Article

DeepColonLab: Attention Guided Separable Receptive Field Block Enhanced Deeplabv3+ Model for Colon Polyp Segmentation

Abduz Zami * and Shadman Sobhan

Posted Date: 5 June 2025

doi: 10.20944/preprints202506.0380.v1

Keywords: colon polyp segmentation; colorectal cancer; DeepColonLab; DeepLabV3+; Receptive Field Block

Article

DeepColonLab: Attention Guided Separable Receptive Field Block Enhanced Deeplabv3+ Model for Colon Polyp Segmentation

Abduz Zami [1] and Shadman Sobhan [2,*]

[1] Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh

[2] Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh

[*] Correspondence: shadmansobhan114@gmail.com

Abstract: Colon polyp segmentation is necessary for early colorectal cancer classification, helping to reduce deaths from one of the most common and deadly cancers worldwide. Accurate segmentation of colon polyp is difficult due to their diverse morphologies and different sizes. Existing models like convolutional neural networks may struggle to preserve fine-grained spatial details and transformer-based architectures may not be computational efficiency for real-time clinical use. So, we introduce DeepColonLab, a modification of the DeepLabV3+ model, specially designed for colon polyp segmentation. Our approach introduces a Separable Receptive Field Block (SRFB), inspired by human visual receptive fields, integrated with a Convolutional Block Attention Module (CBAM) to replace traditional Atrous Spatial Pyramid Pooling (ASPP). DeepLabV3+ is well-suited for colon polyp segmentation due to its encoder-decoder architecture and ASPP module, which enable effective multi-scale feature extraction and precise boundary delineation. It is lighter than many traditional segmentation models but can achieve high accuracy due to its structure. Original DeepLabV3+ with ASPP module lacks sufficient mechanisms for global context awareness and fine boundary refinement. Our proposed design enhances multiscale contextual information capture while preserving spatial resolution, particularly for small and irregularly shaped polyps. It enhanced receptive field's flexibility and better channel wise feature transformation for balancing efficiency and accuracy. Using lightweight EfficientNet encoders, DeepColonLab balances accuracy and computational efficiency. This model also provides better gradient flow and feature retention than base DeepLabV3+. This model outperformed most of the recent and state-of-the-art models on benchmark datasets—Kvasir, CVC-ClinicDB, and CVC-ColonDB— achieving Dice Coefficients of up to 0.9597 ± 0.0060 and Intersection over Union (IoU) scores of up to 0.9314 ± 0.0084. The efficiency of the model supports real-time medical imaging applications, making it a promising tool for clinical deployment in the management of colorectal cancer.

Keywords: colon polyp segmentation; colorectal cancer; DeepColonLab; DeepLabV3+; Receptive Field Block

1. Introduction

Colon polyp segmentation plays a vital role in improving the detection, diagnosis, and treatment of colorectal cancer, one of the most common and deadly cancers worldwide [1]. Polyps are abnormal growths on the inner lining of the colon or rectum, and although many are harmless, certain types, especially adenomatous polyps, can become cancerous over time [2]. Detecting and removing them early is crucial, which is why regular colonoscopy tests are recommended, especially for people over 50 years of age or those with a family history of colorectal disease.

In 2024, the American Cancer Society projected over 150,000 new colorectal cancer cases in the U.S. alone [3]. Studies show that 25–30% of adults over 50 years of age have at least one polyp, and although not all become cancerous, nearly all colorectal cancers begin as polyps[4]. Early removal during colonoscopy can prevent their progression and has been shown to reduce cancer incidence by up to 68% and mortality by nearly 50% [5].

Despite the effectiveness of colonoscopy, polyps—especially small, flat, or hidden ones—can be missed, contributing to post-colonoscopy cancer cases, which account for up to 9% of diagnoses[6]. To address this, automated colon polyp segmentation uses computer vision and AI to provide precise, pixel-level identification of polyps in colonoscopy images. Unlike basic detection, segmentation outlines the exact size and location of polyps, supporting more accurate diagnosis and treatment planning.

The benefits are significant: improved detection accuracy, reduced oversight, quicker decision-making, and automated documentation. These tools assist clinicians during procedures, power real-time systems, and enhance medical education through annotated datasets used to train AI models [7], [8].

Colon polyp segmentation is a critical task in medical imaging, enabling early detection of colorectal cancer, a leading cause of cancer-related mortality worldwide. In recent years, deep learning has significantly advanced the field of medical image segmentation, particularly for tasks like colon polyp detection. Among the most widely used architectures is the U-Net, a convolutional neural network (CNN) designed specifically for biomedical image segmentation [9]. U-Net's encoder-decoder structure allows it to capture both low-level spatial features and high-level semantic context, making it highly effective at producing accurate, pixel-level masks of polyps. Variants like U-Net++ and Attention U-Net further improve upon this design by adding nested skip connections or attention mechanisms, which help the model focus on relevant regions in complex colonoscopy images [10], [11].

Convolutional neural networks, particularly U-Net and its variants, have been the backbone of many colon polyp segmentation systems due to their ability to perform precise pixel-level segmentation. U-Net is especially popular in biomedical applications because of its symmetric encoder-decoder architecture and skip connections, which help retain both spatial and contextual information. However, in the context of polyp segmentation, U-Net exhibits certain limitations. It can struggle with detecting small, flat, or irregularly shaped polyps, especially those located near the boundaries or hidden within folds of the colon [12]. These models also tend to lose fine-grained details when dealing with low-contrast images, which are common in colonoscopy data [13]. Moreover, U-Net relies heavily on local receptive fields due to its convolutional structure, limiting its ability to capture global contextual information, which is essential for accurately segmenting polyps that vary in size and appearance[14].

Transformer-based models, such as Vision Transformers (ViTs), Swin-UNet, and TransUNet, have been introduced to address some of these limitations [15], [16]. These models use self-attention mechanisms to capture long-range dependencies and global context more effectively than traditional CNNs. In polyp segmentation, this helps in distinguishing polyps from surrounding tissue more accurately, even when the polyp shape is irregular or the background is complex. However, Transformer models also come with their challenges. They are computationally expensive, require significantly larger amounts of annotated training data, and may overfit when trained on smaller medical datasets. In addition, their performance may suffer when working with high-resolution endoscopic images unless they are carefully optimized for memory and speed. While transformer-CNN hybrid models show promise, striking the right balance between accuracy and efficiency remains an ongoing challenge in clinical deployment [17].

Existing segmentation models, such as those based on convolutional neural networks (CNNs) or Transformers, often struggle with capturing fine-grained spatial details, handling variable polyp morphologies, or achieving computational efficiency suitable for real-time applications. Motivated by these challenges, there is a pressing need to develop a robust, efficient, and precise segmentation framework that can reliably delineate polyps of diverse characteristics while maintaining practical computational demands for clinical deployment.

DeepLabv3+ is a widely used semantic segmentation model that leverages dilated (atrous) convolutions and Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale contextual information. It is lighter than many traditional segmentation CNNs and transformers, but can achieve high accuracy due to its robust multi-scale feature extraction structure. This makes it well-suited for segmenting polyps of varying shapes and sizes in colonoscopy images. Its ability to model large receptive fields without losing resolution helps detect polyps against complex and low-contrast backgrounds [18], [19]. However, it lacks sufficient mechanisms for global context awareness and fine boundary refinement. To enhance the performance of DeepLabV3+ for colon polyp segmentation, we replaced the standard Atrous Spatial Pyramid Pooling (ASPP) module with the more flexible Receptive Field Block (RFB). While ASPP captures multi-scale features using fixed dilation rates, it lacks the adaptability needed for fine-grained segmentation. RFB's multi-branch structure with more effective convolutions provides better receptive field diversity and feature integration. This modification improves boundary refinement and context awareness, addressing key limitations of ASPP in complex medical segmentation tasks.

The primary objectives of this research are to develop and evaluate DeepColonLab, an enhanced DeepLabV3+ architecture tailored for accurate and efficient colon polyp segmentation. Specifically, we aim to:

- Propose a novel Separable Receptive Field Block (SRFB) integrated with a Convolutional Block Attention Module (CBAM) to replace the traditional Atrous Spatial Pyramid Pooling (ASPP), improving the preservation of spatial details and multi-scale contextual information capture.
- Enhance segmentation accuracy, particularly for small and irregularly shaped polyps, by leveraging lightweight EfficientNet encoders for efficient feature extraction.
- Achieve superior performance compared to state-of-the-art methods through extensive evaluation on benchmark datasets, including Kvasir, CVC-ClinicDB, and CVC-ColonDB, using Dice Coefficient and Intersection over Union (IoU) metrics.
- Ensure computational efficiency suitable for real-time medical imaging applications, facilitating potential clinical deployment.

The rest of the paper is structured as follows: Section 2, *Literature Review* discusses various literature and articles that shaped the way for this work. Section 3 *DeepColonLab* overviews the encoder and decoder architecture proposed in this study. Datasets used in this work and implementation details were addressed in Section 4. Experiments we conducted to evaluate our model are shown in Section 5. This section also provides a comparison between the proposed model and state-of-the-art models. Finally, Section 6 concludes the article.

2. Literature Review

An extensive examination of colorectal polyp segmentation is given in this part, which covers a range of techniques from sophisticated machine learning and deep learning techniques to conventional image processing. It provides a succinct summary of the major methodological and technological developments in the field and highlights how approaches have evolved, especially the contribution of CNNs, Transformers, and Hybrid approaches to improving segmentation accuracy and efficiency. And finally this section discusses a few recent studies that outperformed these methods by utilizing DeepLabV3-based approaches.

2.1. Recent Advances in Colon Polyp Segmentation

Recent advances in colon polyp segmentation have leveraged deep learning, particularly convolutional neural networks (CNNs), transformers, and hybrid models, to improve diagnostic accuracy. CNN-based architectures such as PraNet [13], SANet [20], and HarDNet-MSEG [21] have laid a solid foundation by focusing on efficient and effective local feature extraction. PraNet introduced a parallel reverse attention mechanism to highlight salient polyp regions while suppressing irrelevant background information. SANet followed with a shallow attention design that enabled faster inference with fewer parameters, although it sacrificed deeper semantic understanding. HarDNet-MSEG proposed a lightweight encoder-decoder framework that achieved real-time performance while maintaining strong segmentation metrics, making it suitable for real-world applications where latency is a concern.

Transformer-based models have emerged as an alternative to CNNs by enabling the capture of global dependencies. Polyp-PVT [19] utilized Pyramid Vision Transformers to encode multi-scale features with a broader receptive field. ColonFormer [22] introduced an efficient attention mechanism within a transformer backbone to enhance spatial understanding, while Polyp ViT [23] applied deformable convolutions and positional encoding for improved adaptation to varying polyp morphologies. HiFiSeg [24] pushed this further by focusing on high-frequency edge enhancement through global-local transformer blocks, achieving state-of-the-art accuracy on several datasets. However, these models often involve a large number of parameters and computational overhead.

Hybrid models have been developed to harness the strengths of both CNNs and transformers. PGCFNet [25] integrated a PVT encoder with a CNN decoder and introduced a Progressive Group Convolution Fusion module with differential subtraction to refine boundary features. DLGRAFE-Net [26] combined residual attention and double-loss guidance to improve feature representation, while ContourNet [27] introduced contour supervision to sharpen boundary prediction. Additionally, models like those by Xue et al. [28] and Xu et al. [29] focused on lightweight designs and ensemble learning to maintain accuracy while reducing complexity. These innovations indicate a trend toward achieving both precision and practicality in polyp segmentation tasks.

2.2. *Gaps in Existing Methods*

Despite the remarkable advancements in both CNN and transformer-based architectures for colon polyp segmentation, significant limitations still persist. CNN-based models such as PraNet, SANet, and HarDNet-MSEG have demonstrated strong capabilities in capturing local features and providing efficient inference speeds. However, these models often struggle with generalization across datasets, particularly when dealing with polyps that exhibit low contrast, irregular boundaries, or small sizes. Their reliance on local receptive fields limits their ability to incorporate long-range contextual dependencies, which is critical in differentiating polyps from visually similar background tissue. Even well-established architectures like DeepLabV3+ [18] may face challenges in segmenting polyps with vague edges or under varying illumination conditions, especially without architectural enhancements. Moreover, while DeepLabV3+ is robust in many cases, its performance is not optimal without the integration of additional context-awareness mechanisms or attention modules.

On the other hand, transformer-based approaches such as Polyp-PVT, ColonFormer, and HiFiSeg have shown promise in modeling global dependencies and multi-scale features. Nevertheless, these methods are often computationally intensive and require substantial GPU resources, limiting their practical deployment in real-time or resource-constrained clinical environments. Furthermore, their complex architectures reduce interpretability and make integration with existing medical imaging workflows more challenging. Generalization also remains a concern, as some transformer models show inconsistent performance on external datasets, particularly in cases involving subtle texture differences or challenging imaging conditions.

Overall, the field still lacks a unified solution that offers both high segmentation accuracy and computational efficiency while maintaining generalizability and adaptability across diverse polyp datasets. This gap creates a strong motivation to explore architectures that combine the robustness and modularity of CNNs, like DeepLabV3+, with enhancements that address its limitations—without introducing excessive model complexity.

2.3. *Positioning Our Approach*

To address the above challenges, our work builds upon the DeepLabV3+ framework to propose a solution that is both effective and efficient for colon polyp segmentation as recently many studies has achieved a significant improvement through modifications of DeepLabV3+. For instance, Xiang et al. [30] proposed a lightweight adaptation of DeepLabV3+ aimed at reducing computational complexity while maintaining segmentation performance, making it more viable for real-time applications. Similarly, Gangrade et al. [31] introduced a multi-level context attention mechanism into DeepLabV3+, significantly improving boundary sensitivity and the model's ability to detect subtle polyp regions. Both of them achieved competitive performance compared to other recent works, considering their model sizes. DeepLabV3+ offers an ideal starting point due to its flexible encoder-decoder design and the atrous spatial pyramid pooling (ASPP) module, which facilitates multi-scale feature extraction without reducing spatial resolution. While standard DeepLabV3+ lacks sufficient mechanisms for global context awareness and fine boundary refinement, recent works have explored targeted enhancements.

Inspired by these efforts, we further enhance the DeepLabV3+ architecture by integrating advanced attention-driven modules and feature fusion strategies tailored to the complexities of colonoscopy imagery. Our model is designed to overcome common limitations in both CNN and transformer-based approaches by improving segmentation of small, flat, or low-contrast polyps, while maintaining an efficient, lightweight structure suitable for deployment in clinical environments.

Unlike transformer-based models that are computationally demanding, our approach offers a balance between performance and efficiency. It preserves the modularity and interpretability of DeepLabV3+ while incorporating architectural innovations that expand its capability to capture both global context and fine details. By combining these strengths, our method achieves competitive performance across multiple public datasets, with lower inference times and better generalization than baseline models.

In summary, our work positions itself as a practical and high-performing alternative that bridges the gap between traditional CNN approaches and emerging transformer-based models. Building on the successes of recent DeepLabV3+ enhancements, we offer a refined architecture that meets the dual demands of segmentation accuracy and real-world applicability.

3. DeepColonLab

Figure 1 shows the architecture of our proposed network, DeepColonLab. The network is a modified DeepLabV3+ architecture. The main difference between our DeepColonLab and the DeepLabV3+ is the use of Separable Receptive Field Block with CBAM attention instead of the ASPP block. We will provide detailed descriptions of each component in the sections below.

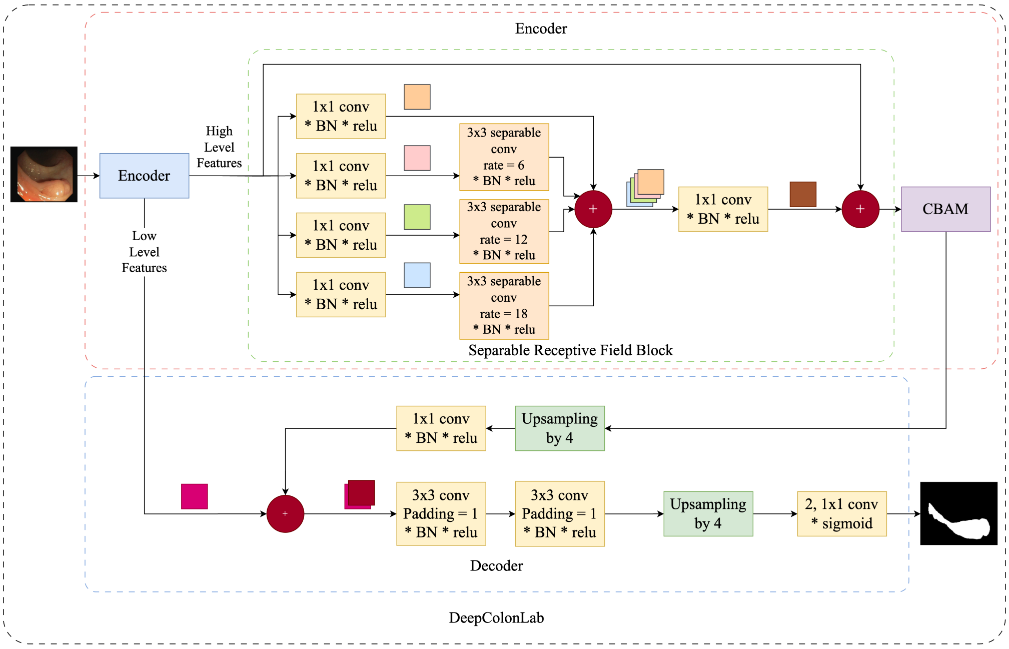


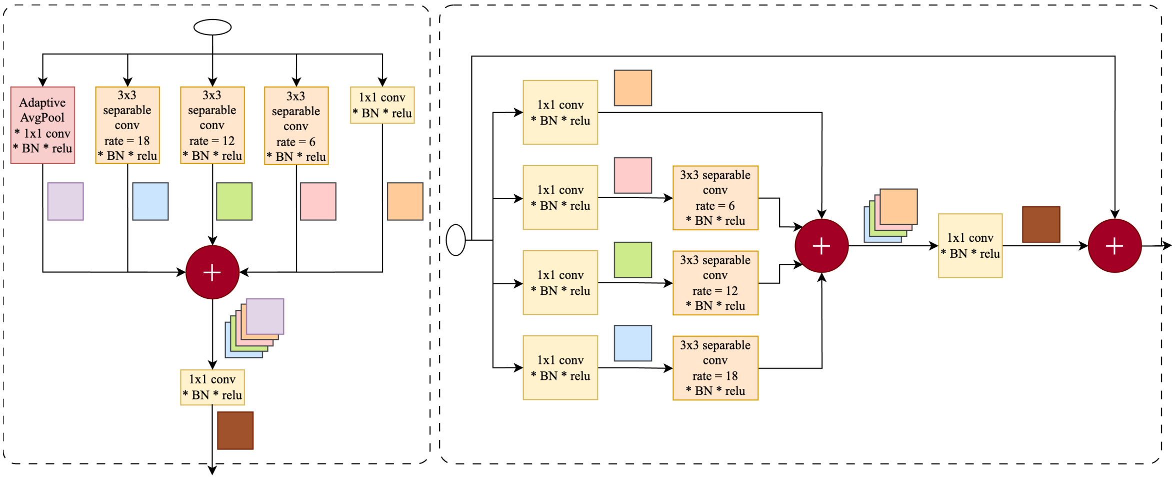
Figure 1. Architecture of our proposed DeepColonLab

3.1. Encoder

DeepLabv3+ consists of an encoder for feature extraction. We tried ResNet, DenseNet, and EfficientNet with different types of them as encoder. But after the trials we proposed EfficientNet as encoder. DeepLabV3+ takes two outputs from this encoder to pass to the next layer. One is called the low-level feature, and the other is a high-level feature. Low level feature is taken at very beginning and high level feature is taken at very end of the encoder.

3.2. Separable Receptive Field Block

This is the most important contribution of our work. We replaced the Atrous Spatial Pyramid Pooling (ASPP) of traditional DeepLabV3+ with a Separable Receptive Field Block (SRFB). This SRFB was inspired by the Receptive Field Block (RFB) originally proposed for object detection in RFBNet [32], which mimics the receptive fields in human vision to enhance feature representation. By adapting and refining this concept, we introduced SRFB as a novel module that leverages separable convolutions and multi-scale receptive field expansion, making it particularly suited for our task. The SRFB builds on the strengths of RFB by combining branches with different dilation rates, each employing separable convolutions instead of standard ones, followed by concatenation and a residual connection. This design emphasized efficient multi-scale contextual information capture—spanning local to medium-scale contexts—while maintaining computational efficiency. Unlike ASPP, which integrates global average pooling to capture image-level features and relies heavily on parallel atrous convolutions for multi-scale aggregation, SRFB was designed to avoid explicit global pooling. Instead, it focused on preserving spatial details through residual connections and optimized receptive field expansion, which is critical for dense prediction tasks like colon polyp segmentation. Figure 2 shows a visual comparison of the architectures of our SRFB and ASPP.


Adaptive AvgPool * 1x1 conv * BN * relu
3x3 separable conv rate = 18 * BN * relu
3x3 separable conv rate = 12 * BN * relu
3x3 separable conv rate = 6 * BN * relu
1x1 conv * BN * relu
+
1x1 conv * BN * relu
1x1 conv * BN * relu
1x1 conv * BN * relu
3x3 separable conv rate = 6 * BN * relu
1x1 conv * BN * relu
3x3 separable conv rate = 12 * BN * relu
1x1 conv * BN * relu
3x3 separable conv rate = 18 * BN * relu
+
1x1 conv * BN * relu
+

(a) ASPP

(b) SRFB

Figure 2. Architecture of ASPP (a) and SRFB (b)

Colon polyp segmentation requires precise localization of polyps, which vary significantly in size, shape, and texture, often blending into surrounding mucosal tissue. While ASPP excels in semantic segmentation tasks by aggregating global context and multi-scale features—making it ideal for scenarios with large, coherent objects—its reliance on global pooling and high dilation rates can dilute fine-grained spatial details. This is a drawback in colon polyp segmentation, where small or irregularly shaped polyps demand high-resolution feature preservation.

SRFB addressed these challenges more effectively. First, its residual connection, inspired by ResNet, ensured that the input features are directly added to the output, preserving spatial information that might otherwise be lost in deep networks. Second, by replacing standard convolutions with separable convolutions, SRFB reduced computational complexity while maintaining the ability to model multi-scale receptive fields. This efficiency was crucial for real-time medical imaging applications. Third, SRFB's focus on local to medium-scale context, rather than global pooling, aligned better with the localized nature of polyps, which do not require extensive image-level context for accurate delineation.

In contrast to ASPP's design, which was tailored for pixel-wise classification across large scenes (e.g., outdoor environments in DeepLab's original use case), SRFB adapted the RFB philosophy to prioritize receptive field diversity without sacrificing resolution. This made it more suitable for the dense, spatially sensitive predictions needed in colon polyp segmentation.

3.2.1. Convolutional Block Attention Module (CBAM)

The Convolutional Block Attention Module (CBAM) enhanced feature maps by emphasizing the most important channels and spatial regions. We added a CBAM attention mechanism after the SRFB block. It starts with channel attention: the input feature map is processed using average and max pooling, then passed through simple layers with ReLU and sigmoid functions to generate channel weights. These weights were applied to the feature map to highlight key channels. Next, it used spatial attention: the adjusted feature map is pooled again, the results were combined, and a convolution with a sigmoid function created spatial weights. These weights further refined the feature map to focus on critical areas. CBAM improved feature quality for tasks requiring precise localization, like ours. Figure 3 shows the CBAM used in our study.

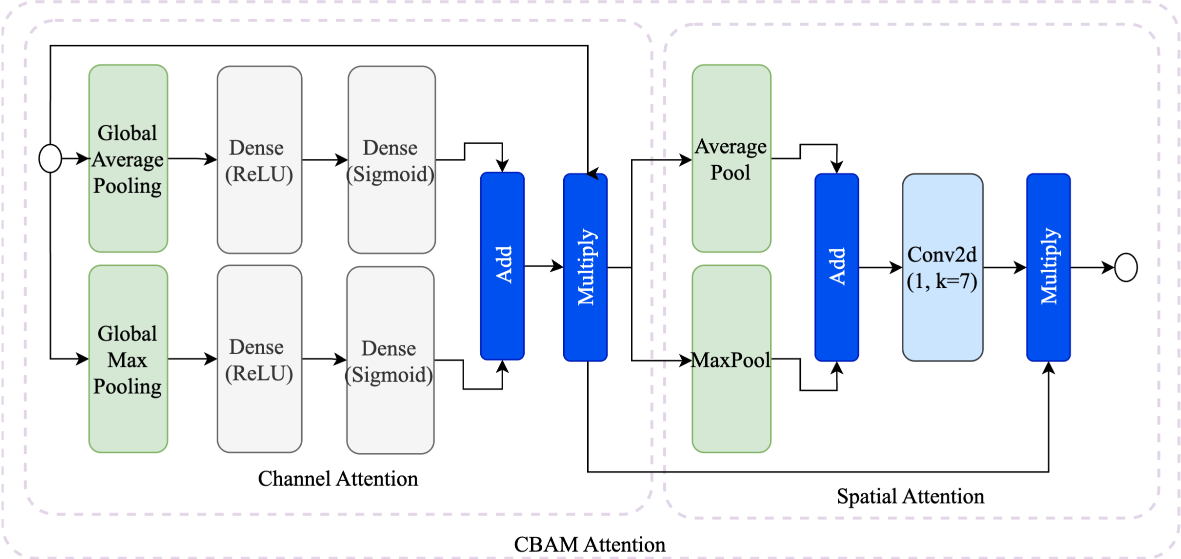


Figure 3. The Convolutional Block Attention Module

3.3. Decoder

The final component of DeepLabV3+ was the decoder, which refined the encoder's output into a high-resolution segmentation map. It upsampled the encoder features using bilinear interpolation, reduced them to 48 channels via a 1x1 convolution, and concatenated them with low-level features. The combined features were processed through two 3x3 convolutions with batch normalization and ReLU, ensuring precise boundary delineation for tasks like colon polyp segmentation.

4. Experimental Setup

4.1. Datasets

We evaluate our approach on three widely used benchmark datasets for polyp segmentation: Kvasir [33], CVC-ClinicDB [34], and CVC-ColonDB [35]. A summary of each dataset is provided below:

- **Kvasir:** This dataset contains 1,000 endoscopic images collected at Vestre Viken Health Trust (VV), Norway. The image resolutions range from 720 × 576 to 1920 × 1072 pixels. All annotations were carefully performed and verified by experienced gastroenterologists from VV and the Cancer Registry of Norway.

- **CVC-ClinicDB:** Composed of 612 image frames with a resolution of 384 × 288 pixels, this dataset was extracted from 31 colonoscopy video sequences. It was used in the MICCAI 2015 Sub-Challenge on Automatic Polyp Detection in Colonoscopy Videos.
- **CVC-ColonDB:** Provided by the Machine Vision Group (MVG), this dataset includes 380 images with a resolution of 574 × 500 pixels. The images were taken from 15 short colonoscopy video sequences.

4.2. Evaluation Metrics and Loss Function

To evaluate the performance of the segmentation model, we employed two widely-used metrics in medical and binary segmentation tasks: the Dice Coefficient and the Jaccard Index, also known as the Intersection over Union (IoU). These metrics were chosen for their effectiveness in quantifying the overlap between predicted and ground truth segmentation masks, especially in cases where class imbalance is present.

The Dice Coefficient measures the harmonic mean of precision and recall, providing a score between 0 and 1, where 1 indicates perfect agreement. Similarly, the Jaccard Index evaluates the similarity between the predicted and true masks by computing the ratio of the intersection to the union of the sets.

For optimization during training, Dice Loss was used as the loss function. Dice Loss is derived from the Dice Coefficient and is particularly effective in handling class imbalance by directly optimizing for better overlap between prediction and ground truth. This made it a suitable choice for our segmentation task, ensuring the model focused on the quality of region overlap rather than just pixel-wise accuracy.

4.3. Implementation Details

The study was conducted using Kaggle's P100 GPU. Images and masks were resized to 256 × 256 pixels, and the masks were one-hot encoded. A five-fold cross-validation strategy was employed to ensure acceptabilty of scores. The dataset split in an 80:20 ratio into training-validation and testing sets. the training-validation was again splitter in 80:20 ratio for training and validation set at each fold. Each fold was trained for 100 epochs using a batch size of 16, and reliability was ensured by calculating mean and standard deviations of metrics across the folds.

A dynamic learning rate schedule was adopted, where training began with a 1-epoch cycle, and the duration of each subsequent cycle was doubled. The learning rate was decayed from the optimizer's initial value to 1×10^{-7} following a cosine annealing pattern, with resets to the initial rate at the start of each new cycle. Early stopping was implemented to stop training after 20 consecutive epochs if there were no improvement in the validation IoU score. Checkpointing was also implemented to save model weights with best validation IoU score.

To prevent overfitting, real-time data augmentation was applied during training, that included transformations - horizontal and vertical flips.

5. Experiments

5.1. Experiment 1

We first trained our model on CVC-ClinicDB database to check the performance comparison of SRFB over the ASPP block. We tried Resnet50, Resnet101, Resnet152, DenseNet121, DenseNet169, DenseNet201, EfficientNetB0, EfficientNetB1, EfficientNetB2 and EfficientNetB3. We found that DeepLabV3+ was getting better results with EfficientNet encoders. They striked an optimal balance between model size, computational efficiency, and accuracy. This can be due to their compound scaling method that uniformly scales depth, width, and resolution, enabling better feature extraction and generalization across tasks while remaining resource-efficient. Our SRFB module was doing better with

doi:10.20944/preprints202506.0380.v1

these EfficientNet encoders. Most importantly, SRFB was doing far more better with the light encoders like EfficientNet, ResNet50, ResNet101 and DenseNet121. Table 1 shows our obtained result for ASPP and SRFB for different encoders. From this result, we conclude that for Colon Polyp Segmentation, light EfficientNet encoders are best, and for these EfficientNet encoders, SRFB performs better than the ASPP.

Table 1. Performance of DeepLabV3+ with different backbones and bridge modules.

Type	Dice (mean ± std)	IoU (mean ± std)	Parameters
Resnet50+ASPP	0.9391 ± 0.0070	0.9012 ± 0.0083	27,026,338
Resnet50+RFB	0.9419 ± 0.0053	0.9036 ± 0.0079	27,110,050
Resnet101+ASPP	0.9354 ± 0.0016	0.8963 ± 0.0019	46,018,466
Resnet101+RFB	0.9394 ± 0.0104	0.9012 ± 0.0141	46,102,178
Resnet152+ASPP	0.9377 ± 0.0086	0.8992 ± 0.0127	61,662,114
Resnet152+RFB	0.9368 ± 0.0102	0.8977 ± 0.0134	61,745,826
Densenet121+ASPP	0.9373 ± 0.0084	0.9000 ± 0.0112	8,986,338
Densenet121+RFB	0.9426 ± 0.0061	0.9072 ± 0.0080	9,097,698
Densenet169+ASPP	0.9413 ± 0.0061	0.9041 ± 0.0082	15,353,442
Densenet169+RFB	0.9412 ± 0.0107	0.9054 ± 0.0139	15,447,522
Densenet201+ASPP	0.9421 ± 0.0049	0.9056 ± 0.0061	21,296,482
Densenet201+RFB	0.9339 ± 0.0014	0.8952 ± 0.0017	21,383,650
EfficientnetB0+ASPP	0.9501 ± 0.0034	0.9180 ± 0.0040	5,000,094
EfficientnetB0+RFB	0.9512 ± 0.0053	0.9202 ± 0.0069	5,130,462
EfficientnetB1+ASPP	0.9455 ± 0.0010	0.9123 ± 0.0014	7,505,730
EfficientnetB1+RFB	0.9558 ± 0.0047	0.9258 ± 0.0060	7,636,098
EfficientnetB2+ASPP	0.9529 ± 0.0050	0.9225 ± 0.0072	8,735,364
EfficientnetB2+RFB	0.9573 ± 0.0053	0.9275 ± 0.0082	8,864,868
EfficientnetB3+ASPP	0.9564 ± 0.0055	0.9266 ± 0.0076	11,781,642
EfficientnetB3+RFB	0.9577 ± 0.0037	0.9288 ± 0.0042	11,910,282

The architectural differences between ASPP and RFB provide insight into why RFB might outperform ASPP in colon polyp segmentation. ASPP consist of four 3x3 separable convolutions with fixed dilation rates i.e. 6, 12, 18 and an adaptive global pooling branch to capture multi-scale features. And this works pretty good, but we thought of enhancing it's performance more. RFB adopts a more dynamic approach with 3x3 separable convolutions at a fixed rate i.e. 6, 12, 18 combined with multiple 1x1 convolutions, enhancing receptive field flexibility and integrating features through a structured multi-branch design, which better handles fine details. It helped encoder to fetch complex patterns by allowing better channel-wise feature transformations. Also, the 1x1 convolution performed as a bottleneck and allowed the model to balance performance and efficiency. Additionally, the residual connection in SRFB improved the gradient flow and feature retention. This enables RFB to stabilize training and preserve critical details, likely contributing to its superior performance over ASPP, especially in complex segmentation tasks.

5.2. Experiment 2

So, we knew from Experiment 1 that lightweight EfficientNet encoders perform best for DeepLabV3+, and the model becomes more refined when using the SRFB module instead of ASPP. To further enhance the representational power of SRFB, we integrated a Convolutional Block Attention Module (CBAM) after the SRFB block. CBAM provides both the channel and spatial dimensions. This helps the model to capture better fine-grained polyp boundaries and subtle texture variations. Table 2 shows the effectiveness of including CBAM with our SRFB block.

Table 2. Cross-validation performance of DeepLabV3+ with RFB_CBAM bridge and various encoders.

Encoder	Dice (mean ± std)	IoU (mean ± std)	Parameters
EfficientNetB0	0.9524 ± 0.0044	0.9209 ± 0.0064	5,138,752
EfficientNetB1	0.9579 ± 0.0030	0.9278 ± 0.0042	7,644,388
EfficientNetB2	0.9590 ± 0.0038	0.9301 ± 0.0046	8,873,158
EfficientNetB3	0.9593 ± 0.0023	0.9304 ± 0.0038	11,918,572
EfficientNetB4	0.9583 ± 0.0029	0.9289 ± 0.0044	18,852,876
EfficientNetB5	0.9591 ± 0.0042	0.9308 ± 0.0051	29,736,180
EfficientNetB6	0.9597 ± 0.0060	0.9314 ± 0.0084	42,213,020
EfficientNetB7	0.9590 ± 0.0026	0.9299 ± 0.0037	65,355,412

Figure 4 shows the difference in output by ASPP, SRFB, and SRFB+CBAM with EfficientNetB6. Outputs are pretty similar, but if it is looked carefully then the difference in the border regions of the polyps can be easily seen. SRFB and SRFB+CBAM configurations can fetch the border details better than the ASPP configuration. As the base DeepLabV3+ obtained very high scores, a little improvement is crucial.

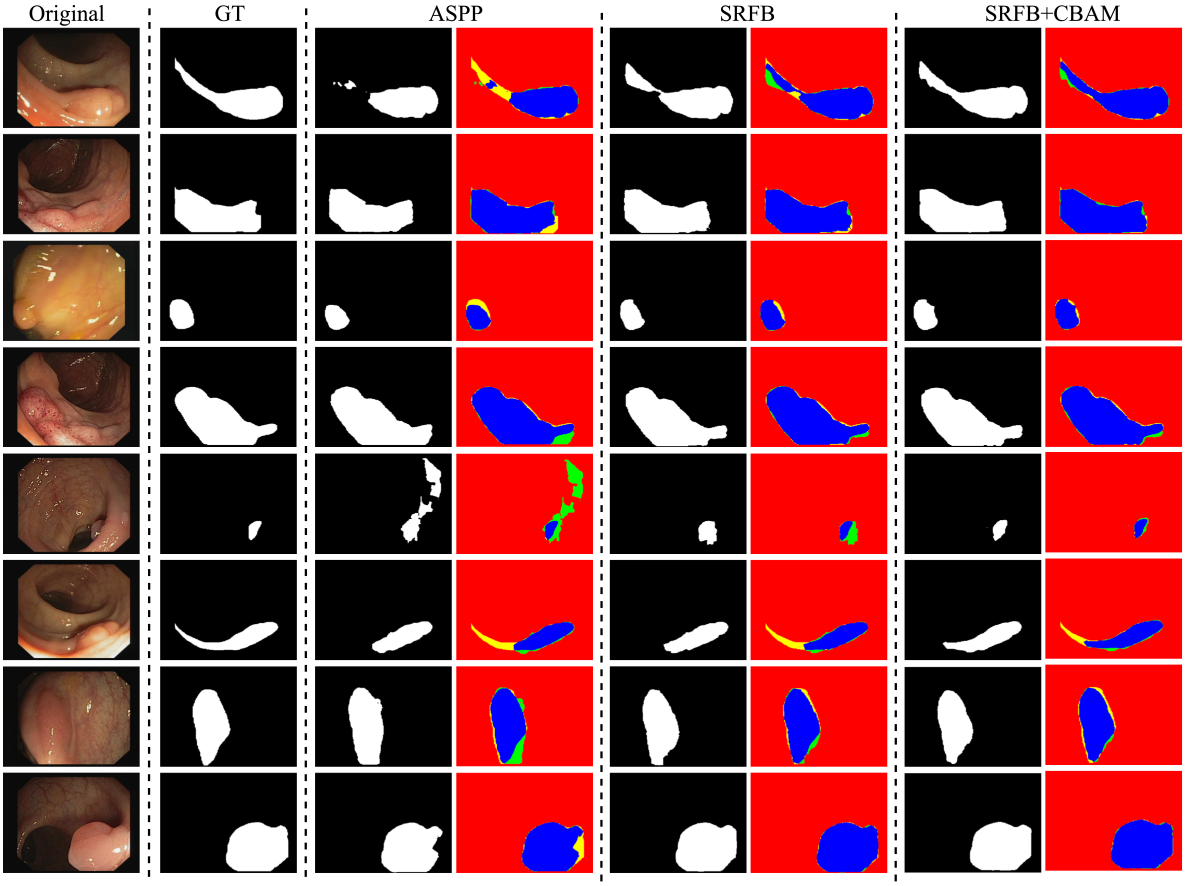


Figure 4. Visual Comparison of the performance of different configurations on CVC-ClinicDB dataset (True Positive = Blue, False Positive = Green, True Negative = Red, False Negative = Yellow)

5.3. Experiment 3

Results of experiment 2 shows that all the EfficientNet backbones from B1 to B6 show almost similar scores. Higher parameters improve the scores only a little. So, as per the results of experiment 2, we selected 3 types of our modified DeepLabV3+ architectures. One is with EfficientNetB2, another with EfficientNetB4, and the other is EfficientNetB6. We are proposing name DeepColonLab(L), DeepColonLab(M), and DeepColonLab(H) respectively for them. This was due to the increment in result with a huge cost in the increment in the number of parameters. EfficientNetB2 is very light but still managed to get a close score with EfficientNetB6. EfficientNetB6 has 40M parameters, but it obtained the highest score. EfficientNetB4 was selected as a middle version of EfficientNetB2 and EfficientNetB6. So we trained and tested all three variations on all three datasets, and the results are significantly very satisfactory. Table 3 shows scores obtained by these three variations of encoders on all three datasets.

Table 3. Cross-validation performance of selected DeepColonLab configurations across different datasets using RFB+CBAM bridge.

Dataset	Configuration	Encoder	Dice (mean ± std)	IoU (mean ± std)
CVC-ClinicDB	DeepColonLab (L)	EfficientNetB2	0.9590 ± 0.0038	0.9301 ± 0.0046
	DeepColonLab (M)	EfficientNetB4	0.9583 ± 0.0029	0.9289 ± 0.0044
	DeepColonLab (H)	EfficientNetB6	0.9597 ± 0.0060	0.9314 ± 0.0084
CVC-ColonDB	DeepColonLab (L)	EfficientNetB2	0.9379 ± 0.0098	0.9019 ± 0.0105
	DeepColonLab (M)	EfficientNetB4	0.9385 ± 0.0189	0.9031 ± 0.0199
	DeepColonLab (H)	EfficientNetB6	0.9397 ± 0.0101	0.9043 ± 0.0112
Kvasir	DeepColonLab (L)	EfficientNetB2	0.9260 ± 0.0064	0.8830 ± 0.0082
	DeepColonLab (M)	EfficientNetB4	0.9318 ± 0.0059	0.8902 ± 0.0075
	DeepColonLab (H)	EfficientNetB6	0.9302 ± 0.0070	0.8885 ± 0.0087

5.4. Comparison with Other Recent Works

Table 4 highlights the superior performance of our proposed DeepColonLab model in comparison to other state-of-the-art methods. The table presents a curated list of prominent works published between 2020 and 2025, showcasing how our approach consistently outperforms existing models across key evaluation metrics. The reported scores are those stated in the respective original papers. While there may be occasional discrepancies or inconsistencies in the reported results, DeepColonLab still demonstrated a clear performance advantage in most cases. This comparative analysis underscored the effectiveness and robustness of our proposed framework.

Table 4. Comparison of our DeepColonLab with other state-of-the-art works on Colon Polyp Segmentation

Work	Year	Dataset	Dice	Jaccard
PraNet [13]	2020	CVC-ClinicDB CVC-ColonDB Kvasir	0.899 0.709 0.898	0.849 0.640 0.840
SANet [20]	2021	CVC-ClinicDB CVC-ColonDB Kvasir	0.916 0.753 0.904	0.859 0.670 0.847
Polyp PVT [19]	2021	CVC-ClinicDB CVC-ColonDB Kvasir	0.937 0.808 0.936	0.889 0.727 0.949
ColonFormer [22]	2022	CVC-ClinicDB CVC-ColonDB Kvasir	0.934 0.811 0.927	0.884 0.733 0.877
HarDNet-MSEG [21]	2021	CVC-ClinicDB CVC-ColonDB Kvasir	0.932 0.731 0.912	0.882 0.660 0.857
Cun Xu et al. [29]	2024	CVC-ClinicDB Kvasir	0.953 0.923	0.975 0.954
He Xue et al. [28]	2024	CVC-ClinicDB Kvasir	0.9471 0.9390	0.9021 0.8907
Shiyu Xiang et al. [30]	2024	CVC-ClinicDB Kvasir	0.9328 0.8882	0.8742 0.7988
Gangrade et al. [31]	2024	CVC-ClinicDB Kvasir	0.955 0.975	0.912 0.962
ContourNet [27]	2024	CVC-ClinicDB CVC-ColonDB Kvasir	0.97 0.99 0.97	0.94 0.98 0.95
Jianuo Liu et al. [26]	2024	CVC-ClinicDB Kvasir	0.9438 0.9166	0.8996 0.8564
HiFiSeg [24]	2025	CVC-ClinicDB CVC-ColonDB Kvasir	0.942 0.826 0.933	0.897 0.749 0.886
DeepColonLab (Ours)		CVC-ClinicDB CVC-ColonDB Kvasir	0.9597±0.0060 0.9397±0.101 0.9318±0.0059	0.9314±0.0084 0.9043±0.112 0.8902±0.0075

6. Conclusions

This study introduced DeepColonLab, a sophisticated enhancement of the DeepLabV3+ architecture tailored for colon polyp segmentation, addressing critical challenges in colorectal cancer prevention. By replacing the conventional Atrous Spatial Pyramid Pooling (ASPP) with a novel Separable Receptive Field Block (SRFB) augmented by a Convolutional Block Attention Module (CBAM), DeepColonLab significantly improved the capture of multi-scale contextual information and the preservation of fine-grained spatial details. The SRFB, inspired by receptive field diversity in human vision, employed separable convolutions and residual connections to enhance feature representation, while CBAM's attention mechanisms focused on salient channels and spatial regions, improving segmentation accuracy for challenging polyps. The adoption of lightweight EfficientNet encoders ensured computational efficiency, making the model viable for real-time clinical applications. Comprehensive experiments across the Kvasir, CVC-ClinicDB, and CVC-ColonDB datasets validated DeepColonLab's superiority, with configurations like DeepColonLab(H) (EfficientNetB6) achieving Dice Coefficients of 0.9597 ± 0.0060 on CVC-ClinicDB, 0.9397 ± 0.0101 on CVC-ColonDB, and 0.9318 ± 0.0059 on Kvasir, alongside IoU scores of 0.9314 ± 0.0084, 0.9043 ± 0.0112, and 0.8902 ± 0.0075, respectively. These results surpassed recent state-of-the-art methods, particularly in handling small, flat, or irregularly shaped polyps that are often missed in colonoscopy. The model's efficiency, with parameter counts as low as 8.87M for DeepColonLab(L) (EfficientNetB2), supports its potential integration into endoscopic systems. Future research will focus on further optimizing DeepColonLab for resource-constrained clinical environments, exploring hybrid CNN-transformer integrations, and extending its applicability to other medical imaging tasks. DeepColonLab represents a significant advancement in AI-driven tools, offering a robust, precise, and efficient solution to enhance early colorectal cancer detection and improve patient outcomes.

References

1. R. L. Siegel, K. D. Miller, N. S. Wagle, A. Jemal, Cancer statistics, 2023, CA: a cancer journal for clinicians 73 (1) (2023) 17–48.
2. J. Bond, Colon polyps and cancer, Endoscopy 35 (01) (2003) 27–35.
3. R. L. Siegel, A. N. Giaquinto, A. Jemal, Cancer statistics, 2024, CA: a cancer journal for clinicians 74 (1) (2024) 12–49.
4. D. A. Lieberman, D. G. Weiss, J. H. Bond, D. J. Ahnen, H. Garewal, W. V. Harford, D. Provenzale, S. Sontag, T. Schnell, T. E. Durbin, et al., Use of colonoscopy to screen asymptomatic adults for colorectal cancer, New England Journal of Medicine 343 (3) (2000) 162–168.
5. A. G. Zauber, S. J. Winawer, M. J. O'Brien, I. Lansdorp-Vogelaar, M. van Ballegooijen, B. F. Hankey, W. Shi, J. H. Bond, M. Schapiro, J. F. Panish, et al., Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths, New England Journal of Medicine 366 (8) (2012) 687–696.
6. S. Sanduleanu, C. M. le Clercq, E. Dekker, G. A. Meijer, L. Rabeneck, M. D. Rutter, R. Valori, G. P. Young, R. E. Schoen, Definition and taxonomy of interval colorectal cancers: a proposal for standardising nomenclature, Gut 64 (8) (2015) 1257–1267.
7. E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature medicine 25 (1) (2019) 44–56.
8. D. A. Hashimoto, G. Rosman, D. Rus, O. R. Meireles, Artificial intelligence in surgery: promises and perils, Annals of surgery 268 (1) (2018) 70–76.
9. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer, 2015, pp. 234–241.
10. Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4, Springer, 2018, pp. 3–11.
11. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas. arxiv, arXiv preprint arXiv:1804.03999 10 (2018).
12. 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.

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. L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587 (2017).
15. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
16. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, arXiv preprint arXiv:2102.04306 (2021).
17. J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, V. M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, in: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24, Springer, 2021, pp. 36–46.
18. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.
19. B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, L. Shao, Polyp-pvt: Polyp segmentation with pyramid vision transformers, arXiv preprint arXiv:2108.06932 (2021).
20. J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, S. Cui, Shallow attention network for polyp segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer, 2021, pp. 699–708.
21. C.-H. Huang, H.-Y. Wu, Y.-L. Lin, Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps, arXiv preprint arXiv:2101.07172 (2021).
22. N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, V. S. Dinh, Colonformer: An efficient transformer based method for colon polyp segmentation, IEEE Access 10 (2022) 80575–80586.
23. M. Y. Sikkandar, S. G. Sundaram, A. Alassaf, I. AlMohimeed, K. Alhussaini, A. Aleid, S. A. Alolayan, P. Ramkumar, M. K. Almutairi, S. S. Begum, Utilizing adaptive deformable convolution and position embedding for colon polyp segmentation with a visual transformer, Scientific Reports 14 (1) (2024) 7318.
24. J. Ren, X. Zhang, L. Zhang, Hifiseg: High-frequency information enhanced polyp segmentation with global-local vision transformer, IEEE Access (2025).
25. Z. Ji, H. Qian, X. Ma, Progressive group convolution fusion network for colon polyp segmentation, Biomedical Signal Processing and Control 96 (2024) 106586.
26. J. Liu, J. Mu, H. Sun, C. Dai, Z. Ji, I. Ganchev, Dlgrafe-net: A double loss guided residual attention and feature enhancement network for polyp segmentation, Plos one 19 (9) (2024) e0308237.
27. S. Pathan, Y. Somayaji, T. Ali, M. Varsha, Contournet-an automated segmentation framework for detection of colonic polyps, IEEE Access (2024).
28. H. Xue, L. Yonggang, L. Min, L. Lin, A lighter hybrid feature fusion framework for polyp segmentation, Scientific Reports 14 (1) (2024) 23179.
29. C. Xu, K. Fan, W. Mo, X. Cao, K. Jiao, Dual ensemble system for polyp segmentation with submodels adaptive selection ensemble, Scientific Reports 14 (1) (2024) 6152.
30. S. Xiang, L. Wei, K. Hu, Lightweight colon polyp segmentation algorithm based on improved deeplabv3+, Journal of Cancer 15 (1) (2024) 41.
31. S. Gangrade, P. C. Sharma, A. K. Sharma, Y. P. Singh, Modified deeplabv3+ with multi-level context attention mechanism for colonoscopy polyp segmentation, Computers in Biology and Medicine 170 (2024) 108096.
32. 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.
33. 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, 2019, pp. 451–462.

34. 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, Computerized medical imaging and graphics 43 (2015) 99–111.
35. N. Tajbakhsh, S. R. Gurudu, J. Liang, Automated polyp detection in colonoscopy videos using shape and context information, IEEE transactions on medical imaging 35 (2) (2015) 630–644.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.