Research Article Vol. 15, No. 4 / 1 Apr 2024 / Biomedical Optics Express 2590
Biomedical Optics EXPRESS

Improved dual-aggregation polyp segmentation network combining a pyramid vision transformer with a fully convolutional network

FENG LI,[1,*] ZETAO HUANG,[1] LU ZHOU,[2] YUYANG CHEN,[1] SHIQING TANG,[1] PENGCHAO DING,[1] HAIXIA PENG,[2] AND YIMIN CHU[2]

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[2]Tongren Hospital, Shanghai Jiao Tong University School of Medicine, 1111 XianXia Road, Shanghai 200336, China
[*]lifenggold@163.com

Abstract: Automatic and precise polyp segmentation in colonoscopy images is highly valuable for diagnosis at an early stage and surgery of colorectal cancer. Nevertheless, it still posed a major challenge due to variations in the size and intricate morphological characteristics of polyps coupled with the indistinct demarcation between polyps and mucosas. To alleviate these challenges, we proposed an improved dual-aggregation polyp segmentation network, dubbed Dua-PSNet, for automatic and accurate full-size polyp prediction by combining both the transformer branch and a fully convolutional network (FCN) branch in a parallel style. Concretely, in the transformer branch, we adopted the B3 variant of pyramid vision transformer v2 (PVTv2-B3) as an image encoder for capturing multi-scale global features and modeling long-distant interdependencies between them whilst designing an innovative multi-stage feature aggregation decoder (MFAD) to highlight critical local feature details and effectively integrate them into global features. In the decoder, the adaptive feature aggregation (AFA) block was constructed for fusing high-level feature representations of different scales generated by the PVTv2-B3 encoder in a stepwise adaptive manner for refining global semantic information, while the ResidualBlock module was devised to mine detailed boundary cues disguised in low-level features. With the assistance of the selective global-to-local fusion head (SGLFH) module, the resulting boundary details were aggregated selectively with these global semantic features, strengthening these hierarchical features to cope with scale variations of polyps. The FCN branch embedded in the designed ResidualBlock module was used to encourage extraction of highly merged fine features to match the outputs of the Transformer branch into full-size segmentation maps. In this way, both branches were reciprocally influenced and complemented to enhance the discrimination capability of polyp features and enable a more accurate prediction of a full-size segmentation map. Extensive experiments on five challenging polyp segmentation benchmarks demonstrated that the proposed Dua-PSNet owned powerful learning and generalization ability and advanced the state-of-the-art segmentation performance among existing cutting-edge methods. These excellent results showed our Dua-PSNet had great potential to be a promising solution for practical polyp segmentation tasks in which wide variations of data typically occurred.

1. Introduction

Colorectal cancer (CRC) was the second most deadly cancer and third most common malignancy, which was estimated to account for approximately 9.4% of cancer-associated mortality worldwide and posed a serious threat to human health [1]. It often stemmed from small benign polyps progressing over time to gradually transform into malignant. Early diagnosis and excision of such diseased polyps could substantially decrease the occurrence and mortality of CRC and had

#510908 https://doi.org/10.1364/BOE.510908
Journal © 2024 Received 31 Oct 2023; revised 26 Feb 2024; accepted 8 Mar 2024; published 26 Mar 2024

Biomedical Optics EXPRESS

become a worldwide public health priority [2]. In clinical practice, colonoscopy was frequently adopted means of examination and considered a gold standard for detecting colorectal lesions known as polyps or adenomas [3]. During the colonoscopy, physicians entailed visual inspection of the bowel by means of an endoscope to access location information and boundary details of polyps, whose accuracy heavily relied on the ability and tedious effort of physicians. Even skilled clinicians may also fail to reach an agreement on the segmentation for the same polyp image. Hence, a viable solution was to propose automated and accurate polyp segmentation strategies, which could provide physicians with great assistance for precisely locating and segmenting polyp areas to make further diagnosis.

Unfortunately, it still faced many challenges as a result of the following factors. First, high variations in polyp's appearance, e.g., color, size, shape, and complex morphological features of polyps, complicated the segmentation task, even if they were of the same type. Second, the fuzzy boundary contrasts between polyps and their surrounding mucosas made the polyps camouflaged against other endoluminal structures, while spot interference occurred when the light source of the lens was reflected on the tissue fluid of intestinal mucosa. Third, the inconsistency in scanning equipments and parameter standards leaded to a domain shift in polyp imaging across different medical institutions. These issues easily contributed to polyp segmentation negatively. Given these challenging factors, there was a growing demand for an automated and precise segmentation method that was ability of identifying nearly all potential polyps at their early stages from colonoscopy images, which could guide physicians to perform quick localization of polyp area and precisely delineate its boundary. To this end, this work intended to establish a polyp segmentation network which could represent polyp features better in a medical scenario for improving polyp segmentation performance and generalization capability with respect to domain shift.

In this context, with the backing of computer vision technologies, numerous automated polyp segmentation methods had been presented and obtained remarkable progress in the past several years. Among diverse polyp segmentation approaches, early learning-based approaches [4–7] resorted to hand-crafted low-level features, including color, shape, texture, appearance, and some combination of these characteristics, etc. Yet, they usually yielded low-quality segmentation results and suffered from poor generalizability to complex scenarios, principally owing to the restricted representation ability of hand-crafted features when handling the high intra-class variations of polyps as well as low inter-class variations between polyps and hard mimics. Benefit from the rapid development of deep learning techniques in the field of medical image analysis, the automated polyp segmentation had progressively evolved from the early learning-based approaches to the deep learning methods, which could roughly fall into deep convolutional neural network (DCNN) based method [8–31], Transformer based technique [32–37], and hybrid architectures of Transformer and DCNN [38–44]. In DCNN-based methods, a fully convolutional network (FCN) with a pre-trained VGG model was utilized for recognizing and segmenting polyp regions through pixel-level predictions, while a modified FCN combining the patch selection strategy was designed to increase the precision in polyp segmentation. Nevertheless, FCN approaches only depended on low-resolution characteristics to perform the final predictions, leading to coarse segmentation results and blurred edges. Currently, the advanced methods were largely based on U-shaped encoder-decoder CNN network, which had become prevalent in the segmentation of polyp images. For instance, the U-Net++ [12] extended the U-Net structure by inserting efficient and densely connected nested-decoder subnetworks while introduced a deep supervision mechanism to enhance the use of aggregative features across multiple semantic scales. Later, ResUNet++ [13] integrated residual computation, squeeze and excitation (SE), and attention mechanism to further boost segmentation ability. In recent studies, SFA [19] adopted a selective feature amalgamation structure along with a boundary-sensitive loss function to recovery the sharp boundary between a polyp and its surrounding mucosa. PraNet [21] merged

Biomedical Optics EXPRESS

high-level characteristic information for generating global feature maps and under their guidance, a parallel reverse attention block was utilized to progressively delve the boundary cues and build the correlations between regions and boundary clues. In spite of improving the segmentation accuracy, most of these methods directly used element-wise addition or concatenation operators to incorporate features at different levels gradually, resulting in diluting really useful features by excessive redundant information and weakening the complementary features among different levels. Moreover, the receptive field of the aforementioned methods [8–31] were restricted and difficult to effectively summarize the global-level contextual information for handling size-varying polyps. Although some methods focused on capturing information from multiple scales and multiple receptive fields on the basis of the atrous spatial pyramid pooling (ASPP) block or DenseASPP, they would produce many extra parameters and computations, not to mention over-fitting. Different from DCNN-based methods, Transformer-based methods such as Polyp-PVT [33], SSFormer [35], and ESFPNet [37] etc., were able to perceive semantic distinctions from a global aspect via adaptively modeling long-distance relationships, and possessed stronger feature representation abilities by way of capturing global contextual information. Notwithstanding demonstrating impressive performance over DCNN-based alternatives, grabbing long-distance dependencies could destroy portion of task-specific critical local features and fail to model interactive features from the neighborhood, which could cause overly smooth segmentation for small/tiny polyps and vague boundaries between polyps. Apart from this, attention dispersion also occurred with the deepening of the Transformer model, whilst they often suffered from a dramatic performance decline on unseen out-of-distribution polyp data when domain shift issues existed, whose generalization capability remained lacked. The hybrid architectures of Transformer and DCNN including TransUNet [38], TransFuse [39], HS-Net [41], FCBFormer [42], TranSEFusionNet [43] and ECTransNet [44] etc., aimed to fuse the advantages of both models in a single architecture so as to strengthen underlying local features and restrict attention dispersion. However, feeding local feature information directly into the Transformer could not precisely deal with local contextual relationships, which could cause the local feature details to be overwhelmed by the dominant global context and generate inferior results in small/tiny polyp segmentation task. Another significant limitation of these hybrid architectures [39] fusing Transformer with DCNN was that the predicted segmentation maps were frequently lower in resolution compared with the input images, namely not full-size segmentation map. While the ECTransNet [44] had shown certain improvements in segmentation accuracy, there was still room for enhancement in building global contextual semantic relationships among pixels in the polyp feature maps.

To address the above-mentioned issues, inspired by encoder-decoder structures [11–14], Transformers [32,45,46], residual learning [47] and the work by Zhang [41], we developed a novel improved dual-aggregation polyp segmentation network, termed as Dua-PSNet, to enhance feature representation ability in network fusion and global information modelling by combining both the Transformer branch and the FCN branch in parallel, thereby achieving automatic and accurate full-size polyp segmentation from colonoscopy images. Our solution adapted to both inductive bias and powerful representation of global context, catching local and long-range characteristics adequately on the polyp. Specifically, in the Transformer branch, we exploited the B3 variant of pyramid vision transformer v2 (PVTv2-B3) [46] as the hierarchical encoder for capturing multi-scale features and establish global correlations between them. Then, we constructed a novel multi-stage feature fusion decoder (MFAD) to put emphasis on important local feature cues and selectively extend them into global features. In the MFAD, we made an attempt to design an adaptive feature aggregation (AFA) block for aggregating adjacent high-level feature representations of different scales derived from PVTv2-B3 encoder in a stepwise adaptive fashion, generating critical global semantic features. After that, we developed the ResidualBlock module to strength local boundary details camouflaged in low-level features while devised the

Biomedical Optics EXPRESS

selective global-to-local fusion head (SGLFH) module to selectively merge them with the global semantic features for eliminating the semantic gap between low-level and high-level features and constraining attention dispersion, enhancing the efficiency of feature fusion across different levels. In the FCN branch, we exploited the designed ResidualBlock module instead of the original residual block (RB) module to refine highly merged multi-scale features at full-size and supplement output of Transformer branch into full-size prediction mask. These two branches influenced and complemented each other for refining the semantic information of the network to improve the recognition of boundary ambiguous features and boost the network's performance in dealing with the scale variations of polyps. Our major contributions were summarized as follows:

1. We newly developed a dual-aggregation polyp segmentation network, named Dua-PSNet for achieving automatic and accurate segmentation of polyp areas at full-size in colonoscopy images, which combined the Transformer branch and the FCN branch in parallel. The fine features extracted by the FCN branch along with important features outputted by Transformer branch complemented reciprocally to enhance the distinctive ability for target information and weaken the background noises.

2. We constructed adaptive feature aggregation (AFA) module to merge and optimize adjacent high-level semantic characteristics at different scales from the hierarchical PVTv2-B3 encoder in a progressive adaptive way, which stimulated the most important global semantic feature extraction for polyp segmentation.

3. We added the ResidualBlock module into low-level feature information processing stage for improving the perception capability of local fine boundary cues, and then the selective global-to-local fusion head (SGLFH) module at the end of the Transformer branch for selectively injecting them into global semantic features and filling semantic gap between low-level and high-level characteristics.

4. Extensive experiments on five challenging polyp segmentation benchmark datasets demonstrated that our Dua-PSNet outperformed other advanced polyp segmentation methods and manifested the new state-of-the-art performance. The cross-dataset generalizability analysis validated its stronger generalization ability than current cutting-edge polyp segmentation approaches. By conducting thorough ablation studies, we confirmed the validity of all critical components in the proposed Dua-PSNet.

2. Related works

In this part, we presented an overview of recent efforts in relation to our work briefly from the following two perspectives: polyp segmentation, as well as multi-scale and multi-level amalgamation.

2.1. Polyp segmentation

Nowadays, polyp segmentation had been progressed by leaps and bounds and numerous exceptional works had paid attention to this area. Generally, the current polyp segmentation approaches were mainly divided into four types, including traditional learning-based approaches [4–7], DCNN-based methods [8–31], Transformer-based techniques [32–37], hybrid architecture combing Transformer and DCNN [38–44].

Traditional learning-based methods. Early polyp segmentation solutions highly depended on hand-crafted low-level features, such as color, shape, texture, appearance, or some combination of these characteristics [4–7]. Following the acquisition of manually designed characteristics, a classifier was often trained for detection or segmentation of polyp region from its surrounding tissue. For instance, Ameling et al. [4] extracted texture features with Grey-Level-Co-occurence

Biomedical Optics EXPRESS

and Local-Binary-Patterns for conducting polyp segmentation. Karkanis et al. [5] leveraged the covariances of the second-order textural measures computed over the wavelet frame decomposition of different color bands to extract color image features for representing different polyp areas, while performing classification using a step-wise linear discriminant analysis (LDA). Mamonov et al. [6] utilized the information about both the texture and the geometry to acquire a binary classification algorithm with pre-selection. Further, Tajbakhsh et al. [7] applied context and shape information to get rid of non-polyp structures and achieve reliably polyp location. Yet, these traditional learning-based methods often suffered from unsatisfactory segmentation results and lacked generalizability to real-world scenarios as a result of the poor representation capacity of hand-crafted characteristics.

DCNN-based methods. As the continuous development of DCNN, it had gained considerable attention with excellent learning ability and greatly promoted the development in polyp segmentation field. The introduction of FCN [8] provided a true solution to the issue of polyp segmentation at the pixel level. For example, Akbari et al. [9] adopted an FCN fusing image patch selection strategy and Otsu thresholding for polyp segmentation, yielding better segmentation results against traditional learning-based solutions. Similarly, Brandao et al. [10] converted VGG and ResNets into FCNs, and fine-tuned them for achieving polyp segmentation, which incorporated depth information restored using shape-from-shading strategy to provide a richer feature representation. However, most of these FCN methods [8–10] only resorted on low resolution feature information to carry out the final prediction of polyp regions, suffering from fuzzy boundaries and rough segmentation results. As a milestone work, U-shaped encoder-decoder structures instead of a single encoder in FCN had gradually come to be the mainstream network architecture with superior performance, e.g., U-Net [11], U-Net++ [12] and ResUNet++ [13]. Yeung et al. [14] introduced a dual attention-based deep neural network called Focus U-Net to promote selective learning on polyp characteristics by way of combining efficiently spatial and channel attentions into the sole focus gate block. Nevertheless, these skip connections operated with feature maps yielded by only a specific level of the encoder, which was not conducive to information transmission between the encoder and decoder and degraded the segmentation performance. To address this problem, Zhou et al. [12] reconstructed the dense skip connections to integrate feature maps between the encoder and the decoder in U-Net and established U-Net++, obtaining promising segmentation performance. Further, Tomar et al. [15] presented a dual decoder attention network built upon ResUNet++ to segment polyp regions. Yet, they mostly exploited a fixed size of kernel at each layer and only considered the feature maps from final decoder layer to conduct the reconstruction, which constrained the model's ability of exploring representational characteristics from different receptive fields. Based on the encoding-decoding network architecture, Sun et al. [16] introduced a dilated convolution for extracting and aggregating high-level semantic characteristics without resolution reduction, thereby improving the encoder network. Banik et al. [17] incorporated 2-D dual tree complex wavelet transform pooling with multiple skip connections to establish an enriched version of CNN named Polyp-Net for automatic polyp segmentation. Mahmud et al. [18] integrated depth dilated inception (DDI) module, deep fusion skip module (DFSM) and deep reconstruction module (DRM) into PolypSegNet for precise automated segmentation of polyp areas. Song et al. [24] proposed a parallel medical segmentation framework-AMNet to gain an improvement in polyp segmentation performance. Lin et al. [25] designed a bit-slicing context attention network (BSCA-Net) for improving performance in polyp segmentation by combining bit slice context attention (BSCA) module, split-squeeze-bottleneck-union (SSBU) block, multipath concatenation attention decoder (MCAD) as well as multi-path attention concatenation encoder (MACE). Shen et al. [26] presented a multi-scale coded colon polyp segmentation network, in which the attention mechanism integrating both spatial and channel dimensions was inserted into the encoding and decoding modules to enable the model to focus more on small polyp segmentation task. Lately,

Biomedical Optics EXPRESS

Li et al. [27] built the multiple feature association network (MFA-Net) with global attention mechanisms for improving the gains in segmentation performance of polyps. Further, more other U-shaped encoder-decoder based works related to polyp segmentation could be found in these literatures [11–24]. Notwithstanding achieving remarkable segmentation effect, the vast majority of these methods [11–22] lacked spatial details following pooling, and modeled long-range interdependencies insufficiently after performing multiple convolution operators. At the same time, most of these methods were prone to overfitting and suffering from weak generalization ability. Besides, there were also some methods [19–22] built upon U-Net, which recovered edge cues for polyp segmentation through establishing the association between the edge and area characteristics. Murugesan et al. [20] designed Psi-Net containing three parallel decoders to segment polyps, one of which was used for prediction, whereas the other two decoders were responsible for detecting contour and estimating distance map for attaining polyp shape and boundary information. But the correlation between them was not fully explored. Fang et al. [19] put forward a selective feature aggregation network (SFA) consisted of a shared encoder together with two reciprocally restrained decoders, to speculate areas and boundaries of polyps. In contrast, Fan et al. [21] developed a parallel reverse attention network (PraNet) for segmenting polyps, which utilized a parallel partial decoder (PPD) to produce a global map to serve as the guidance region along with a reverse attention (RA) block to probe the contour details and build the relationships between areas and edges. More recently, Song et al. [24] used a parallel attention block together with a reverse fusion component to establish associations between areas and edges for refining the edge information and improving polyp segmentation accuracy. Zhou et al. [28] constructed a cross-level feature aggregation network (CFA-Net) incorporating a boundary prediction module using a layer-wise strategy, to enhance hierarchical features to refine segmentation maps on polyps. Li et al. [31] utilized channel and spatial fusion block in conjunction with spatial and channel attention mechanism, feature complementary module, and shape block to construct a multi-scale channel spatial fusion network (MCSF-Net) for real-time polyp segmentation, enhancing lesion boundary feature extraction and the fusion of multi-scale features together with exhibiting excellent segmentation and real-time performance. However, the boundary information derived from these methods [19–22] was often ambiguous, resulting in sub-optimal performance in polyp segmentation task. In addition, these DCNN-based methods [8–31] were good at capturing local neighboring feature details by means of a local receptive field, yet could be powerless to establish long-distance interdependencies effectively, with restricted feature representation capacity to tackle size-varying polyps.

Our network combined both a Transformer branch built upon U-shaped encoder-decoder architecture with a FCN branch in parallel to capture the long-range contexts and local low-level feature details from different viewpoints, which provided a comprehensive insight into polyp characteristics and brought a segmentation performance improvement.

Transformer-based techniques. Different from DCNN, the Transformer consisting of self-attention gained similarities between all pairs of patches via calculating the dot product between their respective vectors so that features between all patches could be adaptively extracted and mixed, which enabled it to possess a strong capability of modeling long-term relations and reduce inductive bias. As the Transformer model emerged, Visual Transformer (ViT) [32] regarded each image as a sequence of patches (tokens) with a fixed size and subsequently forwarded them into multiple Transformer layers for seeking the association between each other. Due to only a single scale of output feature map with low resolution, it was challenging to directly adapt it to polyp segmentation task. Based this consideration, several methods [33–37] incorporated the pyramid structure in CNN into the design of Transformers, presenting a hierarchical Transformer with different stages. Dong et al. [33] utilized a pyramid vision Transformer (PVT) as backbone encoder and presented a polyp segmentation architecture called Polyp-PVT. Tang et al. [34] proposed a Dual-Aggregation Transformer Network (DuAT) to segment polyp regions, which

Biomedical Optics EXPRESS

adapted the PVT as the encoder for capturing richer feature cues. Wang et al. [35] proposed SSFormer comprising a PVT encoder and progressive locality decoder (PLD) to enhance robust and generalization ability for polyp segmentation. Nachmani et al. [36] proposed ResPVT framework to segment polyp, which used Transformer as an encoder to capture more feature representations about polyps. Chang et al. [37] established the ESFPNet architecture for segmenting polyp lesions, which exploited the Mix Transformer (MiT) as the backbone encoder and an efficient stage-wise feature pyramid (ESFP) in the decoder to enhance the usage of high-level semantic information. Despite manifesting notably impressive results, the Transformer network struggled to establish interactive characteristics from the neighborhoods adequately and represent fine-grained feature details accurately, as a result of the low spatial inductive bias and strong global receptive field. In addition, as the Transformer model was deepened, the global feature information was persistently mingled and accumulated, susceptible to attention diffusion. In our network design, we upgraded multi-stage feature aggregation decoder (MFAD) to adapt to the PVTv2-B3 backbone encoder, forming the Transformer branch, together with the FCN to underline local contextual relationship and constrain attention divergence.

Hybrid architectures combing Transformer and DCNN. Aiming at exploiting strengths of both designs, substantial works sought to improve model's ability of global contextual relations while maintaining a strong extraction of local detailed features by integrating Transformer and DCNN. For example, TransUNet [38] was developed by leveraging ResNet50 network and ViT as the encoder of U-Net for feature extraction and global interactions. But, it could result in a complex and bloated architecture which was inclined to overfitting. Zhang et al. [39] constructed TransFuse network to combine the Transformer and CNN parallelly so that global dependencies along with local spatial detailed information could be efficiently collected. Cai et al. [40] combined a shallow CNN encoder with a deep Transformer encoder for extracting richer feature details, and erected the PP-guided self-attention in the decoder so as to guide self-attention with the help of prediction maps, enhancing the model's sensitivity to polyp boundaries. Zhang et al. [41] investigated a hybrid semantic network (HSNet) that incorporated both the Transformer and CNN with a dual-branch framework to improve polyp segmentation, which introduced a hybrid semantic complementary module to capture long-term relationships and local appearance cues. Sanderson et al. [42] combined Transformer and FCN into a single structure to form FCN-Transformer hybrid network (FCBFormer) for achieving polyp segmentation. Zhang et al. [43] took full advantage of the combination of the CNN and Transformer on top of the U-Net structure to build a hybrid CNN-Transformer architecture called TranSEFusionNet, improving accuracy in polyp lesion segmentation. Li et al. [44] presented a novel polyp image segmentation network called ECTransNet by incorporating the local feature extraction ability of DCNN with the global contextual semantic correlation construction capability of Transformer, which achieved promising segmentation accuracy and generalization performance. Yet, there was still a need for further improvement in capturing long-range dependencies among pixels. In these methods [39,41], it could be imprecise to deal with local contextual relations by feeding local features directly into the Transformer, which made the local information be deluged by the dominant global context. Concurrently, they only combined the structures in a simple manner and overlooked the interaction between the two semantic representations, whereas down-sampling operations of these methods could cause the loss of local information, resulting in blurred boundary. What's more, most of these approaches [39] were not able to perform full-size segmentation map prediction. Inspired by this [40,42], we introduced Transformer and FCN in a parallel way to generate dual-aggregation polyp segmentation network (Dua-PSNet) for full-size polyp prediction. In our Dua-PSNet, the Transformer branch was developed to learn critical global semantic information, whereas the FCN branch was used for emphasizing fine local boundary details. These two branches influenced and complemented each other, enhancing the distinguished capability for polyps and achieving full-size segmentation.

Biomedical Optics EXPRESS

2.2. Multi-scale and multi-level amalgamation

Multi-scale feature representation offered a viable mean for tackling variations in scales with respect to segmentation task, whilst multi-level fusion provided multiple granularities for semantic segmentation. For instance, TransFuse [39] created a BiFusion module to incorporate multiple levels of features from both Transformer and CNN branches. SFA [19] embedded selective kernel module (SKM) into convolutional layers for dynamically extracting multi-scale and multi-receptive-field features from kernels of different sizes, and added up-concatenations between encoder and decoder in U-Net structure to aggregate these features. UACANet [23] conducted the atrous convolution with regard to high-level features, which could enlarge the perceptual field and degrade the loss of global spatial feature information to a certain degree, yet the boundary space information of subtle polyps could be lost owing to null convolution. PraNet [21] aggregated multiple high-level features extracted from Res2Net-based backbone network via a PPD component. MSNet [29] used the subtractive units for yielding different characteristics between contiguous levels of the network, which provided different perceptual fields concerning different subtractive units in a feature pyramid way to acquire abundant feature differences of multiple scales. M^2SNet [30] equipped with both the inter-layer and intra-layer subtraction structures to collect multi-scale complementary information among different levels, and effectively aggregated specific level features and multi-path cross-level distinctive features for generating the final prediction, thereby amplifying the perceptions of polyp regions. Yet, it was prone to lose the boundary details of small/tiny polyps and affected segmentation accuracy. AMNet [24] built a multi-scale interactive fusion network for obtaining richer local and global feature details on polyps via merging high-level and low-level characteristics of different scales working together to supplement the position and spatial information, and introduced the parallel network of attention mechanisms for augmenting the model's segmentation capability. MFA-Net [27] integrated a parallelly dilated convolutions arrangement (PDCA) block between the encoder and the decoder to excavate critical feature representations and introduced a multi-scale feature restructuring module (MFRM) to reorganize and merge semantic information at different scales from the encoder, while cascaded global attention stacking (GAS) block in the decoder to enhance global attention perception and guide shallower features through the use of deeper features. Polyp-Net [17] amalgamated 2-D dual-tree complex wavelet transform pooling with local gradient weighting-embedded level-set method to achieve multi-model pixel-level fusion, suppressing high-intensity false area and ensuring smoothness of polyp contour. PolypSegNet [18] utilized deep fusion skip module (DFSM) for fusing different scales of features produced by different levels of the encoder, and created deep reconstruction module (DRM) to restore and optimize decoded feature maps of multiple scales from different levels of decoders, enabling skip inter-connections establishment between encoder and decoder as well as the semantic gap between them reduction. Segformer [48] predicted features of different scales and depths individually by simple up-sampling, and then parallelly fuse them using a multi-stage feature aggregation algorithm. DuAT [34] used global-to-local spatial aggregation (GLSA) module to simultaneously mine local spatial details and global spatial semantic information, and constructed selective boundary aggregation (SBA) module for fine-tuning polyp boundaries. SSFormer [35] used a pyramid Transformer encoder for multi-scale feature extraction and progressive locality decoder for multi-stage feature fusion, enabling characteristics of different depths and representation capabilities to guide reciprocally. The outstanding performance illustrated that the decoder approach of multi-stage feature amalgamation was helpful to boosting the performance of the Transformer in the task of dense prediction. Lately, MCSF-Net [31] introduced the channel and spatial fusion module to amalgamate high-level multi-scale features through using channel and spatial attention mechanisms while developed a feature complementary module to fuse low-level multi-scale feature maps, which not only effectively captured the spatial positional information of polyps but also accurately preserved lesion boundaries. ResPVT [36] implemented PVT as

Biomedical Optics EXPRESS

backbone encoder to extract multi-stage feature information and adopted fusion module (FM) to meld the high-level features. ECTransNet [44] embedded an edge complementary module to fuse and complement polyp feature maps at multiple resolutions for effectively mining crucial edge information of polyps, and employed a residual-based feature aggregation decoder to adaptively merge high-level and low-level features, thereby better retaining spatial position information and reinforcing the model's ability to segment local details. On a similar note as [31,36,44], our Transformer branch employed a stepwise adaptive method to generate global semantic information and focus on local edge details, and then selectively aggregated them, improving the feature information processing capability of different levels and achieving more effective feature fusion. Besides, more studies on multi-scale and multi-level amalgamation could refer to these literatures [8–16,20,22,25,26,28]. The vast majority of these methods combined different level features by utilization of simple element-wise addition or concatenation operations, which tended to create substantially excessive information to attenuate the really useful characteristics and weaken the complementary feature cues between different levels.

In our network design, we utilized the PVTv2-B3 as backbone encoder for capturing different levels of features from input images and explore the correlations between them. In the decoder, we constructed the AFA module to aggregate resultant high-level characteristics from the encoder to form global semantic features in a stepwise adaptive manner, while designed the ResidualBlock module to enhance extraction capability of local boundary details from low-level features. Further, we proposed the SGLFH module to selectively inject local boundary information into global semantic features, which could enable characteristics of different representation abilities to guide mutually and the fusion more efficient.

3. Methodology

In this part, we initially provided an overview of the presented Dua-PSNet for full-size polyp segmentation. Then, we presented two key branches, including the Transformer branch and FCN branch. In the Transformer branch, we introduced PVTv2-B3 encoder and multi-stage feature aggregation decoder (MFAD), respectively, while elaborately depicted the AFA, ResidualBlock and SGLFH components in the MFAD. At last, we developed the entire loss function.

3.1. Overall architecture

The whole architecture of the developed Dua-PSNet for polyp segmentation was illustrated in Fig. 1(a). The model involved two key parallel branches starting from a $H \times W$ input images, one of which was the Transformer branch and returned reduced-size $\frac{H}{4} \times \frac{W}{4}$ feature maps, the other one of which was an FCN branch and returned full-size $H \times W$ feature maps. Specifically, a colonoscopy image with a resolution of $h \times \omega$ was fed into Transformer and FCN branches simultaneously. In the Transformer branch, we leveraged PVTv2-B3 as an image encoder to derive 4 pyramid features $F_i \in R^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i}$ and model long-range dependencies between different features, where $i \in \{1, 2, 3, 4\}$ and $C_i \in \{64, 128, 320, 512\}$. For accurate polyp segmentation, we constructed the MFAD to conduct stepwise adaptive aggregation on high-level features from PVTv2-B3 encoder for extracting the most important global semantic features as well as focus more on local boundary cues from low-level features. In the MFAD, we used convolution units to reduce the number of channels from the last 3 feature maps F_2, F_3 and F_4 to 64, while fed them into the developed AFA module to implement high-level feature fusion in a progressive adaptive way generating global feature map $O_1 \in R^{\frac{H}{4} \times \frac{W}{4} \times 64}$. Meanwhile, was forwarded into the designed ResidualBlock module producing $O_2 \in R^{\frac{H}{4} \times \frac{W}{4} \times 64}$, which enhanced local edge perception ability from low-level features. Afterwards, we presented the SGLFH module to selectively aggregate feature maps O_1 and O_2, resulting in the feature map $F_T \in R^{\frac{H}{4} \times \frac{W}{4} \times 64}$, which could encourage semantic information of different expression powers to guide mutually and bridge the semantic

Biomedical Optics EXPRESS

gap between high-level and low-level features. The resulting feature map $F_T \in R^{\frac{H}{4}\times\frac{W}{4}\times 64}$ was up-sampled into full-size $F_T' \in R^{H\times W\times 64}$ and concatenated with the feature map $F_c \in R^{H\times W\times 32}$ generated by the FCN branch along the channel dimension. Ultimately, the concatenated feature map was tackled into the final full-size polyp segmentation map through the prediction head module.

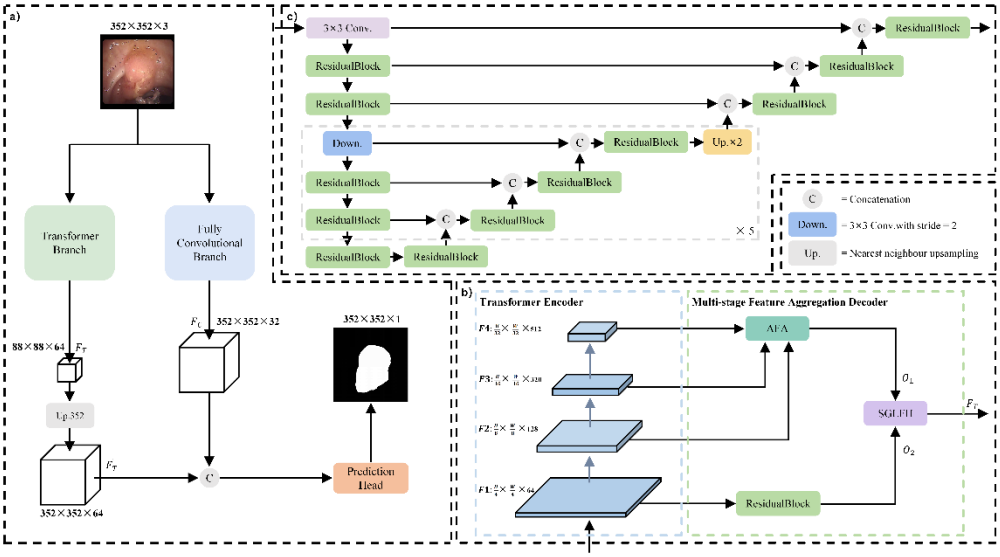


Fig. 1. Overall architecture of the developed Dua-PSNet for polyp segmentation (a), including two critical parts: Transformer branch (b) and FCN branch (c). AFA: adaptive feature aggregation module; SGLFH: selective global-to-local fusion head module.

3.2. *Transformer branch*

Our implementation of the Transformer branch was shown in Fig. 2, which used a pre-trained PVTv2-B3 on ImageNet [49] as an image encoder and designed the AFA, ResidualBlock and SGLFH modules to constitute a MFAD. The PVTv2-B3 encoder produced a 4-level feature pyramid, and then served as the input to the MFAD. In the MFAD, three high levels of feature pyramids were coped with by the AFA module in a stepwise adaptive manner to learn important global semantic context, whereas the remain one low-level pyramid was processed by the ResidualBlock module which improved the ability in representing local fine edge characteristics. In the following, the resulting local features were selectively injected into global semantic features through the SGLFH module, thereby filling the semantic gap between different level features and restricting attention dispersion.

3.2.1. Transformer encoder

To ensure our model enough generalization and powerful multi-scale feature extraction capability in polyp segmentation task, we utilized the PVTv2-B3 (See Fig. 2(a)) built upon a progressive shrinking pyramid structure as the image encoder. In contrast with traditional Transformer, the PVTv2-B3 used a linear spatial reduction attention (SRA) layer for reducing the computational burden, and an overlapping patch embedding via strided convolution to obtain more consistency of spatial information, while introduced zero-padding position encoding into the PVT in convolutional feed-forward network to more easily handle inputs with varying resolutions. In specific, the entire PVTv2-B3 encoder was split into 4 stages, each of which embodied both

Biomedical Optics EXPRESS

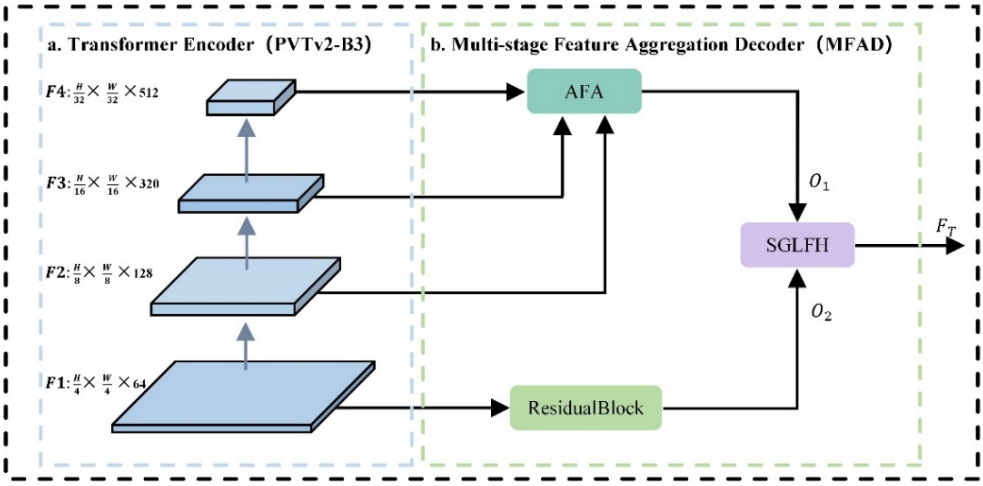


Fig. 2. The architecture of the Transformer branch, including the PVTv2-B3 encoder and multi-stage feature aggregation decoder (MFAD). AFA: adaptive feature aggregation module; SGLFH: selective global-to-local fusion head module.

an overlapping patch embedding layer and a linear-layer Transformer encoder. The feature pyramid with four stages was generated from a rough level (4-stride) to a fine level (32-stride) in a progressive way. In the first stage, given an input image with the size of $H \times W \times 3$, the image was firstly separated into $\frac{HW}{4^2}$ patches, and each of size was $4 \times 4 \times 3$. Then, the flattened patches were provided into a linear projection to acquire embedded patches of size $\frac{HW}{4^2} \times 64$. Subsequently, these embedded patches together with corresponding position embedding were fed into a linear-layer Transformer encoder with $L1$ layers, and the output was reshaped to a feature map $F_1 \in R^{\frac{H}{4} \times \frac{W}{4} \times 64}$. In the same way, taking the feature map resulting from the preceding stage as the input, we could derive the following three feature maps of different scales $F_i \in R^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i}$, where $i \in \{2, 3, 4\}$ and $C_i \in \{128, 320, 512\}$. Among these feature maps, F_2, F_3 and F_4 provided high-level semantic information related with polyps, whilst F_1 contained low-level local details.

3.2.2. Multi-stage feature aggregation decoder (MFAD)

Considering the most existing Transformer models for polyp segmentation lacked local low-level detailed feature processing capability, we proposed a novel MFAD for feature pyramids to stepwise aggregate multiple stages of features. The MFAD mainly consisted of the AFA, ResidualBlock and SGLFH modules, which could be depicted in Fig. 2(b).

Adaptive feature aggregation (AFA) module Taking inspiration from the work [29,30,41], we created the AFA module to gather more varied and distinctive high-level semantic information by stepwise adaptively aggregating three high-level feature maps from Transformer encoder, as shown in Fig. 3. Specifically, we firstly reduced the number of channels to 32 for F_2, F_3 and F_4 by use of local emphasis (LE) module comprising 3×3 convolution operators with padding of 1, group normalization (GN) and SiLU activation functions in sequence, resulting in F_2', F_3' and F_4', respectively. At the same time, our model could also refocus attention on neighboring features for decreasing attention dispersion by utilization of the local receptive field within the LE module, thus amplifying the weights of the neighboring patches associated with the center patch and highlighting critical local characteristics of each patch. Then, we fused the feature maps of F_3' and F_4', as follows: we up-sampled F_4' by 2 and forwarded the resulting feature map through the

Biomedical Optics EXPRESS

LE module, generating F_4''. Meanwhile, F_4'' was multiplied by F_3', and then the resultant feature map was further concatenated with F_4''. After that, we processed the concatenated feature via the LE module and then performed double up-sampled operation, obtaining the aggregated feature map M_2. The above processing steps could be represented by the follows:

$$M_2 = Up_2(LE(Concat(LE(F_3) \otimes LE(Up_2(LE(F_4))), LE(Up_2(LE(F_4))))))) \tag{1}$$

where Up_2 and Up_4 denoted a 2-fold and 4-fold nearest neighbour up-sampling operation, respectively, $\otimes$ stood for element-wise product, *Concat* represented a channel-wise concatenation operation, and *LE* was a sequence of operations comprising a 3×3 convolutional layer, GN, followed by SiLU activation function.

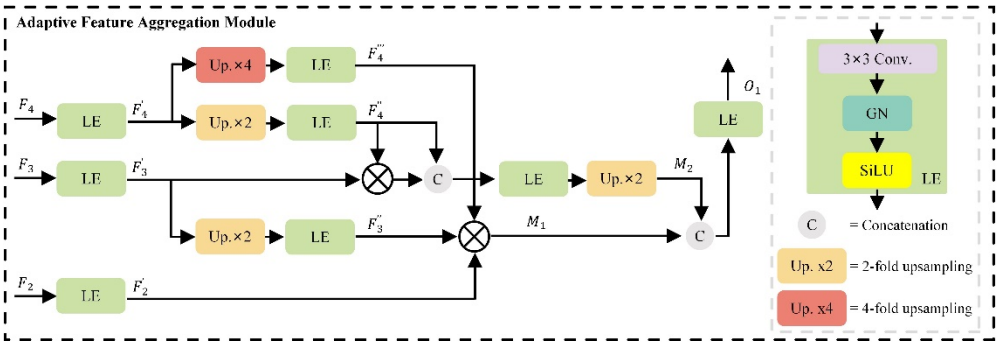


Fig. 3. Detailed structure of the developed adaptive feature aggregation (AFA) component. LE: local emphasis module.

Similarly, we aggregated the feature maps of F_2', F_3' and F_4', as follows: we conducted a 4-fold and 2-fold up-sampling on F_4' and F_3', separately. And then, the resulting feature maps were streamed into the LE module, leading to feature maps F_4''' and F_3'', individually. Further, we carried out an element-wise multiplication operation on F_2', F_3'' and F_4''' to obtain the feature map M_1, thus the fused feature map could be represented as follows:

$$M_1 = LE(F_2) \otimes LE(Up_2(LE(F_3))) \otimes LE(Up_4(LE(F_4))) \tag{2}$$

At last, we concatenated between M_1 and M_2, and the resultant feature map was passed through the LE module for obtaining the fused global feature map O_1, which could be depicted by:

$$O_1 = LE(Concat(M_1, M_2)) \tag{3}$$

Note that through the aforementioned procedures, the AFA module was able to aggregate adjacent high-level features from the Transformer encoder in a stepwise adaptive fashion, generating the most important global semantic information.

ResidualBlock module In order to boost the extraction ability of local low-level feature details for Transformer feature pyramid, we constructed residual blocks named ResidualBlock module through the incorporation of the LE modules as well as a skip connection with convolutional layer with 1×1 kernel (denoted as $C_{1\times1}$) and GN, as displayed in Fig. 4(a). The LE module consisted of 3×3 convolution, GN and SiLU activation function in an order arrangement. Compared to original RB module (See Fig. 4(b)) in FCBFormer network, we made small changes in the order of 3×3 convolution along with adding 1×1 convolution and GN in residual connection. This sequence was designed to optimally leverage the effects of these three operations. We first performed 3×3 convolution operation on high-resolution low-level features from PVTv2-B3

Biomedical Optics EXPRESS

encoder to extract the local boundary information related to polyp lesions. Then, GN was applied immediately to help in normalizing the distribution of the resulting features and mitigating internal covariate shift. Finally, we introduced a SiLU activation function to the normalized features, fostering local edge feature representation power. Through equipping with two units (LE modules) consisting of 3×3 convolution, GN and SiLU activation function in sequence, the boundary details of low-level features could be emphasized. In addition, we placed 1×1 convolution and GN in the residual connection for strengthening cross-channel interaction of boundary information in high-resolution low-level features and scaling their feature channel dimensions. By doing so, it could more efficiently cope with edge details of polyp lesions. Concretely, we first provided the low-level feature map F_1 into two sequential LE modules and an unit with 1×1 convolutional layer and GN, respectively, generating feature maps F_1' and F_1''. Next, we directly implemented an element-wise addition operation on the resulting feature maps followed by a SiLU activation function, obtaining the feature map O_2, which could be expressed as follows:

$$O_2 = S\text{iLU}(LE(LE(F_1)) \otimes GN(C_{1\times1}(F_1))) \tag{4}$$

As such, more fine boundary details from low-level features could be underlined and the detailed information processing capability of our model could be strengthened.

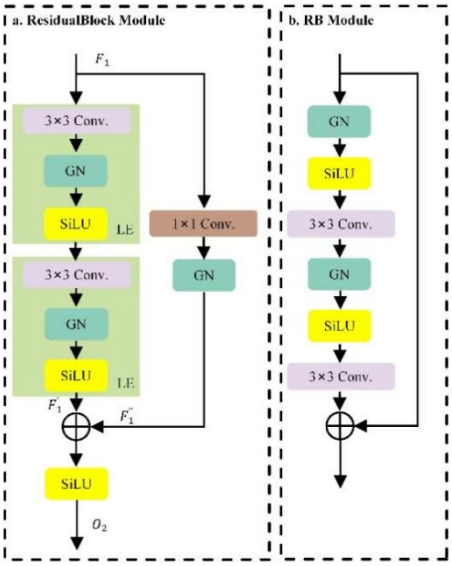


Fig. 4. The proposed ResidualBlock module (a) and the original RB module in FCBFormer (b).

Selective global-to-local fusion head (SGLFH) module With the goal of making effective integration between high-level and low-level characteristics, we implemented selectively fusion on feature maps O_1 and O_2, by designing the selective global-to-local fusion head (SGLFH) module. In the SGLFH module, we utilized two symmetric attention units (AUs) to adaptively pick up mutual representations from two inputs (O_1', O_2) before fusion, thereby exhibiting its selective characteristic, as shown in Fig. 5. The low-level and high-level feature information was fed into the two AUs in different ways for compensating the missing critical spatial edge information of the high-level semantic features and the loss of semantic information of low-level features, followed by concatenation operation to aggregate the outputs of two AUs. As given in

Biomedical Optics EXPRESS

Fig. 5, the AU function *PAU* process could be expressed as follows:

$$T'_H = \sigma(LE(O'_1)), T'_L = \sigma(LE(O_2)) \tag{5}$$

$$PAU(O_2, O'_1) = (T'_L \otimes O_2) \oplus (T'_H \otimes O'_1 \otimes (\Theta T'_L)) \oplus O_2 \tag{6}$$

where $O'_1 = Up_2(O_1)$ and O_2 referred to the input features. Up_2 indicated a 2-fold up-sampling operation. *LE* described a sequence of operations comprising a 3×3 convolutional layer, GN, followed by SiLU activation function. σ denoted sigmoid activation function. $\otimes$ stood for element-wise multiplication. Θ represented the reverse operation by subtracting the feature T'_L, refining the imprecise and coarse estimation into an accurate and complete prediction map. At last, the output features of the two AUs were concatenated along the channel dimension, which could be described by:

$$Z = Concat(PAU(O'_1, O_2), PAU(O_2, O'_1)) \tag{7}$$

where O'_1 contained high-level semantic information while O_2 comprised rich boundary details. The resulting concatenated feature maps were further fed into two sequential units composed of convolutional layer with 1×1 kernel, GN and SiLU activation function for generating ultimate output feature map F_T in the Transformer branch. By designing SGLFH module with attention units, the boundary information and semantic information could be selectively aggregated to depict more fine-grained contours of polyp lesions and recalibrate their locations. The above process could be summarized as:

$$F_T = SiLU(GN(C_{1\times1}(SiLU(GN(C_{1\times1}(Z)))))) \tag{8}$$

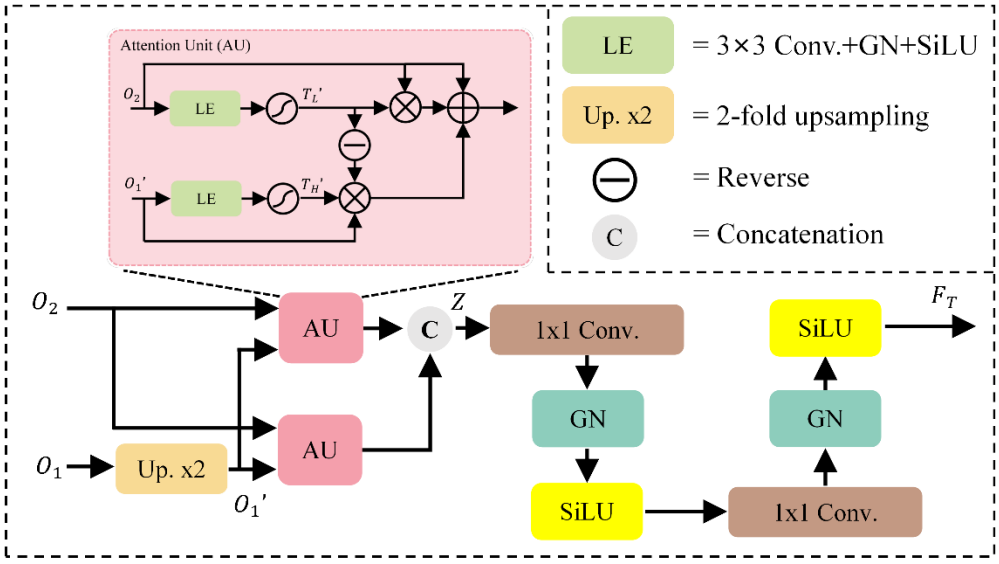


Fig. 5. The flowchart of the proposed selective global-to-local fusion head (SGLFH) module. AU: attention unit.

In addition, the feature map F_T was further implemented 4-fold up-sampling operation to generate F'_T, which maintained consistent size with the input image.

Biomedical Optics EXPRESS

3.3. FCN branch

The FCN branch (See Fig. 1(c)) was characterized as a composition of the ResidualBlock module, strided convolutional layers for down-sampling, nearest neighbour interpolation for up-sampling, and dense U-Net type skip connections. This design encouraged extraction of highly merged features of different scales required for matching outputs of the Transformer branch into full-size segmentation maps. Through the encoder of FCN branch, the feature map $F_j \in R^{\frac{H}{2^j} \times \frac{W}{2^j} \times C_j}$ was attained in sequence from the down-sampling layer 1 to 5, where $j \in \{1, 2, \ldots, 5\}$ and $C_j \in \{32, 32, 64, 128, 512\}$. On the contrary, by way of the decoder of FCN branch, the feature resolution was progressively restored by a factor of 2 from the up-sampling layer 1 to 5, whereas the number of feature channels was reduced consecutively by a factor of 2 ranging from the second to fifth up-sampling layer.

3.4. PredictionHead (PH) module

Compared with the PredictionHead (PH) module in FCBFormer network [42] (See Fig. 6(b)), we made small modification on it including substituting the designed ResidualBlock module for the RB module, as illustrated in Fig. 6(a). The PH module received a full-size feature tensor derived from stitching output tensors from the Transformer branch and FCN branch to perform polyp segmentation mask prediction at full-size. Concretely, the stitched full-size feature tensor was passed through two consecutive ResidualBlock modules followed by 1×1 convolutional layer and SiLU activation function to yield a full-size segmentation map. Each layer of PH module produced 64 channels, apart from the prediction layer (1×1 convolution + SiLU) returning a single channel. Through aggregating complementary features generated by the Transformer branch along with the FCN branch, the PH module was able to achieve accurate prediction of polyp segmentation map.

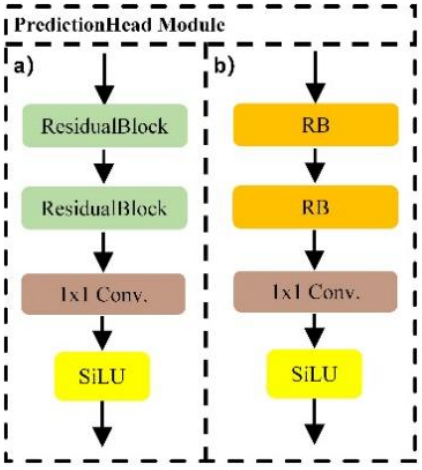


Fig. 6. The PredictionHead (PH) module in the Dua-PSNet (a), and the FCBFormer (b).

3.5. Loss function

It was reported by works [50] that using joint loss functions with adaptive weight coefficients could boost the model's performance with faster convergence speed. Hence, we combined weighted binary cross-entropy loss (L^{ω}_{BCE}) with the weighted Dice loss (L^{ω}_{Dice}) for supervision of the network training. The entire loss function $L(S, G)$ for the proposed Dua-PSNet could be

Biomedical Optics EXPRESS

formulated as follows:

$$L(S,G) = \lambda_1 L^{\omega}_{BCE}(S,G) + \lambda_2 L^{\omega}_{Dice}(S,G) \tag{9}$$

where S referred to the prediction masks and G represented the ground truths, respectively. λ_1 and λ_2 indicated the adjustable weighting coefficients. It was noteworthy that L^{ω}_{BCE} was able to put more emphasis on hard pixels rather than giving all pixel equal weights and L^{ω}_{Dice} could increase weighted coefficients of hard pixels to emphasize their importance, which may tackle the foreground-background imbalance problem.

4. Experiments and results

In this part, we provided the details on datasets and evaluation indicators, as well as implementation details. Then, we performed comparison between the developed Dua-PSNet with the cutting-edge polyp segmentation methods from quantitative and qualitative perspectives. After that, we measured our model's generalization ability across different datasets. Finally, aiming at clarifying the effectiveness of the principal components in the proposed Dua-PSNet, we carried out a series of ablation experiments.

4.1. Datasets and evaluation indicators

4.1.1. Datasets

In order to assess the ability of the presented Dua-PSNet to segment polyps, we conducted a series of experiments on five benchmark datasets, each of which was described in detail as follows:

Kvasir-SEG [51]**:** This dataset was created by the Vestre Viken Health Trust of Norway, and contained 1,000 polyp images gathered from multiple colonoscopy video sequences, where the image resolution distribution varied between 332×487 and 1920×1072. It provided a more fine-grained pixel-level segmentation annotation validated by experienced gastroenterologists.

CVC-ClinicDB [52]**:** This dataset was provided by the Hospital Clinic of Barcelona and comprised 612 images with a resolution of 288×384, which were captured from 23 colonoscopy video sequences of 13 different patients.

CVC-ColonDB [6]**:** It consisted of 380 images with a fixed resolution of 500×570 extracted from 15 colonoscopy videos with a sample of 20 frames at random from each sequence, in which each image was accompanied by a binary mask for putting emphasis on the polyp regions.

ETIS-LaribPolypDB [2]**:** ETIS was an early polyp segmentation dataset, which involved 196 polyp images with a resolution of 966×1225.

CVC-300 [53]**:** It was a cross-domain dataset, and embraced 60 polyp images with a resolution of 500×574.

4.1.2. Evaluation protocols

Following recommendations from common methods for polyp segmentation [21,42], we utilized mean Dice (*mDice*), mean IoU (*mIoU*), mean absolute error (MAE) [54], weighted F-measure (F^w_β) [55], S-measure (S_α) [56], and mean E-measure (mE_ϕ) [57] as our assessment indicators for comprehensively investigating our model's performance. Among these indicators, *mDice* and *mIoU* were regional level similarity measures and mainly highlighted the internal consistency of segmented objects. MAE was a pixel-by-pixel comparison metric, while F^w_β comprehensively considered the precision and recall. S_α concentrated on the structural similarity at the region and object level, and E_ϕ paid attention to the segmentation performance at the pixel and image level.

4.2. Implementation details

The developed Dua-PSNet was built under the PyTorch framework, which was trained utilizing a single NVIDIA RTX3060 GPU with 12 GB memory. Different data augmentation strategies were applied during training, including a gaussian blur, color jitter, horizontal and vertical flips,

Biomedical Optics EXPRESS

and affine transforms. Aiming to assess the learning and generalization power of the proposed Dua-PSNet, following the same settings as in [21,42] we separately partitioned Kvasir-SEG and CVC-ClinicDB into training, validation and test sets with a ratio of 8:1:1 in our experiment A. Similarly, we randomly chose 1,288 and 162 images from CVC-ClinicDB and Kvasir-SEG to form the training and validation sets, respectively, while an overall 798 images (100 images selected from Kvasir-SEG dataset, 62 images chose from CVC-ClinicDB benchmark, 380 images collected from CVC-ColonDB benchmark, 196 images gathered from ETIS-LaribPolypDB dataset, and the remain 60 images from CVC-300) from the used five benchmark datasets were adopted as the test set in our experiment B. The specific data partitioning was illustrated in Table 1. For the ablation studies, we also exploited the Kvasir-SEG and CVC-ClinicDB to evaluate the baseline performance of FCN along with PVTv2-B3, and then investigate the effect with successive additions of PVTv2-B3, AFA, ResidualBlock, and SGLFH modules. Besides, we adjusted the size of the input images into 352×352, whilst normalized the RGB values into a range between -1 and 1. The values of λ_1 and λ_2 for loss function were assigned into 1 and 0.2, empirically. The AdamW [58] optimizer with an initial learning rate of $1e-4$ was adopted for optimization of the proposed model during the training. In order to accelerate the model's convergence, the learning rate was stepped down by a factor of 2 when the *mDice* on the validation set did not further boost over 10 epochs until hitting a minimum of $1e-6$. Eventually, the model with the highest validation *mDice* was considered as the final one. All models were trained with a mini-batch size of 4 for more than 200 epochs, unless otherwise specified. During the testing stage, our model predicted full-size binary segmentation maps for RGB images with $H \times W$ spatial dimensions.

Table 1. Datasets used in this study.

Experiment A					
Dataset	Images	Input size	Train	Valid	Test
Kvasir-SEG [51]	1000	Variable	800	100	100
CVC-ClinicDB [52]	612	288×384	490	61	61
Experiment B					
Dataset	Images	Input size	Train	Valid	Test
Kvasir-SEG [51]	1000	Variable	800	100	100
CVC-ClinicDB [52]	612	288×384	488	62	62
CVC-ColonDB [6]	380	500×570	-	-	380
ETIS-LaribPolypDB [2]	196	966×1225	-	-	196
CVC-300 [53]	60	500×574	-	-	60

4.3. Comparison with state-of-the-art approaches

4.3.1. Comparison approaches

In an effort to investigate the effectiveness of the developed Dua-PSNet, we compared it with a variety of current state-of-the-art counterparts in the domain of polyp segmentation, including U-Net [11], U-Net++ [12], PraNet [21], MCSF-Net [31], SSFormer [35], FCBFormer [42], SFA [19], MSNet [29], SANet [22], M2SNet [30], Polyp-PVT [33], CFA-Net [28], ESFPNet [37], TranSEFusionNet [43] and ECTransNet [44]. Note that, for these methods, we performed retraining and testing using their publicly available codes and recommended parameter settings as well as default data augmentation methods reported in their original literatures. Among data augmentation strategies used in these models and our approach, there only existed slight differences. These slightly different augmentation methods did not significantly affect the comparison of models, which could ensure the relative fairness of comparisons.

Biomedical Optics EXPRESS

4.3.2. Quantitative comparison

Table 2 provided the quantitative comparison results between our model and five advanced methods on both CVC-ClinicDB and Kvasir-SEG benchmark datasets in experiment A. On the CVC-ClinicDB benchmark, we could observe that our model outperformed all compared approaches by a large margin in terms of *mDice* and *mIoU* metrics. Specifically, in comparison with DCNN-based approaches (such as U-Net [11], U-Net++ [12], and PraNet [21]), our model had more obvious advantages. For instance, our model brought considerable performance improvements of 3.7%, 10.6%, 1.6% and 4.3%, 15.2%, 2.2% in *mDice* and *mIoU*, against U-Net, U-Net++ and PraNet, respectively. When compared to other Transformer-based methods (including SSFormer [35] and FCBFormer [42]), our model also manifested excellent learning ability, observing an increase of 2.0% and 3.1% in *mDice* and *mIoU* over SSFormer, and an improvement of 1.5% and 1.4% against FCBFormer. In the Kvasir-SEG dataset, our model also consistently gained the best segmentation results in terms of *mDice*. For example, our model achieved 0.6% and 1.3% gains over SSFormer with respect to *mDice* and *mIoU*, respectively, while performed competitively (*mDice*: 0.9253 vs 0.9231, *mIoU*: 0.8746 vs 0.8752) against FCBFormer. In addition, FCBFormer and SSFormer obtained relatively better segmentation results compared to the other DCNN-based approaches on both datasets. According to these results, the effectiveness of our model could be demonstrated.

Table 2. Quantitative comparison results with different advanced approaches for polyp segmentation on the CVC-ClinicDB and Kvasir-SEG benchmarks in experiment A. The bold fonts denoted the best results.

Datasets	CVC-ClinicDB [52]		Kvasir-SEG [51]	
	mDice	mIoU	mDice	mIoU
U-Net [11]	0.9145	0.8654	0.8629	0.8176
U-Net++ [12]	0.8453	0.7559	0.7475	0.6313
PraNet [21]	0.9358	0.8867	0.9011	0.8403
SSFormer [35]	0.9318	0.8777	0.9196	0.8616
FCBFormer [42]	0.9362	0.8943	0.9231	**0 . 8752**
Ours	**0.9514**	**0.9083**	**0.9253**	0.8746

Table 3 summarized the quantitative comparison results of our model against thirteen cutting-edge methods on the CVC-ClinicDB and Kvasir-SEG benchmarks in experiment B. From the comparison results in Table 3, it could be seen that our model likewise achieved considerable performance gains over other segmentation methods. In particular, comparing to SSFormer, our model attained 3.3%, 4%, 2.3%, 2.2%, and 1.8% score improvements on average in terms of *mDice*, *mIoU*, F_β^w, S_α, and mE_ϕ on the CVC-ClinicDB dataset. Among all evaluation indicators, it performed the best on *mDice*, *mIoU*, F_β^w, and S_α metrics, reaching up to 0.939, 0.895, 0.936 and 0.950, respectively. On the Kvasir-SEG dataset, our model gained overall performance increases of 1.8%, 2.9%, 2.9%, 0.9% and 1.8% against FCBFormer, respectively, concerning *mDice*, *mIoU*, F_β^w, S_α, and mE_ϕ. In contrast, the MAE score of our model reduced by 0.5%. Here, our model manifested best results (*mDice*:0.922 and *mIoU*:0.874) in *mDice* and *mIoU*, exceeding the Polyp-PVT by 0.9% and 1.1%, separately. Even when compared to recently developed state-of-the-art DCNN-based M2SNet, Transformer-based ESFPNet and hybrid CNN-Transformer-based TranSEFusionNet, our model still obtained consistently best segmentation results in all evaluation metrics on both datasets. Notably, our model was slightly inferior to recently proposed CFA-Net (mE_ϕ:0.983 vs 0.989 and MAE:0.008 vs 0.007 on the CVC-ClinicDB dataset; mE_ϕ:0.960 vs 0.962 and MAE:0.024 vs 0.023 on the Kvasir-SEG dataset) in mE_ϕ and MAE indices, but it still maintained advantages over the CFA-Net in *mDice*, *mIoU*,

Biomedical Optics EXPRESS

F_β^w, and S_α. Further, on the CVC-ClinicDB dataset, our model brought significant gains of 1.6% and 1.7% in *mDice* and *mIoU* as comparison with the ECTransNet, respectively, whereas maintained excellent segmentation performance consistent with the MCSF-Net (*mDice*: 0.939 vs 0.941, and *mIoU*: 0.895 vs 0.895). On the Kvasir-SEG dataset, our model consistently surpassed the ECTransNet and the MCSF-Net, with 2.1% and 1.1% increases in *mDice* and 2.7% and 1.3% boosts in *mIoU*, respectively. On the basis of these results, we could conclude that our model had relatively better overall segmentation performance against the highly competitive ECTransNet and MCSF-Net. Again, these results indicated that our model had strong feature learning power, proving its superiority.

Table 3. Quantitative comparison results with different advanced approaches for polyp segmentation on the CVC-ClinicDB and Kvasir-SEG benchmarks in experiment B. The bold fonts denoted the best results.

Datasets	CVC-ClinicDB [52]					
	mDice	mIoU	F_β^ω	S_α	$mE_\emptyset$	MAE
SFA [19]	0.700	0.607	0.647	0.793	0.840	0.042
PraNet [21]	0.899	0.849	0.896	0.936	0.963	0.009
MSNet [29]	0.918	0.869	0.913	0.946	0.973	0.008
SANet [22]	0.916	0.859	0.909	0.939	0.971	0.012
SSFormer [35]	0.906	0.855	0.913	0.928	0.965	0.008
FCBFormer [42]	0.934	0.883	0.932	0.946	0.978	0.010
M^2SNet [30]	0.919	0.870	0.917	0.945	0.974	0.009
Polyp-PVT [33]	0.932	0.889	0.933	0.948	0.982	0.008
CFA-Net [28]	0.933	0.883	0.924	**0 . 950**	**0.989**	**0.007**
ESFPNet [37]	0.908	0.857	0.907	0.931	0.962	0.010
TranSEFusionNet [43]	0.744	0.654	0.712	0.841	0.879	0.033
ECTransNet [44]	0.923	0.878	0.926	**0.950**	0.976	0.011
MCSF-Net [31]	**0.941**	**0.895**	0.925	0.945	0.973	0.010
Ours	0.939	**0.895**	**0.936**	**0.950**	0.983	0.008
Datasets	Kvasir-SEG [51]					
	mDice	mIoU	F_β^ω	S_α	$mE_\emptyset$	MAE
SFA [19]	0.723	0.611	0.670	0.782	0.834	0.075
PraNet [21]	0.898	0.840	0.885	0.915	0.944	0.030
MSNet [29]	0.905	0.849	0.892	0.923	0.947	0.028
SANet [22]	0.904	0.847	0.892	0.915	0.949	0.028
SSFormer [35]	0.917	0.864	**0.916**	0.922	0.958	**0.022**
FCBFormer [42]	0.904	0.845	0.880	0.916	0.942	0.029
M^2SNet [30]	0.908	0.852	0.901	0.922	0.950	0.025
Polyp-PVT [33]	0.913	0.863	0.910	0.924	0.956	0.023
CFA-Net [28]	0.915	0.861	0.903	0.924	**0.962**	0.023
ESFPNet [37]	0.880	0.812	0.870	0.891	0.933	0.040
TranSEFusionNet [43]	0.783	0.686	0.744	0.842	0.878	0.058
ECTransNet [44]	0.901	0.847	0.873	0.913	0.941	0.032
MCSF-Net [31]	0.911	0.861	0.880	0.915	0.943	0.030
Ours	**0.922**	**0.874**	0.909	**0.925**	0.960	0.024

Biomedical Optics EXPRESS

4.3.3. Qualitative comparison

Figure 7 depicted the qualitative comparison results of our model over five cutting-edge polyp segmentation approaches in experiment A. On the basis of the visualizable comparison results, it could be observed that the results of our model were in most agreement with the ground truths, while surpassed the other compared approaches under different challenging conditions. Concretely, for small size of polyps in the first and second rows, our model could perform accurate segmentation on them. Nevertheless, U-Net++ completely failed to segment them, and U-Net produced some over-segmented regions. In this case, FCBFormer, SSFormer and PraNet generated several un-related or over-segmented areas, resulting from the neglect of the edge feature details from the deep characteristics. With respect to polyps with variable shapes and large sizes in the third and fourth rows, U-Net and U-Net++ manifested worse than the other methods, while FCBFormer, SSFormer and PraNet produced some errors with un-related or under-segmented regions. In the case of blurred boundaries between the polyps and background in the fifth and sixth rows making it more challenging to identify them, our model was able to focus more on the boundary details and segment them more accurately by contrast to the other methods. On the contrary, U-Net and U-Net++ conducted the worst among these methods. And yet, the predicted edges by PraNet were overly smooth, whilst FCBFormer and SSFormer missed some fine details. In a nutshell, the qualitative comparison results in experiment A demonstrated that our model had a good capability to deal with different challenging cases with respect to polyp segmentation.

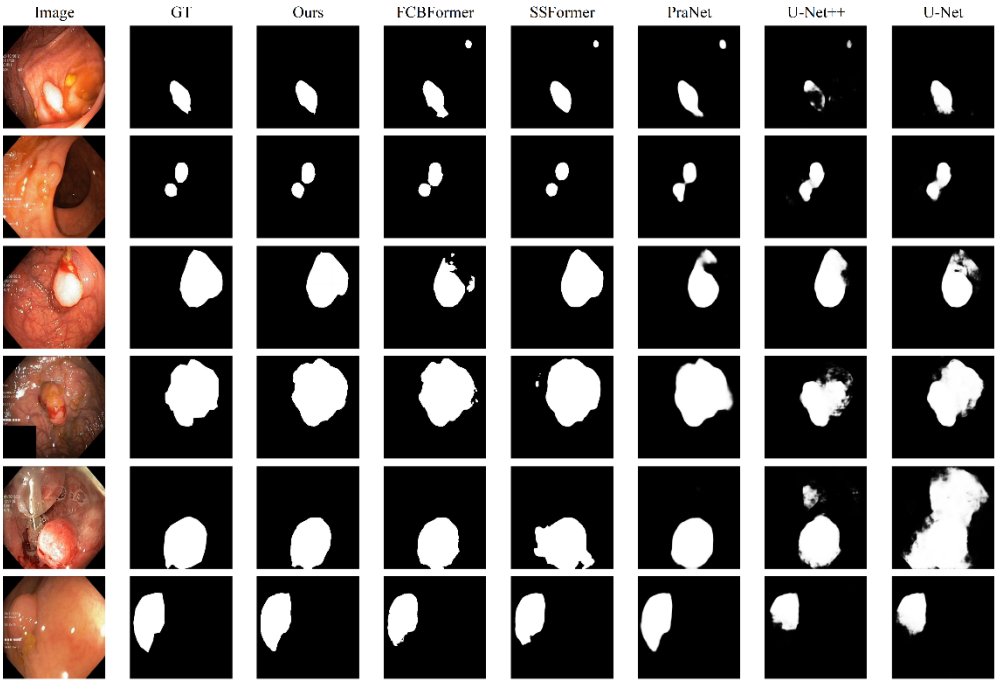


Fig. 7. Qualitative comparisons of polyp segmentation results obtained from different cutting-edge approaches in experiment A, encompassing FCBFormer [42], SSFormer [35], PraNet [21], U-Net++ [12] and U-Net [11].

Moreover, we also displayed prediction results of polyp segmentation from our model and recent cutting-edge methods in experiment B, as shown in Fig. 8. As can be seen in Fig. 8, our model provided consistently best performance under various challenging cases, which kept the polyp boundary segmentation intact and clear while decreasing the miss segmentation inside the

Biomedical Optics EXPRESS

polyp. Among these advanced methods, SFA performed almost the worst and completely failed to segment polyps under different challenges. When polyp regions were small (such as in the second row), TranSEFusionNet, ESFPNet, Polyp-PVT, M^2SNet, FCBFormer, SSFormer and PraNet predicted some un-related segmentation areas. In segmentation tasks of polyps with irregularity shapes and large sizes in the third and fourth rows, SSFormer and SANet completely failed in segmenting them, whereas CFA-Net, Polyp-PVT, M^2SNet, and FCBFormer yielded some incorrect segmentation regions. As for TranSEFusionNet and ESFPNet, there still existed some incomplete and false positive situations. When the edges of the polyp areas were similar to the surrounding tissues in the fifth and sixth rows, ESFPNet encountered some mis-segmentations and PraNet produced over-segmented areas, while CFA-Net, M^2SNet and MSNet experienced fuzzy boundaries resulting from the loss of fine edge feature details. In this context, TranSEFusionNet manifested incomplete segmentation and Polyp-PVT only discriminated a portion of polyp area and suffered from under-segmentation, whilst FCBFormer as well as SSFormer neglected some fine detailed feature information. Conversely, the segmented mask predicted by our model held a best agreement with the ground truth under all these challenging conditions. Simultaneously, we could also notice from Fig. 8 that when coping with extremely small polyp lesions (such as in the second row), the ECTransNet and MCSF-Net produced some un-related segmentation regions. In addition, when facing polyps with irregularity shapes in the third row or accompanied by excessive mucus along with the indistinct demarcation with the surrounding tissues (See the sixth row), the ECTransNet tended to overlook a small portion of polyp lesion area while the MCSF-Net failed to correctly outline polyp samples, leading to incomplete segmentation. In contrast, our model delineated more accurate segmentation regions of polyp lesions with fine boundaries in tackling different challenging scenes, which closely aligned with the ground truth maps and showcased its superior identification ability. In summary, the qualitative results of polyp segmentation in experiment B signified again that our model was owned to the outstanding strength in accurately predicting the polyp segmentation masks under varied size, shape and contrast scenarios.

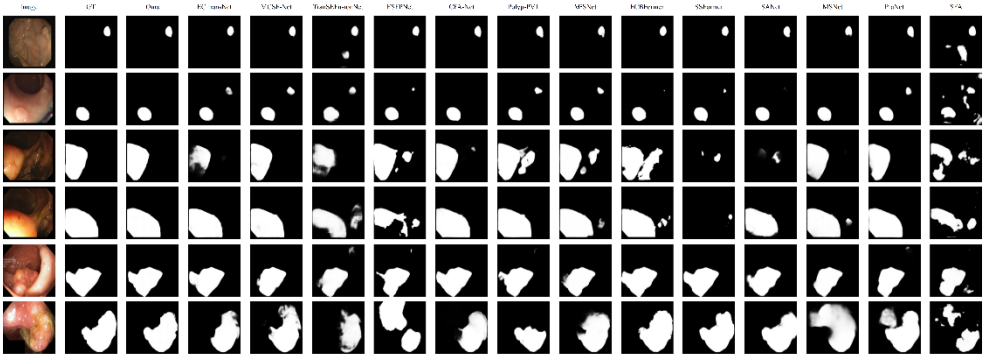


Fig. 8. Qualitative comparisons of polyp segmentation results generated by our model and thirteen different state-of-the-art methods in experiment B, containing ECTransNet [44], MCSF-Net [31], TranSEFusionNet [43], ESFPNet [37], CFA-Net [28], Polyp-PVT [33], M2SNet [30], FCBFormer [42], SSFormer [35], SANet [22], MSNet [29], PraNet [21] and SFA [19].

4.4. Generalization ability

We measured the generalization ability of the proposed Dua-PSNet using two cross-dataset testing schemes. The one was to use the CVC-ClinicDB/Kvasir-SEG training set to train our model and test on the corresponding unseen Kvasir-SEG/CVC-ClinicDB benchmarks, which was referred to

Biomedical Optics EXPRESS

testing scheme A (See Table 4). The other one was to utilize the training set built from part of the CVC-ClinicDB and Kvasir-SEG benchmarks for training, while evaluate the trained model on the unseen CVC-ConlonDB, ETIS, and CVC-300 benchmark datasets, respectively, which was used as testing scheme B (See Table 5). As can be seen in Table 4, U-Net and U-Net ++ owed poor generalization ability on both datasets, especially U-Net++ decreased sharply in *mDice* and *mIoU* metrics. Relative to the SSFormer, our model attained considerable performance improvements, with *mDice* by 7.9% and *mIoU* by 9.5% on the Kvasir-SEG dataset, and *mDice* by 12.4% and *mIoU* by 13.5% on the CVC-ClinicDB benchmark. In addition, our model exhibited the best generalization power on the CVC-ClinicDB dataset, reaching up to 0.9203 *mDice* and 0.8582 *mIoU*. These results illustrated that our model exhibited particular advantages in handling images from a somewhat divergent distribution over that considered during training.

Table 4. Generalizability comparisons of the proposed Dua-PSNet with current mainstream approaches in the testing scheme A. The bold fonts denoted the best results.

TrainDatasets	CVC-ClinicDB [52]		Kvasir-SEG [51]	
TestDatasets	Kvasir-SEG [51]		CVC-ClinicDB [52]	
-	mDice	mIoU	mDice	mIoU
U-Net [11]	0.6222	0.4588	0.7172	0.6133
U-Net++ [12]	0.5926	0.4564	0.4265	0.3345
PraNet [21]	0.7950	0.7073	0.7912	0.7119
SSFormer [35]	0.7790	0.6977	0.7966	0.7229
FCBFormer [42]	0.8514	0.7803	0.9070	0.8470
Ours	**0 . 8576**	**0.7929**	**0.9203**	**0.8582**

In Table 5, it should be noticed that our model consistently outperformed other mainstream models in terms of nearly all indicators on the CVC-ConlonDB, ETIS, and CVC-300 benchmark datasets. For instance, compared with the FCBFormer, it provided 2.3% improvement in *mDice*, 1.8% improvement in *mIoU*, 3.0% improvement in F_β^w, 1.9% improvement in S_α, and 1.4% improvement in mE_ϕ, respectively, on the CVC-ConlonDB dataset. Yet, the MAE of our model reduced by 0.4%. On the CVC-300 dataset, our model exceeded the Polyp-PVT by a large margin, improving *mDice* from 0.899 to 0.910 (1.1% improvement), and *mIoU* from 0.831 to 0.845 (1.4% improvement), respectively. Even in the most challenging ETIS dataset, it also improved the state-of-the-art performance, with a *mDice* score of 0.794, a *mIoU* score of 0.719, a F_β^w score of 0.750 and a S_α score of 0.882. Furthermore, our method was consistently superior to recently proposed best-performing M2SNet, CFA-Net, ESFPNet, and TranSEFusionNet for nearly all evaluation indicators on these unseen datasets. However, another one notable finding was that SFA and TranSEFusionNet exhibited a dramatic decline in segmentation performance on these unseen datasets, partially evidencing that their generalization abilities were weak. When compared with the ECTransNet and the MCSF-Net, our model manifested better generalization ability. For instance, in terms of *mDice* and *mIoU*, our model obtained substantial improvements of 5.5% and 5.4% over the ECTransNet on the CVC-ColonDB dataset, while 6.6% and 6.4% on the ETIS-LabribPolypDB dataset, respectively. Even on the CVC-300 dataset, our model still outperformed the ECTransNet with 0.3% higher *mDice* and 0.5% higher *mIoU*. In contrast, our model significantly exceeded the MCSF-Net by 5.6% *mDice* and 4.9% *mIoU* on the CVC-ColonDB dataset, 2.9% *mDice* and 2.9% *mIoU* for the ETIS-LabribPolypDB dataset, and 0.9% *mDice* and 1.1% *mIoU* for the CVC-300 dataset. These results validated that our model owned excellent segmentation capability even when confronting with different source datasets. Again, these results revealed that by contrast with the current state-of-the-art approaches, our models still manifested remarkable generalization capability in tackling images with a different distribution

Biomedical Optics EXPRESS

Table 5. Generalizability comparisons of the proposed Dua-PSNet with current mainstream approaches in the testing scheme B. The bold fonts denoted the best results.

Datasets	CVC-ColonDB [6]					
	mDice	mIoU	F_β^ω	S_α	$mE_\emptyset$	MAE
SFA [19]	0.456	0.337	0.366	0.628	0.661	0.094
PraNet [21]	0.712	0.640	0.699	0.820	0.847	0.043
MSNet [29]	0.751	0.671	0.736	0.838	0.872	0.041
SANet [22]	0.753	0.670	0.726	0.837	0.869	0.043
SSFormer [35]	0.772	0.697	0.766	0.843	0.880	0.036
FCBFormer [42]	0.798	0.723	0.775	0.850	0.902	0.033
M^2SNet [30]	0.756	0.678	0.737	0.843	0.873	0.038
Polyp-PVT [33]	0.801	0.725	0.791	0.847	0.908	0.031
CFA-Net [28]	0.743	0.665	0.728	0.835	0.898	0.039
ESFPNet [37]	0.767	0.679	0.747	0.837	0.880	0.037
TranSEFusionNet [43]	0.614	0.510	0.572	0.758	0.801	0.052
ECTransNet [44]	0.766	0.687	0.710	0.831	0.846	0.045
MCSF-Net [31]	0.765	0.692	0.700	0.824	0.853	0.041
Ours	**0 . 821**	**0.741**	**0.805**	**0.869**	**0.916**	**0.029**
Datasets	ETIS-LaribPolypDB [2]					
	mDice	mIoU	F_β^ω	S_α	$mE_\emptyset$	MAE
SFA [19]	0.297	0.217	0.231	0.557	0.531	0.109
PraNet [21]	0.628	0.567	0.600	0.794	0.808	0.031
MSNet [29]	0.723	0.652	0.677	0.845	0.875	0.020
SANet [22]	0.750	0.654	0.685	0.849	0.881	0.015
SSFormer [35]	0.767	0.697	0.736	0.863	0.889	0.016
FCBFormer [42]	0.773	0.689	0.723	0.860	0.896	0.017
M^2SNet [30]	0.746	0.668	0.712	0.853	0.880	0.017
Polyp-PVT [33]	0.781	0.705	0.748	0.870	**0 . 901**	0.015
CFA-Net [28]	0.732	0.655	0.693	0.845	0.892	**0.014**
ESFPNet [37]	0.760	0.676	0.723	0.850	0.895	0.019
TranSEFusionNet [43]	0.474	0.393	0.431	0.717	0.731	0.032
ECTransNet [44]	0.728	0.655	0.683	0.847	0.864	0.017
MCSF-Net [31]	0.765	0.690	0.601	0.800	0.837	0.019
Ours	**0.794**	**0.719**	**0.750**	**0.882**	0.894	0.015
Datasets	CVC-300 [53]					
	mDice	mIoU	F_β^ω	S_α	$mE_\emptyset$	MAE
SFA [19]	0.467	0.329	0.341	0.640	0.644	0.065
PraNet [21]	0.871	0.797	0.843	0.925	0.950	0.010
MSNet [29]	0.865	0.799	0.848	0.926	0.945	0.010
SANet [22]	0.888	0.815	0.859	0.928	0.962	0.008
SSFormer [35]	0.887	0.821	0.869	0.929	0.959	**0.007**
FCBFormer [42]	0.897	0.833	0.877	0.931	0.969	0.009
M^2SNet [30]	0.899	0.833	0.881	0.940	0.971	**0.007**
Polyp-PVT [33]	0.899	0.831	0.884	0.932	0.973	0.008
CFA-Net [28]	0.893	0.827	0.875	0.938	**0.978**	0.008
ESFPNet [37]	0.874	0.809	0.858	0.923	0.944	0.009
TranSEFusionNet [43]	0.754	0.649	0.707	0.857	0.902	0.016
ECTransNet [44]	0.907	0.840	0.865	0.937	0.962	**0.007**
MCSF-Net [31]	0.901	0.834	0.859	0.934	0.960	0.008
Ours	**0.910**	**0.845**	**0.890**	**0.941**	0.976	**0.007**

Biomedical Optics EXPRESS

from that utilized for training, which could be an efficient option for real clinical practice with considerable variations in data.

4.5. Ablation studies

In this part, we verified the effectiveness of each critical component in our Dua-PSNet through a series of ablation experiments on the CVC-ClinicDB and Kvasir-SEG datasets, and the quantitative and qualitative results were tabulated in Table 6 and Fig. 9, respectively. In the ablation experiments, we selected the FCN and PVTv2-B3 as the baseline (No.1 and No.2) and used the ResidualBlock module instead of RB module in FCN (No.3) to evaluate the role of ResidualBlock module, and then investigate the effect with successive additions of PVTv2-B3 (No.4), AFA module (No.5), ResidualBlock module (No.6) and SGLFH module (No.7).

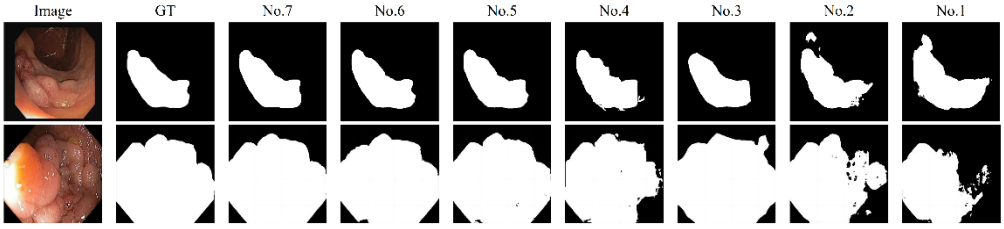


Fig. 9. Visualizable comparison results for verifying the strength of different key components.

Table 6. Quantitative segmentation results from different primary modules in the ablation study utilizing the CVC-ClinicDB and Kvasir-SEG benchmarks.

Datasets	CVC-ClinicDB [52]		Kvasir-SEG [51]	
	mDice	mIoU	mDice	mIoU
No.1 FCN Only with RB	0.8896	0.8123	0.8569	0.7783
No.2 PVTv2-B3 Only	0.9099	0.8488	0.8825	0.8129
No.3 FCN Only with ResidualBlock	0.9002	0.8282	0.8695	0.7992
No.4 FCN + PVTv2-B3	0.9220	0.8701	0.8926	0.8267
No.5 FCN + PVTv2-B3 + AFA	0.9346	0.8861	0.9065	0.8478
No.6 FCN + PVTv2-B3 + AFA + ResidualBlock	0.9438	0.8966	0.9174	0.8556
No.7 FCN + PVTv2-B3 + AFA + ResidualBlock + SGLFH	0.9514	0.9083	0.9253	0.8746

Effectiveness of ResidualBlock Module. We investigated the validity of the ResidualBlock module. As shown in Table 6, compared No.1 with No.3, it could be clearly observed that No.3 using the ResidualBlock module could significantly improve the segmentation performance by 1.1% *mDice* and 1.6% *mIoU* on the CVC-ClinicDB dataset, and 1.3% *mDice* and 2.1% *mIoU* on the Kvasir-SEG dataset. In addition, from No.6 vs No.5, similar trends could be inspected. From the visualizable comparison results in Fig. 9, it could be witnessed that the perception of more local information was enhanced. In short, these results suggested that the ResidualBlock module was benefit for boosting model's performance.

Effectiveness of combination of FCN and PVTv2-B3. In order to evaluate the combination of FCN and PVTv2-B3 branches, we tested the performance of No.4 (FCN + PVTv2-B3). As can be seen from Table 6 and Fig. 9, No.4 brought dramatic performance improvements against No.3 and No.2 on both datasets. Specifically, by comparing No.4 with No.3 in Table 5, the *mDice* and *mIoU* obtained an increment of 2.2% and 4.2% on the CVC-ClinicDB dataset, while 2.3% and 2.8% on the Kvasir-SEG dataset. In contrast to No.2, No.4 attained a performance gain

Biomedical Optics EXPRESS

of 1.2% and 2.1% with respect to *mDice* and *mIoU*, on the CVC-ClinicDB dataset, whilst 1% and 1.4% on the Kvasir-SEG dataset. These improvements confirmed that the work together of both branches rather than only using a single branch structure could enable our model to more accurately recognize true polyp regions, revealing that complementary hierarchical semantic information could be employed for boosting the model's segmentation performance, which was also demonstrated by the visualizable comparison results in Fig. 9.

Effectiveness of Adaptive Feature Aggregation (AFA) Module. We further investigated the importance of AFA module. According to the No.4 vs No.5 in Table 6, we could notice that No.5 showed a significant increase in the performance, improving *mDice* score from 0.9220 to 0.9346 (1.3% ascension) and *mIoU* score from 0.8701 to 0.8861 (1.6% ascension), on the CVC-ClinicDB benchmark, and *mDice* score from 0.8926 to 0.9065 (1.4% ascension) and *mIoU* score from 0.8267 to 0.8478 (2.1% ascension), on the Kvasir-SEG benchmark. This indicated the significance of the AFA module on segmentation performance enhancement. Likewise, the visualizable comparison results in Fig. 9 also implied that the devised AFA block could contribute to the improvement in the performance of polyp segmentation.

Effectiveness of Selective Global-to-Local Fusion Head (SGLFH) Module. Finally, we assessed the contribution of SGLFH component. In Table 6, it was apparent that No.7 integrated the SGLFH module achieved better segmentation performance than No.6 on all the evaluation criterias, especially with *mIoU* metric reaching 0.9083 and 0.8746 on the CVC-ClinicDB and Kvasir-SEG benchmarks, respectively, which were 1.2% and 1.9% higher than those gained with No.6. This demonstrated that the introduction of the SGLFH module helped improve segmentation accuracy a lot, highlighting the effectiveness of the presented SGLFH module, which was achieved by allowing the model to dynamically aggregate local feature details and the global semantic features for narrowing information gap between them. As displayed in Fig. 9, some edge detailed information could not be precisely identified without utilizing the developed SGLFH module.

5. Discussion

In this work, we developed a novel architecture Dua-PSNet for accurate and automated polyp segmentation at full-size from colonoscopy images in many challenging cases. One of the strengths was that it took full use of the advantages of the Transformer with global-receptive-field on modeling long-range relationship and the FCN with local-receptive-field on establishing local spatial correlation in dense prediction. Combining them using two parallel branches allowed our model to pay attention to global semantic information without missing local spatial details. In this way, they could be reciprocally constrained and complemented for enabling more accurate full-size predictions. In the Transformer branch, we constructed the ResidualBlock module to deeper excavate polyp edge details disguised in low-level features and improve the recognition of edge ambiguous features. Further, we proposed two feature fusion modules, AFA module and SGLFH module, which could aggregate high-level features in a stepwise adaptive fashion and selectively incorporate edge cues into global semantic information, bridging semantic gap between high-level and low-level features. The Transformer was a sequence-to-sequence prediction model which attended to global contextual information. However, their predicted segmentation maps were typical types of a lower resolution than the input images, lacking detailed localization information. Our FCN branch allowed for the extraction of highly merged multi-scale fine boundary characteristics at full-size, compensating for output of the Transformer branch into full-size segmentation map. From quantitative and qualitative comparison results with state-of-the-art methods (See Table 2 and Table 3 as well as Fig. 7 and Fig. 8) together with generalizability analysis (See Table 4 and Table 5), it was demonstrated that our model consistently exceeded these advanced methods [11,12,19,21,22,28–31,33,35,37,42–44] and located polyp areas more accurately with clear boundary contours, even though they were varied, which put

Biomedical Optics EXPRESS

emphasis on its powerful strengths in learning and generalization capability. Hence, we could conclude that the proposed Dua-PSNet had great potential to act as an "extra pair of eyes" for endoscopists providing additional objective diagnostic information during colonoscopy, and aid in making a feasible decision for further treatment.

The developed Dua-PSNet performed better than a variety of current advanced methods, comprising DCNN-based methods [11,12,19,21,22,28–31], Transformed-based approaches [33,35,37] and hybrid architecture combining Transformer and CNN [42–44], in the task of polyp segmentation. As reported in Table 1, our model had best segmentation accuracy among all five methods, on the CVC-ClinicDB dataset, with *mDice* score of 0.9514 and *mIoU* score of 0.9083. Despite the *mIoU* value was not highest on the Kvasir-SEG dataset, it was slightly inferior to second-best FCBFormer (*mIoU*: 0.8746 vs 0.8752). In Table 2, our model also attained best scores in nearly all evaluation indicators compared with SFA [19], PraNet [21], MSNet [29], SANet [22], M2SNet [30], SSFormer [35], ESFPNet [37], and TranSEFusionNet [43] on both datasets. In addition to this, the F_β^w value of Polyp-PVT [33] on the Kvasir-SEG was 0.1% marginally higher than our model, whereas the MAE value of Polyp-PVT was 0.1% slightly lower. Through comparison between our model and CFA-Net, a similar observation also existed (mE_ϕ:0.960 vs 0.962; MAE:0.024 vs 0.023) in terms of mE_ϕ and MAE. We believed the reason for this result was due to too small dataset and abundant noise, which made it difficult to learn sufficient feature details and have an adverse impact on our model. In contrast, the *mDice* and *mIoU* scores of our model were higher by 0.9% and 1.1%, 0.7% and 1.3%, respectively. To show our model intuitively, we also provided segmentation results on different testing datasets, and compared testing performance with several representative advanced methods. From visualizable prediction results of different methods in Fig. 7 and Fig. 8, it could also be well visualized that the proposed Dua-PSNet identified a more comprehensive range of polyp regions with smooth and clear boundaries. Even for the challenging scenarios such as irregular shape, varied size as well as blurred boundary between the polyp and its surrounding tissue, it could also deal with well and produce considerably better segmentation mask, proving its strong learning ability. These excellent segmentation results could be attributed to the fact that our model had a powerful ability to fully mine the multi-scale feature representation and dynamically fuse different level of features for tackling the scale variations of polyps, and simultaneously provide fine local edge cues to guide for locating its boundary, thereby boosting the segmentation performance. With the aid of our model, an accurate and operator-independent estimate of the polyp size could be provided to assist in making feasible decisions required during colonoscopy.

Considering stitching features directly using the skip connection of U-Net may cause some relevant information lost, we first restored the feature sizes of three high levels to the same size and then aggregated them in a stepwise adaptive way through the constructed AFA block in the Transformer branch, highlighting important and coarse global contextual features. As listed in Table 6 (No.5 vs No.4), the model adding the AFA component yielded better segmentation results than one without this module, with 1.3% and 1.4% higher *mDice* and 1.6% and 2.1% higher *mIoU* on the CVC-ClinicDB and Kvasir-SEG benchmarks, respectively. Notwithstanding preserving most information in global semantic features, the transmission of these semantic information from high-level to low-level was still weakened by simple up-sampling operation. Accordingly, we constructed ResidualBlock module to emphasize critical local boundary details, and proposed SGLFH module to selectively add these fine local details into global semantic features again at the end of Transformer branch, such that the original characteristics of images could be conserved to the maximum extent. From the visualizable results in Fig. 9, it could be visually perceived that the overall segmentation effect on polyps of Fig. 9 (No.3) with ResidualBlock module was relatively better than that of Fig. 9 (No.1) using only FCN with RB module. From No.1 vs No.3 in Table 6, the same observation was obtained. When further adding the SGLFH block, the identification of focus regions was more precise in Fig. 9 (No.7 vs No.6), and the boundary of Fig. 9 (No.7)

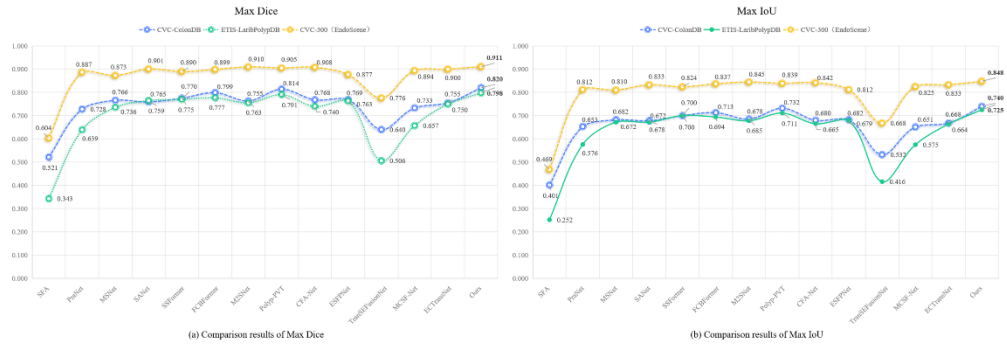
Biomedical Optics EXPRESS

was sharper and smoother compared to Fig. 9 (No.6). In the segmentation task of polyps, low contrast between the target features and the background made it hard to accurately distinguish them, and most Transformer-based methods only predicted a lower resolution segmentation map (not full-size) for the input image. To enhance the discrimination capability and perform full-size prediction, we combined the Transformer and FCN branches in parallel, where the fine-grained features extracted by FCN branch as well as important and rough features outputted by Transformer branch complemented each other to enhance the target information and weaken the background characteristics for dense prediction at full-size. From the segmentation results in Table 6 (No.2 vs No.4, and No.3 vs No.4) together with Fig. 9 (No.2), Fig. 9 (No.3) and Fig. 9 (No.4), we could find that the model combining two branches stimulated segmentation performance gains by a significant increment (On the CVC-ClinicDB dataset, No.4 vs No.2: 1.2% *mDice* and 2.1% *mIoU* improvements, No.4 vs No.3: 2.2% *mDice* and 4.2% *mIoU* improvements; On the Kvasir-SEG dataset, No.4 vs No.2: 1.0% *mDice* and 1.4% *mIoU* boosts, No.4 vs No.3: 2.3% *mDice* and 2.8% *mIoU* boosts). The gradual superposition of each component provided consistent improvement of performance, and combining the presented all components together allowed our model to achieve the highest performance in all evaluation metrics. This justified that all components we devised contributed to performance boost in the entire Dua-PSNet, and excluding any one of them would lead to a decline in the performance of polyp segmentation.

The model's generalizability was also a challenge in the domain of medical image analysis. We argued that the most important angle of the model's generalizability lied in its ability to adapt to different types of datasets. Most of existed methods for polyp segmentation were only evaluated on a single dataset, which was not capable of directly reflecting the model's generalizability. On the contrary, cross-dataset evaluation could be implemented for investigating the generalizability of different networks. In view of the generalizability of U-Net, the Dua-PSNet was also built upon this model for innovation, and the segmentation results on the unseen three datasets implied its superior generalization ability. As can be seen in Table 5, on both the CVC-ColonDB and CVC-300 datasets, our model exhibited best segmentation results, exceeding the second-best Polyp-PVT by 2.0% and 1.6% with respect to *mDice* and *mIoU* for the CVC-ColonDB dataset, while 1.1% and 1.4% for the CVC-300 dataset. Even on the most challenging ETIS benchmark, our model was also better in comparison with the latest Polyp-PVT, FCBFormer and ESFPNet apart from mE_ϕ indicator, which was only slightly lower by 0.7%, 0.2%, and 0.1%, respectively. It was marginally 0.1% higher than CFA-Net in MAE indictor, yet exhibited obviously better in other evaluation metrics. Across all datasets, the proposed Dua-PSNet generalized well with consistently accurate segmentations, which made it more suitable for practical applications in which wide variations of data happened frequently. Furthermore, we also displayed the comparison results on *maxDice* and *maxIoU* in Fig. 10, which further evidenced the stability and leadings of our model. The excellent generalizability may be owed to the successful mixture of the merits of the Transformer and FCN in the proposed Dua-PSNet, resulting in the main region of polyp being handled by the Transformer branch whilst a reliable and fine full-size edge around this main area being guaranteed by the FCN branch.

By comparison with the ECTransNet [44] and MCSF-Net [31], two latest segmentation methods proposed recently, the outstanding segmentation performance of our model was demonstrated as reported in Table 3. With respect to ECTransNet and MCSF-Net, we performed retraining and testing using their publicly available codes and recommended parameter settings as well as default data augmentation methods reported in their original literatures. Among data augmentation strategies used in these models and our approach, there only existed slight differences. Moreover, we also conducted an additional experiment to retrain and test ECTransNet and MCSF-Net based on the released codes and the recommended parameters using our data augmentation techniques in experiment B, and determined the p-values for comparisons of the Dice and IoU metrics, in which five repetitions were adopted for statistical significance evaluation. The

Biomedical Optics EXPRESS



(a) Comparison results of Max Dice (b) Comparison results of Max IoU

Fig. 10. Model generalizability validation using Max Dice (a) and Max IoU (b) metrics. The **bold** fonts denoted the best results.

gains in Dice and IoU for ECTransNet and MCSF-Net using our data augmentation method (compared to counterparts using respective data augmentation strategies reported in their original literatures) were statistically insignificant with p-values far greater than 0.05, which indicated that slightly different data augmentation methods did not significantly affect the comparison of these models. Even in some cases, the Dice and IoU values for ECTransNet and MCSF-Net using our data augmentation method were relatively lower than those using their respective data augmentation techniques. Under this circumstance, we leveraged their best scores attained using their default data augmentation methods to avoid the bias introduced in model re-training, which could guarantee relatively fairness of comparisons of the performance of different models. Concretely, our model gained the highest 0.922 *mDice* and 0.874 *mIoU* scores on the Kvasir-SEG dataset. Notably, on the CVC-ClinicDB dataset, our model achieved a *mDice* value of 0.2% lower slightly than the MCSF-Net, whereas maintaining the same best level as that in the *mIoU* metric. And we believed that this difference was very small and its advantage in polyp segmentation accuracy in the face of challenges arising from variations in size and morphology was still substantiated. Moreover, in the case of the unseen datasets CVC-ColonDB, ETIS-LaribPolypDB, and CVC-300, our model achieved the consistently better segmentation performance across nearly all metrics than the ECTransNet and MCSF-Net, as listed in Table 5. For the CVC-300 dataset, the improvements of our model in *mDice* and *mIoU* values were not obvious, with 0.3% and 0.5% increases over the ECTransNet, and 0.9% and 1.1% improvements over the MCSF-Net. Perhaps, utilizing domain adaptation technique or data augmentation strategy could contribute to further generalization capability improvements. Figure 8 also provided a visual comparison between our model and these two methods. We could see that the proposed Dua-PSNet consistently predicted relatively accurate segmentation maps under various challenges, whilst the ECTransNet and MCSF-Net occasionally generated un-related or incomplete segmentation effects. Different from the ECTransNet and MCSF-Net, our Dua-PSNet combined both Transformer branch and FCN branch in a parallel pattern for modeling multi-scale global contextual semantic relationships and extracting fine low-level features to relieve the loss of useful information. Additionally, the model incorporated the AFA block in the multi-scale fusion process of high-level features to progressively refine global semantic information, thereby enhancing the representation ability of such characteristics and obtaining spatial and location information related to polyps. The ResidualBlock module enabled the extraction of richer local boundary cues of low-level features. The SGLFH module aggregated the resulting local boundary information into global semantic features for delineating more fine-grained contours of polyp lesions and recalibrating their locations. As a result, our Dua-PSNet was able to achieve strong learning and generalization capability.

Biomedical Optics EXPRESS

Despite generating promising performance, this work had also some limitations and rooms for further enhancement. Firstly, during the model training, we needed to resize the input images into a uniform size of 352×352 so that the model's complexity could be decreased, inevitably resulting in the loss of some information in the images and influencing the overall performance of the model. A future study will attempt to explore the data augmentation (such as Generative Adversarial Network (GAN) [59] or Conditional Variational Autoencoders (CVAE) [60] or SAM [61]) and feature extraction technology [62] for alleviating information loss. Secondly, the introducing of the Transformer increased the model's complexity. In the future, we plan to seek more efficient strategies [63] for parameter reduction to reduce the computation complexity of our model without compromising its performance. Finally, our network lacked a comprehensive evaluation under a real-world clinical scenario. The future work is to consider new data from complex real clinical settings to further test the model's generalization ability and thoroughly investigate its potential of practical clinical applications.

In conclusion, we developed a novel improved dual-aggregation polyp segmentation network called Dua-PSNet. One of advantages of this model was the usage of both parallel branches characterized by Transformer and FCN to work together for enhancing polyp segmentation performance. Beyond that, in the decoder (MFAD) of Transformer branch, we also introduced the AFA module to aggregate multi-scale high-level semantic characteristics and added the ResidualBlock module to focus on local fine edge information, whilst designed the SGLFH component to selectively incorporate low-level fine boundary cues with high-level semantic information. Simultaneously, the FCN branch extracted highly merged multi-scale characteristics at full-size, and compensated for outputs of the Transformer branch into full-size prediction. The experimental results on five publicly available benchmark datasets revealed that our model achieved better segmentation performance while exhibiting higher generalization capabilities compared to existing state-of-the-art approaches. We expect that this work will furnish a new perspective on network architecture design for polyp image segmentation and be further extended for other relative fields of medical image segmentation.

Funding. National Key Research and Development Program of China (2021YFB2802303).

Acknowledgements. We thank Tongren Hospital (Shanghai Jiao Tong University School of Medicine) for their invaluable help.

Disclosures. The authors declare no conflict of interest related to this article.

Data availability. The CVC-ClinicDB dataset is accessible at [64]. The Kvasir-SEG dataset is publicly available at [65]. The CVC-ColonDB dataset is acquired by [6]. The ETIS-LaribPolypDB dataset is acquired by [2]. The CVC-300 dataset is acquired by [53].

The source codes will be released at [66].

References

1. Y. Xi and P. Xu, "Global colorectal cancer burden in 2020 and projections to 2040," Transl. Oncol. **14**(10), 101174 (2021).
2. J. Silva, A. Histace, O. Romain, *et al.*, "Towards embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," Int. J. CARS **9**(2), 283–293 (2014).
3. S. Tanaka, Y. Saitoh, T. Matsuda, *et al.*, "Evidence-based clinical practice guidelines for management of colorectal polyps," J. Gastroenterol. **50**(3), 252–260 (2015).
4. S. Ameling, S. Wirth, D. Paulus, *et al.*, "Texture-based Polyp Detection in Colonoscopy," in *Bildverarbeitung für die Medizin 2009: Algorithmen-Systeme-Anwendungen Proceedings des Workshops vom 22. bis 25. März 2009* in Heidelberg. Springer, pp. 346–350, 2009.
5. S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, *et al.*, "Computer-aided tumor detection in endoscopic video using color wavelet features," IEEE Trans. Inform. Technol. Biomed. **7**(3), 141–152 (2003).
6. A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, *et al.*, "Automated polyp detection in colon capsule endoscopy," IEEE Trans. Med. Imaging **33**(7), 1488–1502 (2014).
7. N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information," IEEE Trans Med. Imaging. **35**(2), 630–644 (2015).
8. J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

Biomedical Optics EXPRESS

9. M. Akbari, M. Mohrekesh, E. Nasr-Esfahani, *et al.*, "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, pp. 69–72, 2018.
10. P. Brandao, O. Zisimopoulos, E. Mazomenos, *et al.*, "Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks," J. Med. Robot. Res. **03**(02), 1840002 (2018).
11. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2015: 18th International Conference*, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241, 2015.
12. Z.W. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, *et al.*, "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, pp. 3–11, 2018.
13. D. Jha, P. H. Smedsrud, M. A. Riegler, *et al.*, "ResUNet++: An Advanced Architecture for Medical Image Segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, pp. 225–2255, 2019.
14. M. Yeung, E. Sala, C. B. Schönlieb, *et al.*, "Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy," Comput. Biol. Med. **137**, 104815 (2021).
15. N. K. Tomar, D. Jha, S. Ali, *et al.*, "DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation," in *PATTERN RECOGN. ICPR International Workshops and Challenges: Virtual Event*, January 10-15, 2021, Proceedings, Part VIII. Springer, pp. 307–314, 2021.
16. X.Z. Sun, P.F. Zhang, D.C. Wang, *et al.*, "Colorectal Polyp Segmentation by U-Net with Dilation Convolution," in *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 851–858, 2019.
17. D. Banik, K. Roy, D. Bhattacharjee, *et al.*, "Polyp-Net: A Multimodel Fusion Network for Polyp Segmentation," IEEE Trans. Instrum. Meas. **70**, 1–12 (2020).
18. T. Mahmud, B. Paul, and S. A. Fattah, "PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images," Comput. Biol. Med. **128**, 104119 (2021).
19. Y. Fang, C. Chen, Y. Yuan, *et al.*, "Selective Feature Aggregation Network with Area-Boundary Constraints for Polyp Segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019: 22nd International Conference*, Shenzhen, China, October 13-17, 2019, Proceedings, Part I 22. Springer, pp. 302–310, 2019.
20. B. Murugesan, K. Sarveswaran, S. M. Shankaranarayana, *et al.*, "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, pp. 7223–7226, 2019.
21. D.P. Fan, G.P. Ji, T. Zhou, *et al.*, "PraNet: Parallel Reverse Attention Network for Polyp Segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 263–273, 2020.
22. J. Wei, Y.W. Hu, R.M. Zhang, *et al.*, "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, pp. 699–708, 2021.
23. T. Kim, H. Lee, and D. Kim, "UACANet: Uncertainty Augmented Context Attention for Polyp Segmentation," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2167–2175, 2021.
24. P. F. Song, J. J. Li, and H. Fan, "Attention based multi-scale parallel network for polyp segmentation," Comput. Biol. Med. **146**, 105476 (2022).
25. Y. Lin, J. C. Wu, G. B. Xiao, *et al.*, "BSCA-Net: Bit Slicing Context Attention Network for Polyp Segmentation," Pattern Recogn. **132**, 108917 (2022).
26. T. P. Shen and X. G. Li, "Automatic polyp image segmentation and cancer prediction based on deep learning," Front. Oncol. **12**, 1087438 (2023).
27. Z. X. Li, N. Zhang, H. L. Gong, *et al.*, "MFA-Net: Multiple Feature Association Network for medical image segmentation," Comput. Biol. Med. **158**, 106834 (2023).
28. T. Zhou, Y. Zhou, K. L. He, *et al.*, "Cross-level feature aggregation network for polyp segmentation," Pattern Recogn. **140**, 109555 (2023).
29. X.Q. Zhao, L.H. Zhang, and H.C. Lu, "Automatic Polyp Segmentation via Multi-scale Subtraction Network," in *Medical Image Computing and Computer Assisted Intervention (MICCAI) 2021: 24th International Conference*, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part I 24. Springer, pp. 120–130, 2021.
30. X.Q. Zhao, H.P. Jia, Y.W. Pang, *et al.*, "M^2SNet: Multi-scale in Multi-scale Subtraction Network for Medical Image Segmentation," arXiv, arXiv:2303.10894, (2023).
31. W. K. Liu, Z. G. Li, J. A. Xia, *et al.*, "MCSF-Net: a multi-scale channel spatial fusion network for real-time polyp segmentation," Phys. Med. Biol. **31**(17), 175041 (2023).
32. A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv, arXiv:2010.11929, (2020).
33. B. Dong, W. H. Wang, D. P. Fan, *et al.*, "Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers," CAAI Artificial Intelligence Research **2**, 9150015 (2023).

Biomedical Optics EXPRESS

34. F.L. Tang, Q.M. Huang, J.F. Wang, *et al.*, "DuAT: Dual-Aggregation Transformer Network for Medical Image Segmentation," arXiv, arXiv:2212.11677, (2022).
35. J.F. Wang, Q.M. Huang, F.L. Tang, *et al.*, "Stepwise Feature Fusion: Local Guides Global," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 110–120, 2022.
36. R. Nachmani, I. Nidal, D. Robinson, *et al.*, "Segmentation of polyps based on pyramid vision transformers and residual block for real-time endoscopy imaging," Journal of Pathology Informatics **14**, 100197 (2023).
37. Q. Chang, D. Ahmad, J. Toth, *et al.*, "ESFPNet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video," Medical Imaging 2023: Biomedical Applications in Molecular, Structural, and Functional Imaging **12468**, 1246803 (2023).
38. J.N. Chen, Y.Y. Lu, Q.H. Yu, *et al.*, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv, arXiv:2102.04306, (2021).
39. Y.D. Zhang, H.Y. Liu, and Q. Hu, "TransFuse: Fusing Transformers and CNNs for Medical Image 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, pp. 14–24, 2021.
40. L.H. Cai, M.J. Wu, L.J. Chen, *et al.*, "Using Guided Self-attention with Local Information for Polyp Segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, pp. 629–638, 2022.
41. W. C. Zhang, C. Fu, Y. Zheng, *et al.*, "HSNet: A hybrid semantic network for polyp segmentation," Comput. Biol. Med. **150**, 106173 (2022).
42. E. Sanderson and B. J. Matuszewski, "FCN-Transformer Feature Fusion for Polyp Segmentation," in *Annual Conference on Medical Image Understanding and Analysis (MIUA)*. Springer, pp. 892–907, 2022.
43. Y. Y. Zhang, L. Liu, Z. Y. Han, *et al.*, "TranSEFusionNet: Deep fusion network for colorectal polyp segmentation," Biomed. Signal Proces. **86**, 105133 (2023).
44. W. K. Liu, Z. G. Li, C. Y. Li, *et al.*, "ECTransNet: An Automatic Polyp Segmentation Network Based on Multi-scale Edge Complementary," J. Digit. Imaging **36**(6), 2427–2440 (2023).
45. A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," Advances in Neural Information Processing Systems (NeurIPS) **30**, 1 (2017).
46. W. H. Wang, E. Z. Xie, X. Li, *et al.*, "PVT v2: Improved baselines with Pyramid Vision Transformer," Computational Visual Media **8**(3), 415–424 (2022).
47. K.M. He, X.Y. Zhang, S.Q. Ren, *et al.*, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
48. E. Z. Xie, W. H. Wang, Z. D. Yu, *et al.*, "Segformer: Simple and efficient design for semantic segmentation with transformers," Advances in Neural Information Processing Systems (NeurIPS) **34**, 12077–12090 (2021).
49. O. Russakovsky, J. Deng, H. Su, *et al.*, "ImageNet Large Scale Visual Recognition Challenge," Int. J. Comput. Vision **115**(3), 211–252 (2015).
50. D. D. Yang, Y. Y. Li, and J. K. Yu, "Multi-task thyroid tumor segmentation based on the joint loss function," Biomed. Signal Proces **79**, 104249 (2023).
51. D. Jha, P. H. Smedsrud, M. A. Riegler, *et al.*, "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, pp. 451–462, 2020.
52. J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, *et al.*, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," Comput. Med. Imag. Grap **43**, 99–111 (2015).
53. D. Vázquez, J. Bernal, F. J. Sánchez, *et al.*, "A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images," J. Healthc. Eng. **2017**, 4037190 (2017).
54. F. Perazzi, P. Krähenbühl, Y. Pritch, *et al.*, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 733–740, 2012.
55. D.P. Fan, C. Gong, Y. Cao, *et al.*, "Enhanced-alignment Measure for Binary Foreground Map Evaluation," arXiv, arXiv:1805.10421 (2018).
56. D.P. Fan, M.M. Cheng, Y. Liu, *et al.*, "Structure-Measure: A New Way to Evaluate Foreground Maps," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4548–4557, 2017.
57. D. P. Fan, G. P. Ji, X. Qin, *et al.*, "Cognitive vision inspired object segmentation metric and loss function," Scientia Sinica Informationis **6**(6), 1 (2021).
58. I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," arXiv, arXiv:1711.05101 (2017).
59. A. Creswell, T. White, V. Dumoulin, *et al.*, "Generative Adversarial Networks: An overview," IEEE Signal Proc. Mag. **35**(1), 53–65 (2018).
60. J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning (ICML)*, pp. 5530–5540, 2021.
61. Y.Z. Zhang, T. Zhou, S. Wang, *et al.*, "Input augmentation with SAM: boosting medical image segmentation with segmentation foundation model," arXiv, arXiv:2304.11332, (2023).
62. S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Science and Information Conference (SAI)*, London, UK, pp. 372–378, 2014.

Biomedical Optics EXPRESS

63. B.D. Dinh, T.T. Nguyen, T.T. Tran, *et al.*, "1 M parameters are enough? A Lightweight CNN-based model for medical image segmentation," arXiv, arXiv:2306.16103, (2023).
64. J. Silva, F. J. Sanchez, G. Fernández-Esparrach, *et al.*, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," Computerized Medical Imaging and Graphics, 2015, https://polyp.grand-challenge.org/CVCClinicDB/
65. D. Jha, P. H. Sedsrid, M. A. Riegler, *et al.*, "Kvasir-SEG: A Segmented Polyp Dataset," Lecture Notes in Computer Science, vol. 11962 2020, https://doi.org/10.1007/978-3-030-37734-2_37
66. F. Li, Z. T. Huang, L. Zhou, *et al.*, "Source Code," Github, 2024, https://github.com/Zachary-Hwang/Dua-PSNet