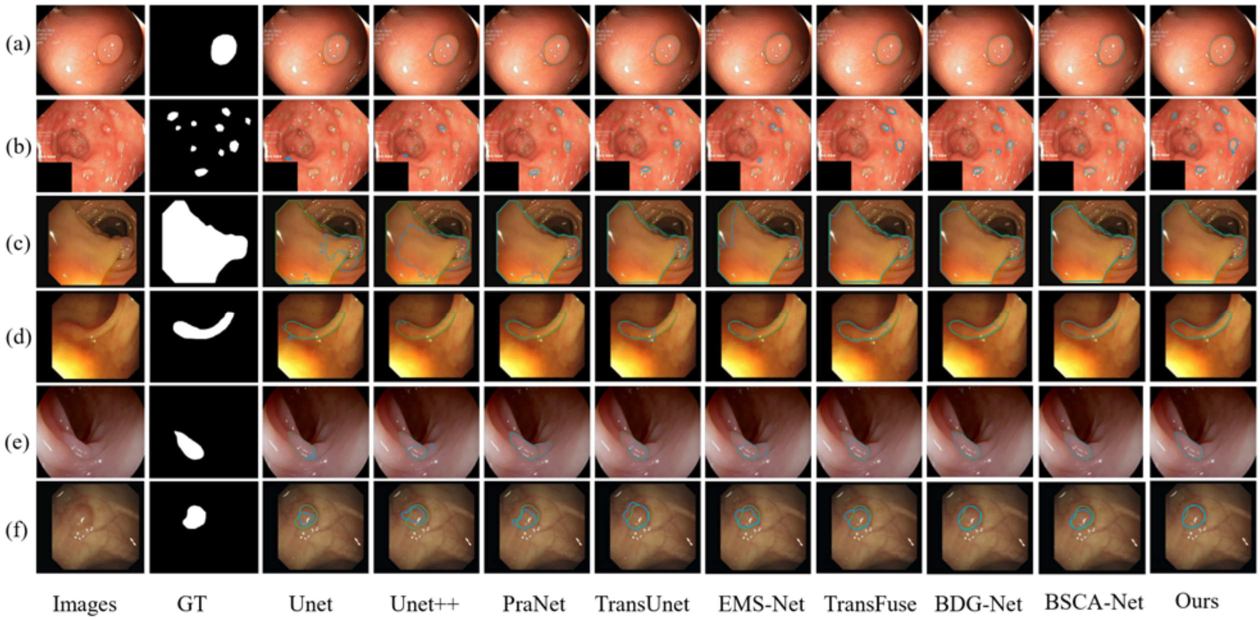


Figure 6. Visualization results for each comparison method, where the green line represents GT, and the blue line represents the segmentation result.

Next, we visually compared the segmentation results obtained with different models. The visualization results clearly demonstrated the superiority of our model for polyp segmentation. Figure 6(a) is an image with easily identifiable polyps, and we found that most methods produced good segmentation results. However, our method exhibited slightly better performance in capturing finer details. Figure 6(b) is an image with small and densely distributed polyps, and we found that some methods failed to identify the correct polyp regions, such as Unet. Others only identified a few polyp regions and have incorrect polyp location information, like Unet++ and EMS-Net. BDG-Net and BSCA-Net identified polyp regions in multiple locations compared to the other methods but still produced incorrect predictions. Our method effectively suppressed non-noise regions of interest, accurately localized polyps and correctly identified most polyp regions. Figure 6(c) is an image with large and irregularly shaped polyps, and our method demonstrated a strong scale adaptation. In contrast, all other methods were negatively affected to some extent and struggled to accurately segment the polyps. Figure 6(d)–(f) are images of polyps with blurred backgrounds that are difficult to distinguish from the direct border of the normal intestinal wall. Many methods struggled to correctly detect the edge region of the polyps. Unet and Unet++ performed the worst and almost failed to recognize the polyp regions. Pranet, TransUnet and EMS-Net had a small number of incorrectly detected regions and missed many target regions that are like the background. BDG-Net and BSCA-Net performed slightly better than Transfuse but still had a small number of under-segmented regions. In contrast, our method excelled in the polyp edge region and accurately detected the polyp boundary with the best segmentation effect.

In summary, the above observations demonstrate that our method outperformed other methods in capturing global context information and local detail information. It performed well on both large polyps in Figure 6(c) and small target polyps in Figure 6(b) and achieves the best detection of edge regions.

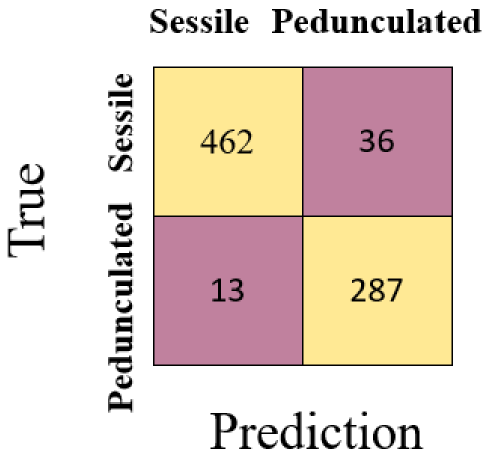
4.2.2. Classification results

For polyp classification, we compared our proposed model with several powerful and effective classification methods based on our dataset classification criteria. These methods included Inceptionv3 [49], MobileNetv3 [50], DenseNet [51], Vit [28], ResNet-50 [52], EfficientNet [53], TransUNet [30], and FusionM4Net [54]. Table 8 presents the average classification results for each metric evaluated on our experiments using the five test datasets. When combining the results from all five datasets, our model achieved the optimal performance with AUC of 0.915, Spe of 0.901, Rec of 0.934, and ACC of 0.937. These results represent 3%, 2%, 3%, and 3% improvements, respectively, over the Second-best results. This indicates that our model possesses strong learning and generalization capabilities. Upon reviewing Table 8, it is evident that the results achieved by other methods on the polyp classification were not very satisfactory. For instance, the AUC of each classification model ranged from approximately 0.813 to 0.855. This suggests that existing classification methods were not effective for recognizing polyps. Particularly, the ViT model and the transformer model performed relatively poorly compared to the other models. This implies that transformer-based classification models struggled to process polyp images with distinct local features despite their advantage in extracting global feature information. However, according to Table 8, our model overcame the performance bottleneck of traditional classification models and achieved more accurate classification results by leveraging the strengths of both CNN and transformer architectures.

Table 8. Average classification results for five datasets.

Methods	AUC	Spe	Rec	ACC
Inceptionv3 [49]	0.820	0.814	0.826	0.859
MobileNetv3 [50]	0.851	0.862	0.892	0.910
DenseNet [51]	0.842	0.846	0.879	0.913
Vit [28]	0.606	0.579	0.633	0.679
ResNet-50 [52]	0.814	0.808	0.822	0.876
EfficientNet [53]	0.859	0.809	0.910	0.895
TransUNet [30]	0.786	0.714	0.865	0.863
FusionM4Net [54]	0.881	0.884	0.901	0.907
Ours	**0.915**	**0.901**	**0.934**	**0.937**

For a clearer understanding of the classification results, we provide a confusion matrix in Figure 7. The horizontal axis represents the predicted classes, which are Sessile polyps and Pedunculated polyps, while the vertical axis represents the actual classes. The figure displays the number of polyp images that were misclassified as other classes for each class. In the confusion matrix, we observed that there were 36 instances where Sessile polyps were misclassified as Pedunculated polyps. On the other hand, there were 13 instances where Pedunculated polyps were misclassified. Therefore, Sessile polyps were more prone to misclassification compared with Pedunculated polyps.



True \ Prediction	Sessile	Pedunculated
Sessile	462	36
Pedunculated	13	287

Figure 7. Confusion matrix for classification results.

4.3. Ablation experiment

To demonstrate the effectiveness of our proposed model for polyp segmentation and classification, we selected three challenging datasets KvasirSEG, CVC-ClinicDB and CVC-ColonDB for ablation experiments. These experiments aimed to showcase the effectiveness of each individual module in our model. The results are presented in Tables 9–11. For the baseline model, we only utilized the transformer encoder (TE) and a simple U-Net decoder. Subsequently, we added the CNN branch (CB) to form a two-branch network. Further, we incorporated the FIM module and the RABE module into the network in sequence to assess the effectiveness of each module. The table displays the results obtained in various cases. It is evident that the model's performance gradually improved as each module was added to the network. Specifically, the inclusion of CB helped in learning local information, resulting in 2%, 1%, and 1% improvement in the segmentation index (DSC), 4%, 2%, and 2% improvement in the classification index (AUC) for datasets KvasirSEG, CVC-ClinicDB, and CVC-ColonDB respectively. The SFEM module enhanced the model's performance by preserving edge detail information, and the FIM module aided in learning by fusing information from the two branches. The data presented in the table demonstrates that each module of our model is effective and contributes to an improvement in model performance.

Figure 8 shows the visualization results of the ablation experimental results for segmentation with different settings of the dual-branch network. From the left to the right, new modules were added to the baseline dual-branch network one by one. The results clearly show that the localization effect of the polyps and the local segmentation effect were gradually improved. After fusing the information of the two branches through FIM, the information of the polyp acquired by the network also became more information, which was more friendly to the segmentation of some details and the segmentation of small targets. In addition, with the addition of the RABE module, the network's ability to detect the boundary region of polyps was improved, and the effect of polyp edge segmentation was more accurate. For the classification task, the first and fourth rows are dedicated to the classification of pedunculated polyps, and the model consistently classified them correctly from start to finish. several experimental processes in the middle of the second and third rows produced some incorrect class predictions, but with the addition of our proposed modules, the learning effect of the network was greatly improved accordingly, and therefore, the correct classification results were finally obtained. The last row is a typical example of misclassification of polyps.

Table 9. Ablation experiments on the KvasirSEG dataset.

Methods	Segmentation			Classification			
	DSC	IoU	HD	AUC	Spe	Rec	ACC
TE	0.870	0.848	14.384	0.842	0.855	0.875	0.891
TE+CB	0.887	0.855	12.874	0.881	0.872	0.901	0.909
TE+CB+FIM	0.907	0.861	9.273	0.913	0.890	0.929	0.918
TE+CB+FIM+RABE	0.920	0.874	7.945	0.927	0.914	0.947	0.936

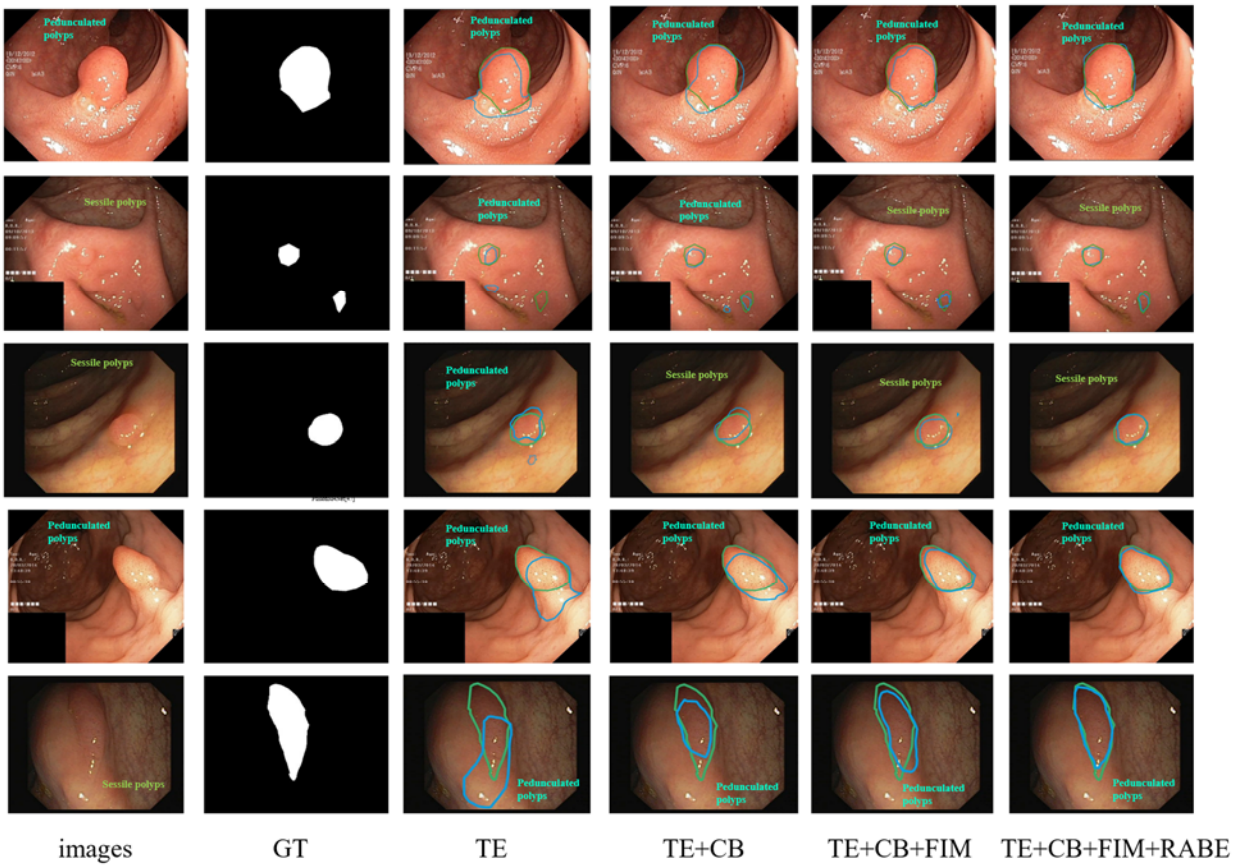


Figure 8. Visualization results of a multi-task ablation study, where the green line represents GT, and the blue line represents the segmentation result.

Table 10. Ablation experiments on the CVC-ClinicDB dataset.

Methods	Segmentation			Classification			
	DSC	IoU	HD	AUC	Spe	Rec	ACC
TE	0.890	0.848	13.047	0.837	0.839	0.852	0.881
TE+CB	0.917	0.862	10.852	0.871	0.862	0.891	0.904
TE+CB+FIM	0.929	0.874	8.409	0.903	0.883	0.920	0.915
TE+CB+FIM+RABE	0.935	0.899	6.235	0.915	0.901	0.934	0.937

Table 11. Ablation experiments on the CVC- ColonDB dataset.

Methods	Segmentation			Classification			
	DSC	IoU	HD	AUC	Spe	Rec	ACC
TE	0.786	0.714	23.047	0.833	0.845	0.782	0.893
TE+CB	0.795	0.722	22.852	0.851	0.873	0.810	0.916
TE+CB+FIM	0.807	0.728	21.324	0.875	0.883	0.822	0.921
TE+CB+FIM+RABE	0.818	0.740	20.423	0.885	0.891	0.834	0.937

4.4. Time complexity and efficiency analysis

The computational complexity and efficiency of deep learning models are crucial indicators for evaluating their prospects in clinical applications. Parameters (Param) and floating-point operations per second (FLOPs) serve as metrics for computational complexity, while frames per second (FPS) is a measure of analysis efficiency. Smaller values for Param and FLOPs indicate lower computational and time complexity, while higher FPS values suggest a faster model. Table 12 displays the values of Param, FLOPs, and FPS for each comparison method. As observed in the table, the Param of our method was approximately 42.4 M, which was lower than that of Unet++ [11] and BSCA-Net [17]. Furthermore, in terms of FPS, the model's efficiency in this paper was notably advantageous, ranking second only to EMS-Net [44] and BSCA-Net [17]. Additionally, our model demonstrated heightened sensitivity to polyp detection accuracy, achieving a commendable trade-off between efficiency, time complexity, and accuracy.

Table 12. Time complexity and efficiency analysis of each method.

Method	Params (M)	FLOPs (G)	FPS
Unet [10]	13.1	21.05	12
Unet++ [11]	48.9	108.76	15
ParNet [14]	30.5	13.1	24
TransUNet [30]	31.2	25.6	35
EMS-Net [44]	31.5	75.75	46
TransFuse [18]	26.2	19.8	37
BDG-Net [45]	32.7	10.84	26
BSCA-Net [17]	64.8	89.5	74
Ours	42.4	74.6	40

5. Conclusions and discussion

Polyp segmentation and classification have important applications in the diagnosis of colorectal cancers. Traditional image segmentation methods [55,56] and traditional image classification methods [57,58] offer low-accuracy performance. Thus, we proposed a multi-task network for polyp segmentation and classification that can better handle segmentation and classification of randomly located polyps with varying sizes and confusing edges and backgrounds. The main structure of the network is a combination of dual-branch encoders, which employ CNN and transformer as its two branches. We also designed several modules to make the model more effective. One module is the feature interaction module (FIM), aiming at eliminating the semantic gap between the two branches, and better fusion of the information obtained by the dual encoder while retaining the information of each branch. Another module is RABE. This module helps the model extract boundary information and enhances the segmentation performance, particularly for small targets and images with fuzzy boundaries.

We performed experiments on five public datasets. Experimental results show that the multi-task network proposed in this paper has high segmentation and classification accuracy and good reliability. However, the network proposed in this paper still has potential for improvement in two key areas: The small target area is small, which is difficult to accurately locate and capture, resulting in inaccurate segmentation. In addition, sessile polyps do not have obvious pedicle features and are not easy to identify, resulting in misclassification. Another possible work is to enhance the inference speed of our network on devices with low computational power. In our future research work, we will focus on addressing these challenges by optimizing the network architecture and reducing redundancy parameters. By doing so, we aim to better meet the requirements of high-precision and real-time clinical applications.

Use of AI tools declaration

The authors declare they have not used artificial intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest. Jinshan Tang is a guest editor for Mathematical Biosciences and Engineering and was not involved in the editorial review or the decision to publish this article.

References

1. A. Leufkens, M. G. H. Van Oijen, F. P. Vleggaar, P. D. Siersema, Factors influencing the miss rate of polyps in a back-to-back colonoscopy study, *Endoscopy*, **44** (2012), 470–475. https://doi.org/10.1055/s-0031-1291666
2. C. F. Chen, Z. J. Du, L. He, Y. J. Shi, J. Q. Wang, W. Dong, A novel gait pattern recognition method based on LSTM-CNN for lower limb exoskeleton, *J. Bionic Eng.*, **18** (2021), 1059–1072. https://doi.org/10.1007/s42235-021-00083-y

3. S. Tian, J. Zhang, X. Y. Shu, L. Y. Chen, X. Niu, Y. Wang, A novel evaluation strategy to artificial neural network model based on bionics, *J. Bionic Eng.*, **19** (2022), 1–16. https://doi.org/10.1007/s42235-021-00136-2
4. L. Xu, R. Maggar, A. B. Farimani, Forecasting COVID-19 new cases using deep learning methods, *Comput. Biol. Med.*, **144** (2022), 105342. https://doi.org/10.1016/j.compbiomed.2022.105342
5. J. X. Xie, B. Yao, Physics-constrained deep active learning for spatiotemporal modeling of cardiac electrodynamics, *Comput. Biol. Med.*, **146** (2022), 105586–105594. https://doi.org/10.1016/j.compbiomed.2022.105586
6. Q. Guan, Y .Z. Chen, Z. H. Wei, A. A. Heidari, H. G. Hu, X. H. Yang, et al., Medical image augmentation for lesion detection using a texture-constrained multichannel progressive GAN, *Comput. Biol. Med.*, **145** (2022), 105444–105449. https://doi.org/10.1016/j.compbiomed.2022.105444
7. R. K. Zhang, Y. L. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. Y. W. Lau, et al., Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain, *IEEE J. Biomed. Health Inf.*, **21** (2016), 41–47. https://doi.org/10.1109/JBHI.2016.2635662
8. M. F. Byme, N. Chapados, F. Soudan, C. Oertel, M. L. Pérez, R. Kelly, et al., Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model, *Gut*, **68** (2017), 94–100. https://doi.org/10.1136/gutjnl-2017-314547
9. F. Younas, M. Usman, W. Q. Yan, A deep ensemble learning method for colorectal polyp classification with optimized network parameters, *Appl. Intell.*, **53** (2023), 2410–2433. https://doi.org/10.1007/s10489-022-03689-9
10. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention*, Springer, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
11. Z. W. 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*, (2018), 3–11. https://doi.org/10.1007/978-3-030-00889-5_1
12. D. Jha, P. H. Smedsurd, D. Johansen, T. D. Lange, H. D. Johansen, P. Halvorsen, et al., A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation, *IEEE J. Biomed. Health Inf.*, **25** (2021), 2029–2040. https://doi.org/10.1109/JBHI.2021.3049304
13. D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, et al., Resunet++: An advanced architecture for medical image segmentation, in 2*019 IEEE International Symposium on Multimedia (ISM)*, (2019), 225–2255. https://doi.org/10.1109/ISM46123.2019.00049
14. D. P. Fan, G. P. Ji, T. Zhou, G. Chen, H. Z. Fu, J. B. Shen, et al., Pranet: Parallel reverse attention network for polyp segmentation, in *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
15. R. F. Zhang, G. B. Li, Z. Li, S. G. Cui, D. H. Qian, Y. Z. Yu, Adaptive context selection for polyp segmentation, in *Medical Image Computing and Computer Assisted Intervention*, Springer, (2020), 253–262. https://doi.org/10.1007/978-3-030-59725-2_25

16. G. P. Ji, G. B. Xiao, Y. C. Chou, D. P. Fan, K. Zhao, G. Chen, et al., Video polyp segmentation: A deep learning perspective, *Mach. Intell. Res.*, **19** (2022), 531–549. https://doi.org/10.1007/s11633-022-1371-y
17. Y. Lin, J. C. Wu, G. B. Xiao, J. W. Guo, G. Chen, J. Y. Ma, BSCA-Net: Bit slicing context attention network for polyp segmentation, *Pattern Recogn.*, **132** (2022), 108917. https://doi.org/10.1016/j.patcog.2022.108917
18. Y. D. Zhang, H. Y. Liu, Q. Hu, Transfuse: Fusing transformers and CNNs for medical image segmentation, in *Medical Image Computing and Computer Assisted-Intervention*, Springer, (2021), 14–24. https://doi.org/10.1007/978-3-030-87193-2_2
19. A. Galdran, G. Carneiro, M. A. G. Ballester, Double encoder-decoder networks for gastrointestinal polyp segmentation, in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event*, Springer, (2021), 293–307. https://doi.org/10.1007/978-3-030-68763-2_22
20. A. Amyar, B. Modzelewski, H. Li, S. Ruan, Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation, *Comput. Biol. Med.*, **126** (2020), 104037. https://doi.org/10.1016/j.compbiomed.2020.104037
21. Z. Wu, R. J. Ge, M. L. Wen, G. S. Liu, Y. Chen, P. Z. Zhang, et al., ELNet: Automatic classification and segmentation for esophageal lesions using convolutional neural network, *Comput. Biol. Med.*, **67** (2021), 101838. https://doi.org/10.1016/j.media.2020.101838
22. C. Chen, W. J. Bai, D. Rueckert, Multi-task learning for left atrial segmentation on GE-MRI, in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop*, Springer, (2018), 292–301. https://doi.org/10.1007/978- 3-030-12029-0_32
23. R. Zhang, X. Y. Xiao, Z. Liu, Y. J. Li, S. Li, MRLN: Multi-task relational learning network for mrivertebral localization, identification, and segmentation, *IEEE J. Biomed. Health Inf.*, **24** (2020), 2902–2911. https://doi.org/10.1109/JBHI.2020.2969084
24. Y. Zhou, H. J. Chen, Y. F. Li, Q. Liu, X. A. Xu, S. Wang, et al., Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images, *Med. Image Anal.*, **70** (2021), 101918–101920. https://doi.org/10.1016/j.media.2020.101918
25. K. Liu, N. Uplavikar, W. Jiang, Y. J. Fu, Privacy-preserving multi-task learning, in *2018 IEEE International Conference on Data Mining (ICDM)*, IEEE, (2018), 1128–1133. https://doi.org/10.1109/ICDM.2018.00147
26. G. K. Zhang, X. A. Shen, Y. D. Zhang, Y. Luo, D. D. Zhu, H. M. Yang, et al., Cross-modal prostate cancer segmentation via self-attention distillation, *IEEE J. Biomed. Health Inf.*, **26** (2021), 5298–5309. https://doi.org/10.1109/JBHI.2021.3127688
27. C. Wang, M. Gan, Tissue self-attention network for the segmentation of optical coherence tomography images on the esophagus, *Biomed. Opt. Express*, **12** (2021), 2631–2646. https://doi.org/10.1364/BOE.419809
28. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. H. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.
29. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *European Conference on Computer Vision*, Springer, (2020), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

30. J. N. Chen, Y. Y. Lu, Q. H. Yu, X. D. Luo, E. Adeil, Y. Wang, et al., TransUNet: Transformers make strong encoders for medical image segmentation, preprint, arXiv:2102.04306.
31. J. W. Wang, S. W. Tian, L. Yu, Z. C. Zhou, F. Wang, Y. T. Wang, HIGF-Net: Hierarchical information-guided fusion network for polyp segmentation based on transformer and convolution feature learning, *Comput. Biol. Med.*, **161** (2023), 107038. https://doi.org/10.1016/j.compbiomed.2023.107038
32. Y. L. Huang, D. H. Tan, Y. Zhang, X. Y. Li, K. Hu, TransMixer: A hybrid transformer and CNN architecture for polyp segmentation, in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, (2020), 1558–1561. https://doi.org/10.1109/BIBM55620.2022.9995247
33. 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), 616–633. https://doi.org/10.1093/jcde/qwac018
34. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, preprint, arXiv:2105.15203.
35. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module, preprint, arXiv:1807.06521.
36. F. I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data, *ISPRS J. Photogramm. Remote Sens.*, **162** (2020), 94–114. https://doi.org/10.1016/j.isprsjprs.2020.01.013
37. Z. Ma, Y. L. Qi, C. Xu, W. Zhao, M. Lou, Y. M. Wang, et al., ATFE-Net: Axial transformer and feature enhancement-based CNN for ultrasound breast mass segmentation, *Comput. Biol. Med.*, **153** (2023), 106533–106545. https://doi.org/10.1016/j.compbiomed.2022.106533
38. D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. D. lange, D. Johansen, et al., Kvasir-seg: A segmented polyp dataset, in *MultiMedia Modeling: 26th International Conference*, Springer, (2020), 451–462. https://doi.org/10.1007/978-3-030-37734-2 37
39. 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 Graphics*, **43** (2015), 99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007
40. J. Bernal, F. J. Sánchez, F. Vilariño, Towards automatic polyp detection with a polyp appearance model, *Pattern Recogn.*, **45** (2012), 3166–3182. https://doi.org/10.1016/j.patcog.2012.03.002
41. 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. Assisted Radiol. Surg.*, **9** (2014), 283–293. https://doi.org/10.1007/s11548-013-0926-3
42. D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, et al., A benchmark for endoluminal scene segmentation of colonoscopy images, *J. Healthcare Eng.*, **2017** (2017). https://doi.org/10.1155/2017/4037190
43. N. Shussman, S. D. Wexner, Colorectal polyps and polyposis syndromes, *Gastroenterol. Rep.*, **2** (2014), 1–15. https://doi.org/10.1093/gastro/got041
44. M. Wang, X. W. An, Y. H. Li, N. Li, W. Hang, G. Liu, EMS-Net: Enhanced Multi- Scale Network for Polyp Segmentation, in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, IEEE, (2021), 2936–2939. https://doi.org/10.1109/EMBC46164.2021.9630787

45. Z. Qiu, Z. H. Wang, M. M. Zhang, Z. Y. Xu, J. Fan, L. F. Xu, BDG-Net: boundary distribution guided network for accurate polyp segmentation, in *Medical Imaging 2022: Image Processing*, SPIE, (2022), 792–799. https://doi.org/10.1117/12.2606785
46. R. Margolin, L. Zelnik-Manor, A. Tal, How to evaluate foreground maps, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, IEEE, (2014), 248–255. https://doi.org/10.1109/CVPR.2014.39
47. D. P. Fan, M. M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, (2017), 4548–4557. https://doi.org/10.1109/ICCV.2017.487
48. D. P. Fan, C. Gong, Y. Cao, B. Ren, M. M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, preprint, arXiv:1805.10421.
49. 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*, IEEE, (2016), 1063–6919. https://doi.org/10.1109/CVPR.2016.308
50. A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. X. Tan, et al., Searching for MobileNetV3, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, IEEE, (2019), 1314–1324. https://doi.org/10.1109/ICCV.2019.00140
51. G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2017), 4700–4708. https://doi.org/10.1109/CVPR.2017.243
52. K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, (2016), 700–778. https://doi.org/10.1109/CVPR.2016.90
53. M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, preprint, arXiv:1905.11946v5
54. P. Tang, X. T. Yan, Y. Nan, S. Xiang, S. Krammer, T. Lasser, FusionM4Net: A multi-stage multimodal learning algorithm for multi-label skin lesion classification, *Med. Image Anal.*, **76** (2022), 102307. https://doi.org/10.1016/j.media.2021.102307
55. J. Tang, S. Millington, S. T. Acton, J. Crandall, S. Hurwitz, Ankle cartilage surface segmentation using directional gradient vector flow snakes, in *2004 International Conference on Image Processing, 2004. ICIP'04*, Singapore, **4** (2004), 2745–2748, https://doi.org/10.1109/ICIP.2004.1421672
56. J. Tang, S. Guo, Q. Sun, Y. Deng, D. Zhou, Speckle reducing bilateral filter for cattle follicle segmentation, *BMC Genomics*, **11** (2010), 1–9. https://doi.org/10.1186/1471-2164-11-S2-S9
57. J. Tang, X. Liu, H. Cheng, K. M. Robinette, Gender recognition using 3-D human body shapes, *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, **41** (2011), 898–908. https://doi.org/10.1109/TSMCC.2011.2104950
58. J. Xu, Y. Y. Cao, Y. Sun, J. Tang, Absolute exponential stability of recurrent neural networks with generalized activation function, *IEEE Trans. Neural Networks*, **19** (2008), 1075–1089. https://doi.org/10.1109/TNN.2007.2000060

©2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)