Ingénierie des Systèmes d'Information
Vol. 31, No. 2, February, 2026, pp. 591-599
Journal homepage: http://iieta.org/journals/isi

A Co-Attentive Pyramid Scene Parsing Network (CA-PSPNet) for Enhanced Colonoscopic Polyp Segmentation at Multiple Scales

E.Nalina[1*], P. Ezhumalai[2]

[1] Department of AIML, R.M.D. Engineering College, Kavaraipettai 601206, India
[2] Department of CSE, R.M.D.Engineering College, Kavaraipettai 601206, India

Corresponding Author Email: nalinasmit@gmail.com

https://doi.org/10.18280/isi.310226

Received: 4 August 2025
Revised: 15 October 2025
Accepted: 16 February 2026
Available online: 28 February 2026

Keywords:
co-attentive Pyramid Scene Parsing Network, polyp segmentation, colonoscopic imaging, multi-scale feature fusion, medical image segmentation

ABSTRACT

Accurate segmentation of polyps from colonoscopic scans is crucial for the timely identification and intervention of colorectal diseases, including cancer and inflammatory sites. However, the significant variability in the appearance, scale, and texture of polyps presents challenges for current segmentation systems. While Pyramid Scene Parsing Networks (PSPNet) use multi-scale contextual attributes via pyramid pooling, their conventional feature extraction methods fail to capture cross-scale dependencies, resulting in blurred edges, missed small polyps, and poor delineation. To address these issues, we propose the CA-PSPNet, an improved model that integrates (i) the Pyramid Pooling Module (PPM) to capture multi-scale features at different spatial resolutions, and (ii) a deep co-attention mechanism that enhances feature fusion and accurately delineates polyp edges. The CA-PSPNet improves polyp recognition by leveraging dynamic focus on the most salient attributes across different scales. Evaluated on the Kvasir-SEG dataset, the model achieves a volume error of 0.852% ± 0.011, a Dice similarity coefficient of 0.977 ± 0.010, and a Jaccard similarity score of 0.956 ± 0.011, while processing each image in just 1.032 ± 0.010 seconds. These results demonstrate the model's efficiency and accuracy in real-world applications, positioning it as a powerful tool for computer-aided diagnosis and real-time disease management.
.

1. INTRODUCTION

Cancer is a deadly disease that poses a serious threat to the healthcare sector because of its high death rate and the complexity of its prevention, diagnosis, and treatment. Globally, in 2022, it was projected that around 9.7 million individuals died from cancer worldwide [1]. Among all malignancies, colorectal cancer (CRC) ranks as the second most commonly identified cancer and the third leading cause of cancer-related fatalities globally [2]. In India, the expected case count of colorectal malignancy was projected to be 44580, with a rough frequency of new cases of 3 per 100,000 individuals. Also, approximately one in every 252 Indians is expected to develop CRC [3]. Projections show that by 2040, the global problem of CRC will increase to 3.2 million incidences and 1.6 million deaths per year. This signifies a 66% rise in incident rate and a 71% rise in mortality relative to figures from 2020 [4]. Presently, digital imaging is deemed the trusted reference for CRC prevention. One of the core tasks in CRC diagnosis is to detect minor anomalous tumors (i.e., polyps), which are recognized as potential precursors to malignant tumors. Therefore, increasing the polyp recognition rate is a significant factor in decreasing death rates. However, such practices are labor-intensive tasks executed by medical professionals and are consequently affected by human aspects, including experience and fatigue.

Since early-stage CRC is typically asymptomatic, it is difficult to achieve timely recognition and exclusion of such polyps at the initial stage. As statistics show, the survival rate is increased significantly if CRC is precisely identified and appropriately treated at an early stage. Occasionally, the 5-year survival rates may even surpass 90% [5]. However, it decreases significantly when the disease is identified at a later stage after metastasis has occurred [6]. When identified earlier, CRC is highly curable, often needs less invasive interventions, and leads to considerably improved prognoses. Since CRC progresses from polyps, timely diagnosis and removal of such tumors at the curable stage can stop their development and decrease related mortality rates. Additionally, early diagnosis decreases the complexity and cost of treatment, thus reducing the burden on the medical industry. However, the accuracy and reliability of polyp identification heavily depend on the skill of the oncologist and the quality of the clinical scans, making diagnostic tools more and more important in disease management systems.

To address the above-mentioned issues, cutting-edge image processing methods are indispensable for increasing recognition rates and patient outcomes. Current developments in Deep Learning (DL) networks have led to substantial improvement in the automatic analysis of clinical image

modalities, mainly using semantic segmentation [7, 8]. The DL networks have developed as effective tools in the classification of CRC by their application in examining colonoscopic images. The diagnostic models using Convolutional Neural Networks (CNNs) have proved notable accuracy in identifying colorectal lesions, isolating polyps, and identifying anomalous tissue with minimal human involvement [9]. Different from conventional image processing methods, DL networks can automatically extract significant attributes from huge databases, enabling them to identify subtle patterns that may be unnoticed by healthcare providers. This aptitude is imperative in CRC detection, where early-stage polyps can be flat, small, and visually unclear. Besides, the incorporation of an attention module and multi-scale attribute extraction methods increase the performance of these networks. Accordingly, DL models not only increase recognition performance but also help physicians by decreasing workload, regulating assessments, and potentially facilitating real-time decision support during colonoscopy screenings [10].

Unconventional DL architectures like U-Net [11], Residual network (ResNet) [12], Region-based CNN (R-CNN) [13], YOLO (You Only Look Once) [14], and Pyramid Scene Parsing Network (PSPNet) [15] have been employed and optimized for tumor segmentation, facilitating accurate identification of disease. Among these models, the PSPNet has developed as a prevailing technique for extracting multi-scale contextual attributes in intricate images. PSPNet is designed to understand the global context of an image by combining attributes at multiple scales for achieving semantic segmentation. The pyramid pooling module is intended to extract contextual data from the input attribute vector by executing many pooling functions with changing filter sizes.

Even though PSPNet is efficient in extracting local and global context using pyramid pooling, it often falls short in clinical settings due to insufficient attribute enhancement across various scales, difficulty in isolating small or low-contrast polyps (i.e., lesions smaller than 10 mm), and poor delineation of edges of polyps in noisy or complex backgrounds. Conventional feature fusion approaches employed in PSPNet may not adequately highlight subtle but medically significant attributes. This restriction hinders its efficiency for polyp isolation in colonoscopic images, where accuracy is critical. Hence, there is a necessity for an architecture that can not only extract contextual information but also effectively combine attributes to highlight clinically significant Regions of Interest (RoI)

This research develops an enhanced segmentation model that integrates the concept of PSPNet with an improved feature fusion mechanism. This cohesive model, the CA-PSP network, targets to combine semantic attributes across several spatial scales more effectively, providing more precise and comprehensive lesion segmentation from colonoscopic images. The core objective of this research is to design a better isolation framework for polyp recognition using an improved version of the PSPNet. The major contributions of this research are threefold.

1. We develop and implement an attribute fusion technique that improves the assimilation of multi-scale pixel-wise attributes in PSPNet for accurate edge delineation of polyp isolation from colonoscopic pictures.
2. We assess the effectiveness of the proposed model using selected performance measures on a benchmark colonoscopy dataset (i.e., Kvasir-SEG).
3. We relate the effectiveness of the proposed CA-PSPNet against other existing segmentation models employed in the healthcare industry.

The remainder parts of the manuscript are structured as follows: We explore similar polyp segmentation models in Section II. We describe the structure and working mechanism of the proposed CA-PSPNet in Section III. Section IV presents the implementation details, dataset preparation, and performance metrics of the intended model. We discuss the experimental outcomes in Section V. Section VI concludes the research.

2. RELATED WORK

The purpose of this research is to develop an in-depth understanding of the existing literature on polyp segmentation, as explored by scholars in the field. Recently, DL networks have led to important growth in polyp isolation. These DL models employ deep neural networks to extract more significant attributes from colonoscopic polyp scans. Gopakumar [16] studied several lesion identification and segmentation models comprehensively. Prasath [17] studied numerous tumor identification models for colonoscopic imaging and presented a systematic review of their merits and limitations. Several scholars have proposed different automated tumor detection models from colonoscopic images. Luca et al. [18] summarized these topical topics in their article. Shin et al. [19] developed a Region-based CNN (R-CNN) for the automated diagnosis of CRC in colonoscopy scans. The authors select Inception ResNet for attribute engineering and combined post- processing methods to achieve more dependable results.

Wang et al. [20] proposed a SegNet structure to identify colorectal tumors with a recognition speed of 25 frames per second. The authors also proved that the proposed model provides high specificity, sensitivity, and storage efficacy. However, this model uses proposal detectors for tumor recognition. Hence, tumor edges are not correctly delineated. To resolve this problem, Kristoffer et al. [21] proposed a Fully Convolutional Network (FCN) with pre-trained models to identify and isolate tumors. These studies [22, 23] introduced an enhanced encoder-decoder design, called PolypSegNet10, for segmenting polyps from colonoscopic pictures. Qadir et al. [24] also developed a method using an FCN model to predict 2D Gaussian shapes, aiming to realize faster polyp recognition speed. This model employs a U-Net to segment polyps from clinical images. This architecture mainly consists of a shrinking path to extract contextual information and a symmetric growing path for accurate diagnosis. Conversely, these models emphasize isolating the polyp but overlook the margin restraints.

Murugesan et al. [22] consider both region and edge constraints for tumor isolation. However, the region-edge relations are not completely explored in this research. To address this issue, Mahmud et al. [23] and Guo et al. [25] developed a confidence-aware resampling technique for handling multiple-scale images and pixel issues in the polyp isolation process. By applying the meta-learning mixup method, the projected framework aims to improve the generality of the model across diverse input images. Additionally, Fan et al. [26] developed the Parallel Reverse Attention Network (PraNet) using the prominent object recognition module. While this model provides better isolation

enactment, its efficiency in handling images with different scales is still confined.

Though the abovementioned models have revealed robust and reliable performance in RoI isolation, their direct application to clinical imaging, particularly colonoscopy, poses unique challenges, including variation in polyps shape and size, low contrast, and complex textures. Due to their reliance on global context information, it is very difficult to identify the malicious potential for smaller lesions. Amalgamation into medical applications also poses challenges, such as a lack of interpretability, regulatory issues, and possible interruption to the developed process. From a technical perspective, dependence on standard evaluation measures such as the Dice Similarity Coefficient (DSC) may obscure clinically significant errors, including missed polyp edges. In addition, the demand for higher processing power for training and real-time implementation presents more difficulties, mainly in settings with inadequate resources. Together, these issues underline the gap between existing research models and medically feasible tools. In this context, we aim to propose an improved segmentation model that uses a CA-PSPNet to extract multi-scale background information and identify lesions accurately.

3. PROPOSED CO-ATTENTIVE PYRAMID SCENE PARSING NETWORK CLASSIFIER

The pyramid scene parsing network is a pixel-wise isolation model introduced by Zhao et al. [15]. It is designed to understand the global context of an image by integrating attributes at multiple scales for achieving pixel-wise isolation [27]. The basic configuration of the CA-PSPNet comprises a backbone attribute network, PPM, and a co-attention module. Figure 1 demonstrates the core architecture of the CA-PSPNet model. The co-attention unit, located between the PPM and the decoder, optimizes multi-scale attributes by cpaturing cross-scale relationship before final segmentation prediction. Given an input colonoscopic image, the ResNet is used as a backbone network to find the attribute vector [28, 29]. The core innovation of ResNet is its residual links, which enable the model to learn identity mappings and alleviate parameter disappearing problems in very deep structures. These links allow the model to send data directly across layers, making it easier to train the framework with multiple layers.

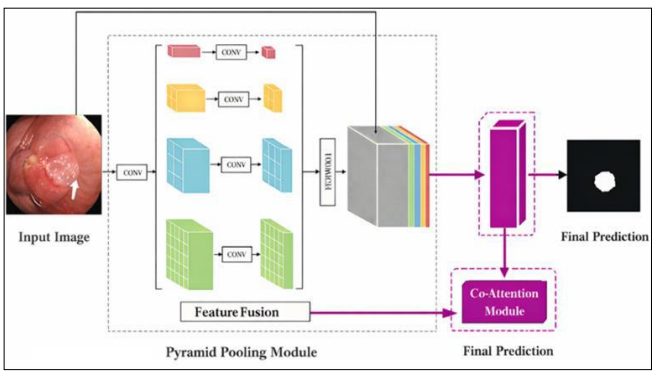

Figure 1. Structure of Co-Attentive Pyramid Scene Parsing Network

In this research, ResNet is used to transform an input colonoscopic image into a rich, high-dimensional attribute vector that transforms spatial and semantic data. This is realized by eliminating the final classification module (i.e., average pooling and fully connected modules) and retaining the convolution layers. Then, this model uses PPM to get different sub-region representations, followed by upsampling and concatenation modules to generate the final attribute maps, comprising both local and global data. Ultimately, this map is transferred to the convolution module to achieve the ultimate semantic segmentation.

3.1 Pyramid pooling module

PPM is developed to extract contextual data from the input attribute vector using several pooling functions with different filter dimensions. This module captures both local and global information at diverse scales. PPM contains four parallel branches, each employing an average pooling module with a distinctive filter dimension. Consequently, a 1 × 1 convolutional operation is employed to decrease the number of channels to a predetermined number (e.g., C/4 in this research). The resulting attribute vectors are then upsampled to their initial spatial size by applying bilinear interpolation. The architecture of PPM used in our proposed model is given in Figure 2. Consider an input attribute vector $A \in \mathbb{R}^{W\times H\times C}$, where C denotes the number of channels; W and H are the width and height of the attribute vector. PPM creates multi-scale contextual attributes through several pooling functions at various scales. Assume the number of pooling levels is L. Every level $l_i \in \{1,2,3 \ldots L\}$ describes a pooling grid of size $P_i \times P_i$, where $P_i$is the spatial size of the pooling for that scale. The pooled attribute for each scale is defined by Eq. (1).

$$A_i = Av.Pool_{Pi}(A) \in \mathbb{R}^{Pi\times Pi\times C} \quad (1)$$

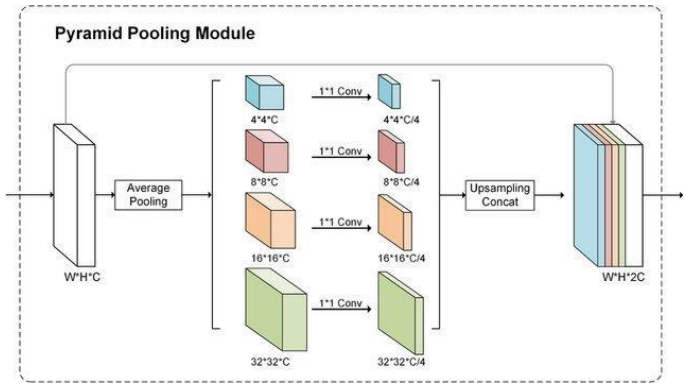

Figure 2. Architecture of pyramid pooling module

All the pooled outputs F_i are sent to a 1 × 1 convolution module for decreasing the channel size as given in Eq. (2).

$$A_i^{'} = Conv_{1\times1}(A_i) \in \mathbb{R}^{Pi\times Pi\times C^{'}} \quad (2)$$

Each $A_i^{'}$ is upsampled back to the initial spatial dimension $W \times H$ through a bilinear interpolation as defined by Eq. (3).

$$A = upsample\,(A', size(W, H)) \in \mathbb{R}^{W\times H\times C^{'}} \quad (3)$$

Then, we apply a 1×1 convolutional operation on the initial input A to decrease it to the identical $C^{'}$ channels as given in Eq. (4).

$$A' = Conv(A) \in \mathbb{R}^{W \times H \times C'} \quad (4)$$

This module concatenates all attributes along the channel size as given in Eq. (5).

$$A = Concat(A') \in \mathbb{R}^{W \times H \times C'(L+1)} \quad (5)$$

Eventually, a 1 × 1 convolution operation is performed to combine the attributes. Eq. (6) defines this operation.

$$A_{out} = Conv(A_{PPM}) \in \mathbb{R}^{W \times H \times C''} \quad (6)$$

The upsampled attribute vectors created from PPM are concatenated with the original input attribute vector, enabling the integration of different contextual data.

3.2 Co-attention module

The co-attention module is a prevailing improvement that allows the model to extract multi-scale attributes (e.g., from diverse scales, modalities, or layers) and inter-attribute interaction. Co-attention module is combined with PSPNet following the attribute concatenation phase to dynamically fine-tune the spatial and channel data. The incorporation of the co-attention module within the decoder further optimizes the attribute pattern, making the model better handle changing RoI shapes and scales. It adaptively focuses on the most discriminative and mutually related regions between these attributes. It receives two attribute vectors (e.g., from different levels) and calculates attention weights for each spatial position or channel based on relative significance. It provides improved attributes by emphasizing relatively discriminative regions.

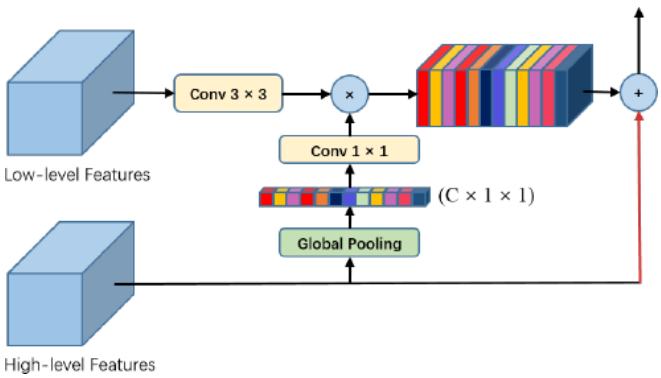

Figure 3. Architecture of the co-attention module

Figure 3 shows the architecture of the co-attention module. For low-level local attributes, a 3 × 3 convolution operation is carried out to decrease the channel dimension of attribute vectors. The global information produced from higher-level attributes is sent to a 1 × 1 convolutional module with batch normalization and ReLU (Rectified Linear Unit) non-linearity, then multiplied by the low-level attributes. To end, the global attributes are combined with the weighted local attributes and upsampled progressively. After applying the co-attention mechanism, the attribute vector is sent to multiple convolution layers to generate the output mask. By assimilating the PSPNet with the co-attention module, the CA-PSPNet efficiently extracts and uses multi-scale data, thus improving the performance of the segmentation process. This model provides more correct and enhanced masks, eventually increasing the overall effectiveness of the model.

Let $F_l \in \mathbb{R}^{C_l \times H \times W}$ represents the low-level local feature vector and $F_h \in \mathbb{R}^{C_h \times H \times W}$ represents the high-level global feature vector captured from diverse phases of the encoder. To align attribute sizes, linear projections are applied using convolutional transformations as given in Eq. (7).

$$Q = \phi_q(F_l), K = \phi_k(F_h), V = \phi_v(F_h) \quad (7)$$

where, $\phi_q(\cdot)$ denotes a 3×3 convolution for local feature enhancement, and $\phi_k(\cdot)$, $\phi_v(\cdot)$ are 1×1 convolutions followed by batch normalization and ReLU activation for encoding global semantic context. The attention matrix is calculated using Eq. (8).

$$A = softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \quad (8)$$

where, d_k is the size of the attribute map, and the softmax function is used along the spatial or channel size to ensure normalized attention weights. The attention-weighted global features are calculated using Eq. (9).

$$F_{att} = A \cdot V \quad (9)$$

These optimized global features are then combined with the local features using element-wise multiplication and residual aggregation as given in Eq. (10).

$$F_{ref} = F_l \odot F_{att} + F_h \quad (10)$$

where, $\odot$ represents element-wise multiplication. Then, the enhanced feature representation F_{ref} is gradually upsampled and transferred via convolutional layers to generate the final segmentation mask as defined by Eq. (11).

$$M = \psi(F_{ref}) \quad (11)$$

where, $\psi(\cdot)$ represents the decoder convolution functions.

4. IMPLEMENTATION OF CA-PSPNET

The proposed model is implemented using Intel® Xeon® Silver 4208 CPU with 64.0 GB RAM, 12 GB video memory capacity, NVidia Titan V graphics card, and Windows 64-bit operating system at 2.10 GHz. The proposed system is trained by TensorFlow and the Adaptive Moment Estimation optimizer. The initial training rate is fixed to 0.001. We use the Kvasir-SEG dataset to evaluate the performance of the intended model. The efficiency of our framework is assessed by relating its numerical results with other existing polyp segmentation models in terms of designated evaluation metrics.

4.1 Kvasir-SEG dataset preparation

The Kvasir-SEG database encompasses 1000 polyp colonoscopic pictures and their analogous reference mask (i.e., gold standard). The image resolution in this dataset differs from 332 × 487 to 1920 × 1072 pixels [30]. The images and their equivalent segmentation maps are achieved in two distinct folders with corresponding filenames. Figure 4

displays some sample input images in the Kvasir-SEG database used for assessment in this research.

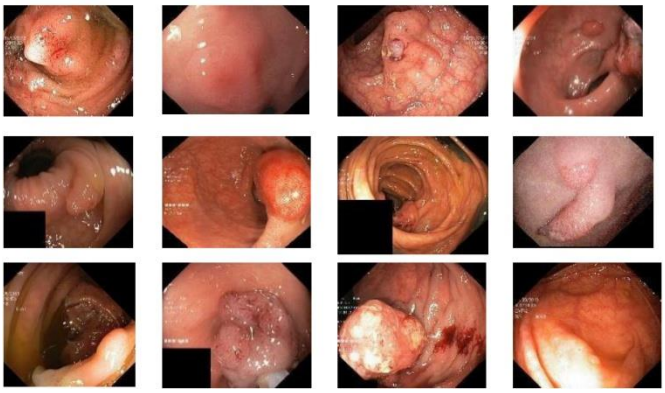

Figure 4. Sample colonoscopic images from the Kvasir-SEG database

4.2 Performance measures

The performance of the CA-PSPNet is comprehensively assessed by some designated metrics, including VE, DSC, and JSS. The metric VE is defined by calculating the deviation between the CA-PSPNet segmentation output and a manually labeled gold standard. The volume error is computed using Eq. (12).

$$VE = \frac{|SR - SG|}{SG} \tag{12}$$

where, S_G is the gold standard and S_R is the segmentation output gained from the proposed model. For clinical applications, a VE of less than 5% is more likely tolerable, particularly for radiotherapy planning [31]. DSC is employed to define the efficiency of the segmentation framework on the input clinical images. It is also called the fitness degree between the initial image and the mask. More precisely, it is a similarity index between two pixels of RoI in the colonoscopic image. The value of DSC is always in [0, 1] and it is calculated by Eq. (13).

$$DSC = \frac{2 \times |SG \cap SR|}{|SG| + |SR|} \tag{13}$$

The JSS is an evaluation metric used to calculate the performance of the isolation framework. Given a dataset, the JSS represents the similarity between the segmentation mask and the gold standard. JSS is computed by Eq. (14).

$$JSS = \frac{|SG \cap SR|}{|SG \cup SR|} \tag{14}$$

To realize more precise segmentation masks, the 10-fold cross-validation (10-f CV) method is applied in this research. Hence, the complete dataset is divided into ten fragments. All images belonging to the same patient were assigned exclusively to a single fold, ensuring that no patient appeared in both the training and testing sets. This splitting strategy prevents data leakage and provides a more realistic assessment of the model's generalization capability in clinical settings. For every fold, one fragment is employed for testing, and the residual fragments are employed to train the model. Then, the average value of all 10 runs is used for analysis. The advantage of this 10-f CV approach is that all the training and testing images are independent, and the fidelity of the outputs could be enhanced. It is important to note that only one-fold may not present accurate outcomes for testing owing to the vagueness of the data samples. As a result, all the outcomes are computed on a mean value of ten runs.

5. RESULT AND DISCUSSION

The proposed segmentation model using the CA-PSPNet is implemented using MATLAB R2018b/deep learning toolbox. A complete study of empirical outputs reveals the strength of the CA-PSPNet. In general, regardless of the dimension of the input colonoscopic scan, the CA-PSPNet segments the lesions efficiently. Figure 5 shows sample colonoscopic pictures used for this research, masks gained by the CA-PSPNet, and their corresponding ground truth segmentation mask. Though the polyps are in diverse arbitrary regions within the colon and look at different sizes, the isolated masks appear to overlap perfectly. The performance of the CA-PSP model is evaluated by relating the numerical outputs with those of 7 similar semantic segmentation models, including U^{2}Net [32], Deeplabv3+ [33], Disentangled non-local neural networks (DnlNet) [34], Feature Augmented Pyramid Network (FAPN) [35], Cross-level Feature Aggregation Network (CFA-Net) [36], and FocusU^{2}Net [37]. All models were trained using the Adam optimizer with an initial learning rate of 1×10^{-4}, batch size of 8, and weight decay of 1×10^{-5}. The learning rate was reduced using a cosine annealing schedule. Binary cross-entropy combined with Dice loss was used as the objective function. Training was performed for 100 epochs with early stopping based on the validation Dice score. Input images were resized to 352 × 352, normalized using ImageNet statistics, and augmented using random flipping, rotation, and scaling.

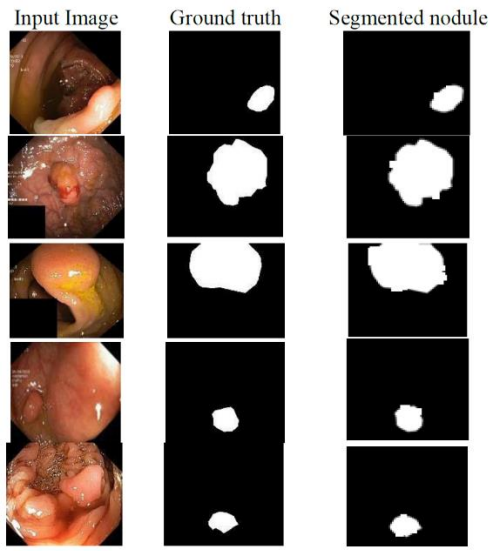

Figure 5. Segmentation results

Table 1 displays the segmentation results gained by the CA-PSPNet segmentation model. The U^{2}Net model achieves 5.645±0.036% VE and 0.846±0.035 DSC. Indeed, the U^{2}Net model can use a nested U-structure with a deep controlling mechanism. Therefore, this model can achieve a 0.787±0.024

JSS. At the same time, the nested architecture and deep controlling mechanism increase the number of variables and processing overhead. Hence, the average processing of the U^{2}Net-based segmentation model for a sample is 4.466±0.002s. Deeplab v3+ provides better results with respect to evaluation metrics since it has better spatial accuracy and better edge detection performance. These attributes make it very efficient for polyp segmentation in colonoscopy datasets. This model provides 4.578±0.031% VE, 0.625±0.072 DSC, and 0.760±0.022 JSS. This model takes 4.083±0.011s for segmenting lesions from colonoscopic images.

Table 1. The results gained by different segmentation networks using the Kvasir-SEG dataset

Model	Criteria	Volume Error (%)	Dice Similarity Coefficient	Jaccard Similarity Score	Average Processing Time (s)
U^{2}Net	Mean	5.645	0.846	0.787	4.466
	SD	0.036	0.035	0.024	0.022
Deeplab v3+	Mean	4.578	0.625	0.760	4.083
	SD	0.031	0.072	0.022	0.011
DnlNet	Mean	3.371	0.812	0.842	3.160
	SD	0.041	0.199	0.020	0.011
FAPN	Mean	2.567	0.897	0.859	3.189
	SD	0.047	0.115	0.017	0.011
CFA-Net	Mean	2.172	0.906	0.904	2.199
	SD	0.025	0.138	0.020	0.013
FocusU^2 Net	Mean	1.332	0.943	0.914	2.168
	SD	0.019	0.017	0.019	0.014
CA-PSPNet	Mean	0.852	0.977	0.956	1.032
	SD	0.011	0.010	0.011	0.010

By separating the content and position attention in attribute vectors, the DnlNet model realizes enhanced outputs related to U^{2}Net and Deeplab v3+ segmentation networks. Also, it enlaces segmentation performance by better modeling structured and long-range relationships. Therefore, it provides a comparatively lower VE (3.371±0.041%), higher DSC (0.812±0.199), and higher JSS (0.842±0.020). Besides, it consumes a moderate mean processing time (3.160±0.011s). However, incorporating disentangled non-local modules into a prevailing structure upsurges model intricacy, making the model more difficult to train and optimize, challenging to correct, and sensitive to weight initialization and training rate. This may hamper reproducibility and utilization in medical edge devices.

The FAPN contains three modules for cross-embedding, predictive regulation, and graded attribute fusion to solve the issues imposed by the multifaceted textures and obscure edges of polyps. By successfully extracting multi-scale attributes and optimizing edges, FAPN enhances the accuracy of polyp segmentation. It provides better results for the segmentation of polyps regarding VE (2.567±0.047%), DSC (0.897±0.115), and JSS (0.859±0.017). For effective polyp isolation, it consumes 3.189 ± 0.011s for each image. The CFA-Net is a DL model developed to improve the isolation of polyps in colonoscopy images. It solves the challenges imposed by various sizes, shapes, and unclear edges, which thwart the accurate isolation of polyps.

By generating boundary-aware attributes, which are vital for segmenting polyps from neighboring healthy tissues, it provides better segmentation performance with 2.172±0.025% VE, 0.906±0.138 DSC, and 0.904±0.020 JSS. This model consumes 2.199±0.013s for segmenting lesions from colonoscopic images. By integrating spatial and channel attention, FocusU2Net provides better performance for accurately segmenting polyps from colonoscopic images. This model realizes performance with 1.332± 0.019%, 0.943± 0.017, 0.914± 0.019, and 2.168 ± 0.014s in VE, DSC, JSS, and average processing time, correspondingly.

Our CA-PSPNet outdoes all other segmentation approaches with respect to measures designated for assessment. The proposed CA-PSPNet leverages the pyramid pooling operation to extract rich contextual information at multiple spatial scales through pyramid pooling and an enhanced feature fusion technique to improve pixel-wise attributes across different scales, providing more accurate edge delineation of polyps.

This study also proves that the pyramid pooling module of the CA-PSPNet ensures effective yet less computationally expensive architecture compared with other segmentation networks. It realizes evaluation metrics of 0.852± 0.011, 0.977±0.010, 0.956±0.011, and 1.032±0.010 in VE, DSC, JSS, and average processing time, respectively. In contrast, early encoder–decoder architectures such as U2Net and DeepLabv3+ exhibit significantly higher VE values (5.645% and 4.578%, respectively), suggesting difficulty in accurately delineating polyp regions, particularly along irregular boundaries. The extensive empirical results reveal that our CA-PSPNet efficiently manages the intricacy of real colonoscopic scans and delivers a favorable solution for polyp segmentation.

The observed performance gains can be directly attributed to the co-attention–based feature fusion strategy introduced in CA-PSPNet. This design allows fine-scale boundary cues to be reinforced by coarse-scale semantic context, leading to improved segmentation of small, flat, and low-contrast polyps. In addition to accuracy, CA-PSPNet demonstrates superior computational efficiency, which is substantially faster than all compared methods. This efficiency stems from the strategic placement of the co-attention module after the PPM, allowing effective feature refinement without introducing excessive computational overhead. Figures 6 and 7 show the performance of various polyp isolation models regarding mean and SD values, respectively.

Figure 8 presents the visualization results of different segmentation models, illustrating the superior performance of CA-PSPNet in terms of sharper boundary delineation and more precise attention to clinically relevant polyp regions compared with other state-of-the-art approaches.

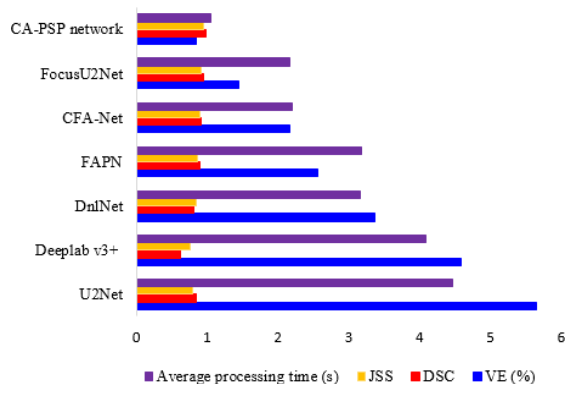

Figure 6. Performance of various polyp isolation models regarding mean values

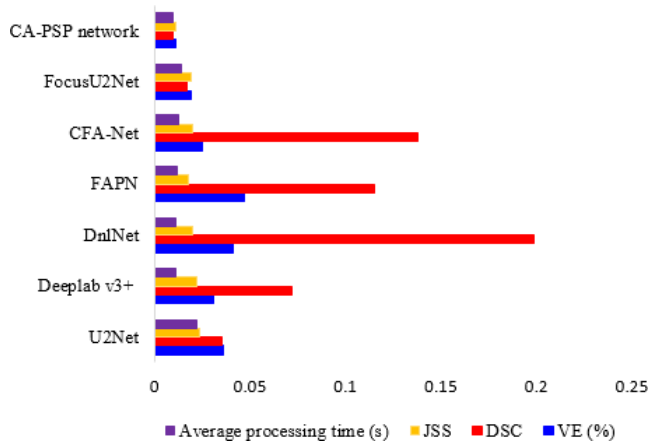

Figure 7. Performance of various polyp isolation models regarding SD values

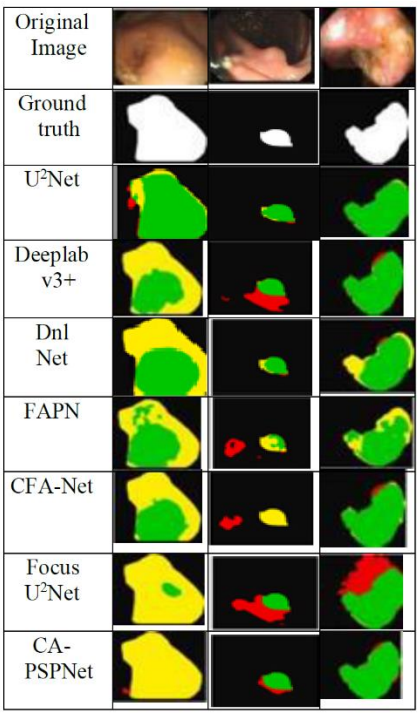

Figure 8. Visualization of segmented polyps

6. CONCLUSION

Exploration of colonoscopic scans plays a substantial role in the timely identification of colorectal cancer. Automatic isolation can be beneficial for polyp identification and characterization, thus providing a significant, correct, and reliable diagnostic tool. Nonetheless, the high variability in appearance, scale, and texture of these RoI poses major challenges for existing polyps segmentation models. This research develops a polyp segmentation model using a pyramid scene parsing model to extract multi-scale image data and identify lesions accurately. The proposed network leverages the PPM to extract rich contextual information at multiple spatial scales by applying pyramid pooling. This model also uses an enhanced feature fusion technique by applying a deep co-attention mechanism. This mechanism improves the pixel-wise attributes across different scales and enables more accurate polyp edge detection. Besides, it adaptively focuses on the most important and relatively germane regions. The proposed model is implemented in a test bed and evaluated on an open-source colonoscopy dataset, called Kvasir-SEG. The extensive empirical results reveal its effectiveness in terms of selected performance indicators. The CA-PSPNet outperforms existing lesion segmentation models in terms of 0.852±0.011% of volume error, 0.977±0.010 of dice similarity coefficient, and 0.956±0.011 of JSS. This model consumes only 1.032±0.010s for processing each image. These results demonstrate the effectiveness of the CA-PSPNet in addressing the complications of colorectal polyp isolation in real-world applications.

REFERENCES

[1] He, C.S., Wang, C.J., Wang, J.W., Liu, Y.C. (2023). UY-NET: A two-stage network to improve the result of detection in colonoscopy images. Applied Sciences, 13: 10800. https://doi.org/10.3390/app131910800

[2] Roshandel, G., Ghasemi-Kebria, F., Malekzadeh, R. (2024). Colorectal cancer: Epidemiology, risk factors, and prevention. Cancers, 16(8): 1530. https://doi.org/10.3390/cancers16081530

[3] Sathishkumar, K., Chaturvedi, M., Das, P., Stephen, S., Mathur, P. (2022). Cancer incidence estimates for 2022 and projection for 2025: Result from national cancer registry programme, India. Indian Journal of Medical Research, 156(4&5): 598-607. https://doi.org/10.4103/ijmr.ijmr_1821_22

[4] Morgan, E., Arnold, M., Gini, A., Lorenzoni, V., Cabasag, C.J., Laversanne, M., Vignat, J., Ferlay, J., Murphy, N., Bray, F. (2023). Global burden of colorectal cancer in 2020 and 2040: Incidence and mortality estimates from GLOBOCAN. Gut, 72(2): 338-344. https://doi.org/10.1136/gutjnl-2022-327736

[5] Kwan, A.S.U., Uwishema, O., Mshaymesh, S., Choudhary, K., Salem, F.K., Sengar, A.S., Patel, R.P., Kazan, Z., Wellington, J. (2025). Advances in the diagnosis of colorectal cancer: The application of molecular biomarkers and imaging techniques: A literature review. Annals of Medicine & Surgery, 87(1): 192-203. https://doi.org/10.1097/MS9.0000000000002830

[6] Brenner, H., Kloor, M., Pox, C.P. (2024). Colorectal cancer. The Lancet, 383(9927): 1490-1502. https://doi.org/10.1016/S0140-6736(13)61649-9

[7] Islam, M.M., Poly, T.N., Walther, B.A., Yeh, C.Y., Seyed-Abdul, S., Li, Y.J., Lin, M.C. (2022). Deep learning for the diagnosis of esophageal cancer in endoscopic images: A systematic review and meta-analysis. Cancers, 14(23): 5996. https://doi.org/10.3390/cancers14235996

[8] Anteby, R., Horesh, N., Soffer, S., Zager, Y., Barash, Y., Amiel, I., Rosin, D., Gutman, M., Klang, E. (2021). Deep learning visual analysis in laparoscopic surgery: A systematic review and diagnostic test accuracy meta-analysis. Surgical Endoscopy, 35(4): 1521-1533. https://doi.org/10.1007/s00464-020-08168-1

[9] Mienye, I.D., Swart, T.G., Obaido, G., Jordan, M., Ilono, P. (2025). Deep convolutional neural networks in medical image analysis: A review. Information, 16(3): 195. https://doi.org/10.3390/info16030195

[10] Alam, M.J., Fattah, S.A. (2023). SR-AttNet: An interpretable stretch-relax attention-based deep neural

network for polyp segmentation in colonoscopy images. Computers in Biology and Medicine, 160: 106945. https://doi.org/10.1016/j.compbiomed.2023.106945

[11] Al Jowair, H., Alsulaiman, M., Muhammad, G. (2023). Multi-parallel U-Net encoder network for effective polyp image segmentation. Image and Vision Computing, 137: 104767. https://doi.org/10.1016/j.imavis.2023.104767

[12] Bouzid, K., Sharma, H., Killcoyne, S., Coelho de Castro, D., Schwaighofer, A., Ilse, M., Salvatelli, V., Oktay, O., Murthy, S., Bordeaux, L. (2023). Enabling large-scale screening of Barrett's esophagus using weakly supervised deep learning in histopathology. Nature Communications, 15: 2026. https://doi.org/10.1038/s41467-024-46174-2

[13] Ren, S., He, K., Girshick, R., Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39: 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[14] Doniyorjon, M., Madinakhon, R., Shakhnoza, M., Cho, Y. (2022). An improved method of polyp detection using custom YOLOv4-tiny. Applied Sciences, 12(21): 10856. https://doi.org/10.3390/app122110856

[15] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 6230-6239. https://doi.org/10.1109/CVPR.2017.660

[16] Gopakumar, G. (2020). A review on polyp detection and segmentation in colonoscopy images using deep learning. International Journal of Engineering Research and Technology, 9(10): 329-335. https://doi.org/10.5281/zenodo.18657958

[17] Prasath, V.B.S. (2017). Polyp detection and segmentation from video capsule endoscopy: A review. Journal of Imaging, 3(1): 1. https://doi.org/10.3390/jimaging3010001

[18] Luca, M., Ciobanu, A., Drug, V. (2019). Deep learning and automatic polyp detection in colonoscopies: A review of recent contributions and future outlook. In 2019 E-Health and Bioengineering Conference (EHB), Iasi, Romania, pp. 1-4. https://doi.org/10.1109/EHB47216.2019.8970041

[19] Shin, Y., Qadir, H.A., Aabakken, L., Bergsland, J., Balasingham, I. (2018). Automatic colon polyp detection using region-based deep CNN and post-learning approaches. IEEE Access, 6: 40950-40962. https://doi.org/10.1109/ACCESS.2018.2856402

[20] Wang, P., Xiao, X., Brown, G. Jr., Berzin, T.M., Tu, M., Xiong, F., Hu, X., Liu, P., Song, Y., Zhang, D. (2018). Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nature Biomedical Engineering, 2: 741-748. https://doi.org/10.1038/s41551-018-0301-3

[21] Kristoffer, W., Michael, K., Robert, J. (2020). Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Medical Image Analysis, 60: 101619. https://doi.org/10.1016/j.media.2019.101619

[22] Murugesan, B., Sarveswaran, K., Shankaranarayana, S.M., Ram, K., Joseph, J., Sivaprakasam, M. (2019). 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), Berlin, Germany, pp. 7223-7226. https://doi.org/10.1109/EMBC.2019.8857339

[23] Mahmud, T., Paul, B., Fattah, S.A. (2021). PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images. Computers in Biology and Medicine, 128: 104119. https://doi.org/10.1016/j.compbiomed.2020.104119

[24] Qadir, H.A., Shin, Y., Solhusvik, J., Bergsland, J., Aabakken, L., Balasingham, I. (2019). Polyp detection and segmentation using Mask R-CNN: Does a deeper feature extractor CNN always perform better? In Proceedings of the 13th International Symposium on Medical Information and Communication Technology (ISMICT), Oslo, Norway, pp. 1-6. https://doi.org/10.1109/ISMICT.2019.8743694

[25] Guo, X., Chen, Z., Liu, J., Yuan, Y. (2022). Non-equivalent images and pixels: Confidence-aware resampling with meta-learning mixup for polyp segmentation. Medical Image Analysis, 78: 102394. https://doi.org/10.1016/j.media.2022.102394

[26] Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L. (2020). Pranet: Parallel reverse attention network for polyp segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 263-273. https://doi.org/10.1007/978-3-030-59725-2_26

[27] Picon, A., San-Emeterio, M.G., Bereciartua-Perez, A., Klukas, C., Eggers, T., Navarra-Mestre, R. (2022). Deep learning-based segmentation of multiple species of weeds and corn crop using synthetic and real image datasets. Computers and Electronics in Agriculture, 194: 106719. https://doi.org/10.1016/j.compag.2022.106719

[28] Lv, Q., Wang, H. (2021). Cotton boll growth status recognition method under complex background based on semantic segmentation. In Proceedings of the 4th International Conference on Robotics, Control and Automation Engineering (RCAE), Wuhan, China, pp. 50-54. https://doi.org/10.1109/RCAE53607.2021.9638864

[29] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

[30] Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P. (2020). Kvasir-SEG: A segmented polyp dataset. In Proceedings of the International Conference on Multimedia Modeling, Daejeon, South Korea, pp. 451-462. https://doi.org/10.1007/978-3-030-37734-2_37

[31] Savjani, R.R., Lauria, M., Bose, S., Deng, J., Yuan, Y., Andrearczyk, V. (2022). Automated tumor segmentation in radiotherapy. Seminars in Radiation Oncology, 32(4): 319-329. https://doi.org/10.1016/j.semradonc.2022.06.002

[32] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaïane, O.R., Jagersand, M. (2020). U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition, 106: 107404. https://doi.org/10.1016/j.patcog.2020.107404

[33] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Lecture

Notes in Computer Science, pp. 833-851. https://doi.org/10.1007/978-3-030-01234-2_49
[34] Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S., Hu, H. (2020). Disentangled non-local neural networks. In European Conference on Computer Vision, pp. 191-207. https://doi.org/10.1007/978-3-030-58555-6_12
[35] Su, Y., Cheng, J., Yi, M., Liu, H. (2022). FAPN: Feature augmented pyramid network for polyp segmentation. Biomedical Signal Processing and Control, 78: 103903. https://doi.org/10.1016/j.bspc.2022.103903
[36] Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., Shen, D. (2023). Cross-level feature aggregation network for polyp segmentation. Pattern Recognition, 140: 109555. https://doi.org/10.1016/j.patcog.2023.109555
[37] Ovi, T.B., Bashree, N., Nyeem, H., Wahed, M.A. (2025). FocusU2Net: Pioneering dual attention with gated U-Net for colonoscopic polyp segmentation. Computers in Biology and Medicine, 186: 109617. https://doi.org/10.1016/j.compbiomed.2024.109617