Displays 88 (2025) 102993

Contents lists available at ScienceDirect

Displays

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

PCRNet: Parent–Child Relation Network for automatic polyp segmentation[☆,☆☆]

Zaka-Ud-Din Muhammad [a,b,c], Zhangjin Huang [a,b,c,*], Naijie Gu [a,b]

[a] School of Computer Science and Technology, University of Science and Technology of China, Huangshan Road, Hefei, 230027, China
[b] Anhui Provincial Key Laboratory of Industrial Internet Intelligent Software and Theory, Huangshan Road, Hefei, 230027, China
[c] Deqing Alpha Innovation Institute, Huzhou, 313299, China

ARTICLE INFO

Dataset link: https://datasets.simula.no/kvasir-seg/, https://www.kaggle.com/datasets/balraj98/cvcclinicdb

Keywords:
Polyp segmentation
Colo-rectal cancer
Medical image segmentation
Gastrointestinal cancer
Abdominal cancer
Lightweight model
Cancer diagnosis

ABSTRACT

Colorectal cancer (CRC) is the third most common cancer worldwide in terms of both incidence and mortality rates. On the other hand, its slow development process is very beneficial for early diagnosis and effective treatment strategies in reducing mortality rates. Colonoscopy is considered the standard approach for early diagnosis and treatment of the disease. However, detecting early-stage polyps remains challenging with the current standard colonoscopy approach due to the diverse shapes, sizes, and camouflage properties of polyps.

To address the issues posed by the different shapes, sizes, colors, and hazy boundaries of polyps, we propose the Parent–Child Relation Encoder Network (PCRNet), a lightweight model for automatic polyp segmentation. PCRNet comprises a parent–child encoder branch and a decoder branch equipped with a set of Boundary-aware Foreground Extraction Blocks (BFEB). The child encoder is designed to enhance feature representation while considering model size and computational complexity. The BFEB is introduced to accurately segment polyps of varying shapes and sizes by effectively handling the issue of hazy boundaries.

PCRNet is evaluated both quantitatively and qualitatively on five public datasets, demonstrating its effectiveness compared to more than a dozen state-of-the-art techniques. Our model is the most lightweight among current approaches, with only (5.0087) million parameters, and achieves the best Dice Score of (0.729%) on the most challenging dataset, ETIS. PCRNet also has an average inference rate of (36.5) fps on an $Intel^{®}$ $Core^{TM}$ i7-10700K CPU with 62 GB of memory, using a GeForce RTX 3080 (10 GB).

1. Introduction

Polyps are abnormal tissue growths in the human digestive system that can develop into cancerous cells, particularly in the colon and rectum, where they can lead to Colorectal Cancer (CRC). CRC is one of the deadliest cancers, with the third-highest fatality rate among all cancer types. Early detection and proper treatment of polyps lowers the likelihood of developing non-tumorous polyps into cancerous cells and thus plays a crucial role in saving human lives. For this purpose, early screening using Colonoscopy is considered the most effective approach to find and track the development of Colorectal polyps; however, the standard colonoscopy presents a miss rate of more than 20% in early-stage polyp detection [1].

In recent years, efforts have been made using computer-based automated techniques to cope with the issue of colonoscopy miss-rate in early-stage polyp detection and segmentation. Earlier efforts have been made using color and textural information-oriented handcrafted feature (HCF) maps [2,3]; and they performs well only in the presence of meaningful color and textural differences between the target object and its background. In contrast, CRC polyps often share similar textures and colors with surrounding tissues, leading to poor performance in early polyp segmentation using HCF.

In recent years, the advent of deep learning (DL) has revolutionized classical computer vision techniques [4]. In the case of DL applications in the medical domain, DL techniques are used to cope with different public health challenges using both imagery and numerical health records datasets. In the realm of medical imaging, vision-based deep learning algorithms are extensively employed for various purposes, including disease diagnosis, foreign object detection, and fetal growth

[☆] This paper was recommended for publication by Guangtao Zhai.
[☆☆] PCRNet for automatic polyp segmentation.
[*] Corresponding author at: School of Computer Science and Technology, University of Science and Technology of China, Huangshan Road, Hefei, 230027, China.
E-mail address: zhuang@ustc.edu.cn (Z. Huang).

https://doi.org/10.1016/j.displa.2025.102993
Received 14 August 2024; Received in revised form 17 October 2024; Accepted 7 February 2025
Available online 14 February 2025

tracking. For instance, Wang et al. [5] developed a polyp warning system for quick polyp edge recognition and polyp shot detection. Shin et al. [6] employed a region-based CNN approach for automatic polyp detection from colonoscopy resources. Later, Shin et al. [7] conducted a study on generative adversarial networks [8] to improve the model performance by increasing the number of training samples using generative approaches. Lee et al. [9] developed a YOLO-v2 [10] based algorithm that presented a higher sensitivity for polyp development detection and localization. The current literature indicates that polyp segmentation, classification, and detection techniques are primarily grounded on three distinct goals. The methods used in the literature might be categorized into three groups based on these goals.

The first category of approaches focuses on achieving high accuracy, while bypassing the importance of model size and generalized performance and are deemed unsuitable for critical real-time applications [11,12]. The second category aims to achieve higher accuracy and generalized performance, bypassing the importance of designing a model with fewer trainable parameters for real-time applications. Approaches of these categories use different combinations of various datasets for training purposes, with examples including PraNet [13], UACANet [14], Polyp-PVT [15], and SegFormer [16]. The third category prioritizes reducing model size by designing models with fewer trainable parameters, typically trained on a single target dataset. However, these methods often compromise on achieving higher accuracy and better generalization. DDANet [17], ColonSegNet [18], and NanoNet [19] are some of the well-known models of this category with fewer trainable parameters.

This study introduces the Parent–Child Relation Network (PCRNet) to address the limitations of existing approaches. PCRNet balances multiple objectives — model size, accuracy, and generalization — rather than focusing on just one aspect. It features a parent–child encoder branch and a decoder branch consisting of Boundary-aware Foreground Extraction Blocks (BFEB). Parent–child encoder design aimed at reducing computational demands while enhancing feature representation. We use the pre-trained EfficientNet-B0 [20] as the parent encoder, and the layers of the child encoder consist of a SE-module [21] similar channel attention block to enhance the feature representations. In addition, the introduction of BFEB seeks to address the camouflage features of the polyps to obtain improved accuracy and more generalized performance. In brief, the primary contributions of PCRNet are as follows:

- We propose a novel model, "PCRNet", aimed at improving generalization and accuracy while minimizing model size and computational complexity.
- PCRNet introduces a parent–child encoder design to produce feature maps with rich information, avoiding the need for computationally intensive transformers or other heavy pre-trained models.
- We present the Boundary-aware Foreground Extraction Block (BFEB) to handle the camouflage properties of polyps, enhancing segmentation accuracy.

The rest of the paper is organized as follows: Section 2 covers the related work, while Section 3 discusses the proposed model and its components. Section 4 explains the experimental setup, and Section 5 presents the experimental results and discussion. Section 6 describes the ablation experiments, and finally, Section 7 encapsulates the concluding remarks.

2. Related work

2.1. Semantic segmentation

Semantic segmentation is a deep-learning-based computer vision technique that assigns a semantic label to every pixel in the image. Fully Convolutional Network (FCN) [22] is the first neural network architecture designed to perform semantic segmentation tasks. Even though the architecture of FCN is limited in its ability to extract information from smaller receptive fields, it is still considered a fundamental basis for modern semantic segmentation approaches. Current methods improve the model performance by mitigating the issues of limited receptive fields through using multi-scale pooling [23], dilated convolution [24,25], and non-local blocks [26]. Some methods also use the border feature extraction strategy to improve the model's effectiveness for accurate object segmentation [27]. Structure modeling strategies, including boundary neural fields [27], affinity fields [28], and random walk [29] are the commonly used methods to improve boundary localization. Instead of these strategies, in current techniques [30], the edge maps are extracted from intermediate layers of CNNs, which aid the model in learning edge information to improve the final prediction accuracy. However, these approaches have two major drawbacks: (1) direct boundary prediction may lack accuracy, and (2) some models require specialized human design, reducing flexibility. To address these issues, we focus on refined foreground extraction rather than solely relying on boundary feature map extraction.

2.2. Attention mechanism

Attention mechanisms, known for their association with complex cognitive functions, are widely used in deep learning to enhance performance. In medical imaging, attention operations are employed to improve model accuracy in disease diagnosis, such as colorectal cancer, and foreign object detection [31]. For instance, ACSNet [32] and DCANet [12] use attention mechanisms to focus on local and global information to enhance the visualization of boundary feature maps to segment polyps of different stages of cancer. Authors in PolypSeg [33] introduced an Adaptive Scaling Context Module (ASCM) for multi-scale contextual information aggregation to improve the segmentation accuracy. PraNet [13] introduced a parallel partial decoder module for aggregating high-level feature information to obtain contextual information. Additionally, a reverse attention operation is also introduced to deal with the issue of hazy boundaries in segmenting early-stage polyps. Authors in UACANet [14] introduced a Parallel Axial Attention Encoder (PAA-e) mechanism to improve the feature representation of the encoder module. They introduced the Parallel Axial Attention decoder (PAA-d) as the improved version of the PraNet decoder module to improve the segmentation accuracy. Even though these approaches presented an improved performance in early-stage polyp segmentation; however, these discussed models contain millions of trainable parameters. In contrast, our proposed PCRNet focuses on minimizing computational cost by optimizing model size while maintaining effective performance in segmenting early-stage polyps.

2.3. Polyp segmentation

The earlier approaches for polyp classification and segmentation were focused on using traditional algorithms to extract color, size, shape, and similar feature information [34,35]. However, these algorithms tend to present lower accuracies for object detection and segmentation with camouflage properties due to the intra-class poor feature representation [36]. However, deep learning (DL) techniques have set out new milestones in computer vision daily life applications, and DL-based techniques present admirable performances to deal with camouflage proprieties of objects in different domains [37,38]. For example, Thambawita et al. [39] investigated several classical and deep learning techniques and proposed five novel models as a potential solution to classify sixteen different classes of GI tract diseases. Ronneberger et al. [40] modified the FCN architecture to UNet by introducing skip connections for medical image semantic segmentation. This U-Net architecture is now used as the baseline of various models [11,41,42] to improve the segmentation accuracies of existing approaches.

Recent advancements in polyp segmentation have significantly improved the U-Net architecture through various techniques, including attention mechanisms, ensemble learning, generative models, multi-level feature manipulation, and multi-scale feature extraction. For instance, Poorneshwaran et al. [43] developed a Generative Adversarial Network (GAN) to address the challenge of limited annotated data. KANG et al. [44] proposed an RCNN-based ensemble approach to enhance segmentation accuracy from colonoscopy images. PolypSegNet [45] introduced a dilated inception (DDI) module for extracting feature maps with broader receptive fields, while SFA [46] embedded Selective Kernel Modules (SKM) between the encoder and decoder to achieve similar goals. SANet [47] employed a color swap strategy to differentiate between the background and target objects. Despite these advancements, many of these models still exhibit suboptimal accuracy in segmenting early-stage polyps. In response, PCRNet focuses specifically on improving accuracy for early-stage polyp segmentation, such as that found in the ETIS dataset.

3. The proposed method

Recent efforts in automatic polyp segmentation face challenges due to the varying properties of polyps and significant noise in colonoscopy images, which hinder the performance of existing methods. To address these issues, recent studies have adopted two main strategies: employing various attention mechanisms or using Transformer models as encoders. However, Transformers are computationally expensive, and multiple attention mechanisms can lead to models with a high number of trainable parameters.

In response to these challenges, we propose PCRNet (Fig. 1), designed to balance model size, computational cost, and performance. PCRNet aims to enhance accuracy in early-stage polyp segmentation while minimizing both model size and computational demands. It features a parent–child encoder branch and a decoder branch. The parent–child encoder branch includes a parent encoder (EfficientNet-B0 [20]) and a child-encoder with five Squeeze-and-Excitation Attention Block (SEAB), enhancing the quality of feature maps without relying on the use of computationally intensive Transformers for feature extraction with better representations. Attention operations in the child-encoder focuses on local information and enable the model to learn global-to-local feature representations to deal with segmenting issues of varying shapes and sizes of polyps. Since boundary information plays a crucial role in accurately segmenting early-stage polyps, we designed Boundary-aware Foreground Extraction Blocks (BFEB) to help the model learn boundary details, enabling the segmentation of polyps with hazy boundaries. The decoder branch of the proposed model integrates layers of BFEB and Convolution Blocks with BatchNorm and ReLU activation (CBR), followed by up-sampling operations to optimize the input for subsequent operations of upcoming layers. BFEB in the decoder part take inputs from two different resources to avoid model over-fitting and also to produce finer foreground maps containing clear boundaries information. Details of both the parent–child encoder branch and the decoder branch are discussed further in the following subsections.

3.1. Encoder branch

Current approaches often utilize Transformer models as the encoder branch to extract both global and local contextual information for precise object segmentation. However, Transformers are computationally intensive and require powerful GPUs for training. On the other hand, traditional CNN models typically focus on local information, which can degrade performance when segmenting objects of varying sizes. To address this, employing multiple attention mechanisms can lead to models with millions of parameters, making them inefficient.

To balance model size with the quality of encoder-extracted feature maps, we adopted a parent–child information transfer strategy in designing our encoder branch. Our approach uses a parent encoder (EfficientNet-B0 [20]) to capture multi-scale information, while the child encoder focuses on local details by leveraging features from the parent encoder. This design helps optimize both the quality of feature extraction and the overall model efficiency.

Parent Encoder: EfficientNet-B0 [20] is utilized as the parent-encoder in our model due to its lightweight architecture and ability to extract multi-scale information efficiently. The EfficientNet architecture employs Mobile Inverted Residual Convolutional Blocks (MBConv) [48], which learn multi-scale feature maps from various receptive fields while maintaining computational efficiency and high performance. The EfficientNet-B0 model begins with a stem layer that reduces the input image dimensions to lower computational complexity. This reduction is achieved through a convolution operation with a (3 × 3) kernel and a stride of (2). MBConvs utilize Depthwise Separable Convolutions (DWSC), which divide standard convolution into depth-wise and point-wise operations, thus reducing computational complexity and model size. Additionally, the squeeze-and-excite operation [21] within MBConvs enhances feature maps by re-weighting the model parameters.

Child-Encoder: To enhance the quality of feature representations from each layer of the parent encoder, we introduce the child encoder, which consists of five Squeeze-and-Excitation Attention Blocks (SEAB), as depicted in Fig. 2. SEAB enhances feature quality through a feature re-calibration mechanism that selectively emphasizes important features while suppressing less useful ones. For an input feature map$X \in R^{H\times W\times C}$, the SEAB first applies a 1 × 1 convolution, represented by the function $f_{1\times1}(X) \rightarrow U \in R^{H\times W\times C}$, to emphasize critical feature representations. Following this, a squeeze operation re-weights the feature maps to further highlight important features and diminish less critical ones. The re-weighted feature maps are then combined with the input maps to refine the emphasis on significant features.

To extract the re-weighted maps $\bar{X}$ from the input map $X \in R^{H\times W\times C}$, the overall process can be expressed as follows:

$$U = f_{1\times1}(X), U \in R^{H\times W\times C}, \tag{1}$$

$$f_{1\times1}(X) = v_c \times X = \sum_{s=1}^{\acute{C}} V_c^S \times X^S, \tag{2}$$

where $U = [u_1, u_2, u_3 ... u_c]$ is the output of the convolution operation on the input map $X = [x_c^1, x_c^2, \dots, x_c^{\acute{C}}]$ using the convolution kernel vector $V = [v_c^1, v_c^2, \dots., v_c^{\acute{C}}]$.

To exploit channel-wise dependencies, global average pooling is applied to the generated feature map U. This operation reduces the spatial dimensions of U to obtain $z \in R^c$. The cth element of z is calculated using the following mathematical expression:

$$z_c = F_{sq}(v_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j), \tag{3}$$

To further capture channel-wise dependencies between the feature maps, two channel-wise convolution operation $f_{1\times1}(.)$ are then performed on these squeezed feature maps z to reduce and increase the number of channel in squeezed maps. These convolution operation is then followed by activation functions and the process can be mathematically expressed as:

$$S = F_{sqex}(z, W) = \sigma F_{ex}(z, W) = \sigma(W_2 \delta(W_1 Z)), \tag{4}$$

where δ denotes the ReLU activation function, σ represents the Sigmoid activation function, and $W_1 \in R^{\frac{c}{R}\times C}$ and $W_2 \in R^{C\times\frac{c}{R}}$ are the weight matrices for this operation. To obtain the final re-weighted output feature maps with enhanced representations, the generated maps S are re-scaled with respect to U to perform the element-wise multiplication operation between both feature maps. The scaling operation $F_{scl}(.)$ is mathematically expressed as follows:

$$\bar{X} = F_{scl}(u_c, s_c) = US, \tag{5}$$

where $\bar{X} = [\bar{x}_1, \bar{x}_2, \dots .\bar{x}_c]$, scalar s_c and the feature map $u_c \in R^{H\times W}$.

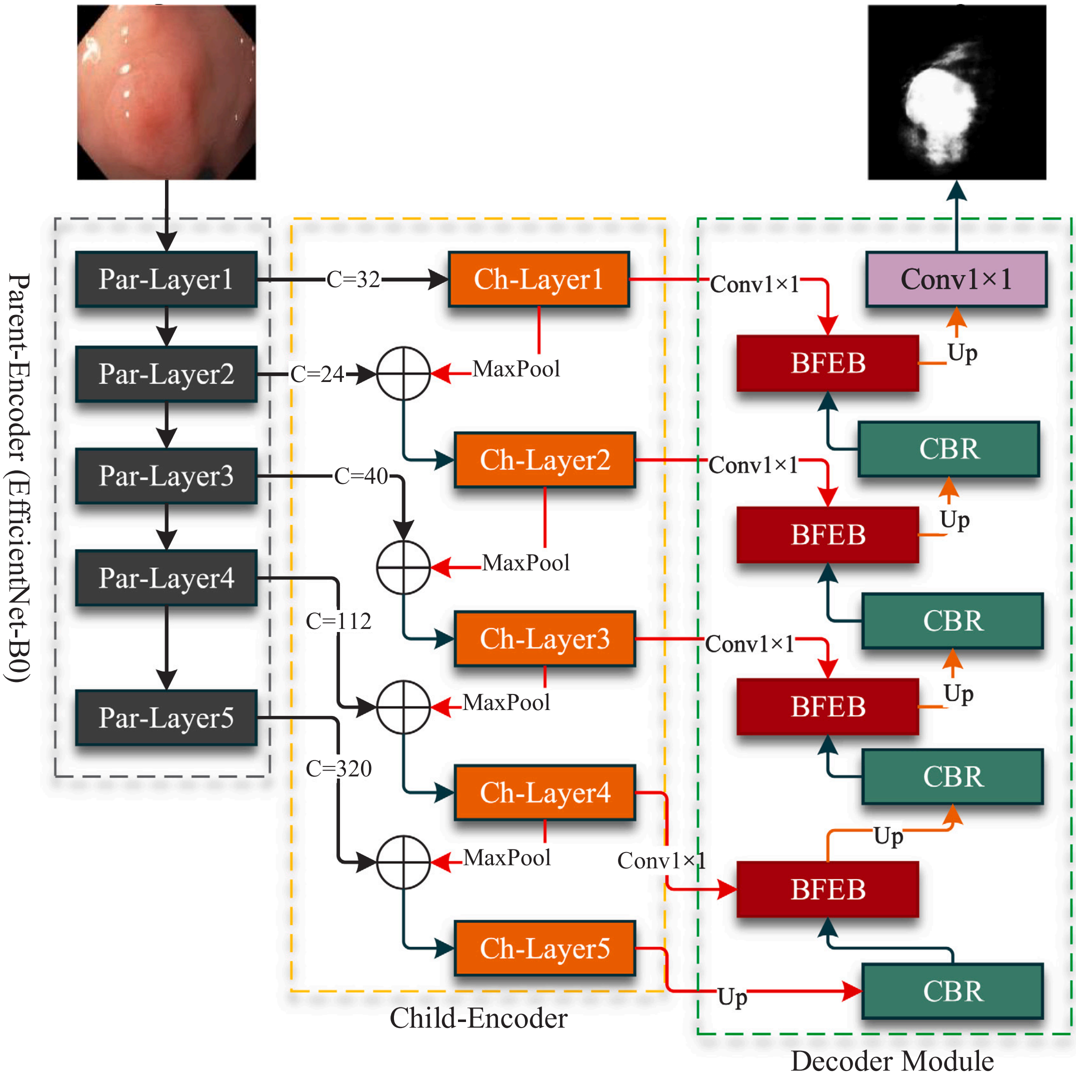

Fig. 1. Graphical representation of the proposed "PCRNet: Parent–Child Relation Network for automatic polyp segmentation". The proposed model consists of three main blocks: the Parent-Encoder, Child-Encoder, and Decoder Module. In the diagram, "Up" represents the up-sampling operation and each layer of the Child-Encoder is consists of a Squeeze-and-Excitation Attention Block (SEAB).

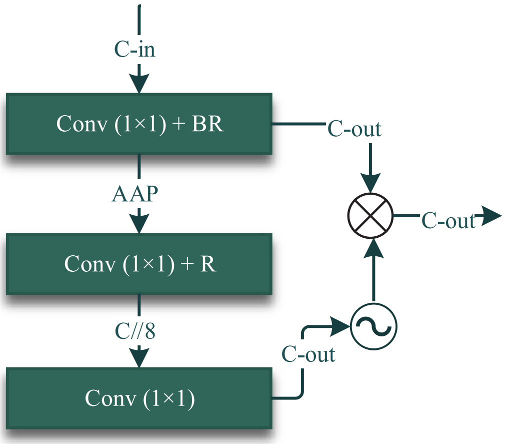

Fig. 2. Squeeze-and-Excitation Attention Block (SEAB).

3.2. Decoder module

In polyp segmentation, deep CNN models often experience significant information loss due to repeated down-sampling, leading to the false prediction of smaller polyp objects. To address this issue, it is crucial to focus on shallow features fs since they contain clear boundary information that can significantly enhance accurate early-stage polyp segmentation. However, these boundaries are submerged in background noise due to the limited receptive field in early layers. In contrast, the deep features fd from the deeper layers have a clean background but contain minimal boundary information.

To effectively utilize both shallow and deep feature maps, models like PraNet [13], CaraNet [49], UACANet [14], DCANet [12], and CCBANet [50] have adopted different strategies to extract uncertain boundary information. Inspired by these, we proposed a Boundary-aware Foreground Extraction Block (BFEB) as shown in Fig. 3. The BFEB in each layer is followed by Convolution, BatchNorm and ReLU (CBR) operation to meet the channels requirements of the following upper layer. In our proposed PCRNet, BFEB is aligns with concepts from the Balancing Attention Module (BAM) [50] and Uncertainty Augmented Context Attention (UACA) [14], but it primarily focuses on foreground extraction using uncertain boundary information. The process of extracting both boundary maps $Batt_i \in R^{1\times H_i\times W_i}$ and foreground

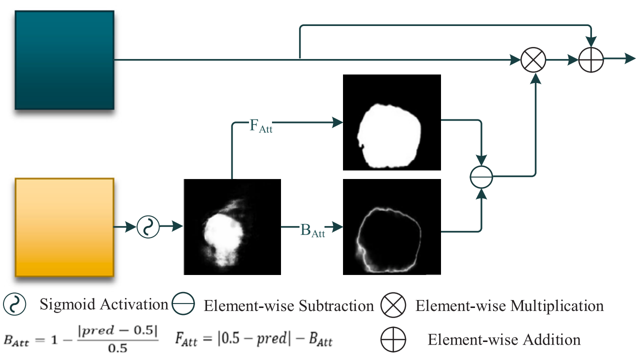
F_Att
B_Att
Sigmoid Activation
Element-wise Subtraction
Element-wise Multiplication
Element-wise Addition
$B_{Att} = 1 - \frac{|pred - 0.5|}{0.5}$
$F_{Att} = |0.5 - pred| - B_{Att}$

Fig. 3. Illustration of the Boundary-aware Foreground Extraction Block (BFEB).

maps $Fatt_i \in R^{1\times H_i \times W_i}$ from f_d (pred) can be mathematically expressed as follows:

$$Batt_i^j = 1 - \frac{|\sigma(fd_{i+1}^j) - T|}{T}, T = 0.5, \tag{6}$$

$$Fatt_i^j = |T - \sigma(fd_{i+1}^j)|, T = 0.5, \tag{7}$$

$$F_i^j = Fatt_i^j - Batt_i^j, \tag{8}$$

where $fd_{i+1}^j \in (0,1)$ is the jth location value of the prediction map $fd_{i+1} \in R^{1\times H_i \times W I i}$ from fd which is generated by the $(i+1)$th decoder block through the sigmoid σ activation function. In this equation, T represents the threshold used to determine whether a specific position belongs to the foreground or the background. The final refined foreground map (F_i^j) is then multiplied and added with the shallow feature maps (fs) through element-wise multiplication and addition operation operations to generate the information-rich feature maps.

$$Fs' = (Fs \times F_i^j) + F_i^j \tag{9}$$

The Fs' is then passed through an up-sampling operation and convolution operation to make the feature maps ready to be used as the input of the next shallow layer.

4. Experimental setup

This section demonstrates the experimental setting to train and test the efficacy of the proposed method, and each component is further described in detail one by one. Section 4.1 provides a brief overview of the benchmark datasets utilized in the experiments. Section 4.2 explains the loss functions employed to train the proposed model. Section 4.3 details the evaluation metrics used, and Section 4.4 describes the specific implementation settings for training the model.

4.1. Benchmark datasets

We used five publicly available benchmark datasets for polyp segmentation to train, validate, and test the proposed model. A brief description of these datasets is provided below:

- **Kvasir-SEG:** Kvasir-SEG [51] is one of the largest polyp segmentation datasets, comprising 1000 images of varying sizes and dimensions, with resolutions ranging from (487 × 332) to (1920 × 1072). These images were collected from video sequences of various cases and provided by the Norwegian Vestre Viken Health Trust. The images were manually labeled by a medical doctor and validated by an experienced gastroenterologists. The dataset includes 700 images of larger polyps, 8 images of smaller polyps, and 323 images of medium polyps. Additionally, Kvasir-SEG contains 196 images of sessile or flattened polyps with diameters of less than ten millimeters.
- **CVC-ClinicDB:** CVC-ClinicDB [52] is a well-known polyp segmentation dataset consisting of 612 images collected from 29 polyp video sequences of 23 patients. The images in this dataset were manually labeled by expert gastroenterologists.
- **CVC-ColonDB:** The CVC-ColonDB [53] dataset consists of 12,000 images extracted from 15 short colonoscopy videos. Out of these 12,000 images, 380 were annotated for segmentation purposes. These images were carefully selected based on color, texture, and feature diversity, making it one of the more challenging datasets for segmentation models.
- **ETIS-LaribPolypDB:** The ETIS [54] dataset is derived from 34 Wireless Capsule Endoscopy (WCE) video sequences and consists of 196 annotated images. Researchers from the Universitat Autonoma de Barcelona collected this dataset. The dataset consists of smaller and varying shapes polyps with camouflage properties, making it one of the challenging datasets for polyp segmentation models.
- **EndoScene:** The EndoScene [55] dataset consisted of 912 manually annotated images extracted from 44 video sequences of 36 patients. The dataset is prepared by combining the CVC-ClinicDB dataset (612 images) and the CVC-300 (300 images) [53] after excluding repeated images.

4.2. Loss function

To train the proposed PCRNet, we employed the structure loss $\mathcal{L}_{Str}$ [13], which combines the weighted intersection over union loss $\mathcal{L}_{IoU}^{w}$ and weighted binary cross-entropy $\mathcal{L}_{BCE}^{w}$. In $\mathcal{L}_{Str}$, the $\mathcal{L}_{IoU}^{w}$ increases the weights of hard pixels to emphasize their significance, while $\mathcal{L}_{BCE}^{w}$ prioritizes hard pixels by assigning them higher weights instead of treating all pixels equally. This combination enhances the model's robustness in learning complex situations and improves performance. The effectiveness of this approach has been validated in various segmentation and salient object detection tasks.

$$\mathcal{L}_{BCE}^{w} = -\tfrac{1}{N} \textstyle\sum_{i\in I} g_m[i] \log\left(p_m[i]\right) + \left(1 - g_m[i]\right) \log\left(1 - p_m[i]\right) \tag{10}$$

$$\mathcal{L}_{IoU}^{w} = 1 - \frac{\sum_{i\in I} g_m[i] p_m[i]}{\sum_{i\in I} g_m[i] + p_m[i] - g_m[i] p_m[i]} \tag{11}$$

$$\mathcal{L}_{Str} = \mathcal{L}_{IoU}^{w}(p_m, g_m) + \mathcal{L}_{BCE}^{w}(p_m, g_m) \tag{12}$$

where N is the total number of image pixels, $i \in I$ refers to a pixel in the output (prediction map) and ground truth, (p_m) represents the prediction, and (g_m) denotes the ground truth maps.

4.3. Evaluation metrics

Six commonly used metrics were utilized to conduct an in-depth performance analysis of the proposed segmentation model. These evaluation metrics are further elaborated below:

The Dice Score (DSc) is a standard metric used to measure pixel-level similarity between the predicted mask and the corresponding ground truth. Mathematically, the Dice Score is defined as follows:

$$Dice = \frac{2|Y \cap X|}{|Y| + |X|} = \frac{2 * TP}{2 * TP + FP + FN}. \tag{13}$$

Intersection over Union (IoU) is another widely used metric for evaluating segmentation tasks. It measures the pixel-level consistency between the ground truth and the predicted segmentation maps. The mathematical representation of IoU is as follows:

$$IoU = \frac{Y \cap X}{Y \cup X} = \frac{TP}{TP + FP + FN}. \tag{14}$$

The mean absolute error (MAE) is the pixel-by-pixel quality evaluation metric calculated between the predicted maps and ground-truth images. It represents the average absolute error between the predicted maps and ground truth pixel values and is presented as follows:

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |X_i - Y_i|. \tag{15}$$

In the above equations, TP, FP, TN, and FN represent the values of true positives, false positives, true negatives, and false negatives cases, respectively, whereas Y represents the model's prediction and X represents the ground truths images of the employed test dataset.

Structural similarity measure (S-measure) [56], denoted by S_α, simultaneously measures region- and object-aware structural similarity between the model predictions and the ground-truth images from the test datasets.

$$S_\alpha = \alpha \times S_o + (1 - \alpha) \times S_r. \tag{16}$$

where S_o stands for object-aware structural similarity, S_r for region-aware structural similarity, and α is the balanced hyper-parameters between them.

The Weighted F-measure (F^w_β) is a metric that considers both recall and accuracy. It calculates the weightage of misclassified pixels by assigning weights to each pixel. It is also interpreted as the harmonic mean of precision and recall. In the (F^w_β), 1 is assigned as positive, while zero is assigned as negative weightage to pixel values. Mathematically (F^w_β) is represented as:

$$(F^i_\beta) = \frac{(1+\beta^2)P^i \times R^i}{\beta^2 \times P^i + R^i}, \tag{17}$$

The variables P^i and R^i denote precision and recall, respectively, for a given threshold value i. Additionally, the parameter β, which represents the trade-off between precision and recall, is often assigned a default value of 0.3.

Enhanced-alignment measure (E-measure)(E_ϕ) is a commonly used segmentation model evaluation metric to assess segmentation results at the pixel and image levels. In mathematical representation, E_α is calculated as:

$$E_\phi = \frac{1}{w \times h} \sum_{b=1}^{w} \sum_{a=1}^{w} \phi(Y, X). \tag{18}$$

where, w and h represent the width and height of the prediction maps, while ϕ denotes the enhanced alignment matrix between them.

4.4. Implementation details

We implemented the proposed model using Python programming and PyTorch framework. Nvidia GeForce RTX 3080 with 10 G memory is used to accelerate the training process. The model is trained and tested using the specified data distribution as adapted in PraNet [13]. In this distribution, the training dataset is the partial combination of Kvasir-SEG and CVC-ClinicDB, and the part of these datasets in the testing set are considered as the seen datasets for the model to evaluate the model performance. Instead of these two datasets, all other datasets were only used to test the model performance and are considered as the unseen for the model to examine the model performance in segmenting polyps of different properties. In the training step, we adopted a multi-scale strategy {0.75, 1, 1.25} to rescale the images to mitigate the size variation issues of polyps and the training images. We employed the widely used AdamW optimizer to update the network weights parameters. For training the model, the initial learning rate is set to 10^{-4}, and the images are resized to 352 × 352. The model is trained using a batch size of 8 for 100 epochs. The combination of weighted IoU loss and weighted binary cross entropy loss assists the training process. The model performance is evaluated using a set of different evaluation metrics, and the testing images are resized to 352 × 352 without any post-processing or optimization techniques.

5. Performance comparison with state-of-the-art methods

We evaluated the performance of the proposed PCRNet both quantitatively and qualitatively against existing methods. The comparison targets two categories of methods based on their encoder architecture:

1. **CNN-Based Pre-Trained Encoders:** This category includes models such as ResUNet [57], SFA [46], R2UNet [58], UNet3+ [59], UNet [40], UNet++ [41], DeepLabV3 [24], HRNet [60], PSPNet [23], HarDNet-MSEG [61], ACSNet [32], SANet [47], PraNet [13], and EU-Net [62].
2. **Transformer-Based Encoders:** This category includes models such as Swin-UNet [63], SETR_Naive_S [64], SETR_PUP_S [64], TransUNet [65], SegFormer [16], UTNet [66], DCR-Net [67], SR-AttNet [68], DilatedSegNet [69], and TranSEFusionNet [70].

In this performance evaluation, Kvasir-SEG and CVC-ClinicDB are considered as seen datasets, while CVC-300, CVC-ColonDB, and ETIS are treated as unseen datasets for PCRNet. Detailed quantitative and qualitative comparisons are provided in the following subsections.

5.1. Quantitative comparison

Tables 1 and 2 present a comprehensive quantitative comparison between our model and other state-of-the-art methods on both seen and unseen datasets. All quantitative measures, including floating-point operations (FLOPs) and model trainable parameters, are sourced from recently published articles.

In terms of trainable parameters, the PCRNet model has a total of 5.0087 million parameters, the smallest number compared to the other models. In contrast, Swin-UNet [63] exhibits the smallest FLOPs value of 5.85786G. Additionally, PCRNet demonstrated superior overall performance across various datasets compared to existing approaches. In Tables 1 and 2, the best values are highlighted in bold, while a "–" indicates unreported results.

5.1.1. Results on the Kvasir-SEG dataset

The Kvasir-SEG dataset is considered a seen dataset, as the model was trained using the data distribution specified in [13]. The model's performance on the Kvasir-SEG dataset is evaluated using the test set from the same distribution. As shown in Table 1, the performance of the proposed PCRNet on Kvasir-SEG is comparable to that of EU-Net [62]. Notably, PCRNet has significantly fewer trainable parameters than EU-Net.

In contrast, PCRNet demonstrates superior performance compared to recently published transformer-based models, such as Swin-UNet [63], SETR_Naive_S [64], SETR_PUP_S [64], TransUNet [65], SegFormer [16], UTNet [66], DCR-Net [67], SR-AttNet [68], DilatedSegNet [69], and TranSEFusionNet [70]. The performance gap is nearly 20% in some cases, particularly in terms of Dice, IoU, and $S\alpha$. Despite this, PCRNet consistently presents the fewest trainable parameters, validating the effectiveness of our strategy in extracting information-rich feature maps. Even though, PCRNet presents promising performance in segmenting varying polyps; however, PCRNet fails to segment polyps in the cases when there is a higher illumination difference and also in the presence of occlusions.

5.1.2. Results on the CVC-ClinicDB dataset

Similar to Kvasir-SEG, the CVC-ClinicDB dataset was also part of the combined training datasets, and PCRNet's performance was evaluated using the test subset of this dataset. As reported in Table 1, PCRNet achieves the highest Dice score (0.921%) and IoU (0.869%) among all existing methods. While PSPNet and HarDNet-MSEG demonstrate leading performance in $S\alpha$ and MAE, they also have several times more trainable parameters compared to PCRNet.

Compared to the base model PraNet [13], which uses the same training and testing data distribution, PCRNet shows a 2.2% improvement in Dice and a 2% improvement in IoU. Notably, PCRNet has six times fewer trainable parameters than PraNet, highlighting the efficiency of our model.

Table 1
Quantitative performance comparison of our PCRNet against other state-of-the-art approaches using the CVC-ClinicDB and Kvasir-SEG datasets. All results are copied from recently published manuscripts, and "–" indicates the unreported results. Boldface letters denote the best outcomes in this comparison among all these methods.

Methods	Flops (G)↓	Parameters (M)↓	Kvasir-SEG				CVC-Clinic-DB			
			Dice↑	IoU↑	$S\alpha$ ↑	MAE↓	Dice↑	IoU↑	$S\alpha$ ↑	MAE↓
ResUNet [57]	153.013	13.043	0.511	0.436	–	–	0.451	0.457	–	–
SFA [46]	–	–	0.723	0.611	0.782	0.075	0.700	0.607	0.793	0.042
R2UNet [58]	288.917	39.091	0.731	0.638	0.791	0.072	0.711	0.631	0.814	0.033
UNet3+ [59]	377.502	26.971	0.810	0.721	0.842	0.056	0.866	0.795	0.901	0.018
UNet [40]	103.382	31.037	0.818	0.746	0.858	0.055	0.823	0.755	0.889	0.019
UNet++ [41]	65.925	9.163	0.821	0.743	0.862	0.048	0.794	0.729	0.873	0.022
DeepLabV3 [24]	60.408	58.625	0.835	0.761	0.871	0.047	0.902	0.843	0.928	0.011
HRNet [60]	42.831	29.530	0.849	0.773	0.879	0.045	0.888	0.827	0.927	0.013
PSPNet [23]	87.203	46.706	0.896	0.838	0.908	0.028	0.918	0.869	0.938	0.010
HarDNet-MSEG [61]	11.376	17.423	0.897	0.839	0.912	0.028	0.909	0.864	**0.938**	**0.007**
ACSNet [32]	21.726	29.450	0.898	0.838	0.920	0.032	0.882	0.826	0.927	0.011
PraNet [13]	13.078	30.498	0.898	0.840	0.915	0.03	0.899	0.849	0.936	0.009
SANet [47]	11.274	23.899	0.904	0.847	0.915	0.028	0.916	0.859	0.939	0.012
EU-Net [62]	23.128	31.358	**0.908**	**0.854**	0.917	**0.028**	0.902	0.846	0.936	0.011
Swin-UNet [63]	**5.8576**	27.12	0.608	0.494	0.727	0.096	0.686	0.586	0.799	0.038
SETR_Naive_S [64]	41.682	86.173	0.649	0.541	0.748	0.095	0.746	0.661	0.832	0.030
SETR_PUP_S [64]	43.343	85.839	0.650	0.544	0.747	0.097	0.751	0.676	0.845	0.032
TransUNet [65]	61.559	66.782	0.715	0.593	0.664	0.771	0.828	0.753	0.882	0.023
SegFormer [16]	21.999	57.769	0.854	0.778	0.883	0.039	0.898	0.838	0.927	0.014
UTNet [66]	81.607	14.406	0.854	0.784	0.879	0.047	0.880	0.828	0.920	0.018
DCR-Net [67]	–	28.7	0.886	0.825	0.911	0.028	0.896	0.844	0.933	0.010
SR-AttNet [68]	–	37.1	0.871	0.806	–	–	0.786	0.693	–	–
TranSEFusionNet [70]	124.43	127.74	0.845	0.781	–	–	0.864	0.790	–	–
DilatedSegNet [69]	27.1	18.11	0.895	0.833	–	–	0.827	0.754	–	–
PRCNet [71]	31.17	9.60	0.799	0.716	0.837	0.086	0.925	0.869	0.943	0.009
CSCAU-Net [72]	–	35.27	0.903	0.846	**0.918**	0.031	0.915	0.864	0.942	0.010
PCRNet (Ours)	16.941	**5.0087**	0.902	0.840	0.903	0.030	**0.921**	**0.869**	0.936	0.011

Table 2
The quantitative performance comparison of PCRNet with recent methods on the CVC-ColonDB, ETIS, and CVC-300 datasets. Similar to the Table 1, all the results are copied from recently published articles, and "–" indicates the unreported results. Boldface letters denote the best outcomes in this comparison among all these methods.

Methods	Flops (G)↓	Parameters (M)↓	CVC-ColonDB				ETIS				CVC-300			
			Dice↑	IoU↑	$S\alpha$ ↑	MAE↓	Dice↑	IoU↑	$S\alpha$ ↑	MAE↓	Dice↑	IoU↑	$S\alpha$ ↑	MAE↓
SFA [46]	–	–	0.469	0.347	0.634	0.094	0.297	0.217	0.557	0.109	0.467	0.329	0.341	0.065
R2UNet [58]	288.917	39.091	0.449	0.384	0.669	0.060	0.405	0.335	0.663	0.035	0.620	0.551	0.782	0.018
ResUNet [57]	153.013	13.043	0.489	0.381	0.673	0.061	0.354	0.266	0.622	0.051	0.545	0.422	0.731	0.030
ResUNet++ [11]	134.109	14.482	0.417	0.341	0.646	0.059	0.244	0.196	0.579	0.042	0.344	0.263	0.623	0.035
UNet [40]	103.382	31.037	0.512	0.444	0.712	0.061	0.398	0.335	0.684	0.036	0.710	0.627	0.843	0.022
UNet++ [41]	65.925	9.163	0.483	0.410	0.691	0.064	0.401	0.344	0.683	0.035	0.707	0.624	0.839	0.018
UNet3+ [59]	377.502	26.971	0.603	0.520	0.749	0.052	0.483	0.404	0.708	0.034	0.734	0.649	0.838	0.016
DeepLabV3 [24]	60.408	58.625	0.658	0.573	0.784	0.046	0.462	0.386	0.696	0.039	0.780	0.684	0.867	0.017
HRNet [60]	42.831	29.530	0.627	0.546	0.773	0.045	0.384	0.350	0.676	0.033	0.801	0.710	0.883	0.012
HarDNet-MSEG [61]	11.376	17.423	0.735	0.666	0.834	0.038	0.700	0.630	0.828	**0.015**	0.874	0.804	0.924	0.009
ACSNet [32]	21.726	29.450	0.716	0.649	0.829	0.039	0.578	0.509	0.754	0.059	0.863	0.787	0.923	0.013
PraNet [13]	13.078	30.498	0.709	0.640	0.819	0.045	0.628	0.567	0.794	0.031	0.871	0.797	0.925	0.010
EU-Net [62]	23.128	31.358	0.756	0.681	0.831	0.045	0.687	0.609	0.793	0.067	0.837	0.765	0.904	0.015
SETR_Naive_S [64]	41.682	86.173	0.307	0.226	0.568	0.110	0.291	0.219	0.577	0.106	0.381	0.274	0.632	0.069
SETR_PUP_S [64]	43.343	85.839	0.283	0.213	0.556	0.127	0.288	0.222	0.577	0.112	0.221	0.163	0.548	0.124
Swin_UNet [63]	**5.8576**	27.12	0.353	0.263	0.599	0.101	0.259	0.191	0.557	0.107	0.404	0.312	0.651	0.070
TransUNet [65]	61.559	66.782	0.441	0.339	0.647	0.064	0.325	0.241	0.608	0.045	0.506	0.397	0.712	0.028
SegFormer [16]	21.999	57.769	0.683	0.591	0.794	0.047	0.550	0.473	0.737	0.041	0.801	0.700	0.868	0.018
UTNet [66]	81.607	14.406	0.702	0.620	0.805	0.061	0.485	0.414	0.655	0.173	0.806	0.720	0.882	0.023
DCRNet [67]	–	28.7	0.704	0.631	0.821	0.038	0.556	0.496	0.736	0.015	–	–	–	–
HardNet-CPS [73]	–	–	0.729	0.658	0.829	0.037	0.690	0.619	0.822	0.014	0.891	0.826	**0.938**	0.008
SR-AttNet [68]	–	37.1	0.665	0.539	–	–	0.476	0.355	–	–	–	–	–	–
BLE-Net [74]	–	–	0.731	0.658	0.832	0.044	0.673	0.594	0.810	0.032	0.879	0.805	0.928	0.009
PRCNet [71]	31.17	9.60	0.719	0.642	0.824	0.041	0.611	0.549	0.781	0.036	0.883	0.814	0.927	0.007
CSCAU-Net [72]	–	35.27	**0.788**	**0.703**	**0.857**	0.036	0.688	0.608	0.814	0.026	–	–	–	–
PCRNet(Ours)	16.941	**5.0087**	0.766	0.684	0.840	**0.035**	**0.729**	**0.647**	**0.833**	0.026	**0.895**	**0.826**	0.932	**0.006**

5.1.3. Results on the CVC-ColonDB dataset

The CVC-ColonDB dataset is an unseen dataset for most methods in this comparison, including the proposed PCRNet. We evaluated the responsiveness and generalization ability of PCRNet on CVC-ColonDB and compared its performance against other approaches. The results are reported on the left side of Table 2.

PCRNet achieved the highest scores across all metrics, including Dice, IoU, $S\alpha$, and MAE. Specifically, compared to the vital baseline PraNet [13], PCRNet shows a 5.7% improvement in Dice, a 4.4% improvement in IoU, a 2.1% improvement in $S\alpha$, and a 1.1% improvement in MAE. When compared to transformer-based models, PCRNet demonstrates a minimum of 4.6% improvement in Dice, 3.5% in IoU, and 1.5% in $S\alpha$. In overall performance comparison on CVC-ColonDB dataset, CSCAU-Net [72] presented a bit better performance in terms of Dice Score and IoU but on the other hand it has about 35.27 million parameters, while the proposed approach contains only 5.0087 million parameters.

Table 3
Quantitative analysis of the conducted ablation experiments to evaluate the effectiveness of individual modules on the ETIS dataset.

Baseline	Pre-trained Encoder	Child Encoder	BFEB	Dice	IoU	F_{β}^{W}	$S\alpha$	E_{ϕ}^{max}	MAE
✓				0.398	0.335	0.366	0.684	0.740	0.036
✓	✓			0.676	0.591	0.630	0.810	0.876	0.020
✓	✓	✓		0.687	0.608	0.650	0.815	0.876	0.026
✓	✓		✓	0.702	0.616	0.662	0.819	0.861	0.020
✓	✓	✓	✓	0.729	0.647	0.690	0.833	0.881	0.026

5.1.4. Results on the ETIS dataset

The ETIS dataset is another unseen dataset and is considered the most challenging among the five due to significant variations in image color, resolution, polyp size, and polyp shape. These factors contribute to its difficulty in achieving high segmentation accuracy. The evaluation results for the ETIS dataset are presented in the middle of Table 2.

The comparison reveals that most baseline methods, including UT-Net [66] and ACSNet [32], performed poorly on the ETIS dataset, despite showing strong performance on other datasets with larger polyps. This diminished performance indicates these methods' limited learning and generalization capabilities for segmenting polyps with diverse shapes and sizes.

In contrast, the proposed PCRNet outperformed all existing systems, including transformer-based models, on this challenging dataset. PCRNet achieved a 2.9% improvement in Dice, a 1.7% improvement in IoU, and a 1.1% improvement in $S\alpha$ compared to the highest scores achieved by existing approaches. This exceptional performance in segmenting early-stage small polyps with blurry boundaries underscores the model's greater generalization and learning capabilities.

5.1.5. Results on the CVC-300 dataset

Similar to the CVC-ColonDB and ETIS datasets, the CVC-300 dataset is also unseen for most models listed in this comparison. The images in CVC-300 are similar to those in the training datasets, making segmentation tasks somewhat less challenging compared to CVC-ColonDB and ETIS. The quantitative results for this dataset are presented on the right side of Table 2.

The proposed PCRNet outperformed all other approaches in terms of Dice and IoU scores. Compared to the best-performing method, EU-Net [62], PCRNet achieved a 5.8% improvement in Dice, a 6.1% improvement in IoU, a 2.8% improvement in $S\alpha$, and a 6% improvement in MAE. Additionally, PCRNet demonstrated a notable performance improvement over the baseline PraNet [13] on this dataset.

5.2. Qualitative comparison

Fig. 4 illustrates the qualitative evaluation of the proposed PCRNet compared to seven state-of-the-art CNN-based polyp segmentation approaches. In the first seven rows, the polyps are either from ETIS or from CVC-ColonDB datasets, and the images from these datasets are unseen for our model. The comparison is organized into three groups based on polyp size:

1. **Small Polyps (≤5% of the overall image):** The first four rows ($s_1 - s_4$) display tiny polyps in various scenarios. In the first three rows ($s_1 - s_3$), polyps closely resemble the surrounding tissue due to illumination differences during image capture. The fourth row (s_4) shows polyps with indistinct, blurred boundaries caused by camera shake. In all these cases, PCRNet outperformed the other models, with segmented maps closely matching the ground truths.
2. **Medium-Sized Polyps (5%–20% of the overall image):** The middle four rows ($m_1 - m_4$) depict medium-sized polyps. Except for the polyps in row (m_3) with unclear boundaries, the polyps in this group are larger and show varied colors within the polyp area. CNN-based methods often struggle to accurately segment these polyps, while PCRNet successfully segments all except the polyp in row (m_1).
3. **Large Polyps (≥20% of the overall image):** The final group ($l_1 - l_4$) includes large polyps that cover a significant portion of the image. CNN models often produce incorrect predictions due to inaccurate estimation of polyp regions, especially if they overlook the importance of extracting global information. For instance, in row (l_4), where the upper part of the polyp is similar to normal tissue, PCRNet, like most models, fails to capture the entire affected area.

Overall, PCRNet demonstrates exceptional performance by effectively integrating both local and global information. It is achieved due to its learning strategy, which emphasizes learning from global to local information, allows it to capture larger receptive fields and improve segmentation accuracy.

6. Ablation experiment

We conducted a series of ablation experiments to assess the effectiveness of each component in the proposed network. These experiments included testing a simple UNet [40] as the baseline, a UNet with a pre-trained EfficientNet-B0 as the backbone, and a UNet with the addition of a Child encoder module. We also evaluated the impact of incorporating the Boundary-aware Foreground Extraction Block (BFEB) and examined the performance with the combination of all these modules. For each configuration, except for the baseline UNet, we used the same training datasets and conditions as those employed for training the proposed PCRNet. The performance of these models is evaluated on the ETIS dataset, which is known for its complexity due to varying polyp shapes, sizes, and illumination conditions, making it particularly difficult for accurate segmentation of tiny polyps. The results of these ablation experiments are detailed in Table 3.

The first row of Table 3 shows the evaluation results of the simple UNet on the ETIS dataset, with values sourced from a recently published article. These results illustrate the limited learning capability of the basic UNet architecture and its poor performance in segmenting tiny polyps. To enhance performance while managing model size and computational complexity, we employed a pre-trained EfficientNet-B0 [20] as the encoder, leading to a notable improvement, as demonstrated in the second row of Table 3.

Instead of using computationally expensive transformer models, we introduced a child encoder module featuring four Squeeze and Excite Attention Blocks (SEAB) to extract more informative semantic feature maps. This module, which learns from the parent encoder and its previous layer feature maps, provides refined outputs, as seen in the third row of Table 3.

The Boundary-aware Foreground Extraction Block (BFEB) was specifically designed to address the challenges of segmenting tiny polyps with indistinct boundaries. The impact of BFEB on segmenting small polyps is evident in the fourth row of Table 3.

To evaluate the combined effectiveness of all these components, we designed our proposed PCRNet. The final row in Table 3 shows the performance of PCRNet, highlighting the significant contribution of the child encoder module. The results indicate that, without the child encoder module, the Dice score was 70%. However, with the addition of the child encoder, this score improved to 72.9%. These experiments demonstrate that each module plays a crucial role in achieving higher accuracy and generalizability while maintaining a minimal model size.

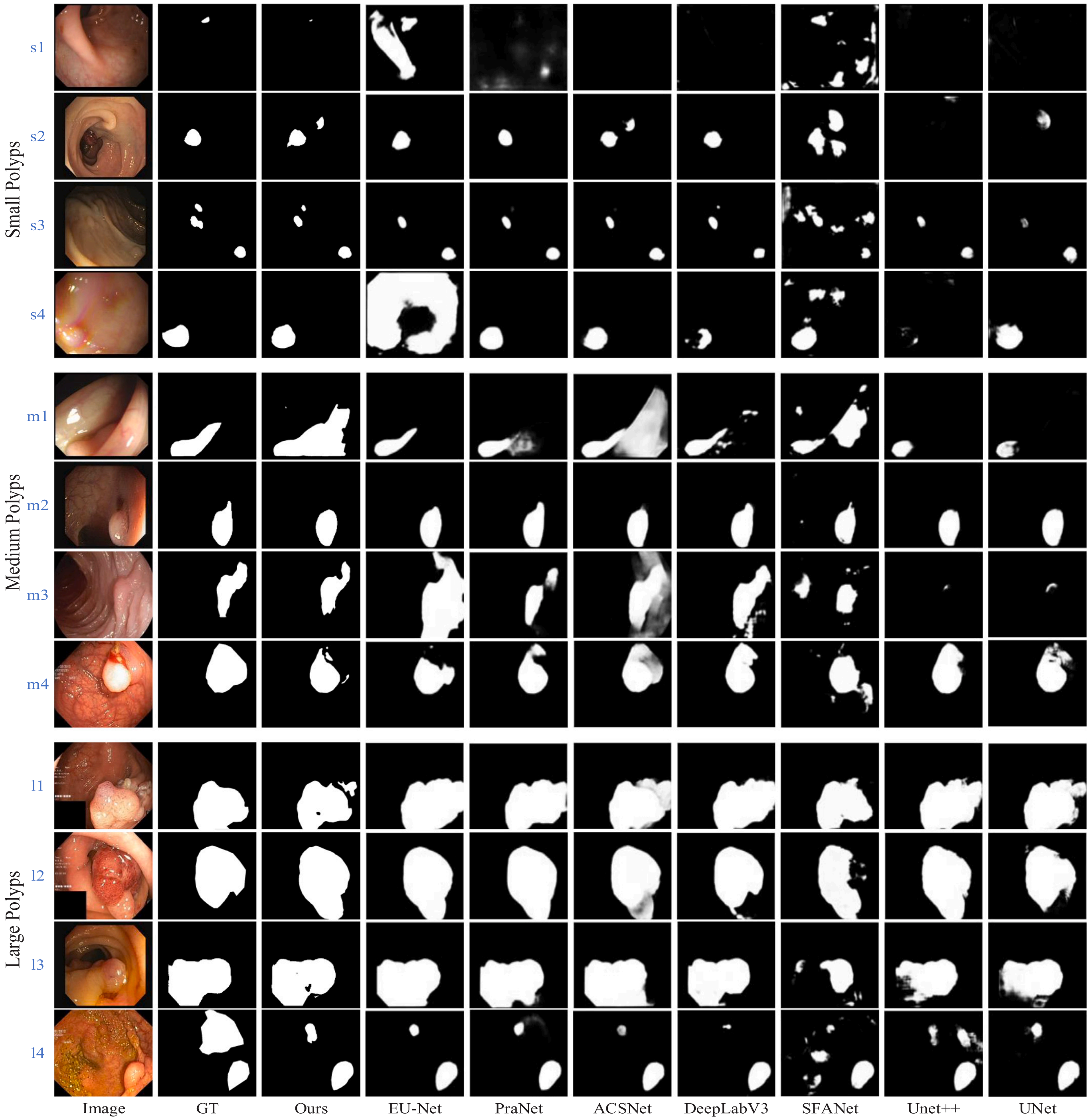

Fig. 4. Visualization of qualitative comparison of the proposed PCRNet with other CNN-based models. The initial four rows depict the presence of tiny polyps (covering ≤5% of the whole image) in various settings, such as darker targets (s_1, s_2, s_3) and indistinct boundaries (s_4). The polyps in the m_1 - m_4 rows are of medium size, covering about 5% to 20% of the image in size. The boundaries of the polyps in the seventh row m_3 are blurred, while the polyps in the m_1, m_2, and m_4 have various internal colors. Polyps in the final rows ($l_1 - l_4$) comprise a larger proportion (≥ 20%) of the image. In this comparison, our proposed approach outperformed all cases, and the prediction maps are identical to the ground truth maps.

7. Conclusion

This paper introduces “PCRNet”, a lightweight model designed for segmenting polyps of varying shapes, sizes, and with hazy boundaries. PCRNet incorporates a child encoder to enhance feature map representation and a Boundary-aware Foreground Extraction Block (BFEB) to address hazy boundaries in early-stage polyp segmentation. The analysis across five challenging datasets demonstrates the model’s superior performance compared to existing approaches, effectively segmenting various types of polyps.

PCRNet stands out with its efficiency, featuring just 5.0087 million parameters and achieving an average inference rate of 36.5 frames per second (fps). On the ETIS dataset, PCRNet shows improvements of 2.9% in Dice, 1.7% in IoU, and 1.1% in $S\alpha$ compared to the highest scores of existing methods. These outstanding features of the proposed model could be considered as a starting step towards the development of an embedded program for screening devices to help the oncologist in detecting polyps of varying shapes and sizes during the screening procedure.

Despite these advancements, the model faces challenges in segmenting scattered polyps with subtle reflections and distinguishing between polyps and background areas. Future work will focus on addressing these limitations by developing methods to extract multi-scale global feature maps. Additionally, exploring PCRNet's performance on video data will be a key area of research to enhance real-time segmentation efficiency.

CRediT authorship contribution statement

Zaka-Ud-Din Muhammad: Writing – original draft, Methodology, Formal analysis, Visualization, Conceptualization. **Zhangjin Huang:** Writing – review & editing, Supervision. **Naijie Gu:** Writing – review & editing, Investigation.

Informed consent

This articles does not contain patient data.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Declaration of Generative AI and AI-assisted technologies in the writing process

Authors declares that there is no AI tool has been used in writing or formatting this manuscript.

Declaration of competing interest

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

Acknowledgments

This work was supported in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2023C01143), the Anhui Provincial Major Science and Technology Project (No. 202203a05020016), and the National Key R&D Program of China (No. 2022YFB3303400).

Data availability

The datasets generated during and/or analyzed during the current study are available in the Kvasir SEG repository https://datasets.simula.no/kvasir-seg/ and CVC-CilinicDB https://www.kaggle.com/datasets/balraj98/cvcclinicdb.

References

[1] J.C. Van Rijn, J.B. Reitsma, J. Stoker, P.M. Bossuyt, S.J. Van Deventer, E. Dekker, Polyp miss rate determined by tandem colonoscopy: a systematic review, Off. J. Am. Coll. Gastroenterol.| ACG 101 (2) (2006) 343–350.

[2] S.A. Karkanis, D.K. Iakovidis, D.E. Maroulis, D.A. Karras, M. Tzivras, Computer-aided tumor detection in endoscopic video using color wavelet features, IEEE Trans. Inf. Technol. Biomed. 7 (3) (2003) 141–152.

[3] S. Ameling, S. Wirth, D. Paulus, G. Lacey, F. Vilarino, Texture-based polyp detection in colonoscopy, in: Bildverarbeitung FÜR Die Medizin 2009, Springer, 2009, pp. 346–350.

[4] K. Mendel, H. Li, D. Sheth, M. Giger, Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography, Academic Radiol. 26 (6) (2019) 735–743.

[5] Y. Wang, W. Tavanapong, J. Wong, J.H. Oh, P.C. De Groen, Polyp-alert: Near real-time feedback during colonoscopy, Comput. Methods Programs Biomed. 120 (3) (2015) 164–179.

[6] Y. Shin, H.A. Qadir, L. Aabakken, J. Bergsland, I. Balasingham, Automatic colon polyp detection using region based deep cnn and post learning approaches, IEEE Access 6 (2018) 40950–40962.

[7] Y. Shin, H.A. Qadir, I. Balasingham, Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance, IEEE Access 6 (2018) 56007–56017.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[9] J.Y. Lee, J. Jeong, E.M. Song, C. Ha, H.J. Lee, J.E. Koo, D.-H. Yang, N. Kim, J.-S. Byeon, Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets, Sci. Rep. 10 (1) (2020) 1–9.

[10] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

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

[12] Z.-U.-D. Muhammad, Z. Huang, N. Gu, U. Muhammad, DCANet: deep context attention network for automatic polyp segmentation, Vis. Comput. (2022) 1–13.

[13] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 263–273.

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

[15] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, L. Shao, Polyp-pvt: Polyp segmentation with pyramid vision transformers, 2021, arXiv preprint arXiv:2108.06932.

[16] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.

[17] N.K. Tomar, D. Jha, S. Ali, H.D. Johansen, D. Johansen, M.A. Riegler, P. Halvorsen, DDANet: Dual decoder attention network for automatic polyp segmentation, 2020, arXiv preprint arXiv:2012.15245.

[18] D. Jha, S. Ali, N.K. Tomar, H.D. Johansen, D. Johansen, J. Rittscher, M.A. Riegler, P. Halvorsen, Real-time polyp detection, localization and segmentation in colonoscopy using deep learning, IEEE Access 9 (2021) 40496–40510.

[19] D. Jha, N.K. Tomar, S. Ali, M.A. Riegler, H.D. Johansen, D. Johansen, T. de Lange, P. Halvorsen, NanoNet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems, CBMS, 2021, pp. 37–43, http://dx.doi.org/10.1109/CBMS52027.2021.00014.

[20] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[21] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[22] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

[25] S. Liu, D. Huang, et al., Receptive field block net for accurate and fast object detection, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 385–400.

[26] X. Li, L. Zhang, A. You, M. Yang, K. Yang, Y. Tong, Global aggregation then local distribution in fully convolutional networks, 2019, arXiv preprint arXiv:1909.07229.

[27] G. Bertasius, J. Shi, L. Torresani, Semantic segmentation with boundary neural fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3602–3610.

[28] T.-W. Ke, J.-J. Hwang, Z. Liu, S.X. Yu, Adaptive affinity fields for semantic segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 587–602.

[29] G. Bertasius, L. Torresani, S.X. Yu, J. Shi, Convolutional random walk networks for semantic image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 858–866.

[30] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, L. Lin, Instance-level human parsing via part grouping network, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 770–785.

[31] J. Li, X. Feng, H. Fan, Saliency-based image correction for colorblind patients, Comput. Vis. Media 6 (2) (2020) 169–189.

[32] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, Y. Yu, Adaptive context selection for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 253–262.

[33] J. Zhong, W. Wang, H. Wu, Z. Wen, J. Qin, PolypSeg: An efficient context-aware network for polyp segmentation from colonoscopy videos, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23, Springer, 2020, pp. 285–294.

[34] H.J. Vala, A. Baxi, A review on Otsu image segmentation algorithm, Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) 2 (2) (2013) 387–389.

[35] M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active contour models, Int. J. Comput. Vis. 1 (4) (1988) 321–331.

[36] J. Li, X. Feng, Z. Hua, Low-light image enhancement via progressive-recursive network, IEEE Trans. Circuits Syst. Video Technol. 31 (11) (2021) 4227–4240.

[37] Y. Liu, L. Zhou, G. Wu, S. Xu, J. Han, Tcgnet: Type-correlation guidance for salient object detection, IEEE Trans. Intell. Transp. Syst. (2023).

[38] K. Yang, J. Han, G. Guo, C. Fang, Y. Fan, L. Cheng, D. Zhang, Progressive adapting and pruning: Domain-incremental learning for saliency prediction, ACM Trans. Multimed. Comput. Commun. Appl. (2024).

[39] V. Thambawita, D. Jha, M. Riegler, P. Halvorsen, H.L. Hammer, H.D. Johansen, D. Johansen, The medico-task 2018: Disease detection in the gastrointestinal tract using global features and deep learning, 2018, arXiv preprint arXiv:1810.13278.

[40] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[41] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.

[42] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39 (6) (2019) 1856–1867.

[43] J. Poorneshwaran, S.S. Kumar, K. Ram, J. Joseph, M. Sivaprakasam, Polyp segmentation using generative adversarial network, in: 2019 41St Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, IEEE, 2019, pp. 7201–7204.

[44] J. Kang, J. Gwak, Ensemble of instance segmentation models for polyp segmentation in colonoscopy images, IEEE Access 7 (2019) 26440–26447.

[45] T. Mahmud, B. Paul, S.A. Fattah, PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images, Comput. Biol. Med. 128 (2021) 104119.

[46] Y. Fang, C. Chen, Y. Yuan, K.-y. Tong, Selective feature aggregation network with area-boundary constraints for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 302–310.

[47] J. Wei, Y. Hu, R. Zhang, Z. Li, S.K. Zhou, S. Cui, Shallow attention network for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 699–708.

[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[49] A. Lou, S. Guan, M. Loew, CaraNet: Context axial reverse attention network for segmentation of small medical objects, 2021, arXiv preprint arXiv:2108.07368.

[50] T.-C. Nguyen, T.-P. Nguyen, G.-H. Diep, A.-H. Tran-Dinh, T.V. Nguyen, M.-T. Tran, Ccbanet: Cascading context and balancing attention for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 633–643.

[51] D. Jha, P.H. Smedsrud, M.A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H.D. Johansen, Kvasir-seg: A segmented polyp dataset, in: International Conference on Multimedia Modeling, Springer, 2020, pp. 451–462.

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

[53] J. Bernal, J. Sánchez, F. Vilarino, Towards automatic polyp detection with a polyp appearance model, Pattern Recognit. 45 (9) (2012) 3166–3182.

[54] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado, Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer, Int. J. Comput. Assist. Radiol. Surg. 9 (2) (2014) 283–293.

[55] D. Vázquez, J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, A.M. López, A. Romero, M. Drozdzal, A. Courville, A benchmark for endoluminal scene segmentation of colonoscopy images, J. Heal. Eng. 2017 (2017).

[56] M.-M. Cheng, D.-P. Fan, Structure-measure: A new way to evaluate foreground maps, Int. J. Comput. Vis. 129 (2021) 2622–2638.

[57] F.I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data, ISPRS J. Photogramm. Remote Sens. 162 (2020) 94–114.

[58] M.Z. Alom, M. Hasan, C. Yakopcic, T.M. Taha, V.K. Asari, Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation, 2018, arXiv preprint arXiv:1802.06955.

[59] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, Unet 3+: A full-scale connected unet for medical image segmentation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 1055–1059.

[60] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3349–3364.

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

[62] K. Patel, A.M. Bur, G. Wang, Enhanced u-net: A feature enhancement network for polyp segmentation, in: 2021 18th Conference on Robots and Vision, CRV, IEEE, 2021, pp. 181–188.

[63] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 205–218.

[64] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.

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

[66] Y. Gao, M. Zhou, D.N. Metaxas, Utnet: a hybrid transformer architecture 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 III 24, Springer, 2021, pp. 61–71.

[67] Z. Yin, K. Liang, Z. Ma, J. Guo, Duplex contextual relation network for polyp segmentation, in: 2022 IEEE 19th International Symposium on Biomedical Imaging, ISBI, IEEE, 2022, pp. 1–5.

[68] M.J. Alam, S.A. Fattah, SR-AttNet: An interpretable stretch–relax attention based deep neural network for polyp segmentation in colonoscopy images, Comput. Biol. Med. 160 (2023) 106945.

[69] N.K. Tomar, D. Jha, U. Bagci, Dilatedsegnet: A deep dilated segmentation network for polyp segmentation, in: International Conference on Multimedia Modeling, Springer, 2023, pp. 334–344.

[70] Y. Zhang, L. Liu, Z. Han, F. Meng, Y. Zhang, Y. Zhao, TranSEFusionNet: Deep fusion network for colorectal polyp segmentation, Biomed. Signal Process. Control. 86 (2023) 105133.

[71] J. Li, J. Wang, F. Lin, A.A. Heidari, Y. Chen, H. Chen, W. Wu, PRCNet: A parallel reverse convolutional attention network for colorectal polyp segmentation, Biomed. Signal Process. Control. 95 (2024) 106336.

[72] X. Shu, J. Wang, A. Zhang, J. Shi, X.-J. Wu, CSCA U-Net: A channel and space compound attention CNN for medical image segmentation, Artif. Intell. Med. 150 (2024) 102800.

[73] T. Yu, Q. Wu, HarDNet-CPS: Colorectal polyp segmentation based on harmonic densely united network, Biomed. Signal Process. Control. 85 (2023) 104953.

[74] N. Ta, H. Chen, Y. Lyu, T. Wu, BLE-Net: boundary learning and enhancement network for polyp segmentation, Multimedia Syst. (2022) 1–14.